THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En informatique

École doctorale :
I2S - Information, Structures et Systèmes (ED 166)

Unité de recherche :
LIRMM - Laboratoire d'informatique, de Robotique et de Micro-électronique de Montpellier

Méthodes et Modèles pour l'élaboration automatisée de Graphes de Connaissances dans le domaine juridique : Application aux Ressources Juridiques et Juridico-Pratiques des Collectivités Locales et Territoriales

Présentée par Camille GOSSET

Sous la direction de Mathieu LAFOURCADE et Mokhtar Boumedyen BILLAMI

Devant le jury composé de :

Mokhtar Boumedyen BILLAMI, Berger-Levrault	**Co-encadrant**
Sandra BRINGAY, Professeur, Université Paul Valéry, LIRMM	**Examinatrice**
Marianne HUCHARD, Professeur, Université de Montpellier, LIRMM	**Examinatrice**
Mathieu LAFOURCADE, Université de Montpellier	**Directeur de thèse**
Didier SCHWAB, Professeur des universités, LIG, UMR 5217	**Rapporteur**
Cassia TROJAHN, Professeur des universités, Université Grenoble Alpes	**Rapporteur**

UNIVERSITÉ DE MONTPELLIER

Remerciements

Avant tout, un immense merci aux membres du jury d'avoir accepté de prendre le temps de lire et d'évaluer mon travail. Votre présence et vos retours me touchent beaucoup. Vos remarques et suggestions ont vraiment enrichi cette thèse, et je suis très reconnaissante pour l'attention que vous y avez portée.

Un grand merci à mon directeur de thèse, Mathieu Lafourcade, pour m'avoir guidée et accompagnée tout au long de cette aventure. Je n'oublie pas non plus mon co-encadrant, Mokhtar BILLAMI, dont les conseils, l'expertise, et les encouragements ont été essentiels dans la construction de ce travail. Merci pour votre patience, vos conseils et surtout votre soutien, qui m'ont été précieux à chaque étape. Merci d'avoir été là pour éclairer ma démarche avec vos idées et votre enthousiasme.

Merci également à Berger-Levrault de m'avoir permis de mener cette thèse. Je suis très reconnaissante d'avoir pu compter sur le soutien de toute l'équipe et d'avoir eu accès à toutes les ressources nécessaires pour mener à bien ce projet. Je voudrais également faire une mention spéciale à la team des bébous de BL : vous avez rendu les journées de travail bien plus agréables et souvent même inoubliables ! Merci pour votre bonne humeur, votre soutien et les fous rires partagés. Travailler avec vous a été une vraie source de motivation et de Workaholisme.

Je pense aussi aux membres du LIRMM avec qui j'ai eu la chance de travailler. Merci pour votre accueil chaleureux, pour les discussions toujours enrichissantes, et pour les moments partagés qui ont rendu ces années encore plus agréables. Un grand merci aussi à mon équipe un peu spéciale du LIRMM : rICARd.TXT ! Votre soutien et votre amitié ont été essentiels tout au long de cette thèse. Merci pour les idées partagées, les échanges inspirants, et les bons moments passés ensembles. Je n'oublierai jamais nos pauses cafés, nos restos ainsi que les apéros LIRMM ! Vous avez rendu l'expérience au laboratoire unique et plus riche.

Et enfin, merci du fond du cœur à ma famille et mes proches. Vous avez été là à chaque instant, dans les moments de doute comme dans les petites victoires. Merci pour votre patience, vos encouragements, et pour avoir toujours cru en moi, même quand je n'y croyais plus trop. Grosse mention spéciale à Mei-Li qui en plus d'être ma star a relu mon manuscrit. Également à tous mes proches qui m'ont soutenu lors de ma soutenance et sont restés jusqu'au bout alors qu'ils ne sont pas du domaine. Petit clin d'œil à Peyrou de m'avoir tout de suite accueillie et m'avoir fait sentir chez moi durant chacun de mes déplacements. Je suis rentrée à Toulouse avec une douce chaleur envahissant mon cœur ! Petite dédicace à ces 2 mois à Mahebourg pendant la rédaction, le manuscrit aura survécu à la chaleur et les tempêtes tropicales. Évidemment, je n'oublie pas ma Toulouse qui m'a vue grandir avec les bro' de Toulouse (pas que Vincent) qui m'ont épaulé et notamment quand ça n'allait pas. Même loin d'eux, je pense à ma FieroTeam : on a évolué ensemble et j'ai hâte de vous retrouver prochainement.

Merci à chacun d'entre vous que j'ai côtoyé, vous avez tous contribué de près ou de loin à l'accomplissement de cette thèse et personnel.

Ringraziamenti

Innanzitutto, un enorme ringraziamento ai membri della commissione per aver dedicato il loro tempo a leggere e valutare il mio lavoro. La vostra presenza e il vostro feedback hanno significato molto per me. I vostri commenti e suggerimenti hanno davvero arricchito questa tesi e vi sono molto grato per l'attenzione che mi avete dedicato.

Ringrazio il mio relatore di tesi, Mathieu Lafourcade, per avermi guidato e accompagnato in questa avventura. Vorrei anche ringraziare il mio co-supervisore, Mokhtar BILLAMI, i cui consigli, la competenza e l'incoraggiamento sono stati essenziali nella costruzione di questo lavoro. Grazie per la vostra pazienza, i vostri consigli e soprattutto il vostro sostegno, che mi sono stati preziosi in ogni fase. Grazie per avermi illuminato con le vostre idee e il vostro entusiasmo.

Desidero inoltre ringraziare Berger-Levrault per avermi permesso di svolgere questa tesi. Sono molto grata di aver potuto contare sul sostegno di tutto il team e di aver avuto accesso a tutte le risorse necessarie per portare a termine questo progetto. Vorrei anche fare una menzione speciale al team dei bébous a BL : avete reso le giornate di lavoro molto più piacevoli e spesso anche indimenticabili ! Grazie per il vostro buon umore, il vostro sostegno e le risate che avete condiviso. Lavorare con voi è stata una vera fonte di motivazione e di Workaholism.

Penso anche ai membri del LIRMM con cui ho avuto la fortuna di lavorare. Grazie per la vostra calorosa accoglienza, per le discussioni sempre stimolanti e per i momenti condivisi che hanno reso questi anni ancora più piacevoli. Un grande ringraziamento va anche alla mia squadra speciale al LIRMM : rICARd.TXT ! Il vostro sostegno e la vostra amicizia sono stati essenziali per tutta la durata di questa tesi. Grazie per le idee condivise, gli scambi stimolanti e i bei momenti trascorsi insieme. Non dimenticherò mai le nostre pause caffè, i ristoranti e gli aperitivi del LIRMM ! Avete reso l'esperienza del laboratorio unica e più ricca.

E infine, un sentito ringraziamento alla mia famiglia e ai miei amici. Siete stati presenti in ogni momento del percorso, nei momenti di dubbio e nelle piccole vittorie. Grazie per la vostra pazienza, il vostro incoraggiamento e per aver sempre creduto in me, anche quando non credevo davvero in me stessa. Una menzione speciale a Mei-Li che, oltre a essere la mia Star, ha corretto il mio manoscritto. Vorrei anche ringraziare tutti i miei amici e la mia famiglia che mi hanno sostenuto durante la difesa e sono rimasti con me fino alla fine, anche se non sono nel campo. Vorrei anche ringraziare Peyrou per avermi accolto subito e per avermi fatto sentire a casa ovunque andassi. Sono tornata a Tolosa con un dolce calore che mi ha invaso il cuore ! Una piccola dedica ai due mesi trascorsi a Mahebourg durante la stesura del libro, il cui manoscritto è sopravvissuto al caldo e alle tempeste tropicali. Naturalmente, non dimentico la mia città natale, Tolosa, dove sono cresciuta e i fratelli di Tolosa (non solo Vincent) che mi hanno sostenuto, soprattutto quando le cose non andavano bene. Anche lontano da loro, penso al mio FieroTeam : siamo cresciuti insieme e non vedo l'ora di rivedervi presto.

Grazie a ognuno di voi con cui ho fatto amicizia e/o ho collaborato, tutti avete contribuito in un modo o nell'altro alla realizzazione di questa tesi e alla mia vita personale.

Abstract

This thesis examines the construction of knowledge graphs from texts, focusing primarily on information extraction from unstructured texts. The objective of this work is to explore various aspects of information extraction from specialized corpora in the legal domain. To this end, we divide this study into two subtasks : terminology extraction and relation extraction.

Terminology extraction aims to automatically identify relevant terms in a given corpus of texts. Subsequently, from these terms, we extract the relations that link them. For relation extraction, two approaches are conceivable : either by determining in advance the types of relations that structure the terms, or by using the context, notably the verbs or other actions that link these terms (in the field of OpenIE). Thus, we address three main sub-problems.

We introduce the terminology extraction system InfoGlean KeyTerms, composed of three modules : one for named entity recognition (NER), one for the extraction of relevant terms/text segments (KPE), and a final one for the extraction of legal entities. Expert annotations were provided in addition to this system to constitute a terminological base.

After building this terminological base, we implemented two relation extraction systems : Relational Embeddings Model (REM) and *GPT Open Relation EXtraction* (GOREX). REM identifies typed relations between the extracted terms using the lexical network rezo-JDM. REM represents the relation pairs using a Word2Vec model, then classifies the types of relations. GOREX, on the other hand, exploits the principle of OpenIE by focusing on the verbs or action terms in the local context of the terms. GOREX uses LLMs to perform this task.

The analysis of the results revealed promising research avenues to be explored in future work for all systems. More specifically, the implementation of a hybrid relation extraction system could be an interesting path to explore.

Keywords

•AI•Natural Language Processing•NLP•Knowledge Representation•Automatic Construction of Lexico-Semantic Knowledge Graphs from Texts•Knowledge Graph•Information Extraction•Terminology Extraction•Named Entity Recognition (NER)•KeyPhrase Extraction (KPE)•Legal Entities•Relation Extraction•Word Embeddings•Open Information Extraction (OpenIE)•LLM •Domain-Specific Texts•Legal Corpus•Unstructured Texts

Résumé

Cette thèse examine la construction de graphes de connaissances à partir de textes, en se concentrant principalement sur l'extraction d'informations à partir de textes non structurés. L'objectif de ce travail est d'explorer différents aspects de l'extraction d'informations à partir de corpus spécialisés dans le domaine juridique. Pour ce faire, nous divisons cette étude en deux sous-tâches : l'extraction terminologique et l'extraction de relations. L'extraction terminologique vise à identifier automatiquement les termes pertinents dans un corpus de textes donné. Ensuite, à partir de ces termes, nous extrayons les relations qui les lient. Pour l'extraction de relations, deux approches sont envisageables : soit en déterminant au préalable des types de relations qui structurent les termes, soit en utilisant le contexte, notamment les verbes ou autres actions qui relient ces termes (dans le domaine de l'OpenIE).

Nous abordons donc trois sous-problématiques principales. Nous introduisons le système d'extraction terminologique InfoGlean KeyTerms, composé de trois modules : un pour la reconnaissance d'entités nommées (NER), un pour l'extraction de termes/segments textuels pertinents (KPE) et un dernier pour l'extraction d'entités juridiques. Des annotations d'experts ont été fournies en complément à ce système pour constituer une base terminologique. Après la construction de cette base terminologique, nous avons mis en place deux systèmes d'extraction de relations : *Relational Embeddings Model* (REM) et *GPT Open Relation EXtraction* (GOREX). REM identifie des relations typées entre les termes extraits en utilisant le réseau lexical rezoJDM. REM représente les paires de relations à l'aide d'un modèle Word2Vec, puis classe les types de relations. GOREX , quant à lui, exploite le principe de l'OpenIE en se concentrant sur les verbes ou termes d'action dans le contexte local des termes. GOREX utilise les LLMs pour réaliser cette tâche. L'analyse des résultats a révélé des pistes de recherche prometteuses à approfondir dans les travaux futurs pour l'ensemble des systèmes. Plus spécifiquement, la mise en place d'un système hybride d'extraction de relations pourrait être un chemin intéressant à explorer.

Mots-clefs

•IA•TALN•Représentation des connaissances•Construction automatique de graphes de connaissances lexico-sémantiques à partir de textes•Graphes de connaissances•Extraction d'informations•Extraction terminologique•Reconnaissance d'entités nommées (NER)•Extraction de termes pertinents (KPE)•Entités juridiques•Extraction de relations•Plongements lexicaux•Open Information Extraction (OpenIE)•LLM•Textes de spécialité•Corpus juridique•Textes non structurés

Résumé étendu

Cette thèse, intitulée « Méthodes et Modèles pour l'élaboration automatisée de Graphes de Connaissances dans le domaine juridique : Application aux Ressources Juridiques et Juridico-Pratiques des Collectivités Locales et Territoriales », explore la construction automatique de bases de connaissances à partir de ressources textuelles, en se concentrant sur les données de Berger-Levrault, éditeur de contenus juridiques et pratiques. Le corpus utilisé provient de la base éditoriale de Berger-Levrault, qui contient des centaines d'ouvrages et une dizaine de milliers d'articles juridiques et pratiques couvrant divers domaines du secteur public. La nature évolutive de ce corpus, enrichi constamment de nouveaux articles et domaines, complique encore cette tâche. En effet, la réglementation d'expertise juridique et pratique évolue quasi-quotidiennement au fil des décisions politiques, de l'évolution de la société, et des publications au journal officiel de la république française. Les ouvrages et articles évoluent donc en permanence. La représentation des connaissances qui en découle doit donc évoluer en fonction. Une base de connaissances construite et figée s'avèrerait très rapidement obsolète. Pour cette raison, il est essentiel dans cette thèse de travailler à la méthode permettant de faire évoluer les bases de connaissances qui seront produites : c'est-à-dire, permettre l'évolution des graphes de connaissances produits avec un apprentissage dynamique.

Les connaissances pouvant être extraites de la base éditoriale française de Berger-Levrault sont très utiles pour l'amélioration d'applications, en passant de l'extraction d'information et de la recherche de documents vers l'enrichissement de connaissances d'agents conversationnels afin de satisfaire au mieux les besoins des utilisateurs. La structuration de la connaissance est donc essentielle pour faciliter l'accès et l'utilisation de ces informations dans diverses applications, telles que les agents conversationnels et les systèmes de recherche documentaire. La problématique principale de cette thèse réside dans la difficulté de structurer des données textuelles brutes en informations exploitables de manière automatique. L'objectif principal de la thèse est de créer automatiquement un graphe de connaissances lexico-sémantique à partir de texte, similaire à celui créé dans le cadre du projet JeuxDeMots. Ce graphe permet de structurer les informations textuelles. Les informations extraites pour alimenter le graphe proviennent de diverses méthodes comme la classification de documents, l'extraction d'informations et de relations, et l'extraction terminologique.

Les graphes de connaissances lexico-sémantiques sont des structures permettant de représenter et de manipuler des connaissances issues de textes. L'intérêt d'employer un graphe réside dans le fait de pouvoir appliquer des algorithmes de graphe dessus, ainsi que de représenter l'incertitude et l'ambiguïté de façon pratique. La construction de ces graphes re-

pose sur des techniques d'extraction de termes et de relations, en utilisant des ressources linguistiques existantes pour enrichir le contenu du graphe. La méthodologie adoptée débute donc par des processus et méthodes d'extraction terminologique, notre système *InfoGlean KeyTerms* pour l'extraction automatique de termes, et une évaluation rigoureuse des sous-systèmes d'extraction terminologique. Notre système *InfoGlean KeyTerms* exploite diverses méthodes d'extraction terminologique et se divise en plusieurs sous-programmes : un module pour la reconnaissance d'entités nommées (NER), un module pour l'extraction de termes pertinents (KPE) et enfin un système d'extraction d'entités juridiques. Les résultats de cette évaluation montrent une amélioration notable des performances grâce à l'intégration de connaissances sémantiques.

Après avoir construit une base terminologique solide, cette thèse propose un algorithme d'extraction par classification de relations et un modèle d'extraction de relations ouvertes utilisant des techniques de clustering et des modèles de langage de grande taille (LLM). Notre première méthode intitulée *Relational Embeddings Model* (REM), permet d'identifier des relations typées entre les termes extraits en utilisant le réseau lexical rezoJDM. Les principaux défis abordés incluent la gestion des listes de relations prédéfinies et l'amélioration de la spécificité et de la pertinence des relations extraites en filtrant les domaines des relations sans prendre en considération le contexte local. Afin de pallier le problème de manque de spécificité des relations extraites par le système REM, des techniques d'extraction de relations ouvertes, exploitant les capacités des LLMs, ont été proposées. Contrairement aux méthodes traditionnelles basées sur une liste prédéterminée de types de relations, notre approche, intitulée *GPT Open Relation EXtraction* (GOREX), offre une plus grande flexibilité en focalisant l'extraction sur les verbes d'action spécifiques aux textes, offrant ainsi une extraction plus précise et adaptée aux besoins spécifiques des corpus étudiés. En effet, nous nous intéressons à pouvoir retrouver le sens du texte cible avec un (sous-) graphe qui puisse être représentatif du texte donné. En l'occurrence, les corpus étudiés étant de spécialité du domaine public, les relations de types génériques limitent la compréhension d'un texte donné. GOREX est une méthode du domaine de l'OpenIE. L'évaluation de ces méthodes révèle des critères essentiels pour la résolution des cas complexes. En effet, un cas comme (Tribunal administratif, statuer, sous trois jours) n'aurait pu être extrait à l'aide de type prédéfinis et génériques. L'évaluation offre également des perspectives pour les travaux futurs dans l'amélioration des systèmes d'extraction et de classification de textes spécialisés.

Acronymes

RELATIVE *RELations as LATent dIscourse VEctors*. 124

API *Application Programming Interface*. xiii–xv, 83, 84, 86–89, 91, 94, 95, 102–106, 139–141, 143, 146, 153

BALD *Bayesian Active Learning by Disagreement*. 33

BERT *Bidirectional Encoder Representations from Transformers*. 34, 37–40, 43, 44, 50, 52–54, 57, 63, 66, 69, 83, 127, 128, 130

BiLSTM Bidirectional *Long Short Term Memory*. 33, 35, 36

BoW *Bag of Words*. vii, 26, 128, 130

CBOW Continuous *Bag of Words*. 26, 120

CNN Réseaux Neuronaux Convolutifs (*Convolutional Neural Networks*). 35, 36, 57

CRE *Continual Relation Extraction*. 54

CRF Champs Aléatoires Conditionnels (*Conditional Random Fields*). 28, 30, 31, 33, 35, 36, 53, 61

DBN *Deep Belief Network*. 67

DCNN Réseaux Neuronaux Convolutionnels Dilatés (*Dilated Convolutional Neural Networks*). 35

DEFT DÉfi Fouille de Textes. 66

DNN Réseaux Neuronaux Profonds - *Deep Neural Network*. 33

DT *Decision Tree*. 123, 124, 132

EE Extraction d'Évènement. 25

EXACT *Entity eXtraction And Classification Tools*. 82

EXTERMO *EXtraction of TERM using OpenIE*. 83

FCL Couche entièrement connectée (*Full Connected Layer*). 36

GCN réseaux de neurones convolutionnels (*graph convolutional networks*). 24

GCRA Réseau neuronal récurrent convolutionnel gated avec attention (*Gated Convolutional Recurrent Neural Network with Attention*). 36

GOREX *GPT Open Relation EXtraction*. i, ii, iv

GRN *Graph Recurrent Network*. 53

GRU *Gated Recurrent Units*. 35

GWAP Jeux avec buts (*Games With A Purpose*). 10–12

HMM Modèle de Markov Caché (*Hidden Markov Model)*. 28, 33

IDF *Inverse Document Frequency*. ix, 26, 42, 67

IE *Information Extraction*. 25, 39, 44, 51, 53, 61, 63–65, 69

IGL Étiquetage itératif de la grille (*Iterative Grid Labeling*). 63

JDM JeuxDeMots. 10

JUREX *Juridicial entity Regular Expression eXtractor*. 84

JURID'OCC *JURIdical DOCument Classification*. xi, 117, 125, 128

k-NN k plus proches voisins (*K-Nearest Neighbors*). 39, 123, 124, 131, 132

KB Base de Connaissances (*Knowledge Base*). 64

KDE Estimation par noyau (*Kernel Density Estimation*). 29

KG Graphe de connaissance (*Knowledge Graph*). 23, 24

KPE Extraction de termes ou phrases clés (*Keyphrases extraction*). i, ii, iv, xvii, 44, 69, 91–94, 102, 104, 105, 110, 162

LC Least Confidence. 33

LLM Modèle de langue à large échelle (*Large Language Model*). i, ii, iv, xi, 2, 6, 7, 23, 24, 34, 37–40, 44, 45, 48, 50, 51, 53–56, 59, 63, 64, 69–71, 83, 84, 87, 93, 96, 97, 105, 106, 111, 113, 137–142, 144–148, 150, 152–156, 158–162, 164, 165

LM Modèle de langue (*Language Model*). 56

LRA *Latent Relational Analysis*. 56

LSTM *Long Short Term Memory*. vii, 33, 35, 36, 57

MAML modèle de méta-apprentissage agnostique (*model-agnostic meta-learning*). 38

MeSH *Medical Subject Headings*. 66, 67

MLM *Masked Language Model*. 37

MLN *Markov logic networks*. 62

MNB Multinomial Naïve Bayes. 67

MNLP *Maximum Normalized Log-Probability*. 33

MR Mention Replacement. 32

MRC *Machine Reading Comprehension*. 37

MRDC Critère de discrimination relative multivariée - (*Multivariate Relative Discrimination Criterion*). 67

MUC *Message Understanding Conference*. xv, 83

MWU Multi-mots (*multi-word units*). 41

NER Reconnaissance des entités nommées (*Named Entity Recognition*). i, ii, iv, xiii, xv, xvii, 25–40, 44–46, 48, 50, 51, 54, 58, 59, 61, 69, 73, 82–84, 86–97, 110, 114, 115, 162

NoTA None-of-The-Above. 54, 55

OOV *Out of Vocabulary*. 36

OpenIE *Open Information Extraction*. i, ii, iv, 7, 24, 40, 60–65, 70, 71, 137, 138, 148, 152, 162

OpenRE *Open Relation Extraction*. 70

PMR *Pre-trained Machine Reader*. 37

POS Classe ou catégorie grammaticale (*Part-Of-Speech (Tagging)*. 12, 13, 16, 56, 61

QA *Question Answering*. 51, 64

RE Extraction de relations (*Relation Extraction*). 25, 56

REGEX Expression régulière (*Regular expression*). 27, 50, 84, 108–111

REM *Relational Embeddings Model* : Notre modèle d'extraction de relations. i, ii, iv, xi, 7, 117, 118, 124, 127, 132

rezoJDM Réseau Lexico-Sémantique construit à partir du GWAP JeuxDeMots. i, ii, iv, xiii, xvii, 7, 10–14, 17, 18, 119–122, 137, 162

RF Forêt aléatoire (*Random Forest*). 124, 131, 132

RL Reinforcement Learning. 31

RLS Réseau Lexico-Sémantique. 10, 11

RMDL *Random Multimodel Deep Learning*. 67

RNN Réseau Neuronal Récurrent (*Recurrent Neural Network*). 35, 36, 57

Seq2Seq *Sequence to Sequence*. 36

SiS *Shuffle within Segments*. 32

SLM *Small Language Model*. 34, 37–39, 50, 54–56, 83, 113

SoTA *State-of-The-Art*. 36, 38, 55

SR Synonym Replacement. 32

SRL Étiquetage des rôles sémantiques (*Semantic role labeling*). 62, 64

SVM Machines à Vecteurs de Support (*support-vector machine*). 28, 35, 53, 66, 120, 123, 124, 131, 132

TALN Traitement Automatique du Langage Naturel. xxiv, 1, 3–5, 23, 29, 36, 38, 44, 51, 54, 162

TF-IDF *Term Frequency - Inverse Document Frequency*. 25, 26, 42–44, 128, 130, 212

TREC *Text REtrieval Conference*. 66, 67

Table des matières

Avant-Propos xxix

1 Introduction **1**
1.1 Introduction générale 1
1.2 Motivations et Contexte de travail 2
1.2.1 Présentation des données 2
1.2.2 Structurer la connaissance 3
1.2.3 Problématique générale 3
1.3 Axe de recherche 5
1.4 Plan 6

2 Graphe de connaissances lexico-sémantiques **9**
2.1 Définition générale 9
2.2 Ressource exploitée : le réseau lexico-sémantique issu de JeuxDeMots 10
2.2.1 Présentation générale du graphe de connaissances lexico-sémantique rezoJDM 11
2.2.2 Nœuds 12
2.2.3 Relations 13
2.3 Conclusion 17

3 État de l'art : Construction automatique de bases de connaissances à partir de ressources textuelles **19**
3.1 Définitions et différentes structures de graphe 19
3.2 Construction de graphes à partir de textes 22
3.3 Extraction d'information 25
3.3.1 Extraction terminologique 25
3.3.2 Extraction de relations 52
3.3.3 Open Information Extraction 60
3.4 Classification de documents 65
3.5 Contributions : Résumé des choix d'orientation et d'implémentation 68

4 Extraction terminologique **73**
4.1 Processus et Méthodes d'annotations des Experts 76
4.2 Système InfoGlean KeyTerms pour l'Extraction Automatique de Termes . . . 82
4.3 Méthodologie et Résultats de l'évaluation des sous-systèmes d'extraction terminologique 85
4.3.1 Structuration et Critères de l'évaluation des sous-systèmes 85
4.3.2 Présentation et Analyse des Résultats Obtenus par les Sous-Systèmes d'Extraction 94
4.3.3 Discussion critique sur les Résultats et Identification des Limites . . . 112
4.4 Synthèse des Contributions et Perspectives pour les Travaux Futurs 115

4.4.1 Bilan des Contributions Principales en Extraction Terminologique . . . 115
4.4.2 Perspectives d'Amélioration et Suggestions pour les Recherches Futures 115

5 Extraction de relations à l'aide du réseaux lexical JeuxDeMots 117
5.1 Algorithme d'extraction par classification de relations . . . 117
5.1.1 *Relational Embeddings Model* (REM) : Description du système . . . 118
5.1.2 Évaluation . . . 121
5.2 Classification de domaine . . . 123
5.2.1 Analyse des données pour la classification de domaine . . . 125
5.2.2 *JURIdical DOCument Classification* (JURID'OCC) : Description du système . . . 125
5.2.3 Évaluation . . . 129
5.3 Discussion . . . 133
5.3.1 Évaluation . . . 133
5.3.2 Relations dans le graphe de connaissances lexico-sémantique . . . 134
5.3.3 Relations en non-contextuel . . . 134

6 Extractions de relations ouvertes à l'aide des LLMs 137
6.1 Extraction de relations ouvertes par LLM . . . 139
6.2 Clustering d'instances similaires et découverte de relations . . . 142
6.3 Évaluation . . . 145
6.3.1 Extraction de relations ouvertes par LLM . . . 146
6.3.2 Clustering de triplets de relations . . . 153
6.3.3 Bilan . . . 158
6.4 Discussion . . . 159
6.4.1 Bilan . . . 159
6.4.2 Discussion . . . 159

7 Conclusion et perspectives 161
7.1 Bilan général . . . 161
7.2 Perspectives . . . 163
7.2.1 Extraction terminologique . . . 163
7.2.2 Extraction de relations à l'aide du réseau lexical issu de JeuxDeMots . 163
7.2.3 Extractions de relations ouvertes à l'aide des LLMs . . . 164
7.2.4 Graphes de connaissances lexico-sémantiques . . . 165

A Annexes 203
Données . . . 203
Résultats classification de domaine . . . 205

Liste des exemples

1.1 Exemple d'un paragraphe issu d'un texte de Berger-Levrault 3
2.1 Liste des différents raffinements sémantiques et morphologiques pour le mot « avocat » 13
2.2 Exemple d'imbrication de raffinements sémantiques pour le mot avocat 14
2.3 Exemples de nœuds d'information sémantique dans rezoJDM 14
2.4 Représentation de la relation « *le manchot ne vole pas* ». 16
2.5 Représentation de la relation « *Le scorpion de mer n'est pas un scorpion* ». 16
2.6 Liste de différents exemples de relations associées aux terme « code civil » tirés de rezoJDM 17
3.1 Exemples d'extraction d'information d'un texte issu des rapports du fil d'actualité sur les fusions d'entreprises. 25
4.1 Exemple d'un paragraphe avec les candidats puis les filtrages associés 75
4.2 Exemple d'un paragraphe avec les annotations qui lui sont associées 76
4.3 Exemple d'un paragraphe avec des termes extraits du texte confrontés aux annotations des experts 78
4.4 Exemple d'un paragraphe avec les trois différentes extractions de InfoGlean KeyTerms superposées aux annotations des experts 82
4.5 Exemple de *prompt* pour l'extraction des termes pertinents et des NER. 84
4.6 Exemple de *prompt* pour l'extraction de termes pertinents. 85
4.7 Liste de différents exemples pour chacun des types d'entités juridiques 86
4.8 Prompt utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'obtenir les termes pertinents d'un texte donné. 88
4.9 Exemple de prompt utilisé dans l'API de ChatGPT afin d'obtenir les termes pertinents d'un paragraphe issu du corpus de Berger-Levrault. 89
6.1 Prompt générique utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'obtenir triplets de relations à partir des termes pertinents d'un texte 140
6.2 Exemple de prompt utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'obtenir les triplets de relations à partir des termes pertinents d'un paragraphe issu du corpus de Berger-Levrault 141
6.3 Prompt générique utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin de regrouper les triplets de relations en cluster à partir d'un ensemble de triplets lié à un domaine 143
6.4 Prompt générique utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'évaluer les triplets de relations extraits 146

6.5 Prompt générique utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'évaluer les clusters des triplets de relations d'un domaine donné 153

Liste des figures

1.1 Exemple d'un graphe de connaissances lexico-sémantique obtenu à partir du texte de l'exemple 1.1. 4
1.2 Système général de construction de graphe de connaissances lexico-sémantique 6

3.1 Millefeuille d'apprentissage 23
3.2 Schéma de conceptualisation de l'état de l'art : Du corpus jusqu'au graphe lexico-sémantique. 24
3.3 Exemple d'une extraction d'information dans le cadre de l'OpenIE sur la phrase « Un rapport a été dressé par la Banque de France au sujet du surendettement ». 61
3.4 Schéma des différents blocs d'orientations suite à l'état de l'art. 68

4.1 Texte étiqueté par un système de reconnaissance d'entités nommées utilisé lors de la campagne d'évaluation MUC 83
4.2 Sortie de Gemini pour la tâche de NER 88
4.3 Pourcentage multimots parmi l'ensemble des termes pertinents extraits : l'ensemble ainsi que la liste unique 98
4.4 Nombre moyen de mots contenus dans les termes NER extraits 99
4.5 Nombre moyen de mots contenus dans les termes extraits 100
4.6 Nombre moyen de termes par texte 101
4.7 Score moyen attribué par les LLMs pour chacun des domaines pour le NER . 102
4.8 Score de confiance moyen attribué par les LLMs pour chacun des domaines pour le NER 103
4.9 Statistiques sur les termes pertinents extraits par l'API d'OpenAI et Gemini . 105
4.10 Statistiques sur les termes pertinents extraits par l'API d'OpenAI et Gemini . 106
4.11 Score de pertinence moyen attribué par les LLMs pour chacun des domaines pour le KPE 107
4.12 Score de confiance moyen attribué par les LLMs pour chacun des domaines pour le KPE 108
4.13 Statistiques sur les entités juridiques extraites par REGEX 109
4.14 Statistiques sur les entités juridiques extraites par REGEX 110
4.15 Score moyen attribué par les LLMs pour chacun des domaines pour le système des entités juridiques 111
4.16 Score de confiance moyen attribué par les LLMs pour chacun des domaines pour le système d'extraction d'entités juridiques 112

5.1 Schéma (général) du système d'extraction de relations à l'aide du réseau lexical JeuxDeMots . . . 119
5.2 Matrice de corrélation de *Pearson* entre les différents domaines pour les annotations unifiées des experts . . . 126
5.3 Schéma (général) du système d'extraction de relations en utilisant la classification de documents . . . 127
5.4 Schéma de l'architecture du système de classification de documents . . . 128
5.5 Répartition des documents en fonction des différents domaines dans le jeu de données complet . . . 129

A.1 Matrice de confusion - TF-IDF – CLASSIFICATION MULTINOMIALNB – (ACC = 0,88) . . . 208
A.2 Matrice de confusion - TF-IDF – CLASSIFICATION LinearSVC – (ACC = 0,88) 210
A.3 Matrice de confusion - TF-IDF – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,92) . . . 212
A.4 Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - TF-IDF – CLASSIFICATION LINEARSVC (ACC = 0.94) . . . 214
A.5 Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - TF-IDF – CLASSIFICATION SGD (ACC = 0.94) . . . 216
A.6 Matrice de confusion - ClassicBOW – CLASSIFICATION MultinomialNB (ACC = 0.94) . . . 218
A.7 Matrice de confusion - ClassicBOW – CLASSIFICATION LINEARSVC (ACC = 0.94) . . . 219
A.8 Matrice de confusion - ClassicBOW – CLASSIFICATION SGDClassifier (ACC = 0.94) . . . 221
A.9 Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - ClassicBOW – CLASSIFICATION LINEARSVC (ACC = 0.95) . . . 222
A.10 Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - ClassicBOW – CLASSIFICATION MultinomialNB (ACC = 0.90) . . . 224
A.11 Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - ClassicBOW – CLASSIFICATION SGDClassifier (ACC = 0.95) . . . 225
A.12 Matrice de confusion - Sentence CamemBERT – CLASSIFICATION LINEARSVC (ACC = 0.86) . . . 228
A.13 Matrice de confusion - Sentence CamemBERTsur un dataset sans prétraitement – CLASSIFICATION LINEARSVC (ACC = 0.94) . . . 230
A.14 Matrice de confusion - Sentence CamemBERT sur un Dataset sans pré-traitement – CLASSIFICATION SGDClassifier (ACC = 0.91) . . . 232
A.15 Matrice de confusion - Sentence CamemBERT – Représentation n°2 – CLASSIFICATION LINEARSVC (ACC = 0.91) . . . 234
A.16 Matrice de confusion - Sentence CamemBERT – Représentation n°2 – CLASSIFICATION SGDClassifier (ACC = 0.90) . . . 236
A.17 Matrice de confusion - Sentence FlauBERT – CLASSIFICATION LINEARSVC (ACC = 0.87) . . . 238
A.18 Matrice de confusion - Sentence FlauBERT – CLASSIFICATION SGDClassifier (ACC = 0.87) . . . 240
A.19 Matrice de confusion - WORD2VEC Représentation N°1 – VOCABULAIRE ENTRAINE SUR LE DATASET ENTIER EN GARDANT LES APOSTROPHES - CLASSIFICATION LINEARSVC - (ACC = 0,16) . . . 242

A.20 Matrice de confusion - WORD2VEC Représentation N°2 – VOCABULAIRE ENTRAINE SUR LE DATASET ENTIER EN GARDANT LES APOSTROPHES - CLASSIFICATION LINEARSVC - (ACC = 0,35) 245
A.21 Matrice de confusion - Doc2Vec – MostSimilar – (ACC = 0,89) 248
A.22 Matrice de confusion - Doc2Vec - Représentation en gardant les apostrophes – MostSimilar – (ACC = 0,87) 250

Liste des tableaux

2.1 Liste non exhaustive des différents types de nœuds présents dans rezoJDM . 12
2.2 Principaux types de relations avec leur nom et consignes associés (Liste non-exhaustive) 15

3.1 Liste des jeux de données regroupés dans des catégories et sous-catégories (domaine par exemple) avec leurs descriptions et langues associées. 49

4.1 Échantillon de calcul d'occurrences pour différents ouvrages 79
4.2 Échantillon de calcul d'occurrences pour différents articles et différents domaines 79
4.3 Calculs totaux et moyens des occurrences, nombre d'annotations et termes annotés 80
4.5 Nombre d'annotations pour un paragraphe donné (associé à son id de brique) 80
4.4 Annotations des experts par domaine 81
4.6 Nombre d'occurrences et nombre de paragraphes des annotations 81
4.7 Nombre moyen de mots par domaine 92
4.8 Différents prompts utilisés pour les différentes évaluations 94
4.9 Résultats des différents modèles sur des corpus standards de NER 94
4.10 Tableau récapitulatifs des différents calculs effectués sur les NER extraits avec les différents modèles donnés 97
4.11 Résultats de différents systèmes pour l'extraction de termes pertinents (KPE) 104
4.12 Dataset face aux prédictions réalisées 115

5.1 Tableau comparatif des performances des modèles lors de la classification de types de relations avec Relative puis notre système REM 124
5.2 Répartition des termes annotés des experts par des termes mono-domaines et des termes multi-domaines 125
5.3 Résultats de classification de documents 130
5.4 Tableau comparatif des performances des classificateurs / Résultat de l'extraction de relations à la fois lors de la classification de domaine et sans classification de domaine 132
5.5 Statistiques effectuées sur les termes des instances de relations extraites . . . 133
5.6 Statistiques réalisées sur les relations extraites (Echantillon pour les 4 types de relations : *syn, isa, hypo, has_part*) 135
5.7 Détail des instances de relations de même brique 136

6.1 Exemple de clustering réalisé à partir de quelques relations d'un fichier JSON du domaine de l'urbanisme 145
6.2 Exemple de scores de pertinence et de confiance moyen attribués par les LLMs pour un triplet donné en fonction du domaine d'urbanisme avec la justification associée 147
6.3 Scores de pertinence et de confiance moyen attribué aux différents triplets relationnels par les LLMs pour chacun des domaines 148
6.4 Statistiques effectuées sur les triplets extraits en fonction du domaine 150
6.5 Exemple de scores de pertinence et de confiance moyen attribués par un LLM pour un cluster de relations donné en fonction du domaine *comptabilité et des finances locales* avec la justification associée 154
6.6 Scores de pertinence et de confiance moyen attribué aux différents clusters par les LLMs pour chacun des domaines 155
6.7 Statistiques sur les différents clusters pour chacun des domaines 157

A.1 Tableau des 8 domaines avec un exemple réel de texte. Exemples tirés des données disponibles sur legibase.fr 205
A.2 Résultats de Classification Zero-Shot Learning en utilisant le modèle « camembert-base-xnli » de Hugging Face 206
A.3 Résultats de Classification Zero-Shot Learning en utilisant le modèle « distilcamembert-base-nli » de Hugging Face 206
A.4 Résultats de Classification de domaine TF-IDF – CLASSIFICATION MULTINOMIALNB – (ACC = 0,88) 207
A.5 Résultats de Classification de domaine TF-IDF – CLASSIFICATION LINEARSVC – RESULTATS (ACC = 0,93) 209
A.6 Résultats de Classification de domaine TF-IDF – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,92) 211
A.7 Résultats de Classification de domaine PRETRAITEMENT EN GARDANT LES APOSTROPHES - TF-IDF – CLASSIFICATION LINEARSVC (ACC = 0.94) 213
A.8 Résultats de Classification de domaine PRETRAITEMENT EN GARDANT LES APOSTROPHES - TF-IDF – CLASSIFICATION SGD (ACC = 0.94) 215
A.9 Résultats de Classification de domaine CLASSICBOW– CLASSIFICATION MULTINOMIALNB (ACC = 0,94) 217
A.10 Résultats de Classification de domaine CLASSICBOW– CLASSIFICATION LINEARSVC – (ACC = 0,94) 219
A.11 Résultats de Classification de domaine CLASSICBOW– CLASSIFICATION SGDCLASSIFIER – (ACC = 0,94) 220
A.12 Résultats de Classification de domaine CLASSICBOW - Représentation en gardant les apostrophes – CLASSIFICATION LinearSVC – (ACC = 0,95) . . . 222
A.13 Résultats de Classification de domaine CLASSICBOW - Représentation en gardant les apostrophes – CLASSIFICATION MultinomialNB – (ACC = 0,90) 223
A.14 Résultats de Classification de domaine CLASSICBOW - Représentation en gardant les apostrophes – CLASSIFICATION SGDClassifier – (ACC = 0,95) . 225
A.15 Résultats de Classification de domaine SENTENCE CAMEMBERT BASE – CLASSIFICATION LINEARSVC – (ACC = 0,86) 227
A.16 Résultats de Classification de domaine FLAUBERT-BASE-UNCASED-XNLI-STS – CLASSIFICATION LINEARSVC – (ACC = 0,87) 229

A.17 Résultats de Classification de domaine sur les textes SANS PRETRAITEMENT : SENTENCE CAMEMBERT BASE – CLASSIFICATION LINEARSVC – (ACC = 0,92) 229
A.18 Résultats de Classification de domaine sur les textes SANS PRETRAITEMENT : SENTENCE CAMEMBERT BASE – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,91) 231
A.19 Résultats de Classification de domaine SENTENCE CAMEMBERT BASE – Représentation n°2 – CLASSIFICATION LINEARSVC – (ACC = 0,91) 233
A.20 Résultats de Classification de domaine SENTENCE CAMEMBERT BASE – Représentation n°2 – CLASSIFICATION SGDClassifier – (ACC = 0,90) 235
A.21 Résultats de Classification de domaine FLAUBERT-BASE-UNCASED-XNLI-STS – Représentation n°2 – CLASSIFICATION LINEARSVC – (ACC = 0,87) . 237
A.22 Résultats de Classification de domaine FLAUBERT-BASE-UNCASED-XNLI-STS – Représentation n°2 – CLASSIFICATION SGDClassifier – (ACC = 0,87) . 239
A.23 Résultats de Classification de domaine WORD2VEC Représentation N°1 – VOCABULAIRE ENTRAINE SUR LE DATASET ENTIER - CLASSIFICATION LINEARSVC - (ACC = 0,19) 241
A.24 Résultats de Classification de domaine WORD2VEC Représentation N°1 – VOCABULAIRE ENTRAINE SUR LE DATASET ENTIER EN GARDANT LES APOSTROPHES - CLASSIFICATION LINEARSVC - (ACC = 0,16) 241
A.25 Résultats de Classification de domaine WORD2VEC MOYENNE DES VECTEURS DE MOTS (REP N°2) – TRIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,16) 243
A.26 Résultats de Classification de domaine WORD2VEC MOYENNE DES VECTEURS DE MOTS (REP N°2) – TRIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION SGDCLASSIFIER – En gardant les apostrophes (ACC = 0,35) 243
A.27 Résultats de Classification de domaine WORD2VEC MOYENNE DES VECTEURS DE MOTS (REP N°2) – TRIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION LINEARSVC – (ACC = 0,2) 243
A.28 Résultats de Classification de domaine WORD2VEC MOYENNE DES VECTEURS DE MOTS (REP N°2) – TRIGRAMS – VOCABULAIRE DATASET ENTIER EN GARDANT LES APOSTROPHES – CLASSIFICATION LINEARSVC – (ACC = 0,35) 244
A.29 Résultats de Classification de domaine WORD2VEC Représentation N°3 : MOYENNE DES VECTEURS DE MOTS – BIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,19) 246
A.30 Résultats de Classification de domaine WORD2VEC Représentation N°3 : MOYENNE DES VECTEURS DE MOTS – BIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION LinearSVC – (ACC = 0,17) 246
A.31 Résultats de Classification de domaine WORD2VEC Représentation N°2 : MOYENNE DES VECTEURS DE MOTS – TRIGRAMS – VOCABULAIRE DATASET ENTIER – COSINUS SIMILARITY – (ACC = 0,19) 246
A.32 Résultats de Classification de domaine Doc2VEC sans les apostrophes – Most Similar – (ACC = 0,89) 248
A.33 Résultats de Classification de domaine Doc2VEC en gardant les apostrophes – Most Similar – (ACC = 0,87) 249

Liste des définitions

2.1 Graphe de connaissances lexico-sémantique 9
2.2 Nœud 12
2.3 Relation 13
3.1 Extraction d'information 25
4.1 Extraction terminologique 73

Avant-propos

Contexte industriel

Berger-Levrault est un éditeur de logiciels métiers de renommée internationale, initialement spécialisé dans l'édition et l'impression, avec une expertise particulière dans les métiers des services publics. Aujourd'hui, l'entreprise se concentre sur le développement de logiciels destinés aux administrations publiques locales, notamment dans les secteurs de l'éducation, de la santé, du sanitaire, du social et de la gestion des territoires. La Direction de la Recherche et de l'Innovation Technologique (DRIT) de Berger-Levrault réunit chercheurs, doctorants, étudiants et ingénieurs pour collaborer à des projets de recherche et d'innovation. Ces efforts collectifs visent à explorer divers domaines tels que les territoires, le médico-social, l'industrie, l'éducation et d'autres métiers spécifiques de Berger-Levrault.

Dans le cadre de ses activités, la DRIT se focalise sur l'exploitation des collections d'ouvrages et d'articles disponibles dans la base éditoriale française de Berger-Levrault. Ces collections sont thématiques, partiellement annotées, et résultent d'une collaboration étroite entre l'entreprise et de nombreux experts. En intégrant ces ressources éditoriales riches et variées, Berger-Levrault s'engage à renforcer ses capacités en matière de recherche et d'innovation, notamment dans le domaine de la construction (semi)-automatique d'ontologies spécifiques aux métiers du secteur public. Cette approche reflète non seulement l'engagement continu de l'entreprise envers l'amélioration de ses produits et services, mais aussi sa volonté de demeurer à la pointe de l'innovation technologique au service des administrations publiques locales.

Contexte scientifique

Le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), une unité mixte de recherche associant l'Université Montpellier et le CNRS (UMR 5506), se distingue par ses activités de recherche variées et ses collaborations internationales. Avec une équipe de plus de 400 chercheurs et étudiants, le LIRMM s'investit dans plusieurs domaines clés de la science informatique. Les axes de recherche du LIRMM englobent la conception et la vérification de systèmes intégrés, mobiles et communicants, ainsi que la modélisation de systèmes complexes à base d'agents. L'unité se spécialise également dans l'algorithmique, la bio-informatique, les interactions homme-machine, la robotique, le traitement du langage naturel et l'intelligence artificielle. Ces domaines sont explorés à travers

des approches novatrices telles que les méthodes par graphes, les approches logiques et statistiques.

Au sein du LIRMM, l'équipe de recherche *Exploration et exploitation de données textuelles* (TEXTE[1]), fondée en 1998 par Mathieu Lafourcade en collaboration avec Jacques Chauché, se concentre spécifiquement sur le Traitement Automatique du Langage Naturel (TALN). Une partie de l'équipe se concentre dans l'analyse sémantique des textes et la représentation des connaissances à travers les réseaux lexico-sémantiques. Leur domaine d'expertise englobe la désambiguïsation lexicale, l'utilisation du crowdsourcing pour l'apprentissage automatique, ainsi que l'exploitation des jeux en ligne comme plateforme de supervision.

Objectif industriel

Le Traitement Automatique de la Langue (TALN) moderne offre une multitude d'applications potentielles, révolutionnant divers secteurs avec des outils comme les assistants conversationnels, les moteurs de recherche avancés, et les systèmes de fouille de textes. Ces technologies jouent un rôle crucial dans l'amélioration de l'accès à l'information non structurée et dans l'automatisation des tâches fastidieuses de saisie et d'extraction. Les avancées récentes en matière de plongements lexicaux, tels que Word2Vec, Glove, et BERT, ont considérablement enrichi l'analyse du langage naturel. Ces modèles prêts à l'emploi sont efficaces pour les applications grand public mais peinent à interpréter avec précision des termes spécifiques et des expressions complexes propres à des domaines spécialisés comme l'administration publique, l'industrie, l'éducation, et la santé. Cette lacune est principalement due à leur entraînement sur des corpus génériques tels que Wikipédia, ne captant pas la spécificité du vocabulaire technique utilisé dans ces secteurs.

Pour remédier à ce défi, l'utilisation d'ontologies spécialisées et de graphes de connaissances se révèle essentielle. Ces ressources permettent une interprétation plus fine et précise des termes hyper-spécialisés. Berger-Levrault se distingue dans ce domaine grâce à son héritage en édition juridique et pratique, enrichi par une expertise centenaire dans les métiers de l'administration publique française. Leur vaste base de connaissances, incluant plus de 200 ouvrages et des centaines d'articles sur le portail Légibases, couvre des domaines tels que l'état civil, les marchés publics, et la santé. En effet, les connaissances pouvant être extraites de la base éditoriale française de Berger-Levrault sont très utiles pour l'amélioration d'un ensemble d'applications, en passant de l'extraction d'information et de la recherche de documents vers l'enrichissement de connaissances d'agents conversationnels afin de satisfaire au mieux les besoins des utilisateurs.

L'approche de Berger-Levrault repose sur une représentation structurée des connaissances, soutenue par des annotations expertes sur chaque paragraphe de leurs documents. Cette méthodologie permet la création de graphes de connaissances lexico-sémantiques de haute qualité, cruciales pour améliorer les performances des agents conversationnels et autres applications basées sur le TALN. Cependant, la gestion de ces connaissances éditoriales est complexe et nécessite une évolution constante, en raison des changements fréquents dans la réglementation et les pratiques administratives. Une ontologie statique serait rapidement obsolète, justifiant ainsi la nécessité d'un apprentissage dynamique permettant d'adapter en permanence les bases de connaissances aux nouvelles informations et réglementations.

En conclusion, l'approche de Berger-Levrault s'inscrit dans une démarche scientifique et outillée visant à exploiter pleinement les techniques avancées du TALN pour répondre aux

1. Lien : https://www.lirmm.fr/equipes/TEXTE/

besoins spécifiques et évolutifs des métiers de l'administration publique et autres secteurs spécialisés. Ce partenariat entre édition et technologie permet de transformer efficacement l'information textuelle en connaissances exploitables, enrichissant ainsi les services offerts aux utilisateurs finaux.

Objectif scientifique

Dans le cadre de cette recherche, notre objectif principal est d'explorer et de mettre en œuvre des méthodes avancées pour la construction automatique de graphes de connaissances lexico-sémantiques à partir de textes. L'automatisation de ce processus consiste à extraire des concepts et des relations spécifiques à des domaines de spécialité, en l'occurrence le secteur public.

Notre approche se fonde sur plusieurs axes stratégiques :

1. **Utilisation des ressources lexico-sémantiques :** Nous exploitons JeuxDeMots, une ressource riche comprenant 14 millions de nœuds et 320 millions de relations, pour établir des fondations solides d'ontologies. Cette base permet de capturer le sens commun et spécialisé des termes français polysémiques.
2. **Méthodologies et techniques d'apprentissage :** Nous adoptons une approche intégrant diverses techniques issues de l'intelligence artificielle, de l'apprentissage automatique, de la fouille de textes, et du traitement automatique du langage naturel. Ces techniques incluent l'analyse lexicale, la catégorisation grammaticale, l'analyse des cooccurrences, et l'inférence logique pour extraire et structurer les ontologies.
3. **Processus d'acquisition et vérification :** Nous mettons en œuvre un processus itératif de construction de graphes de connaissances, débutant par l'identification des termes clés et de leurs synonymes, puis la formation de concepts et l'extraction de relations taxonomiques et non-taxonomiques. L'accent est mis sur la validation humaine pour assurer la qualité des ontologies produites, tout en explorant des méthodes d'autosupervision basées sur l'inférence.
4. **Perspectives d'évaluation et amélioration continue :** Nous prévoyons deux types d'évaluation des graphes de connaissances : intrinsèque, impliquant des experts pour évaluer la qualité des ressources ontologiques, et extrinsèque, évaluant l'impact des graphes de connaissances dans des applications de moteur de recherche sémantique. Cette démarche vise à garantir la pertinence et l'utilité des graphes de connaissances dans des contextes réels.
5. **Application spécifique au secteur public :** Notre approche est spécifiquement adaptée aux besoins du secteur public, où la précision et la contextualisation des ontologies sont cruciales pour améliorer la gestion des connaissances et des informations.

En conclusion, cette recherche s'inscrit dans le cadre d'un effort multidisciplinaire visant à automatiser et à améliorer la construction de graphes de connaissances à partir de textes, en intégrant des méthodologies avancées et en se basant sur des ressources lexicales de pointe. L'objectif ultime est de faciliter la création de modèles de graphes de connaissances précis et adaptés, essentiels pour une gestion efficace des informations dans le domaine complexe du secteur public.

CHAPITRE 1

Introduction

Ce premier chapitre introduit les concepts fondamentaux nécessaires à la compréhension des problématiques et des approches abordées dans cette thèse. L'objectif est de définir les bases théoriques et contextuelles qui serviront de fondation pour les chapitres suivants. Les défis majeurs de ce chapitre incluent l'identification des concepts clés et la structuration de la connaissance initiale.

Sommaire du chapitre

1.1 Introduction générale **1**
1.2 Motivations et Contexte de travail **2**
1.2.1 Présentation des données 2
1.2.2 Structurer la connaissance 3
1.2.3 Problématique générale 3
1.3 Axe de recherche **5**
1.4 Plan **6**

1.1 Introduction générale

Dans un contexte où la disponibilité des données textuelles explose, la capacité à structurer et à interpréter ces informations devient essentielle. Les graphes lexico-sémantiques, qui modélisent les relations entre les termes dans un domaine spécifique, émergent comme un outil puissant pour organiser et exploiter ces vastes quantités de données. Les textes de spécialité, souvent riches en termes techniques et en relations complexes, représentent une source précieuse pour la construction de graphes lexico-sémantiques. Ces graphes, qui capturent les relations sémantiques entre les termes et concepts dans un format structuré, facilitent la recherche d'information, la compréhension des domaines spécifiques et l'extraction de connaissances implicites. Cette thèse se concentre sur la construction de tels graphes à partir de textes de spécialité du domaine public, en utilisant des techniques avancées du traitement automatique des langues (TALN).

Importance de la Recherche : Les textes de spécialité du domaine public, tels que les articles scientifiques, les rapports techniques, et les publications légales, contiennent une richesse d'informations précises et détaillées. Exploiter ces ressources pour construire des graphes lexico-sémantiques permet de transformer des données non structurées en connaissances structurées et accessibles. Ces graphes offrent une représentation claire des relations sémantiques entre les termes, facilitant ainsi la recherche d'informations et la compréhension des connaissances spécialisées.

Problématiques et Défis La construction de graphes lexico-sémantiques à partir de textes spécialisés pose plusieurs défis. Premièrement, il faut extraire et normaliser les termes techniques de manière précise. Deuxièmement, il est nécessaire de détecter et de modéliser correctement les relations entre ces termes, ce qui requiert une compréhension approfondie du contexte et des nuances sémantiques. Enfin, les graphes doivent être régulièrement mis à jour pour refléter l'évolution des connaissances et des terminologies.

Cependant, ces défis s'accompagnent d'opportunités considérables. Les avancées dans ce domaine fournissent de nouveaux outils pour surmonter ces obstacles. Les larges modèles de langue (LLM), par exemple, peuvent être utilisés pour comprendre et générer du texte de manière contextuellement pertinente et ainsi être utilisés par prompting afin de détecter la terminologie et les triplets relationnels d'un texte, tandis que les réseaux lexicaux peuvent fournir une structure initiale pour l'organisation des termes et des relations.

Objectifs de la Thèse L'objectif principal de cette recherche est de développer des méthodes et des systèmes pour la construction de graphes lexico-sémantiques à partir de textes de spécialité du domaine public. Les sous-objectifs incluent :

1. **Extraction et Normalisation des Termes Techniques :** Identifier et normaliser les termes techniques dans les textes de spécialité.
2. **Extraction des Relations Sémantiques :** Développer des algorithmes pour extraire les relations entre les termes, en utilisant les contextes d'apparition dans les textes.
3. **Construction et Mise à Jour des Graphes :** Structurer les termes et relations dans des graphes lexico-sémantiques, puis l'évaluation de leur pertinence et de leur exhaustivité. Après la construction, le développement de mécanismes pour leur mise à jour continue est attendu.

1.2 Motivations et Contexte de travail

1.2.1 Présentation des données

Historiquement, Berger-Levrault est un éditeur d'ouvrages et d'articles juridiques et pratiques. Les articles sont en ligne et issus des Légibases [1], une Légibase pour chacun des huit domaines. Légibase est une base de données d'articles réglementaires annotés et commentés par les experts que Berger-Levrault met à disposition à ces clients sous la forme d'un site web. Le corpus extrait en 2021 était constitué de 211 ouvrages et de 13085 d'articles répartis dans 8 sous-domaines de spécialité du secteur public. La table A.1 en annexe regroupe ces 8 sous-domaines associés à un exemple de début d'un des articles du domaine. Ce corpus est en constante évolution. En effet, en 2022, un neuvième domaine a ouvert et est visible actuellement dans la portail des Légibases. Le nombre d'ouvrages et d'articles ne

1. Disponibles à l'adresse `https://www.legibase.fr/`

cesse d'augmenter au fil du temps. En effet, en 2020, le corpus comptait 172 ouvrages. Notre travail doit donc prendre en considération ces données en constante évolution et non des données statiques. De plus, l'expertise évolue et doit être prise en considération. Cela signifie que l'information extraite est potentiellement obsolète et doit être mise à jour dans nos connaissances. L'exploitation de ce corpus est naturellement recherchée car il met à notre disposition une grande richesse de connaissances d'expertise. Cependant, nous constatons sur la table A.1 en annexe que ce corpus de spécialité est écrit au format brut et non structuré, ce qui rend son exploitation plus difficile. En effet, il est compréhensible qu'un humain recherche l'information dans un livre ou un article en le parcourant quasi intégralement. Ce même principe s'applique aux systèmes exploitant des textes bruts.

1.2.2 Structurer la connaissance

L'exploitation de la connaissance présente dans les textes brutes est difficile. C'est pourquoi il est nécessaire de structurer la connaissance afin de pouvoir y accéder rapidement. Berger-Levrault, spécialisée dans les métiers des services publics, se consacre à la conception et à la réalisation de logiciels destinés aux administration publiques locales, notamment dans les domaines de l'éducation, de la santé, du sanitaire, du social et de la gestion territoriale. Ces logiciels consistent à simplifier les processus administratifs par le biais de la mise à disposition de solutions numériques à l'usage des agents publics et des citoyens. Berger-Levrault développe des solutions logicielles spécialisée en Traitement Automatique du Langage Naturel (TALN) comme un moteur de recherche, un moteur de réponse, une automatisation de la veille juridique et d'autres encore. Ces solutions logicielles ont besoin d'exploiter les connaissances des données textuelles de Berger-Levrault. Mais jusqu'à présent, elles étaient limitées par le parcours intégral de la connaissance brute. Il est donc nécessaire de structurer cette connaissance afin de pouvoir l'exploiter pour enrichir les applications et ainsi les améliorer. Pour cela, deux aspects sont mis en évidence : (1) Représenter l'ensemble des connaissances au travers d'une structure de connaissances ; et (2) Utiliser ces connaissances pour améliorer les performances de plusieurs de ses produits.

1.2.3 Problématique générale

À partir d'un texte donné comme présenté sur l'exemple 1.1, le but est de structurer l'information comme présenté sur la figure 1.1.

En décembre 2014, un rapport a été dressé par la Banque de France au sujet de la question préoccupante du surendettement. Justement intitulé « Étude des parcours menant au surendettement », il nous livre une analyse en amont des dépôts de dossiers auprès des commissions de surendettement.

Exemple 1.1 – Exemple d'un paragraphe issu d'un texte de Berger-Levrault

Nous nous intéressons donc à cette question de structurer les informations présentes dans nos textes juridiques qui plus est. Nous en venons donc à nous demander si l'état de l'art du domaine du TALN est capable de construire une structure de connaissances à partir de textes juridiques ? Nous devons séparer d'une part cette problématique en deux axes : la notion de structure de connaissance et la notion d'adaptation aux textes juridiques.

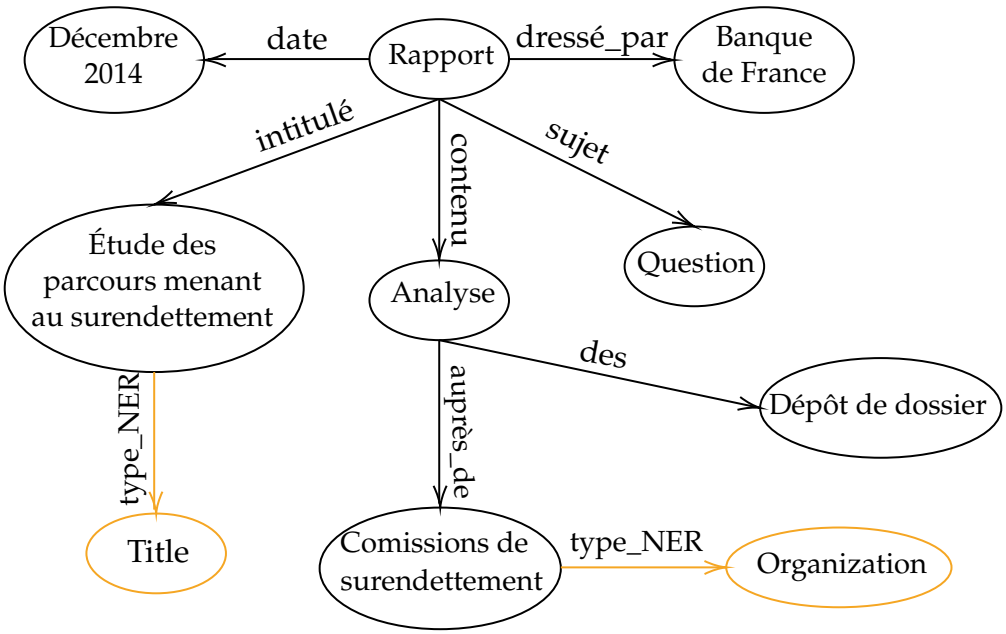

FIGURE 1.1 – Exemple d'un graphe de connaissances lexico-sémantique obtenu à partir du texte de l'exemple 1.1.

Comme présenté sur la figure 1.1, une structure de graphe ressort très fréquemment lorsque nous souhaitons extraire l'information d'un texte et ainsi la structurer. Plusieurs grandes catégories de structures ressortent de l'état de l'art à savoir les ontologies, les graphes de connaissances et les thésaurus. Dans le domaine de l'ingénierie des connaissances, des définitions formelles sont données pour les structures de graphes. Or, deux paramètres ont besoin au préalable d'être fixés car nous nous intéressons à des données textuelles. D'une part, étant donné qu'aucune structure de connaissance n'est préétablie, nous partons des textes bruts et sommes donc tenus de construire initialement notre structure à partir de la terminologie. Avec le temps, cette structure nous permettra de faire la distinction entre la terminologie de base et les concepts. D'une autre part, nous cherchons à créer un graphe à visée lexicale et sémantique. En effet, la structure de connaissances va être remployée par Berger-Levrault pour être utilisée dans des applications de TALN et permettra de les guider. La notion de sens ou concepts dans les nœuds donne la notion de sémantique. L'aspect lexical quant à lui demande à la structure de contenir des relations lexicales et sémantiques entre les mots comme des relations de taxonomie par exemple, à l'instar de *Wordnet*[2] faisant référence dans le domaine (Fellbaum, 1998; Harabagiu et al., 1999). Quelques différenciations peuvent donc être présentes dans la définition formelle de l'objet.

Nous nous intéressons donc à construire un graphe de connaissances lexico-sémantique. Nous rapprochons la définition d'un **graphe de connaissances lexico-sémantique** à celui du réseau lexical mentionné dans (Lafourcade, 2011) étant lui-même une structure de graphe. Un graphe de connaissance est une base de connaissances sous forme de graphe. Un tel graphe est composé de faits et de connaissances que l'on peut voir comme des classes, sous-classes et instances. Les **faits** sont des segments textuels. Il peut s'agir de termes pouvant être simples ou composés. Un label du nœud est un terme dit préférentiel d'un texte à savoir les termes sous une forme privilégiée par rapport à toutes les autres présentes dans le texte. Il peut s'agir aussi de raffinements sémantiques (ex : $souris > informatique$) ou de locutions. De nouvelles connaissances sont inférées grâce à des règles. Un tel graphe contient des **pro-**

2. Wordnet disponible au lien suivant :`http://wordnetweb.princeton.edu/perl/webwn`

priétés entre les entités qui constituent ce graphe. Ces propriétés sont des relations lexicales ou d'ordre sémantique (ontologiques, associatives, prédicatives ou de typicalité). Les relations dites **lexicales** (synonymie, antonymie, locution, famille) portent sur les termes, leurs dérivés, synonymes ou encore de sens contraires. Les relations **ontologiques** (ex : générique comme l'hyperonymie, spécifique comme l'hyponymie ou encore partie comme la méronymie) portent sur des connaissances liées à des objets du monde. Les relations **associatives** (ex : association libre, sentiment associé, signification) s'apparentent à de la connaissance subjective et/ou globale. Les relations **prédicatives** (ex : agent, patient, instrument) sont associées à un verbe ou un prédicat et aux valeurs de ses arguments de ses arguments au sens large. Enfin, les relations de **typicalité** représentent des lieux, des moments ou encore des caractéristiques typiques. L'ensemble de ces relations peuvent s'apparenter de deux façons différentes. Il peut s'agir de relations bien définies telles que la synonymie ou équivalence, la notion de hiérarchie, voire d'association. Des relations implicites tirées du textes peuvent aussi être pertinentes comme un verbe ou un nom qui lie deux entités entre elles dans un texte. Un tel graphe embarque la notion de **domaine** qui est primordiale dans notre cas. En effet, comme explicité précédemment, nos textes appartiennent à 8 sous-domaines différents. Une définition plus étoffée est donnée dans le chapitre 2.

1.3 Axe de recherche

Cette thèse se concentre principalement sur les défis du TALN, bien qu'elle soit influencée par les problématiques de l'ingénierie des connaissances. Dans l'examen des méthodes de construction de structures de connaissances à partir de textes, nous pouvons identifier une approche que nous pouvons ajuster selon nos besoins et notre perspective. En effet, il est clair qu'un certain nombre d'étapes sont requises pour élaborer une structure de connaissances à partir de textes. Nous pouvons cibler trois grandes étapes de construction dépendantes les unes des autres.

1. La première consiste à extraire la terminologie qui deviendra par la suite un ensemble de concepts.
2. La deuxième consiste à extraire les relations. Plusieurs types de relations peuvent être identifiés. Afin de regrouper les termes entre eux et ainsi en faire des concepts, nous souhaitons identifier les relations de hiérarchie et de synonymie entre nos termes. Puis, nous souhaitons extraire des relations sémantiques. Il peut s'agir d'une même approche si les relations sont bien définies, c'est à dire que nous avons une connaissance préalable des types de relations et cet ensemble de types est fini. Il s'agit souvent d'extraire des candidats de paires de termes et de les classifier dans un type donné parmi ceux précisés en amont. Dans le cadre de l'extraction de relations en monde ouvert - c'est à dire que nous extrayons la relation depuis le texte - une approche différente devra être abordée. Ici, la liberté est totale et rend le nombre de possibilités très grand puisque l'on peut avoir un nombre de types maximal avoisinant le nombre de termes de liaison dans un texte.
3. Finalement, une étape d'application de règles est nécessaire afin de faire des inférences ou pour valider la cohérence du graphe de connaissance. En effet, un graphe se doit de maintenir la correction et la complétude lorsque l'on ajoute de l'information. Cela signifie qu'un graphe doit contenir une information correcte qui ne se contredit pas au sein du graphe.

Nous pouvons donc dessiner un système dit général dont l'architecture est présentée sur la figure 1.2. En effet, à partir des textes bruts issus des documents, nous classifions les do-

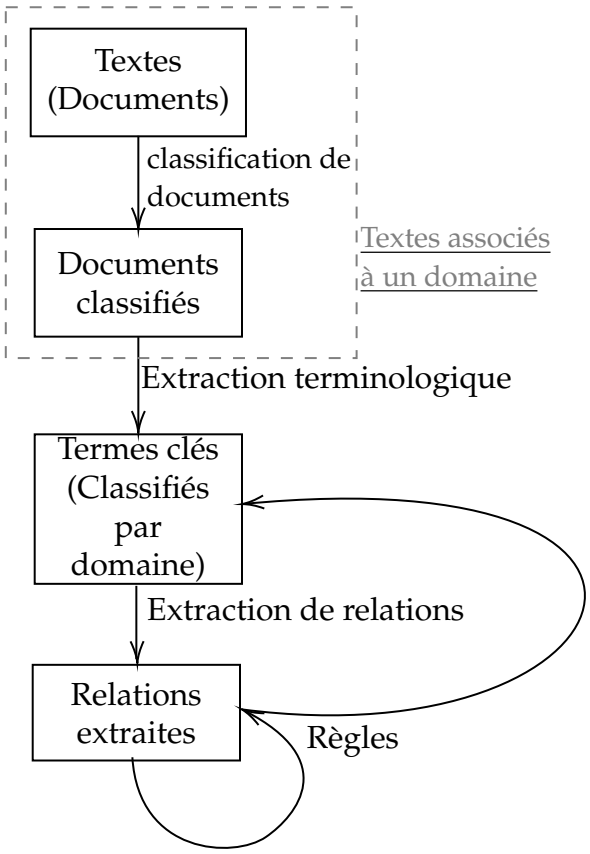

FIGURE 1.2 – Système général de construction de graphe de connaissances lexico-sémantique

cuments dans l'un des 8 sous-domaines. Ensuite, nous parcourons chacun des paragraphes afin d'en extraire les termes clés et ainsi de pouvoir les rattacher à un domaine donné. Cela permet d'une part de le renseigner dans le graphe comme explicité précédemment mais aussi de filtrer les relations par domaine et ainsi de se concentrer sur de l'information pertinente adaptée à un domaine fixé. Ensuite, entre ces termes clés pourront être extraites les relations qui les unissent qu'il s'agisse de relations d'un type déterminé ou à partir de texte. Et pour finir, les règles pourront mettre à jour l'ensemble des termes clés et des relations en supprimant les informations erronées ou obsolètes ou en ajoutant des informations inférées.

1.4 Plan

Après avoir présenté et exploré les graphes de connaissances lexico-sémantiques dans un chapitre dédié (c.f. chapitre 2), nous allons exposer un état de l'art (c.f. chapitre 3) récapitulant les différentes approches afin de construire des graphes à partir de textes et leurs limites, tout en identifiant les lacunes que cette thèse vise à combler. Ce chapitre est composé d'une présentation des méthodes traditionnelles et modernes pour l'extraction d'information composée de l'extraction terminologique ainsi que l'extraction de relations en général (relations typées ou relations ouvertes), en mettant l'accent sur leurs avantages et inconvénients. Une partie est également consacrée aux LLMs avec leur rôle dans l'extraction de relations, ainsi que les défis associés à leur utilisation.

Ensuite, l'extraction terminologique (c.f. chapitre 4) a été explorée comme une étape cru-

ciale préalable à l'extraction de relations. Le chapitre sur l'extraction terminologique présente dans un premier temps les annotations manuelles des experts utilisées comme terminologie de base ainsi que le système InfoGlean KeyTerms. Le système InfoGlean KeyTerms a été présenté et évalué, montrant son efficacité pour identifier des termes pertinents à partir de corpus textuels.

Ensuite, l'intégration du réseau lexical rezoJDM issu du projet JeuxDeMots dans le processus d'extraction de relations a été détaillée (c.f. chapitre 5). L'algorithme Relational Embeddings Model (REM) a été développé pour la classification et l'extraction des relations, avec des résultats prometteurs. REM permet d'extraire et classifier les relations à partir des termes identifiés dans des corpus textuels. REM utilise des embeddings relationnels, une technique qui permet de représenter les relations entre termes dans un espace vectoriel de manière à capturer la sémantique et les connexions implicites entre eux. Le système a été amélioré en utilisant des techniques de classification de domaines pour améliorer la spécificité et la pertinence des relations extraites que nous détaillerons dans une section du chapitre dédiée.

Enfin, le dernier chapitre de contributions (c.f. chapitre 6) présente un nouveau système d'extraction de relations ouvertes après avoir mis en évidence les limites liées au typage de relations lors de l'évaluation de l'efficacité de l'approche proposée en utilisant rezoJDM. Ce chapitre propose une nouvelle technique d'extraction de relations, visant à surmonter ces limitations en utilisant des approches dynamiques telles que l'OpenIE en utilisant les larges modèles de langues (LLMs) et en focalisant l'extraction sur les verbes d'action spécifiques aux textes, offrant ainsi une extraction plus précise et adaptée aux besoins spécifiques des corpus étudiés. Pour conclure cette thèse (c.f. chapitre 7), nous allons résumer les principales contributions et proposer des perspectives pour des travaux futurs.

CHAPITRE 2

Graphe de connaissances lexico-sémantiques

Le chapitre précédent a présenté les concepts de base et le contexte de cette thèse. Dans ce chapitre, nous nous focalisons sur les graphes de connaissances lexico-sémantiques, une structure permettant de structurer et d'exploiter la connaissance textuelle. Les défis incluent la définition et l'implémentation de cette structure ainsi que l'exploitation de ressources existantes pour enrichir notre graphe. Afin de structurer la connaissance d'un texte, nous choisissons de nous orienter vers une structure de graphes intitulée **graphe de connaissances lexico-sémantiques**. Un graphe de connaissance permet de structurer de la connaissance afin d'être utilisé par des applications diverses, dont l'analyse sémantique de texte qui permet de structurer la connaissance contenue dans un texte. Ce chapitre sert à introduire et définir cette notion, puis à détailler une ressource existante que nous avons cherché à exploiter dans cette thèse.

Sommaire du chapitre

2.1 Définition générale **9**
2.2 Ressource exploitée : le réseau lexico-sémantique issu de JeuxDeMots . . . **10**
2.2.1 Présentation générale du graphe de connaissances lexico-sémantique rezoJDM 11
2.2.2 Nœuds 12
2.2.3 Relations 13
2.3 Conclusion **17**

2.1 Définition générale

Definition 2.1 (Graphe de connaissances lexico-sémantique)**.** *La définition que nous avons de la notion de* ***graphe de connaissances lexico-sémantiques*** *s'inspire fortement de celle du réseau lexico-sémantique mentionné dans (Lafourcade, 2007, 2011), lui-même une structure de graphe. Un* ***graphe de connaissances*** *est une base de connaissances sous forme de graphe, composée d'un ensemble de nœuds et de relations. Les nœuds représentent des faits ou des connaissances, qui, dans notre cas, sont des segments textuels. Il peut s'agir de termes simples ou composés. Le label d'un*

nœud est un terme préférentiel d'un texte, ce qui signifie qu'un terme privilégié sert d'étiquette par rapport à toutes les autres présentes dans le texte. Il peut également s'agir de raffinements sémantiques (par exemple : souris $>$ informatique) ou de locutions. De nouvelles connaissances sont inférées grâce à des règles. Un tel graphe contient des ***relations****, également appelées* ***propriétés****, entre les entités qui constituent ce graphe. Ces propriétés sont des relations lexicales ou sémantiques (ontologiques, associatives, prédicatives ou de typicalité).*

Les relations dites **lexicales** (synonymie, antonymie, locution, famille) portent sur les termes, leurs dérivés, synonymes ou encore leurs sens contraires. Les relations **ontologiques** (par exemple : générique comme l'hyperonymie, spécifique comme l'hyponymie, ou encore « partie de » comme la méronymie) portent sur des connaissances liées à des objets du monde. Les relations **associatives** (par exemple : association libre, sentiment associé, signification) s'apparentent à de la connaissance subjective et/ou globale. Les relations **prédicatives** (par exemple : agent, patient, instrument) sont associées à un verbe ou à un prédicat, ainsi qu'aux valeurs de ses arguments dans un sens large. Enfin, les relations de **typicalité** représentent des lieux, des moments ou encore des caractéristiques typiques. L'ensemble de ces relations peut se découper en deux caractéristiques différentes. Il peut s'agir de relations bien définies telles que la synonymie ou l'équivalence, la notion de hiérarchie, voire d'association. Il peut également s'agir de relations implicites tirées d'un texte, telles qu'un verbe ou un nom qui lie deux entités entre elles dans un texte. Un tel graphe inclut la notion de **domaine**, qui est primordiale dans notre cas d'étude.

2.2 Ressource exploitée : le réseau lexico-sémantique issu de JeuxDeMots

Mathieu Lafourcade a fondé en 2007 le projet JeuxDeMots (JDM) au sein de l'équipe TEXTE du LIRMM. JDM est un jeu sérieux appartenant à la famille des jeux à but (GWAP - *Games With A Purpose*). Ce projet vise la construction d'un Réseau Lexico-Sémantique (RLS) pour le français que nous appellerons **rezoJDM**. Il est conçu de manière à pouvoir être utilisé pour l'analyse sémantique de texte. Il a été construit grâce aux réponses des joueurs mais aussi grâce aux inférences faites par déduction, induction et abduction. La plateforme invite le joueur à expliciter les relations qui relient les différents mots du réseau ou les différents termes associés par la relation à un terme donné. Cette approche permet la collecte de données, engendrant ainsi un RLS comprenant plus de 6 millions de termes pour 20 millions de nœuds, et plus de 518 millions de relations réparties sur 164 types de relations. Les données mentionnées sont celles recueillies entre le 24 août 2023 et le 23 décembre 2023 et sont en constante évolution [1]. En effet, le 24 décembre 2024, le réseau lexico-sémantique comprenait plus de 7 millions de termes pour 22 millions de nœuds, et plus de 545 millions de relations lexicales et sémantiques dont près de 49 millions de relations nulles ou négatives réparties sur 167 types. Nous constatons donc un accroissement à chacune des mises à jour. Nous nous intéressons à un tel réseau pour son utilité dans l'analyse sémantique de textes.

1. Pour les valeurs exactes, se référer aux statistiques sur l'adresse suivante : `https://www.jeuxdemots.org/jdm-about.php`

2.2.1 Présentation générale du graphe de connaissances lexico-sémantique rezoJDM

Dans cette partie, nous présentons le graphe de connaissances lexico-sémantiques rezoJDM exploité en tant que ressource linguistique et sémantique. rezoJDM est un Réseau Lexico-Sémantique (RLS). Un RLS est une structure sous forme de graphe orienté permettant de représenter les informations lexicales et sémantiques par le biais de relations entre les termes d'une langue qu'il contient. Dans une telle structure, les nœuds du graphe désignent les segments textuels et les arcs sont des relations lexicales d'ordre sémantique comme explicité dans la section précédente (2.1) concernant la définition d'un graphe de connaissances lexico-sémantiques. Dans la suite du manuscrit, nous utiliserons la notion de graphe de connaissances lexico-sémantiques. Les arcs et par conséquent les relations sont typés, pondérés et potentiellement annotés. La pondération permet de créer des relations d'inhibition. Par exemple, $voler \xrightarrow{\text{agent}} oiseau$ est une relation de pondération positive mais $voler \xrightarrow{\text{*agent}} autruche$ a une pondération négative signalée par l'astérisque. Les annotations permettent également de donner une valeur symbolique à un arc en ajoutant des méta-informations comme "rare", "pertinent", "non pertinent", etc. Cela facilite l'exploitation et le parcours par heuristique du graphe. Grâce à de nombreux outils exploitant rezoJDM, il est également plus facile de requêter le réseau que ce soit humainement (par exemple, Diko[2]) ou dans les applications (par exemple, rezo-dump[3]).

Afin d'assurer l'engagement des participants et de garantir un intérêt accru, le jeu doit susciter l'intérêt des joueurs, favorisant ainsi une participation active et volontaire, ce qui contribue à la qualité des données collectées. En effet, des participants engagés sont plus susceptibles de fournir des réponses réfléchies et pertinentes, ce qui améliore la fiabilité des informations obtenues. Dans le déroulement du jeu GWAP JeuxDeMots, chaque joueur est confronté à un mot cible sélectionné dans une base de mots, avec une consigne associée pour un type de compétence ou de relation. Le joueur fait alors une dizaine de propositions possibles. Pour valider une proposition, il lui suffit d'appuyer sur le bouton « ok », ou de cliquer dessus. Il a également la possibilité de se rétracter et de supprimer une ou plusieurs propositions. Cette même consigne est ensuite proposée un certain nombre de fois à d'autres joueurs, garantissant ainsi une collecte de données basée sur l'accord des joueurs, renforçant la validité et la diversité des informations recueillies. La validation indirecte des relations proposées par un joueur est effectuée par d'autres joueurs, dans un processus de covalidation. Ce mécanisme s'opère sous forme de jeu, où les joueurs proposent des associations et la concordance des propositions entre paires de joueurs permet de valider indirectement les relations. En effet, si plusieurs joueurs ($\geqslant 2$) donnent une réponse commune, le réseau est modifié : si la relation n'existe pas, elle est créée avec un poids nul (proche de 0), sinon son poids est renforcé par incrémentation. Il est important de conserver tous les poids du réseau car dans le cas où la relation aurait été ajoutée avec un poids négatif (en tant que fausse), cela signifie qu'un traitement a été fait de manière humaine ou algorithmique et n'est plus à refaire (ce qui est coûteux en termes de calcul ou de travail humain). Cela peut être utile lors d'une analyse des liens entre les mots d'un texte. Les joueurs, cependant, ne sont pas conscients que leur activité de jeu génère des relations, ajoutant une dimension subtile au processus.

Toutefois, il convient de noter qu'il existe un biais introduit dans le processus, nécessitant une prise en compte attentive. Les réponses des joueurs peuvent être influencées par divers

2. Diko accessible au lien suivant : `https://www.jeuxdemots.org/diko.php`
3. Rezo-dump accessible au lien suivant : `https://www.jeuxdemots.org/rezo-dump.php`

facteurs tels que leurs expériences personnelles, leurs connaissances préalables ou même des biais culturels. De plus, le vocabulaire actif des joueurs, notamment utilisé pour s'exprimer, peut être fortement orienté par la consigne donnée dans le jeu GWAP JeuxDeMots. Il est également important de noter que certains joueurs, par facilité, peuvent consulter des sources externes telles que Wikipédia ou des dictionnaires, introduisant ainsi un vocabulaire plus soutenu qui compense le vocabulaire actif initial.

Enfin, les associations liées à l'actualité constituent un aspect dynamique et évolutif des relations sémantiques. Elles sont étroitement liées au contexte temporel, évoluant avec le temps. Par exemple, pour un pays donné, le président en exercice est fortement associé pendant la période de son mandat. Cette association peut être significative à un instant t, mais elle n'est pas nécessairement valide dans le passé ni garantie après la fin de son mandat. Ainsi, ces relations éphémères reflètent la nature changeante des liens sémantiques en fonction des événements actuels, soulignant la nécessité de prendre en compte la temporalité dans l'analyse des associations conceptuelles.

2.2.2 Nœuds

Definition 2.2 (Nœud)**.** *Un **nœud** d'un graphe lexico-sémantique représente un objet lexical qui contient un segment textuel ainsi qu'un poids pour déterminer son importance en usage. Cet objet lexical est le plus souvent un terme dit **standard**, mais il existe d'autres types de nœuds particuliers comme des raffinements sémantiques ou morphologiques, des méta-informations ou encore des locutions. Un segment textuel peut être un mot unique ou bien un terme composé.*

L'intérêt d'avoir différents types de nœuds possibles est de pouvoir associer les termes de la langue à d'autres informations de type lexicales ou sémantiques. Cela permet donc d'éliminer toute ambiguïté sur la nature du segment textuel. Le tableau 2.1 présente certains types de nœuds présents dans rezoJDM. Nous omettons tous les types de nœuds locaux au réseau. Nous allons détailler ici uniquement les différents types manifestés dans le tableau, utiles à l'analyse textuelle pour notre cas d'usage.

Id nœud (ntid)	Nom	Information
1	n_term	Nœud de terme standard
4	n_pos	Nœud de Part of Speech
18	n_data	Nœud d'informations lexicales/syntaxiques
36	n_data_pot	Nœud d'informations sémantiques

TABLE 2.1 – Liste non exhaustive des différents types de nœuds présents dans rezoJDM

Nous pouvons associer à un terme dit standard sa ou ses classes grammaticales ou POS qui lui conviennent. C'est pour cela que nous avons des nœuds de **POS**. Un terme peut avoir plusieurs POS car il peut changer de sens en fonction de son contexte. Privé de ce dernier, il devient ambigu. On dit alors que notre terme est polysémique. Un mot peut comporter plusieurs sens (gloses), ce qui met en difficulté le système pour l'analyse du texte. Par exemple, le mot « avocat » peut avoir deux gloses possibles : le fruit ou bien la fonction en justice. Il est donc important d'expliciter les différentes gloses pour un mot donné.

Afin d'expliciter les différents sens d'un terme donné, nous utilisons le raffinement d'un terme. Les **raffinements** sont des représentations des différents sens ou usages d'un terme, sous la forme *terme>glose*. Il s'agit de nœuds de type standard mais ayant un label particulier. Il existe deux types de raffinements : le raffinement **morphologique** et le raffinement **sémantique**. Le **raffinement morphologique** permet de représenter la forme des mots. Il

illustre l'organisation des catégories grammaticales. Pour cela, nous associons au terme sa classe grammaticale ou POS. Par exemple, pour le mot « avocat », les raffinements morphologiques sont notés « avocat>Nom : » (pour le fruit et la profession) et « avocat>Adj : » (pour la couleur). Le **raffinement sémantique** permet de donner une signification au terme. Par exemple, les raffinements sémantiques du mot précédent sont « avocat>justice » et « avocat>fruit ». Le raffinement est géré par une relation qui relie le nœud du terme vers l'un de ses raffinements (pouvant être unique si le mot n'est pas polysémique). L'exemple 2.1 présente les raffinements sémantiques et morphologiques de l'exemple précédent. Le nœud

— avocat $\xrightarrow[p=68]{\text{r_raff_sem \#1}}$ avocat>justice
— avocat $\xrightarrow[p=56]{\text{r_raff_sem \#1}}$ avocat>fruit
— avocat $\xrightarrow[p=49]{\text{r_raff_sem \#1}}$ avocat>médiateur
— avocat $\xrightarrow[p=43]{\text{r_raff_sem \#1}}$ avocat>couleur
— avocat $\xrightarrow[p=48]{\text{r_raff_morpho \#2}}$ avocat>Nom :
— avocat $\xrightarrow[p=42]{\text{r_raff_morpho \#2}}$ avocat>Adj :

Exemple 2.1 – Liste des différents raffinements sémantiques et morphologiques pour le mot « avocat » tirés de rezoJDM. Requête en utilisant l'outil rezo accessible sur le lien suivant : Lien Rezo. Le premier terme avant la flèche est le terme source et celui d'après est le terme cible dans une relation. La flèche exprime le type de la relation (ex : r_raff_morpho) avec son identifiant précédé par un # (ex : #1) et le poids de la relation renseigné par p (ex : p=42).

représentant le raffinement (ex : avocat>justice) contient une relation d'hyperonymie (r_isa) vers les autres raffinements associés au terme général (ex : avocat), mais le poids est négatif ($p \leq 0$). La notion de relation est expliquée dans la sous-section suivante (2.2.3).

Le raffinement sémantique d'un terme peut s'avérer être lui-même polysémique. Dans ce cas, nous aurons une imbrication de gloses sous la forme *terme>glose>sous-glose*. L'exemple 2.2 présente une illustration de cette imbrication de raffinement sémantique pour le mot « avocat ».

Comme expliqué précédemment, un nœud est un objet textuel mais n'est pas forcément un terme. Certains nœuds de rezoJDM sont dits de concept ou de **données** (data et data_pot) car ils contiennent des informations conceptuelles qui sont non ambiguës. Les nœuds de données (data) fournissent des informations principalement lexicales telles que la langue employée (ex : français), le niveau de langue (courant, soutenu, . . .), *etc.* Les nœuds d'informations sémantiques (data_pot), quant à eux, donnent des informations sémantiques qui sont des symboles ne faisant pas partie du lexique. Leur information est non ambiguë. Par exemple, il s'agit d'un terme « animé », « non animé », . . . D'autres cas sont donnés dans l'exemple 2.3 afin de présenter la forme associée à l'explication.

2.2.3 Relations

Definition 2.3 (Relation)**.** *Une **relation** dans le réseau rezoJDM est un lien typé, orienté et pondéré entre deux nœuds du graphe. Plus formellement, une relation est un quadruplet des composantes*

— avocat $\xrightarrow[p=68]{\text{r_raff_sem \#1}}$ avocat>justice
— avocat $\xrightarrow[p=35]{\text{r_raff_sem \#1}}$ avocat>justice>individu
— avocat>justice $\xrightarrow[p=35]{\text{r_raff_sem \#1}}$ avocat>justice>individu
— avocat $\xrightarrow[p=26]{\text{r_raff_sem \#1}}$ avocat>justice>métier
— avocat>justice $\xrightarrow[p=26]{\text{r_raff_sem \#1}}$ avocat>justice>métier

Exemple 2.2 – Exemple d'imbrication de raffinements sémantiques pour le mot « avocat » tirés de rezoJDM. Requête en utilisant l'outil rezo accessible sur le lien suivant : Lien Rezo. Le premier terme avant la flèche est le terme source et celui d'après est le terme cible dans une relation. La flèche représente le type de la relation (ex : r_raff_sem) avec son identifiant précédé par un # (ex : #1) et le poids de la relation renseigné par p (ex : p=42).

— _INFO_SEM_ABSTRACT : entité abstraite
— _INFO_SEM_PERS : personne
— _INFO_SEM_ALIVE : entité vivante
— _INFO_SEM_PLACE : lieu
  — _INFO_SEM_PLACE_HUMAIN : lieu humain
  — _INFO_SEM_PLACE_GEO : lieu géographique

Exemple 2.3 – Exemples de nœuds d'information sémantique dans rezoJDM

suivantes : <$chaine_1$, $chaine_2$, type, poids> (Lafourcade & Brun, 2020). Les deux éléments ***chaine***$_1$ *et* ***chaine***$_2$ *représentent respectivement le nœud source et cible de la relation. Il s'agit ici des labels des nœuds. Comme explicité précédemment, il est représenté par une chaîne de caractères ou un segment textuel qui peut être un terme ou une autre information lexicale en fonction des différents types. Le* ***type*** *d'une relation permet de donner un nom identifiant la relation dans un ensemble de types fini*[4]*. Par exemple, la relation générique ou d'hyperonymie (« est un type de ») est représentée par le symbole « r_isa ». Afin de nommer l'inverse d'une relation donnée, le nom contient « -1 » à la fin, dans le cas où le type n'est pas connu. Par exemple, la relation « r_carac » possède la relation opposée « r_carac-1 », tandis que la relation « r_isa » a pour relation opposée « r_hypo », car l'hyponymie est un opposé reconnu dans la littérature (nom attribué au type). Il est à noter que les noms des types ont souvent des alias. Ainsi, une même relation peut avoir plusieurs noms différents. Par exemple, « r_conseq » et « r_has_conseq » font tous deux référence au même type de relation. Le tableau 2.2 présente les principaux types de relations auxquels nous nous intéressons, avec leurs noms et consignes associés. Le* ***poids*** *est un réel positif ou négatif représentant la pertinence du lien entre les deux termes pour un type donné.*

Comme explicité précédemment, les poids des relations du réseau sont mis à jour par le jeu des utilisateurs. Le poids est initialisé à zéro puis est incrémenté si la relation est validée par plusieurs joueurs. Il est important de conserver tous les poids du réseau car dans le cas où la relation détient un poids négatif, cela signifie qu'un traitement a été effectué de façon humaine ou algorithmique et n'est plus à refaire (ce qui est coûteux au niveau com-

4. Le détail de tous les types de relations de rezoJDM est accessible ici : `https://www.jeuxdemots.org/jdm-about-detail-relations.php`

Type de relation	Nom identifiant la relation	Consigne de jeu associé
Idées associées	r_associated	Donner des idées associées au terme cible
Thèmes ou domaines	r_domain	Donner des noms de domaines relatifs au terme cible (par exemple, rougeole relève du domaine de la médecine)
Synonymes	r_syn	Donner des synonymes du terme cible
Contraires	r_anto	Donner des contraires (antonymes) du terme cible
Génériques	r_isa	Donner des génériques (hyperonymes) du terme cible (par exemple animal et mammifère sont des génériques de chat)
Spécifiques	r_hypo	Donner des spécifiques (hyponymes) du terme cible (par exemple, mouche, abeille, moustique, sont des hyponymes d'insecte)
Parties, constituants ou éléments	r_has_part	Donner des parties (méronymes) du terme cible (par exemple, roues, volant, siège, moteur, pour automobile)
Caractéristique et Caractéristique-1 (relation inverse de Caractéristique)	r_carac / r_carac-1	Donner des caractéristiques typiques du terme cible (par exemple, froide, chaude, liquide, limpide , pour eau) / Donner des termes qui peuvent avoir la caractéristique-cible (par exemple, original peut caractériser un dessert, un projet, un homme. . .)

TABLE 2.2 – Principaux types de relations avec leur nom et consignes associés. Liste non-exhaustive pour limiter la surcharge et ne présenter que les exemples intéressants pour notre cas d'usage mais le détail complet est disponible sur la page JeuxDeMots : `https://www.jeuxdemots.org/jdm-about-detail-relations.php`

putationnel ou humain). Cela peut être utile lors d'une analyse des liens entre les mots d'un texte. Il peut s'agir d'informations négatives tout comme des erreurs d'inférence créées par le système. De plus, un poids négatif permet de représenter des exceptions. Par exemple, le manchot est un oiseau et pourtant il ne vole pas. L'exemple 2.4 présente l'exception précédente sous forme de relations. La construction des relations négatives se manifeste de trois manières.

1. La première est construite à partir d'erreurs dans le graphe. En effet, il s'agit d'une relation avec un poids positif mais fausse en réalité. Manuellement, son poids est ainsi multiplié par (-1).
2. La deuxième est intégrée volontairement avec un poids négatif comme pour l'exemple précédent (c.f. Exemple 2.4).
3. La dernière permet de modéliser une relation dont l'un des mots composant un terme ne fait pas référence à la signification du mot. Par exemple, un « scorpion de mer » n'est pas un scorpion mais un poisson. L'exemple 2.5 illustre cette exception sous forme de relations.

— voler>(déplacement aérien) $\xrightarrow[p<0]{\text{r_agent \#13}}$ manchot

— manchot>(oiseau de mer) $\xrightarrow[p<0]{\text{r_agent-1 \#24}}$ voler

Exemple 2.4 – Représentation de la relation « *le manchot ne vole pas* ».

— scorpion de mer $\xrightarrow[-248]{\text{r_isa \#6}}$ scorpion

— scorpion $\xrightarrow[-219]{\text{r_hypo \#8}}$ scorpion de mer

— scorpion de mer $\xrightarrow[59]{\text{r_isa \#6}}$ poisson

Exemple 2.5 – Représentation de la relation « *Le scorpion de mer n'est pas un scorpion* ».

Les relations peuvent être d'ordre lexical, syntaxique ou sémantique (ontologique, associative, prédicative ou de typicalité). Les relations dites **lexicales** (synonymie, antonymie, locution, famille) portent sur les termes, leurs dérivés, synonymes ou encore leurs sens contraires. Les relations **syntaxiques** donnent la syntaxe d'un terme comme sa nature morpho-syntaxique, sa classe grammaticale (POS) ou encore son lemme. Les relations **ontologiques** (ex : générique comme l'hyperonymie, spécifique comme l'hyponymie ou encore « partie de » comme la méronymie) portent sur des connaissances liées à des objets du monde. Les relations **associatives** (ex : association libre, sentiment associé, signification) s'apparentent à de la connaissance subjective et/ou globale. Les relations **prédicatives** (ex : agent, patient, instrument) sont associées à un verbe ou un prédicat et aux valeurs de ses arguments au sens large. Enfin, les relations de **typicalité** représentent des lieux, des moments ou encore des caractéristiques typiques. L'ensemble de ces relations peut se catégoriser de deux façons différentes. Il peut s'agir de relations bien définies telles que la synonymie ou l'équivalence, la notion de hiérarchie, voire d'association. Des relations implicites tirées du texte peuvent aussi être pertinentes, comme un verbe ou un nom qui lie deux entités entre elles dans un texte.

La figure 2.6 décrit des exemples tirés d'une requête sur rezoJDM concernant le terme « Code Civil ». Il peut s'agir de relations entrantes (le terme est la cible de la relation) ou sortantes (le terme est la source de la relation). Dans l'exemple donné, le premier terme avant la flèche désigne le terme source, et celui après la flèche est le terme cible dans une relation. La flèche représente la relation par son type (ex : *r_associated*), son identifiant précédé par un # et enfin son poids renseigné par *p*.

— Code civil $\xrightarrow[p=80]{\text{r_associated \#0}}$ Texte de loi

— Décret $\xrightarrow[p=78]{\text{r_lieu \#3}}$ Code civil

— Code civil $\xrightarrow[p=80]{\text{r_pos \#4}}$ Nom :

— Code civil $\xrightarrow[p=25]{\text{r_isa \#6}}$ Code juridique

— Code civil $\xrightarrow[p=30]{\text{r_domain \#3}}$ Droit (juridique)

Exemple 2.6 – Liste de différents exemples de relations tirés de rezoJDM. Requête réalisée sur le terme « code civil » en utilisant l'outil rezo accessible sur le lien suivant : Lien Rezo

Une forme un peu particulière de relations existe. Il s'agit des **annotations**. En effet, il est possible d'ajouter des annotations aux relations. Une annotation est une relation qui relie une relation avec un terme du réseau. Le nœud terme représentant l'annotation peut être n'importe quel type de mot. En effet, nous pouvons annoter une relation avec n'importe quel mot. L'intérêt est de pouvoir annoter avec des mots du type « fréquent », « possible », « factice », *etc.* Il est à noter que si une relation est annotée par « non pertinente », cela signifie que la relation est certes vraie mais n'apporte aucune information pertinente.

Les annotations sont intéressantes à utiliser dans le cadre de l'inférence. Lorsque des inférences déductives sont effectuées, les prémisses annotées comme « non pertinentes » peuvent amener à ne pas sélectionner cette inférence et à opter pour une autre. Par exemple, le fait qu'un animal est composé d'atomes pourrait conduire à déduire cette même relation pour tous les hyponymes du terme « animal ». Or, cette déduction ne présente pas d'intérêt. Annoter la relation comme « non pertinente » empêche donc de saturer la base de connaissances ou d'introduire un biais dans le graphe construit.

2.3 Conclusion

Dans ce chapitre, nous avons établi les bases fondamentales du concept de graphe de connaissances lexico-sémantique, en mettant l'accent sur la définition, la structure, et les relations spécifiques qui le constituent. Le graphe de connaissances lexico-sémantique, inspiré du réseau lexico-sémantique décrit par Lafourcade (Lafourcade, 2007, 2011), se caractérise par des nœuds représentant des segments textuels et des relations définissant les liens sémantiques entre ces segments textuels.

Un graphe de connaissances est présenté comme une base de connaissances sous forme de graphe, composée de nœuds et de relations. Les nœuds représentent des faits ou des connaissances sous forme de termes simples ou composés, tandis que les relations englobent des propriétés lexicales et sémantiques telles que les relations ontologiques, associatives, prédicatives, et de typicalité. Le graphe de connaissances lexico-sémantique vise à structurer

et exploiter la connaissance textuelle, ce qui est primordial pour diverses applications telles que l'analyse sémantique de textes, la classification de documents, et l'extraction d'informations. Les inférences et la mise à jour des relations permettent de maintenir la pertinence et la cohérence des informations contenues dans le graphe, rendant ce dernier adaptable et évolutif pour des usages spécifiques. Le chapitre introduit et détaille l'exploitation du réseau lexico-sémantique issu de JeuxDeMots (rezoJDM). Cette ressource peut être utilisée pour l'analyse textuelle et ainsi enrichir le texte en fournissant une structure orientée de relations lexicales et sémantiques pondérées et annotées, facilitant ainsi l'exploitation et le parcours heuristique du graphe

Les bases théoriques et la définition du graphe de connaissances lexico-sémantique posées dans ce chapitre servent de fondation pour les chapitres suivants. Le chapitre suivant aborde la construction automatique de bases de connaissances à partir de ressources textuelles, mettant en application les concepts et structures définis dans ce chapitre. La méthodologie et les techniques d'extraction d'information détaillées dans les chapitres suivants s'appuient sur la construction des nœuds et des relations présentées ici, renforçant ainsi la cohérence et la continuité de l'étude.

En résumé, ce chapitre établit les éléments essentiels pour la conception et l'implémentation d'un graphe de connaissances lexico-sémantique, servant de socle théorique et pratique pour les développements et applications ultérieurs explorés dans cette thèse.

CHAPITRE 3

État de l'art : Construction automatique de bases de connaissances à partir de ressources textuelles

Après avoir introduit les concepts fondamentaux et le cadre théorique de cette thèse, ce chapitre fournit une vue d'ensemble de l'état de l'art en matière de construction automatique de bases de connaissances à partir de ressources textuelles. Il s'agit de comprendre les méthodes et approches existantes, ainsi que leurs avantages et limitations. Nous nous intéressons à explorer et à détailler un état de l'art des solutions existantes pour répondre au problème de la construction automatique de graphes à partir de textes spécialisés du secteur public en français. Il s'agit d'ouvrages et d'articles juridiques et pratiques de divers domaines. Les défis incluent la comparaison des différentes techniques, l'identification des lacunes dans la recherche actuelle, et la mise en perspective des contributions potentielles de cette thèse.

Sommaire du chapitre

3.1 Définitions et différentes structures de graphe **19**
3.2 Construction de graphes à partir de textes **22**
3.3 Extraction d'information **25**
3.3.1 Extraction terminologique 25
3.3.2 Extraction de relations 52
3.3.3 Open Information Extraction 60
3.4 Classification de documents **65**
3.5 Contributions : Résumé des choix d'orientation et d'implémentation **68**

3.1 Définitions et différentes structures de graphe

La modélisation de l'information au niveau sémantique est l'un des principaux objectifs de l'utilisation des structures de graphe, telles que les ontologies, par exemple (Guarino et al., 2009). La définition originale de l'**ontologie** dans le domaine de l'informatique a été

donnée par Gruber (1995), la définissant comme une « spécification explicite d'une conceptualisation », et a été raffinée plus tard par Vrandečić (2009). Au fil du temps, la définition d'ontologie a évolué. L'ontologie, selon Borst (1997), se définit comme une « spécification formelle d'une conceptualisation partagée ». Cette définition est renforcée par Taniar & Rahayu (2006), qui la décrivent comme une « conceptualisation d'un domaine de connaissance dans un format informatiquement exploitable qui modélise les entités, les attributs et les axiomes ». Fondamentalement, une ontologie se compose d'un vocabulaire et des relations entre les concepts. Selon Antoniou & Harmelen (2004), les ontologies comprennent des propriétés de concepts, des déclarations de disjonction, des restrictions de valeur et la spécification de relations logiques entre les objets. En fin de compte, l'ontologie est un outil précieux pour formaliser la structure d'un système en se basant sur les relations qui émergent de son observation.

En résumé, une ontologie se présente comme une structure formelle complexe, constituée d'un ensemble d'axiomes logiques définissant des relations entre des concepts formels. Son idée fondamentale réside dans la capacité à encoder la connaissance, permettant ainsi de la rendre accessible et compréhensible aux machines telles que les ordinateurs ou les processus de calcul. Plus formellement, l'ontologie constitue une formalisation précise d'un ensemble partagé de termes destinés à décrire et à représenter un domaine spécifique de connaissances. Elle agit comme un outil fondamental pour délimiter et structurer ce domaine, en fournissant une base sémantique solide pour la communication et la compréhension. Au cœur de cette formalisation résident les termes eux-mêmes, organisés en classes et interconnectés par des relations qui reflètent les liens conceptuels inhérents au domaine étudié. Outre ces relations de base, les ontologies peuvent également intégrer des propriétés, permettant ainsi de décrire en détail les caractéristiques et les attributs des concepts, voire de les associer de manière plus complexe. Ces ontologies englobent donc non seulement des classes et des relations, mais aussi des règles logiques formelles et des instances de classe, ainsi que les relations et propriétés qui les relient. En somme, l'ontologie ne se limite pas à une simple structure de classification ; elle s'étend à l'instanciation et au raisonnement, offrant ainsi un cadre complet pour la représentation et l'exploitation de la connaissance dans un domaine donné.

L'ontologie et la **taxonomie** sont deux approches distinctes dans l'organisation des connaissances. La taxonomie est la science de la classification des choses ou des concepts, caractérisée par l'inclusion de relations hiérarchiques dans ses classifications. Elle se concentre sur la création de systèmes de catégorisation où les termes ou concepts sont organisés dans une hiérarchie, souvent selon leurs similitudes ou leurs attributs communs. Son objectif principal est de classifier les éléments dans des catégories discernables. En revanche, l'ontologie va au-delà de la simple classification en introduisant une structure formelle complexe basée sur des axiomes logiques définissant les relations entre les concepts formels. Elle vise à capturer la signification des concepts et leurs interrelations de manière plus précise et exhaustive, permettant une représentation sémantique plus profonde des connaissances. Ainsi, tandis que la taxonomie se concentre sur la classification, l'ontologie vise à modéliser la réalité dans sa complexité en fournissant un cadre conceptuel riche pour la représentation et la compréhension des connaissances.

L'ontologie et le **thésaurus** sont tous deux des outils utilisés dans la gestion des connaissances, mais chacun a des objectifs et des fonctionnalités distincts. Le thésaurus peut être considéré comme une extension des taxonomies, offrant une organisation hiérarchique des termes ou concepts tout en ajoutant la possibilité de définir d'autres descriptions à ces termes, comme des libellés, des traductions, des synonymes et d'autres éléments de définition. Il vise principalement à indexer du contenu en associant des mots-clés pertinents et

à faciliter l'accès à des contenus spécifiques en utilisant une terminologie contrôlée. En revanche, l'ontologie va au-delà de cette fonctionnalité en introduisant une structure formelle complexe basée sur des axiomes logiques. Elle modélise les concepts et leurs relations de manière plus précise et exhaustive, permettant une représentation sémantique plus riche des connaissances. Alors que le thésaurus est principalement axé sur l'indexation et la recherche de contenu, l'ontologie vise à fournir un cadre conceptuel robuste pour la représentation et la compréhension des connaissances dans divers domaines d'application.

L'ontologie et le **graphe de connaissances** représentent deux approches distinctes mais complémentaires dans la gestion et la représentation des connaissances. L'ontologie se concentre sur la formalisation précise d'un ensemble de termes et de relations dans un domaine spécifique, fournissant ainsi un cadre sémantique pour décrire et structurer la connaissance de manière logique et explicite. Elle met l'accent sur la définition formelle des concepts et de leurs interrelations à travers des axiomes logiques, ce qui facilite la manipulation et l'exploitation de la connaissance par des machines ou des processus informatiques. En revanche, le graphe de connaissances adopte une approche plus souple et interconnectée, où les connaissances sont représentées sous forme de nœuds et de liens dans un réseau complexe. Ce réseau permet de capturer la richesse des relations entre les entités de manière plus dynamique, en intégrant souvent des données provenant de différentes sources et en reflétant la nature interconnectée du savoir humain. Alors que l'ontologie se concentre sur la formalisation et la structure logique, le graphe de connaissances met en avant la représentation contextuelle et la flexibilité, offrant ainsi une approche plus holistique pour organiser et exploiter la connaissance dans des contextes variés.

Comme exprimé dans les chapitres précédents, nous nous intéressons aux graphes de connaissances car il s'agit d'une approche plus flexible et adaptable, capable de gérer des données non structurées et hétérogènes tout en capturant efficacement les relations complexes entre les entités présentées dans les textes ciblés. Il est également important de noter que le choix entre un graphe de connaissances et une ontologie dépendra toujours des besoins spécifiques de la tâche, et la construction d'ontologies à partir de textes a été profondément explorée au sein de la communauté de l'extraction d'informations.

La diversité des structures de connaissances englobe plusieurs catégories :

— Les structures de connaissances génériques ou de haut niveau : elles fournissent une abstraction universelle en décrivant des concepts fondamentaux, souvent en se basant sur des perspectives philosophiques ou logiques plutôt qu'applicatives. Ces structures sont conçues pour être réutilisables dans différents contextes.

— Les structures de connaissances spécifiques à un domaine : celles-ci se limitent à des concepts et des relations pertinents à un domaine particulier. Par exemple, la *Gene Ontology*[1] dans le domaine biologique ou une ontologie juridique dans le domaine du droit. Elles offrent une modélisation détaillée du monde dans des contextes d'application spécifiques.

— Les structures de représentation : elles visent à décrire des méta-structures génériques, en se concentrant sur des aspects logiques plutôt que sur des domaines spécifiques. *Ontolingua*[2] en est un exemple, fournissant un cadre pour la représentation sémantique de connaissances à un niveau abstrait.

— Les structures d'application : ces structures combinent souvent une ontologie de domaine avec des composantes spécifiques à une application donnée, permettant ainsi une représentation précise et fonctionnelle de la connaissance dans un contexte applicatif particulier.

1. Gene ontology : `https://www.geneontology.org/`
2. Ontolingua : `https://ksi.cpsc.ucalgary.ca/KAW/KAW96/farquhar/`

— Et bien d'autres encore, chacune avec ses propres caractéristiques et avantages selon les besoins spécifiques de la modélisation de la connaissance et de l'application envisagée.

Les structures de connaissances, au-delà de leur utilisation dans les ontologies, sont largement impliquées dans une gamme variée de domaines et d'applications. Voici quelques exemples significatifs :

— **Sciences de la vie** : Les structures de connaissances sont cruciales pour l'intégration de nouvelles données et la vérification des connaissances dans des domaines comme la médecine. Des ontologies médicales telles que GALEN sont des exemples notables.
— **Recherche d'informations** : Les moteurs de recherche sémantiques utilisent des structures de connaissances pour améliorer la précision et la pertinence des résultats. Ils exploitent des ontologies pour comprendre le sens des requêtes et des documents, facilitant ainsi une recherche plus contextuelle et précise (Panovski & Titus, 2015; Reymonet et al., 2010; Vallet et al., 2007).
— **Domaine juridique** : Les structures de connaissances sont utilisées pour modéliser les concepts et les relations dans le domaine juridique, facilitant la recherche, l'analyse et la compréhension des textes juridiques complexes (El Ghosh et al., 2017; Ma et al., 2018; G et al., 2019; Castano et al., 2022).
— **Systèmes de questions-réponses** (*QA System : Question Answering System*) : Ces systèmes exploitent des structures de connaissances pour comprendre et répondre aux questions posées par les utilisateurs, en s'appuyant sur des ontologies spécifiques au domaine (par exemple, (Emani & Haralambous, 2017) ont mis un système de questions-réponses dans le domaine du légal) pour extraire et présenter des informations pertinentes (Emani & Haralambous, 2018; Rani et al., 2022).
— **Systèmes d'aide à la décision** : Les structures de connaissances sont utilisées pour représenter les connaissances et les règles de décision dans les systèmes d'aide à la décision, facilitant ainsi l'analyse et la prise de décision dans des contextes complexes (Sanchez-Pi et al., 2016).
— **Classification de texte** : Les structures de connaissances sont employées pour classer et organiser automatiquement de vastes ensembles de texte, améliorant ainsi la recherche et la navigation dans des corpus de documents variés (Sanchez-Pi et al., 2016; Elhadad et al., 2017).
— **Recommandation de contenu personnalisé** : Les structures de connaissances sont utilisées pour modéliser les préférences et les intérêts des utilisateurs, permettant ainsi de recommander du contenu adapté à leurs besoins spécifiques (Vallet et al., 2007; Nilashi et al., 2018).
— **Extraction d'information** : Les structures de connaissances sont essentielles pour extraire et organiser des informations pertinentes à partir de sources de données hétérogènes, facilitant ainsi la recherche et l'analyse d'informations dans divers domaines (Wimalasuriya & Dejing Dou, 2010; Mannai et al., 2018; Hogan et al., 2022).

Ces exemples illustrent la polyvalence et l'importance des structures de connaissances dans de nombreux domaines et applications, offrant des outils puissants pour la représentation, l'organisation et l'exploitation efficaces de la connaissance.

3.2 Construction de graphes à partir de textes

Dans cet état de l'art, nous nous intéressons à la construction automatique d'une structure de données. En effet, il peut s'agir d'un graphe de connaissances mais aussi d'autres

types de structures comme les ontologies principalement. Il est vrai que les deux types de structures sont différents mais les méthodologies de construction se ressemblent et l'état de l'art concernant la construction automatique d'une base ontologique à partir de textes a permis de proposer une méthodologie avec un ensemble de méthodes et de techniques pour la construction des ontologies à partir de textes (Asim et al., 2018). L'ensemble du processus d'acquisition des ontologies se divise en plusieurs étapes, et est connu sous le nom de mille-feuille d'apprentissage de l'ontologie (*Ontology Learning Layer Cake*), initialement proposé par Buitelaar et al. (2003), et repris par Cimiano et al. (2005).

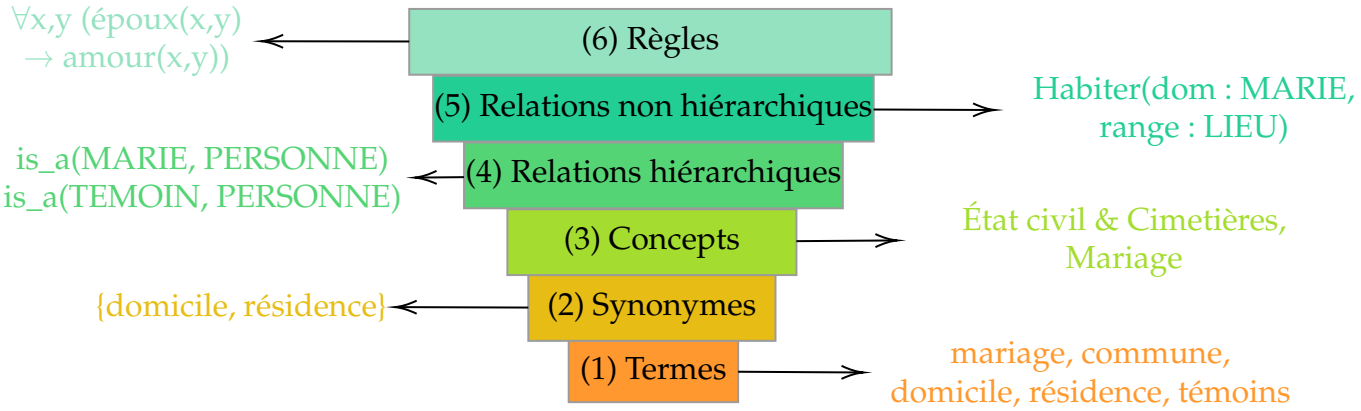

FIGURE 3.1 – Millefeuille d'apprentissage

La figure 3.1 présente le schéma du mille-feuille d'apprentissage de l'ontologie. Le modèle de représentation connu sous le nom de mille-feuille est un diagramme en couches conçu pour illustrer les relations conceptuelles inhérentes aux diverses étapes de l'apprentissage d'ontologies. Chaque strate de ce schéma repose sur les étapes précédentes, où les résultats des tâches effectuées dans les couches inférieures servent généralement de données d'entrée pour les couches supérieures. Le processus commence par l'identification des termes-clés et de leurs synonymes (étapes 1 et 2), suivie de la combinaison de ces termes et synonymes pour former des concepts (étape 3). Ensuite, des relations, tant taxonomiques que non-taxonomiques, entre ces concepts sont extraites (étapes 4 et 5). Enfin, l'instanciation des schémas d'axiomes et la déduction des axiomes généraux constituent la dernière étape de ce processus complexe (étape 6). Ce modèle séquentiel offre une méthodologie structurée pour l'acquisition progressive de connaissances ontologiques, facilitant ainsi la modélisation conceptuelle et la représentation sémantique dans le domaine de l'intelligence artificielle.

L'état de l'art de Khorashadizadeh et al. (2024) explore la relation synergique entre les larges modèles de langue (LLM) et les graphes de connaissances (KG). Il est question des rôles significatifs des LLMs, comme la série GPT d'OpenAI[3], dans la génération de texte ressemblant à celui des humains, la réponse à des questions et la création de contenu, tandis que les KGs organisent et intègrent l'information de manière structurée, permettant aux machines de comprendre et d'inférer des relations. La combinaison des LLMs et des KGs ouvre de nouvelles frontières pour les applications du TALN en améliorant la capacité des machines à traiter, interpréter et générer des informations avec contexte et précision. L'état de l'art vise à répondre à plusieurs questions de recherche et notamment « Comment les LLMs peuvent-ils être utilisés pour la génération d'ontologies ? », « Comment les LLMs peuvent-ils aider à détecter des incohérences au sein des KGs ? » et « Comment les LLMs peuvent-ils améliorer la précision, la cohérence et l'exhaustivité des KGs par la vérification des faits ? ».

3. Lien vers le site d'OpenAI avec l'ensemble des produits : `https://openai.com/`

Les LLM sont utilisés pour améliorer et affiner les fonctionnalités des KG. Et notamment, l'état de l'art explore le rôle des LLMs dans la création et l'enrichissement d'ontologies pour les KGs, essentielles pour la récupération d'informations et le raisonnement. Il souligne le potentiel transformateur de l'intégration des LLMs avec les KGs et suggère que les recherches futures devraient se concentrer sur les lacunes et les défis identifiés, tels que l'intensité computationnelle, les biais dans les LLMs, et une évaluation plus large des méthodes de génération d'ontologies.

L'article de Şükrü Demir İnan Özer et al. (2024) présente une étude expérimentale comparative de méthodes de prédiction de liens adaptées aux réseaux bipartites. L'article propose une classification des méthodes de prédiction de liens spécifiquement conçues pour les réseaux bipartites. Cette taxonomie est accompagnée de représentations visuelles pour aider les lecteurs à comprendre les différences entre ces méthodes. Les auteurs adaptent des modèles de recommandation personnalisée, basés sur les réseaux de neurones convolutionnels (GCN), pour la prédiction de liens dans les réseaux bipartites. Ces modèles, initialement conçus pour des systèmes de recommandation, montrent des résultats prometteurs dans ce nouveau contexte. L'article étudie l'efficacité de l'utilisation des mesures de centralité des nœuds, telles que le degré, la centralité de proximité, et la centralité de vecteur propre, comme caractéristiques pour l'entraînement de modèles supervisés de prédiction de liens. Une comparaison expérimentale de 19 méthodes de prédiction de liens est réalisée sur trois jeux de données réels avec des tailles et des densités variées. Cette évaluation permet de déterminer quelles méthodes sont les plus efficaces dans différents contextes de réseaux bipartites.

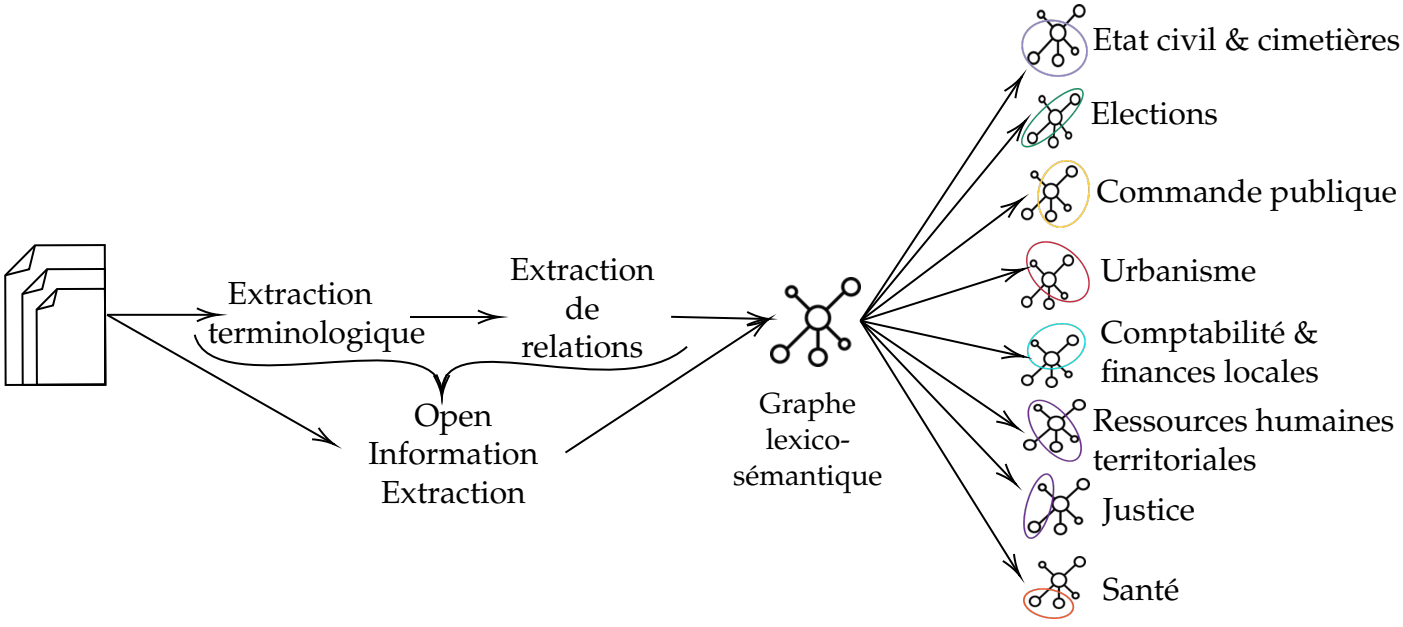

FIGURE 3.2 – Schéma de conceptualisation de l'état de l'art : Du corpus jusqu'au graphe lexico-sémantique.

La figure 3.2 présente le schéma expliquant le cheminement de l'état de l'art. En effet, afin de construire un corpus de textes en graphe lexico-sémantique, deux étapes ont été relevées par l'état de l'art : soit l'*Open Information Extraction* (OpenIE), soit une étape d'extraction terminologique suivie d'une extraction de relations. L'OpenIE extrait à la fois des termes appelés arguments et les relations. Enfin, ce graphe peut être filtré en fonction des différents domaines des données à analyser.

Ce chapitre s'intéresse à l'extraction d'informations pour la construction de graphes à partir de textes.

3.3 Extraction d'information

L'extraction d'information (*Information Extraction* - IE) sert de base fondamentale pour des tâches en aval telles que la construction de graphes de connaissances (*knowledge graph construction*) (Zhong et al., 2023), le raisonnement sur les connaissances (*knowledge reasoning*) (Fu et al., 2019) et la réponse aux questions (*question answering*) (Srihari & Li, 1999).

Definition 3.1 (Extraction d'information)**.** *L'extraction d'information est une tâche automatique qui consiste à extraire des informations structurées à partir de documents non structurés dans un domaine restreint. Il s'agit donc de convertir du texte brut en une forme structurée de connaissances.*

L'extraction d'information à partir d'un texte brut vers des formats de données structurées inclut trois types de sous-tâches : l'extraction de triplets entité-relation-entité (RE), l'extraction de termes clés et plus précisément la reconnaissance d'entités nommées (*named entity recognition - NER*) et, enfin, l'Extraction d'Évènement (EE) dont nous ne parlerons pas dans ce manuscrit (Wang et al., 2023b). La figure 3.1 présente un exemple d'extraction d'information. À partir d'un texte issu des rapports du fil d'actualité sur les fusions d'entreprises, sont extraits deux types d'informations : la relation « FusionEntre » ainsi que les entités qui composent cette relation : les entreprises (extraction terminologique).

« Hier, la société new-yorkaise Foo Inc. a annoncé l'acquisition de Bar Corp. »

Informations extraites : FusionEntre($entreprise_1$, $entreprise_2$, *date*)
$entreprise_1$ = Foo Inc.
$entreprise_2$ = Bar Corp.
date dépend de la date de l'article

Exemple 3.1 – Exemples d'extraction d'information d'un texte issu des rapports du fil d'actualité sur les fusions d'entreprises.

Nous nous intéressons donc aux deux sous catégories de l'extraction d'informations à savoir l'extraction terminologique et l'extraction de relations.

3.3.1 Extraction terminologique

L'extraction terminologique consiste à extraire automatiquement les termes pertinents d'un corpus de textes donné. Parfois, dans l'état de l'art, il est mentionné l'identification de la terminologie comme étant une extraction des termes pertinents pour caractériser des concepts, puis leur organisation en groupes de mots synonymes. Une définition plus explicite est donnée dans le chapitre associé (chapitre 4 sur l'extraction terminologique).

Le travail mené par Balachandran & Ranathunga (2016) s'intéresse à l'extraction terminologique dans le cadre de la construction d'ontologie. Il note que de nombreuses recherches existantes considèrent exclusivement les noms comme des termes simples et les mots-clés comme étant des mots complexes de 2-grams exclusivement (nom + adjectif, par exemple). Cependant, d'autres instances sont possibles, telles que les N$\geqslant$3-grams (par exemple, nom + adjectif + déterminant + nom ou adjectif + nom + adjectif), qui peuvent également être des

termes complexes et pertinents. Une des techniques d'extraction terminologique les plus simples consiste à compter la fréquence des termes dans un ensemble donné de documents (par exemple, *term frequency–inverse document frequency* - TF-IDF (Salton & Buckley, 1988; Rajaraman & Ullman, 2011)). Ces systèmes reposent sur l'hypothèse que si un terme est fréquent dans un ensemble de textes spécifiques au domaine, cela indique qu'il est probablement pertinent. Cependant, certaines techniques se sont avérées potentiellement plus efficaces. Nous pouvons diviser l'extraction terminologique en deux sous-catégories : l'extraction des entités nommées et l'extraction de termes pertinents. Leurs définitions sont également données dans le chapitre 4.

Représentation terminologique

Les représentations terminologiques sont généralement constituées des termes d'un ou plusieurs corpus, indiqués par leurs indices associés à leur place dans un dictionnaire. Les représentations de mots habituellement utilisées sont l'encodage One-Hot (Harris & Harris, 2022) (*One-Hot encoding*), le *Count Vectorizer* (de *Scikit-Learn* (Pedregosa et al., 2011)), le *Term Frequency - Inverse Document Frequency* (TF-IDF) (Rajaraman & Ullman, 2011), et les méthodes Word2vec (Mikolov et al., 2013) (skip-gram ou Continuous *Bag of Words*). L'encodage One-Hot dans la Reconnaissance des entités nommées (*Named Entity Recognition*) utilise la construction automatique à partir de données d'apprentissage, par exemple à partir d'applications de messagerie de médias sociaux (Lee & Ko, 2020). Cette méthode est très facile à comprendre et à utiliser. Cependant, elle ne peut pas hiérarchiser les mots en fonction de leur importance et peut entraîner la formation d'une matrice peu dense qui est coûteuse en termes de calcul et de mémoire.

Le *Count Vectorizer* convertit une collection de documents textuels en une matrice de fréquence d'apparition d'un *token* ou terme dans un document. Le *Term Frequency - Inverse Document Frequency* (TF-IDF) (Upendraa & Sudheer, 2016; Rajaraman & Ullman, 2011) est une mesure de l'importance d'un mot pour un document dans une collection ou un corpus, ajustée pour tenir compte du fait que certains mots apparaissent plus fréquemment en général.

Word2Vec est une classe de réseaux neuronaux peu profonds à deux couches, formés pour récupérer des contextes de mots à partir de données linguistiques. Un vaste corpus de mots est utilisé pour récupérer des paires de mots contexte/cible pour l'apprentissage, où le mot cible est un mot contexte, et les termes sélectionnés de manière aléatoire se trouvent dans une fenêtre spécifiée entourant le mot cible (Mikolov et al., 2013). Dans le cas d'un Word2Vec CBOW, les mots du contexte sont introduits dans le modèle sous leur forme d'intégration (d'abord initialisée de manière aléatoire), et le modèle est anticipé pour produire le mot cible à l'aide de *softmax*. Le skip-gram (Hsieh et al., 2017) fait exactement l'inverse : il faut produire les mots contextuels en même temps que le mot cible.

Extraction des entités nommées

Définition : Token classification ou sequence labeling L'extraction d'entités nommées consiste à tagger des entités dans le texte avec l'étiquette associée. Nous pouvons donc distinguer deux sous-tâches : l'identification d'entités (*Entity Identification*) et l'attribution d'étiquettes aux entités (*Entity Typing*). Par exemple, dans un texte, nous allons identifier en premier lieu l'entité « Dadju Djuna » et dans un second temps, nous allons lui attribuer l'étiquette « PERSONNE ». Il est important d'extraire les entités nommées d'un texte car elles représentent plus de 20 % des termes clés du corpus. En effet, nous constatons bien souvent une grande présence de noms propres dans les textes.

Les travaux de l'état de l'art ont conduit à formuler le problème de la NER comme un problème d'étiquetage de séquence (*sequence labeling*), plus connu sous le nom de *sequence tagging* ou *token classification* (Hochreiter & Schmidhuber, 1997; Sang & Meulder, 2003; Yang & Katiyar, 2020; Devlin et al., 2019a; Xia et al., 2023). Selon la revue de Jehangir et al. (2023a), nous pouvons diviser l'état de l'art de la NER en trois grandes catégories : (1) NER utilisant des approches basées sur des règles ; (2) NER à l'aide de l'apprentissage supervisé ; (3) NER utilisant l'apprentissage non supervisé ; et (4) NER utilisant le *deep learning* et les modèles de langues (*Small* et *Large Language Models*). Néanmoins, la section sur les différentes orientations à prendre dans l'avenir de cette revue souligne également que de nouveaux modèles et technologies ont émergé et que leur intégration nécessite une exploration. Nous prévoyons également de consacrer une partie aux ensembles de données faiblement annotés, notre corpus spécialisé n'étant pas annoté pour les entités nommées. Nous réserverons également une section pour l'évaluation et les ensembles de données actuellement proposés dans l'état de l'art.

Approches basées sur des règles Ces approches consistent à concevoir un ensemble de règles souvent syntaxiques manuellement pour un domaine bien particulier. L'intérêt de ces approches est d'obtenir des systèmes plus précis et efficaces, mais elles ont l'inconvénient de ne pas être généralisables et nécessitent l'intervention humaine pour la conception des règles. Les expressions régulières consistent à faire correspondre une suite de chaînes de caractères avec des motifs (*patterns*) simples ou complexes. Elles peuvent également être utilisées pour découvrir et récupérer des motifs, ainsi que pour remplacer les motifs correspondants d'une chaîne par d'autres motifs.

Plusieurs systèmes fondés sur des règles ont été développés, chacun présentant des caractéristiques spécifiques. Parmi les premiers, FASTUS (*Finite State Automaton Text Understanding System*) (Hobbs et al., 1993) se distingue en se révélant particulièrement adapté aux tâches d'extraction d'informations plutôt qu'à une interprétation exhaustive du texte. Il s'agit d'un système permettant d'extraire des informations d'un texte libre en anglais qui fonctionne essentiellement comme un ensemble d'automates à états finis non déterministes en cascade. Il fonctionne mieux pour des situations où seule une petite partie du texte est significative, où l'information doit être mise en correspondance avec une représentation cible prédéfinie et simple, et où le public visé par l'auteur n'a pas d'importance.

D'autres systèmes, comme LaSIE (*Large Scale Information Extraction System*) (Humphreys et al., 1998), ainsi que LTG (*Language technology group*) (Mikheev et al., 1998), sont utilisés pour l'extraction d'informations dans divers domaines, offrant des outils tels que des lemmatiseurs, des étiqueteurs, ou toute autre annotation liée au texte, des collecteurs de données et FACILE (Black et al., 1998), un outil de catégorisation multilingue. Ces trois systèmes ont été évalués dans le cadre de la compétition MUC-7 (Chinchor, 1998).

Pour le domaine particulier du médical, Korkontzelos et al. (2015) cherchaient à reconnaître les noms de médicaments, même en l'absence ou avec peu de données d'apprentissage standard annotées manuellement. Pour atteindre cet objectif, l'équipe propose une méthode qui utilise la similarité de chaînes basée sur des motifs de suffixes fréquents des noms de médicaments, tirés de la base de données *Drug Bank*. Cette méthode utilise également des étiquetages d'éléments du dictionnaire qui permettent de créer des expressions régulières (REGEX). Le système drNER (Eftimov et al., 2017) consiste à extraire des informations diététiques basées sur des preuves (*Evidence-based dietary information*). DrNER est une méthode basée sur des règles qui se compose de deux phases. La première concerne la détection et la détermination des entités mentionnées, et la seconde concerne la sélection et l'extraction des entités. L'approche est testée en utilisant des corpus de textes provenant de

différentes sources telles que des sites web scientifiques et des publications de recherche.

Des méthodes hybrides ont été mises en place. En effet, une approche basée sur des règles peut également être utilisée avec d'autres techniques dans le cadre d'un processus en deux étapes. Toute approche basée sur des règles effectue l'étape initiale d'extraction des données, et la classification ultérieure est effectuée par un algorithme supervisé, semi-supervisé ou non supervisé. Par exemple, Sari et al. (2010) ont créé un système hybride afin d'extraire des motifs à partir d'un petit corpus et d'identifier les entités déjà définies dans un ensemble d'enregistrements d'accidents. Cette approche combine l'extraction de motifs (*patterns*) basée sur des règles et une technique de reconnaissance d'entités nommées (NER) semi-supervisée. L'article présente un modèle d'extraction basé sur des règles, qui s'appuie sur l'analyseur syntaxique *Link Grammar* (Grinberg et al., 1995) et le marqueur *Stanford Part-of-Speech* (Toutanova et al., 2003) pour identifier les entités nommées et construire des modèles d'extraction. Ces modèles sont ensuite utilisés dans le cadre d'une approche NER semi-supervisée pour classer les entités dans des catégories spécifiques prédéfinies.

Bien que ces systèmes reposent sur des règles précises établies par des experts maîtrisant la linguistique, la syntaxe et le vocabulaire liés à un domaine spécifique, ils présentent des limites. Il reste intéressant de noter que ces règles sont souvent spécifiques à un seul domaine, ce qui, dans notre cas, nous intéresse puisque nous souhaitons restreindre leur application au domaine du juridique (contextes particuliers). Néanmoins, leur mise en œuvre nécessite une implication humaine pour le développement des règles.

Modèles supervisés Les approches d'apprentissage supervisé sont utilisées pour les données étiquetées et sont entraînées sur ces dernières. Chaque entité dans la séquence textuelle peut être assignée à une étiquette d'entité prédéfinie. Entraîner un système NER supervisé nécessite de nombreuses données d'entraînement étiquetées. Les approches les plus populaires des systèmes NER sont les modèles de Markov caché (HMM) (Zhou & Su, 2002), les champs aléatoires conditionnels (CRF) (Patil et al., 2020), et les machines à vecteurs de support (SVM) (Singh et al., 2009).

Dans le domaine biomédical, Zhang et al. (2004) ont réalisé un NER basé sur les HMMs pour le NER en cascade et le NER des abréviations (*cascaded and abbreviation NER*). Le système proposé intègre plusieurs caractéristiques (morphologie, caractéristiques de déclenchement du nom principal, orthographe et partie de discours) et techniques (y compris deux NER en cascade et une méthode d'algorithme de reconnaissance des abréviations).

Il est utile de s'intéresser à Patil et al. (2020) qui ont mis en place un système pour la catégorisation et l'identification des entités nommées pour la langue régionale marathi. Le marathi étant une langue riche sur le plan morphologique, une méthode d'apprentissage automatique est utilisée pour l'identification et la catégorisation des entités nommées présentes dans la langue. Les techniques statistiques produisent des résultats précis en matière d'identification et de classification des entités nommées. L'utilisation de la CRF permet de localiser les entités nommées dans le système et de les classer ensuite. Les méthodes statistiques permettent d'identifier et de classer avec succès les entités nommées dans cette langue morphologiquement complexe, bien que des informations supplémentaires puissent améliorer la précision. Des études sur le corpus FIRE-2010 démontrent que des connaissances supplémentaires peuvent améliorer la précision. Ces travaux montrent que la construction du corpus doit être prise en compte et que le choix du système doit tenir compte de l'ensemble du corpus, notamment de son domaine et de sa langue.

Nothman et al. (2013) ont construit automatiquement un grand jeu de données d'apprentissage annoté pour la NER. Ce jeu d'entraînement est multilingue et relie des textes aux structures de Wikipédia.

Modèles non supervisés & *Few-Shot NER* L'étiquetage d'un grand corpus de tokens nécessite beaucoup de connaissances, ce qui rend la création d'un tel corpus coûteuse en temps et en argent. De plus, la mise en place de systèmes de NER à grande échelle avec des annotations riches pour différents scénarios d'application dans le monde réel est un travail laborieux. Il existe potentiellement des centaines de nouveaux domaines pour les cas d'utilisation en entreprise, sans parler des différentes langues. C'est pourquoi de nombreux systèmes se sont tournés vers des approches *few-shot* (Huang et al., 2020) ou non-supervisées. Les approches d'apprentissage non supervisé sont utilisées pour les données qui ne sont pas étiquetées. Les approches courantes de l'apprentissage non supervisé sont l'association et le regroupement.

Les avancées significatives dans le domaine de la NER ont conduit à l'émergence de méthodes novatrices, notamment dans le contexte des réseaux sociaux et de la NER biomédicale. Ces méthodes ont également dû faire face à la gestion du décalage de distribution des domaines, ainsi qu'à l'évaluation non supervisée des entités.

Au sein de l'analyse des réseaux sociaux, une méthodologie novatrice a été développée pour surmonter les défis posés par les *tweets* bruyants, courts et insuffisants (Liu & Zhou, 2013). Cette approche particulière englobe une reconnaissance d'entités nommées en deux étapes, débutant par le pré-étiquetage de chaque *tweet* au moyen d'un étiqueteur séquentiel. Les *tweets* sont ensuite regroupés sur la base de leur similarité, et les étiquettes au sein de chaque grappe sont affinées grâce à l'utilisation d'un modèle de champs aléatoires conditionnels amélioré.

Dans le domaine de la NER biomédicale, la méthode de Zhang & Elhadad (2013), adaptée par Śniegula et al. (2019), se distingue par sa capacité à s'adapter aisément à différentes catégories sémantiques et genres de textes. En se basant sur la terminologie, la compréhension syntaxique superficielle et les statistiques du corpus, elle élimine la nécessité d'instances ou de règles d'entités créées manuellement, et démontre son efficacité sur des ensembles de données biomédicales variées.

Peng et al. (2021) ont proposé un modèle interdomaine non supervisé pour la NER, visant à résoudre les problèmes de décalage de la distribution des domaines. En utilisant l'entraînement contradictoire avec une attention consciente des entités, ce modèle a surpassé d'autres approches récentes, comme l'indiquent les résultats expérimentaux sur divers ensembles de données.

Concernant le classement non supervisé des entités, Dutta & Gupta (2022) ont exploré l'impact du bruit en développant un nouvel algorithme basé sur l'estimation de la densité du noyau (*kernel density estimation* - KDE). Leur méthode, qui commence par rassembler une liste de noms avec des caractéristiques basées sur le contenu, illustre l'efficacité du classement non supervisé malgré les défis liés au bruit.

Dans le panorama global des méthodes non supervisées pour la NER, il est crucial de souligner le développement continu de techniques d'apprentissage non supervisé pour les applications de traitement du langage naturel (TALN). Cette approche tire avantage de la construction de modèles exclusivement à partir de données non étiquetées, exploitant ainsi la disponibilité répandue de textes non étiquetés sous forme numérique (Vlachos, 2011). Cependant, l'évaluation de ces méthodes demeure un défi, nécessitant l'exploration de métriques alternatives et d'approches spécifiques pour évaluer leur efficacité dans la reconnaissance d'entités nommées.

Les approches *Few-Shot* consistent à utiliser un petit nombre de données d'apprentissage étiquetées. L'objectif de l'apprentissage *Few-shot* est de permettre aux modèles de généraliser les résultats obtenus de nouveaux échantillons de données inédites sur la base d'un petit nombre d'échantillons fourni au cours du processus d'apprentissage. Nous pouvons séparer

les modèles *Few-shot NER* en deux axes : les centrés modèle et les centrés données (Moscato et al., 2023). Parmi les centrés modèle, deux types se dégagent : l'apprentissage par transfert (*Transfer Learning*) et le *Meta Learning*. Le *transfer learning* est un outil qui n'a pas encore été complètement exploré. Les modèles linguistiques, en particulier, sont fréquemment utilisés pour l'apprentissage par transfert via le pré-entraînement. Leur importance peut être due à plusieurs facteurs : premièrement, il s'agit de modèles non supervisés qui s'entraînent sur des données textuelles brutes, une ressource abondante, et deuxièmement, ils s'entraînent à construire une connaissance intuitive de la syntaxe et de la sémantique. Le *Transfer Learning* désigne l'ensemble des méthodes qui permettent de transférer les connaissances acquises à partir de la résolution de problèmes donnés pour traiter un autre problème. Nous pouvons découper les différentes stratégies du *transfer learning* en trois axes : le *cross-domain*, le *cross-language* ou le *Prompt*.

En ce qui concerne le *cross-domain*, un système basé sur un mécanisme de transfert de dépendance a été implémenté (Hou et al., 2019). Il consiste à apprendre les modèles abstraits de transition d'étiquettes des domaines sources et à généraliser ces modèles dans le domaine cible afin de faciliter la prédiction d'une séquence d'étiquettes. Ce système utilise également un réseau de *matching* de séquences en adaptant le réseau de *matching* au cas d'étiquetage de séquences. Ce système d'étiquetage de séquences *few-shot* est donc basé sur le CRF afin d'intégrer à la fois le mécanisme de transfert de dépendances et le réseau de *matching* de séquences. Un autre système de *cross-domain* s'est intéressé à l'apprentissage par double transfert en tenant compte des étiquettes (*label-aware double transfer learning* : La-DTL) (Wang et al., 2018). Ce système cherche à minimiser la différence entre les représentations des tokens. Pour le *cross-language*, le concept est de chercher des données supplémentaires dans des langues qui sont plus riches. Cette méthode peut causer des problèmes en raison des différences de mots et d'ordre des mots entre les langues. Certains systèmes ont donc cherché à utiliser des *embeddings* multilingues (Xie et al., 2018). Un système pour le transfert multilingue à partir de zéro (*zeroshot*) apprend la qualité du transfert et les erreurs de transfert spécifiques à une langue donnée afin de déduire le meilleur étiquetage dans la langue cible, dans le cadre d'un modèle graphique bayésien. Cela permet de faire des déductions sans supervision explicite en minimisant le bruit (Rahimi et al., 2019).

Dans la même optique, le système de Ni et al. (2017) consiste à projeter des représentations distribuées de mots (*word embeddings*) d'une langue cible vers une langue source, de sorte que le système de NER de la langue source puisse être appliqué à la langue cible sans réentraînement. Les auteurs ont également cherché à créer des données de NER automatiquement étiquetées pour une langue cible via la projection d'annotations sur des corpus comparables, où un schéma heuristique sélectionne efficacement des données de bonne qualité étiquetées par projection à partir de données bruyantes. Un schéma d'apprentissage par transfert a été développé (Cotterell & Duh, 2017) dans lequel des réseaux de neurones CRF au niveau des caractères sont formés pour prédire les entités nommées à la fois pour les langues à ressources élevées et pour les langues à ressources faibles. Pour les langues pauvres en ressources, un système utilise des informations interlinguistiques (Dandapat & Way, 2016). Il utilise un système de traduction automatique en ligne et une phase distincte d'alignement des mots pour trouver la projection de chaque mot de la langue donnée dans la phrase anglaise traduite. Le système estime les caractéristiques interlinguistiques à l'aide d'un système de reconnaissance d'entités nommées en anglais et des informations d'alignement. Ensuite, ces caractéristiques multilingues sont utilisées dans un classificateur basé sur une machine à vecteurs de support.

Intéressons-nous aux modèles basés sur du Prompt. Pattern-Exploiting Training for NER (PETER), une extension de Pattern-Exploiting Training (PET), est une méthode *prompt-based*

conçue à l'origine pour la classification de phrases. PETER, lui, reformule la tâche de *fine-tuning* en utilisant des questions fermées pour fournir des descriptions de tâches qui permettent au modèle d'exploiter les connaissances qu'il a acquises pendant la phase de pré-entraînement. Une approche similaire à PETER a été proposée mais traitant également en étendant aux *n-grams* d'entités (Cui et al., 2021). Il s'agit de voir le problème de NER comme étant un problème de classement en utilisant du *sequence-to-sequence*, où les phrases originales et les candidats sont considérés comme la séquence source et la séquence cible, respectivement. Le modèle de langue utilisé est BART et différents patterns sont expérimentés. LightNER (Chen et al., 2022b) cherche à corriger un problème de transfert de classe et de domaine. En effet, le domaine cible possède des ensembles de catégories d'entités différents de ceux d'un domaine source riche en ressources. Pour cela, au lieu d'aborder le problème en formant des classifieurs discriminants spécifiques à l'étiquette, le système convertit l'étiquetage des séquences pour générer la séquence d'index de pointeur d'entité et les catégories d'entités sans aucun classifieur spécifique à l'étiquette. De plus, il propose une attention guidée par *prompt* en incorporant des *prompts* continus dans la couche d'auto-attention (*self-attention*) pour remoduler l'attention et adapter les poids pré-entraînés.

Pour le *meta-learning*, les auteurs se sont intéressés à trois points : *Model-Agnostic Meta Learning* (MAML), les réseaux prototypiques (*prototypical networks*), qui représentent une technique d'apprentissage de métriques, et les réseaux relationnels (*relation networks*). Le système MetaNER (Li et al., 2020) adapte la notion de MAML avec un *sequence encoder* et un *tag decoder*. Les réseaux relationnels (*relation networks*) sont étendus en incluant un module d'apprentissage de l'attention (Oguz & Vu, 2021), notamment en tirant parti d'*embeddings* contextuels pré-entraînés tels que ELMO et BERT et en utilisant le mécanisme d'attention. Finalement, les réseaux prototypiques ont été adaptés par plusieurs systèmes (Fritzler et al., 2019; Hou et al., 2020; Krone et al., 2020). Par exemple, ProtoNER (Fritzler et al., 2019) les adapte directement pour la tâche de NER. Il apprend des représentations intermédiaires des mots qui se regroupent bien dans des classes d'entités nommées. Cette propriété du modèle permet de classer les mots avec un nombre extrêmement limité d'exemples d'apprentissage et peut potentiellement être utilisée comme méthode d'apprentissage *Zero-Shot*. Le système L-TapNet (Hou et al., 2020), quant à lui, exploite les dépendances des étiquettes grâce à un mécanisme de transfert de dépendances effondré dans le CRF afin de transférer des modèles abstraits de dépendances entre les étiquettes sous forme de scores de transition. Finalement, le système de Krone et al. (2020) démontre que l'adaptation ne nécessite pas de modèle distinct pour chaque type d'entités.

Les systèmes centrés données se divisent en quatre catégories : l'augmentation des données (*data augmentation*), la supervision à distance (*distant supervision*), l'apprentissage actif (*active learning*) et l'auto-apprentissage (*self-learning*). L'une des façons courantes de traiter le manque de données est l'augmentation des données, qui consiste à accroître la taille de l'ensemble de données disponible avec de nouveaux échantillons générés au moyen d'heuristiques ou de sources de données externes. Data Boost (Liu et al., 2020a) explore la capacité de génération de texte des modèles de langage pour générer des échantillons augmentés. En particulier, GPT-2 (Radford et al., 2019) est utilisé comme générateur conditionnel et guidé vers des étiquettes de classes spécifiques au moyen d'une approche d'apprentissage par renforcement (*Reinforcement Learning* - RL). Cette méthode a été conçue pour l'augmentation des données textuelles et ne s'applique pas directement à la NER, car elle ne fournit pas le mappage de la mention de l'entité de la phrase originale à la phrase augmentée.

La méthode *Counterfactual Generator* (Zeng et al., 2020) génère des exemples contrefactuels par des interventions sur les exemples basés sur des observations existantes afin d'améliorer l'ensemble de données d'origine. En effet, certains systèmes présentent des difficultés

à être généralisés lorsque des entités sont faussement corrélées avec leurs contextes. Par exemple, dans la phrase « Jean vit à Toulouse », l'entité « Toulouse » et son contexte (« Jean vit à ») sont fortement corrélés, mais il n'est pas vrai que l'un cause l'autre. Ainsi, en se basant sur l'idée que les entités et les contextes ne sont pas en relation de cause à effet, *Counterfactual Generator* permet de générer des exemples contrefactuels faiblement étiquetés.

Dai & Adel (2020) ont étudié l'adaptation de plusieurs méthodes simples d'augmentation des données aux problèmes de NER telles que le remplacement de tokens par classe (*Label-wise Token Replacement - LwTR*), le remplacement de synonymes (*Synonym replacement* - SR), le remplacement de mentions (*Mention replacement* - MR) et le mélange au sein de mêmes segments (*Shuffle within segments* - SiS). Ils ont montré que ces techniques d'enrichissement de données étudiées sont efficaces dans des contextes où les données sont peu nombreuses, mais qu'elles échouent souvent sur des ensembles de données complets en raison des effets dégradants du bruit. COSINER (Bartolini et al., 2022) est une variante plus élaborée du remplacement des mentions (Dai & Adel, 2020). Elle utilise la similarité cosinus pour remplacer les mentions par celles qui sont les plus similaires, au lieu de les remplacer de manière aléatoire. StyleNER (Chen et al., 2021) apprend des *patterns* (par exemple, le style, le bruit, les abréviations) qui différencient les données des domaines à ressources élevées et à ressources faibles, ainsi que les éléments qu'elles ont en commun. L'idée principale est que les données textuelles peuvent être différentes (par exemple, des phrases longues et complexes dans des résumés de recherche par rapport à des données de médias sociaux), mais que leur sémantique est transférable. Dans les scénarios *Few-shot*, les données étiquetées peuvent être récupérées à partir d'heuristiques, de domaines ou de langages différents, de bases de connaissances externes ou d'ontologies. La supervision à distance (*distant supervision*) (Mintz et al., 2009) vise à tirer parti de ces ressources pour annoter de manière heuristique les données d'apprentissage. SwellShark (Fries et al., 2017) vise à automatiser la génération de candidats et d'étiquettes bruyantes sans données étiquetées manuellement.

AutoNER (Shang et al., 2018) est un modèle neuronal qui marque des entités potentielles grâce à des schémas de type « *tie and break* ». En effet, le système cherche à exploiter les entités qui n'ont pas pu être automatiquement annotées et donc ignorées. Après avoir extrait les phrases non présentes dans le dictionnaire comme « entités potentielles », AutoNER marque deux *tokens* adjacents comme (1) égalité s'ils appartiennent au même type d'entité, (2) inconnu si au moins l'un d'entre eux appartient à une phrase de type inconnu, et (3) rupture dans le cas contraire. Ce schéma est utilisé pour la détection de l'étendue des entités, les types d'entités étant ensuite identifiés sur la base des vecteurs de caractéristiques des candidats. En prédisant si deux *tokens* adjacents font référence à la même entité ou non, AutoNER construit des modèles supervisés distants plus robustes (puisqu'il est souvent bruyant sur les limites des mentions d'entités mais pas dans leurs liens internes).

Yang et al. (2018) abordent également les problèmes d'annotations incomplètes et de bruit. Pour traiter le problème des annotations incomplètes, les phrases sont autorisées à avoir des annotations partielles : certaines plages de *tokens* sont annotées avec des étiquettes définies, tandis que toutes les autres sont associées à toutes les étiquettes possibles. Le problème des annotations bruyantes est traité par une approche d'apprentissage par renforcement qui suit celle de Feng et al. (2018) pour obtenir des instances propres à partir de données de NER issues de la supervision à distance (*distant supervision*).

Knowledge-Augmented Language Model (KALM) (Liu et al., 2019) enrichit un modèle de langue traditionnel à l'aide d'une base de connaissances sans nécessiter de composant supplémentaire. De plus, il apprend à reconnaître les entités de manière totalement non supervisée en utilisant les informations sur les types d'entités qui sont latentes dans le modèle. Cao et al. (2019) exploitent les données du web afin d'étiqueter automatiquement, comme

les ancres de Wikipédia par exemple. Pour cela, il est nécessaire de traiter les problèmes de bruit dans les annotations ainsi que le fait que ces dernières soient incomplètes. Afin d'obtenir un rappel élevé, le système génère autant de données faiblement étiquetées que possible avec une approche d'induction d'étiquettes attribuant des étiquettes aux mots sur la base des ancres et de la taxonomie de Wikipédia, et un schéma de sélection de données qui calcule une fonction de notation pour distinguer les données de haute qualité des phrases faibles. Ensuite, un modèle neuronal est formé sur ces données.

Graph Attention Model (Lou et al., 2020) utilise un dictionnaire spécifique à un domaine pour créer un graphe de groupes de mots qui capture les variantes des entités et découvre ainsi autant de nouvelles mentions d'entités que possible. En particulier, l'ensemble des sommets contient tous les mots du dictionnaire, qui sont reliés par des arêtes non dirigées s'ils apparaissent dans le même type d'entité. *Linked HMM* (Safranchik et al., 2020) affirme que l'approche standard consistant à générer des portées candidates puis à étiqueter indépendamment chaque candidat, comme dans *SwellShark* (Fries et al., 2017), limite l'applicabilité aux tâches pour lesquelles il existe des générateurs de candidats et accroît les efforts humains. En effet, les générateurs de candidats doivent identifier toutes les portées d'entités possibles, car les erreurs se propagent dans le *pipeline*. *Linked HMM* permet donc, à l'aide de différentes règles données (soit des règles d'étiquetage, soit des règles de liaison), de fournir des étiquettes partielles aux séquences, dont la précision est estimée à l'aide d'un modèle génératif probabiliste identifiable sans données d'apprentissage étiquetées.

Enfin, *Consensus Network (ConNet)* (Neelakantan et al., 2022) peut être entraîné sur des annotations imparfaites provenant de sources multiples (par exemple, *crowd annotations*, *cross-domain data*) en apprenant des représentations pour chaque source et en les agrégeant dynamiquement par un mécanisme d'attention contextuelle (*context-aware attention mechanism*). Cette méthode repose sur l'intuition que les différentes sources de supervision peuvent avoir des forces différentes en fonction des scénarios dans lesquels elles sont appliquées. Le système utilise d'abord un schéma d'apprentissage multitâche basé sur un réseau Bidirectional *Long Short Term Memory* (BiLSTM)-CRF pour dissocier les paramètres du modèle en un ensemble de paramètres de modèles invariants pour l'annotateur et un ensemble de représentations spécifiques à l'annotateur, puis forme un module d'attention contextuelle pour une représentation basée sur le consensus en combinant les prédictions sur les données cibles. Ensuite, l'apprentissage actif (*active learning*) vise à sélectionner des ensembles d'exemples informatifs pour la phase d'apprentissage, en demandant activement à l'utilisateur de fournir des étiquettes.

Shen et al. (2017) sont les premiers à explorer les méthodes d'apprentissage actif sur les réseaux neuronaux profonds (*Deep Neural Network* - DNN) pour la tâche NER. En raison de la complexité des réseaux neuronaux profonds, le modèle n'est pas ré-entraîné à chaque cycle d'apprentissage actif, mais il est entraîné de manière incrémentale à chaque série de nouvelles étiquettes. Le travail utilise des stratégies d'échantillonnage basées sur l'incertitude en considérant trois méthodes de classement : *Least Confidence (LC)* (Culotta & McCallum, 2005), *Maximum Normalized Log-Probability* (MNLP), et *Bayesian Active Learning by Disagreement* (BALD) (Gal et al., 2017).

Enfin, l'auto-apprentissage (*Self-Learning*) est une approche similaire à la supervision à distance, où le modèle est formé sur des exemples étiquetés par le modèle lui-même. La différence est que l'heuristique d'étiquetage est remplacée par le modèle lui-même. LM-LSTM-CRF (Liu et al., 2017) utilise un modèle de langue non supervisé et un co-apprentissage (*co-training*). Il exploite les connaissances obtenues avec son modèle de langue pour améliorer les performances d'étiquetage des séquences. En outre, il tire parti des informations au niveau des mots et des caractères des échantillons d'entrée dans le cadre d'un co-entraînement,

où chaque entrée est représentée par différents ensembles de caractéristiques, chacun fournissant des informations complémentaires. Chen & Zhang (2018) utilisent un apprentissage par renforcement pour apprendre à sélectionner les instances d'un ensemble non étiqueté à ajouter à l'ensemble d'apprentissage dans les scénarios d'auto-apprentissage. En effet, ils ont constaté que les performances de l'auto-apprentissage dépendent fortement de la manière dont les nouvelles données sont sélectionnées. Le modèle NER, qui a été précédemment formé avec peu d'échantillons, classera les nouveaux échantillons sélectionnés et sera formé à nouveau.

L'approche d'auto-apprentissage est considérée comme un processus de décision, décrit comme une fonction qui reçoit les instances auto-étiquetées comme entrées et qui produit l'acceptation ou le rejet de ces instances dans l'ensemble d'apprentissage. La récompense est définie comme la différence de performances en NER lors de l'ajout d'un nouvel échantillon (c'est-à-dire l'action) à l'ensemble de données d'apprentissage. RDANER (Yu et al., 2020) utilise une approche de *bootstrapping* pour obtenir des prédictions de modèles sur des données non étiquetées faciles à obtenir et réentraîner le modèle avec l'ensemble de données faiblement augmentées (*augmented weak dataset*). Un modèle de langue pré-entraîné pour un domaine général est *finetuné* sur le corpus annoté (*few-shot*), puis il est utilisé pour annoter un ensemble de données non supervisé. Cependant, leurs expériences ont montré que cette approche donne de moins bons résultats que les modèles linguistiques pré-entraînés spécifiques à un domaine.

BOND (Liang et al., 2020) combine une approche d'auto-apprentissage avec une supervision à distance. Au cours de la première étape, des étiquettes distantes sont générées à l'aide de bases de connaissances externes, et un modèle BERT pré-entraîné est adapté à la tâche de NER supervisée à distance avec un arrêt précoce. Au cours de la deuxième étape, les étiquettes distantes sont ensuite utilisées dans l'approche d'auto-apprentissage pour améliorer davantage les performances du modèle. De cette manière, les pseudo-étiquettes générées par l'auto-apprentissage sont progressivement affinées par les premières étiquettes générées. Enfin, Huang et al. (2021) utilisent des modèles de langues pré-entraînés auto-supervisés (*Transformer-based self-supervised pre-trained language models*) basés sur les *Transformers* et généralisent le modèle par 3 schémas : (1) le méta-apprentissage pour construire des prototypes pour différents types d'entités, (2) le pré-entraînement supervisé spécifique à une tâche sur des données web bruyantes pour extraire des représentations liées aux entités et (3) l'auto-entraînement pour exploiter des données non étiquetées dans le domaine.

Deep Learning, *Small* & *Large Language Models* Dans cette partie, nous allons explorer les travaux d'apprentissage profond (*Deep Learning*) en passant par deux grandes catégories qui ont explosé récemment, à savoir les modèles dits *Small Language Model* (SLM) et les modèles de langue à grande échelle (*Large Language Models* - LLMs). Lorsque les modèles de langues, tels que BERT et ses dérivés par exemple, sont impliqués dans les systèmes NER, nous parlons de SLM. Nous parlerons de LLM lorsque les systèmes emploient des modèles plus larges tels que GPT, par exemple. Il s'agit des modèles qui ont connu une explosion depuis l'apparition de ChatGPT[4]. Les premiers travaux sur l'apprentissage profond pour la NER ont été réalisés par Collobert & Weston (2008), qui ont utilisé pour la première fois un réseau neuronal profond pour réaliser les techniques de NER sur les données textuelles. Ils ont défini une architecture de réseau convolutionnel simple afin d'effectuer diverses tâches, dont celle du NER. Ces tâches ont été, pour la première fois, réalisées automatiquement par un réseau neuronal profond plutôt que d'utiliser des caractéristiques extraites manuellement.

4. ChatGPT disponible au lien suivant : `https://chat.openai.com/`

Akbik et al. (2018a) ont proposé un modèle de *sequence labeling* en utilisant des *embeddings* de chaînes contextuelles basés sur la modélisation des chaînes au niveau du caractère. Ils ont évalué leur modèle sur CoNLL03 pour la NER en anglais et en allemand.

Les méthodes les plus utilisées pour la NER avec l'apprentissage profond sont les réseaux neuronaux récurrents (*Recurrent Neural Network* - RNNs) tels que le BiLSTM, les réseaux neuronaux convolutifs (*Convolutional Neural Networks* - CNNs), leur combinaison hybride, et leur combinaison avec d'autres algorithmes d'apprentissage supervisé tels que les SVMs, les réseaux neuronaux *Bootstrapped*, et bien d'autres encore. Les RNNs sont des réseaux de neurones qui utilisent la mémoire interne pour stocker de longues séquences d'entrée. Le problème du *vanishing gradient* affecte principalement le modèle de base RNN, le rendant incapable de représenter avec précision les relations à long terme dans des situations du monde réel.

Les LSTMs (*Long Short Term Memory*) et les GRUs (*Gated Recurrent Units*) ont été développés pour résoudre ce problème. Le CNN (Alzubaidi et al., 2021) est un type de réseau neuronal largement accepté qui a fait l'objet de recherches approfondies pour traiter la classification des relations entre entités nommées. Le CNN extrait les caractéristiques en effectuant une opération de convolution sur les données d'entrée. En règle générale, les CNNs comprennent un ensemble de filtres convolutifs. Pour créer une caractéristique basée sur les n-grammes, chaque filtre convolutionnel effectue des opérations de convolution sur x mots continus spécifiés par la matrice de poids.

Utilisés à l'origine dans le traitement des images, les CNNs sont également appliqués aux tâches de NER. Pour créer efficacement une intégration de caractéristiques combinatoires de haut niveau dans la NER biomédicale, le modèle proposé par Cho et al. (2020) combine un CNN et une mémoire à long terme bidirectionnelle (BiLSTM) pour la représentation au niveau des mots et des caractères. Le format de jeton (*token*) inclut à la fois les caractéristiques globales et locales au niveau du caractère de l'intégration en combinant les caractéristiques au niveau du caractère des réseaux neuronaux convolutifs et de la BiLSTM. Afin d'apprendre la combinaison de chaque représentation, les représentations CNN au niveau des mots et BiLSTM au niveau des caractères sont ensuite enchaînées et introduites dans un réseau entièrement connecté. Afin de se concentrer sur les *tokens* qui sont liés dans la phrase et de prévoir l'étiquette du *token* actuel, le travail utilise également une technique d'attention au modèle de BiLSTM - CRF pour calculer la ressemblance entre les *tokens* qui sont en entrée. Le modèle a été validé sur deux ensembles de données standardisés : NCBI Disease et JNLPBA.

D'autres variantes des réseaux neuronaux convolutionnels dilatés (*Dilated Convolutional Neural Networks* - DCNN), les *gated CNNs*, et les RNNs sont également utilisées dans plusieurs applications NER. Par exemple, dans le cadre du NER biomédical, Zhang et al. (2022) ont utilisé des BiLSTM et un DCNN pour l'encodage hiérarchique et ont exploité les avantages du DCNN pour enregistrer efficacement des quantités massives de données. Pour améliorer les performances du NER médical, plusieurs caractéristiques de mots sont prises simultanément en compte dans les données du texte médical. Sur trois ensembles de données réelles, des tests approfondis ont été menés et ont montré que leur approche était meilleure que les modèles de comparaison.

Dans la même optique, ASTRAL (Wang et al., 2020) est un système LSTM-CNN à entraînement adversarial visant à exécuter les techniques de NER sur des données non structurées. Pour exploiter les informations spatiales entre les mots adjacents, le *Gated*-CNN est introduit pour fusionner les informations des mots adjacents. Une méthode d'apprentissage adversarial spécifique est proposée pour traiter le problème de surajustement dans le NER. Des perturbations sont ajoutées à des variables arbitraires afin de créer une capacité de gé-

néralisation dans le modèle. Ensuite, le modèle est contraint par la variable cible, ce qui lui permet d'être appliqué à n'importe quelle variable. Le modèle est évalué sur trois *benchmarks* de référence : CoNLL-03 (Sang & De Meulder, 2003), OntoNotes 5.0 (Pradhan et al., 2013a) et WNUT-17 (Derczynski et al., 2017a). Le CNN est intéressant pour extraire des caractéristiques à l'intérieur des phrases et représente le lien entre les phrases. Un CNN 3D a été utilisé pour l'extraction de caractéristiques au niveau du document au lieu d'un BiLSTM (Chang & Han, 2023).

Étant donné qu'un BiLSTM est incapable d'identifier plusieurs phrases à la fois, il est impossible d'obtenir des informations complètes. Il est donc nécessaire d'utiliser un CNN. Afin de traiter le problème du OOV (*Out of vocabulary*), une extension d'un CRF BiLSTM a été implémentée en étudiant la représentation basée sur les caractères (Na et al., 2019). Cette extension implique le déploiement d'une représentation hybride incorporant des vecteurs de morphèmes compositionnels basés sur LSTM et CNN. Les résultats expérimentaux ont montré que la représentation hybride LSTM-CNN proposée apportait des améliorations par rapport à chaque représentation individuelle.

En ce qui concerne le NER, les LSTMs sont devenues le modèle le plus largement utilisé (Gasmi et al., 2018) et ont atteint des performances de pointe. La caractéristique importante de ce modèle est le mécanisme de porte (*gate mechanism*) du modèle LSTM de Hochreiter & Schmidhuber (1997), qui comprend la porte d'entrée (*input gate*), la porte d'oubli (*forget gate*) et la porte de sortie (*output gate*) et apprend la dépendance à long terme (*long-term dependence*).

Dans le travail de Jin et al. (2019), un nouveau réseau neuronal récurrent convolutionnel gated avec attention (*Gated Convolutional Recurrent Neural Network with Attention* - GCRA) pour le travail NER chinois est basé sur les caractères. Un CNN hybride avec un mécanisme de filtre *gating* pour collecter des informations contextuelles locales et un *highway neural network* suivant un LSTM pour choisir des caractères d'intérêt sont introduits. Une méthode d'auto-attention est ajoutée pour collecter les dépendances globales de nombreux sous-espaces et de caractères aléatoires proches. Yan et al. (2021) ont formulé la tâche de NER comme une tâche de génération de séquences d'entités qui peut être résolue en utilisant un modèle *Sequence to Sequence* (Seq2Seq) unifié.

Les BiLSTMs sont fréquemment utilisées dans les applications TALN pour extraire les caractéristiques contextuelles du texte en entrée dans les deux sens (sorties de la LSTM avant et de la LSTM arrière). La concaténation est la sortie de la mémoire à court terme bi-longue (*Bi-Long Short-Term Memory's output*). La représentation des caractéristiques est précédée d'une couche entièrement connectée (*Fully Connected Layer* - FCL), comme dans le modèle CNN, et une fonction *softmax* est employée pour effectuer la classification finale. Pour une période donnée, les RNNs bidirectionnels utilisent efficacement les données passées (par le biais des états avant) et les données futures (par le biais des états arrière). L'un des premiers travaux à avoir utilisé le modèle BiLSTM-CRF pour l'étiquetage des séquences est celui de Huang et al. (2015), qui propose divers modèles, y compris des combinaisons de LSTM, CRF et BiLSTM, et a donné quelques résultats *State-of-The-Art* (SoTA)[5]. Il a donc été comparé à plusieurs autres modèles sur différentes tâches.

Dic-Att-BiLSTM-CRF (Xu et al., 2019) est un réseau neuronal qui consiste à faire correspondre les entités avec un dictionnaire de maladies. Il construit un mécanisme d'attention au niveau du document et une méthode de mise en correspondance avec le dictionnaire. Ensuite, un réseau BiLSTM et un CRF combinent le dictionnaire des maladies pour développer un NER. De manière similaire, Lample et al. (2016) est un système de *sequence labeling*

5. Lorsqu'un algorithme d'apprentissage profond est dit SoTA, cela signifie qu'il s'agit de l'algorithme le plus performant pour une tâche ou un ensemble de données spécifique.

multilingue considéré comme SoTA et présente deux architectures neuronales : l'une basée sur des BiLSTM et des champs aléatoires conditionnels, et l'autre qui construit et étiquette des segments en utilisant une approche basée sur les transitions inspirée par les analyseurs syntaxiques à réduction de décalage (*shift-reduce parsers*). Deux sources d'informations sont utilisées : des représentations de mots basées sur les caractères apprises à partir du corpus annoté et des représentations de mots non supervisées apprises à partir de corpus non annotés.

Ensuite, nous allons détailler les systèmes utilisant des SLMs. GLiNER (Zaratiana et al., 2023) est un modèle NER *zero-shot* qui utilise un encodeur transformateur bidirectionnel (*bidirectional transformer encoder*) semblable à BERT. Ce système date de 2023 et a donc été développé après la montée en puissance des LLMs. Cependant, selon les auteurs, les modèles NER traditionnels sont limités aux entités prédéfinies et les LLMs, malgré leur flexibilité, sont coûteux et volumineux pour les scénarios contraints en ressources. Le système a été entraîné sur des entités multi-mots à partir de deux jeux de données : NuNER[6] (Bogdanov et al., 2024) pour l'anglais et Pile-NER-type pour le multilingue[7]. COPNER (Huang et al., 2022) est une méthode de NER en *Few-Shot* qui combine à la fois le *contrastive learning* et le guidage par *prompt* dans le processus d'apprentissage vers le domaine cible. Cette méthode a été implémentée dans le but de corriger le problème de référents des différentes classes qui ne sont pas correctement représentées en raison du manque de données. L'objectif est d'améliorer les représentations des entités afin de mieux les discriminer grâce à l'apprentissage contrastif. FactMix (Yang et al., 2022) est un modèle en trois composants pour le NER *Cross-domain*. Le premier composant s'appelle *context-level semi-fact generation*, qui consiste à prélever des *tokens* de chaque phrase et à en masquer certains à l'aide d'un *Masked Language Model* (MLM) de type BERT (Devlin et al., 2019a) pour les faire deviner par un MLM. Ensuite, un filtre est appliqué pour supprimer les exemples bruités. Le deuxième composant s'appelle *entity-level semi-fact instances generations* et s'appuie sur les termes entités déjà existants. Pour chaque échantillon, un terme entité est sélectionné et remplacé par un jeton spécifique qui constitue son type. Enfin, le dernier composant s'appelle *augmented data selection and mixing* et consiste à combiner les deux types de données automatiquement générées en privilégiant le *entity-level semi-fact instances generation*.

Xu et al. (2022b) ont proposé un *Pre-trained Machine Reader* (PMR) à partir d'une adaptation d'un MLM pré-entraîné (ici, RoBERTa) pour extraire des séquences de jetons (*tokens*), incluant le NER. Le PMR part d'un *Machine Reading Comprehension* (MRC) mais effectue une étape complète de pré-entraînement de données de type MRC à grande échelle. Ici, un grand volume de données d'entraînement générales et de haute qualité de type MRC, basées sur les ancres de Wikipédia (c'est-à-dire les textes hyperliés), a été créé. L'intérêt d'un PMR est de faciliter le transfert des connaissances de manière discriminante. Afin d'extraire les NER, une étape d'unification du pré-entraînement et du *fine-tuning* est réalisée au sein du même processus de MRC. EntLM (Ma et al., 2021) est une méthode de *prompting* pour le NER mais sans utiliser de *template*. La tâche de NER a été reformulée comme une tâche de SLM mais orientée vers l'entité. Sans modifier la tête de sortie, les SLMs pré-entraînés (ici, *bert-base-cased*) sont *finetunés* pour prédire les mots pivots liés à la classe (ou mots étiquettes) au lieu des mots originaux aux positions de l'entité, tout en continuant à prédire le mot original aux positions non liées à l'entité. ProML (Chen et al., 2022c) est une méthode d'apprentissage de métrique basée sur le *Prompt* pour le NER *Few-Shot*. Plus précisément, la méthode emploie ce qu'on appelle des *prompts* dits *mask-reducible*, qui sont une classe

6. Disponible sur *HuggingFace* au lien suivant : `https://huggingface.co/datasets/numind/NuNER`

7. Disponible sur *HuggingFace* au lien suivant : `https://huggingface.co/datasets/Universal-NER/Pile-NER-type`

spéciale de *prompts*. En effectuant une moyenne pondérée masquée sur les représentations obtenues à partir de plusieurs *prompts*, ProML accepte plusieurs possibilités de *prompts* tant qu'ils sont dits *mask-reducible*. Cette méthode permet d'insérer des annotations sémantiques dans les entrées textuelles. Finalement, le modèle est composé de deux *prompts* préfixes : un *prompt* de préfixe d'options des différentes étiquettes candidates est fourni au modèle et un *prompt* de préfixe d'étiquette permet d'associer chaque entité à son type d'entité en entrée. Pareillement, Li et al. (2022) est un modèle NER de faible ressource basé sur un *prompt* BERT (Devlin et al., 2019a). CONTaiNER (Das et al., 2021) est une approche (quasi SoTA) pour le NER en *Few-Shot* qui exploite l'intégration gaussienne dans des *embeddings* pour l'apprentissage contrastif (*contrastive learning*). Les représentations contextuelles des *tokens* sont générées à l'aide de BERT (Devlin et al., 2019a) comme SLM pré-entraîné. DecomposedMetaNER (Ma et al., 2022) est une méthode de méta-apprentissage décomposée pour le NER en *Few-Shot*, traitant séquentiellement la détection de séquences en *Few-Shot* et le typage d'entités en *Few-Shot* grâce au méta-apprentissage. Pour la partie de détection de séquences, un modèle de méta-apprentissage agnostique (*model-agnostic meta-learning* - MAML) permet de trouver une bonne initialisation des paramètres du modèle qui pourrait s'adapter rapidement à de nouvelles classes d'entités. Quant à la partie du typage d'entités, appelée MAML-ProtoNet, il s'agit de réseaux prototypiques améliorés par MAML pour trouver un bon espace d'*embeddings* qui peut mieux distinguer les représentations de séquences textuelles des différentes classes d'entités. T-NER (*Transformer-based Named Entity Recognition*) (Ushio & Camacho-Collados, 2021; Ushio et al., 2022) est un outil Python pour le *finetuning* de modèles de langue (notamment il est mis, dans le *repository*, à disposition plusieurs modèles disponibles comme RoBERTa, Twitter-BERT, et d'autres) implémenté dans PyTorch. Le *framework* encapsule des modèles pré-établis avec T-NER. Tous les modèles et ensembles de données sont partagés via le groupe T-NER sur HuggingFace [8]. Les jeux de données proposés par T-NER ont ensuite été étendus à la NER dans Twitter. Ils ont mis en place un jeu de données NER, TweetNER7, qui contient sept types d'entités annotées sur 11 382 tweets de septembre 2019 à août 2021.

Le NER en français (ou dans d'autres langues) a souvent des ressources relativement limitées, ce qui implique qu'il n'est pas toujours possible de bénéficier des approches proposées pour des langues comme l'anglais en raison d'un manque de grands ensembles de données. Choudhry et al. (2023) ont cherché à atténuer les effets de cette pénurie de grands ensembles de données étiquetées en proposant une approche NER pour le français basée sur un transformeur, qui utilise l'adaptation contradictoire à des domaines similaires ou à des corpus généraux pour améliorer l'extraction des caractéristiques et permettre une meilleure généralisation. Cette approche utilise des corpus non étiquetés à grande échelle du même domaine ou de domaines mixtes afin d'introduire plus de variations pendant l'entraînement et de réduire le surajustement.

Pour finir, nous allons détailler les systèmes utilisant les LLMs dans leurs systèmes. L'avènement des modèles à grande échelle a récemment transformé de nombreux paradigmes dans le domaine du TALN grâce à l'utilisation de *prompts* en langage naturel (Min et al., 2022). Agrawal et al. (2022) est un modèle utilisant les LLMs comme InstructGPT. L'intérêt est de pouvoir l'utiliser en *zero-shot* et *few-shot* pour l'extraction d'informations, comme ils l'ont démontré pour le domaine clinique, étant un domaine spécifique. PromptNER (Ashok & Lipton, 2023) est une méthode de *Few-Shot* et *cross-domain* NER basée sur le *prompting*. La méthode s'appuie principalement sur un LLM pré-entraîné sur des tâches de génération conditionnelle. Cependant, la méthode contient d'autres composants pour compléter cette méthode. En effet, elle contient une brique de définitions modulaire, c'est-à-dire

8. Le groupe T-NER est disponible sur HuggingFace sur le lien suivant : `https://huggingface.co/tner`

un document contenant des descriptions en langage naturel de l'ensemble des types d'entités, explicitant ce qui constitue ou ne constitue pas une entité. Un autre composant consiste à fournir quelques exemples du domaine cible. Enfin, un format précis pour la sortie des entités est communiqué au modèle via le formatage des exemples *few-shot*. L'intérêt de cette méthode est de pouvoir s'adapter aux domaines cibles en modifiant les définitions des entités et les exemples fournis.

Li & Zhang (2023) ont démontré par le biais de leur modèle RT (*Retrieving and Thinking*) que les LLMs ont une performance supérieure par rapport aux modèles SLMs, surtout lorsque les LLMs sont fournis avec des instructions de haute qualité. Il est également mentionné le transfert de modèles de NER du domaine général au domaine médical, soulignant ainsi la flexibilité et l'adaptabilité de ces modèles. De plus, la quantité d'échantillons impacte davantage les performances des SLMs que celles des LLMs, mettant en évidence la robustesse des LLMs face à des ensembles de données plus petits. RT combine un LLM par défaut avec une méthode de recherche k-NN (*K-Nearest Neighbors* - k plus proches voisins) pour identifier les classes d'entités présentes dans une phrase donnée, suivie d'un processus de reconnaissance logique des entités pour identifier ces dernières de manière efficace et progressive. Cette approche permet la sélection d'exemples appropriés et une reconnaissance précise des entités.

NERDA-Con (Vijay & Priyanshu, 2022) est une *pipeline* qui intègre l'apprentissage continu (*continual learning*) dans l'entraînement des NER en utilisant des LLMs basés sur les transformeurs. L'apprentissage continu consiste à apprendre en permanence en intégrant de nouvelles connaissances tout en conservant les connaissances acquises lors des expériences précédentes.

L'article de Xie et al. (2023) se concentre sur l'exploration de la performance des LLMs dans le domaine du NER en *zero-shot*, avec une application particulière à ChatGPT. Les auteurs décomposent la tâche de NER en sous-problèmes plus simples par étiquettes, adoptant ainsi un paradigme de question-réponse décomposée (*Decomposed-QA*). De plus, ils introduisent une augmentation syntaxique sous forme d'incitation syntaxique, incitant le modèle à analyser la structure syntaxique du texte donné en entrée, et d'augmentation de l'outil, fournissant au modèle des informations syntaxiques via un outil d'analyse syntaxique. Enfin, les auteurs adaptent le concept d'autoconsistance (*self-consistency* - SC) à la tâche de NER en proposant une stratégie de vote majoritaire en deux étapes, garantissant ainsi une cohérence optimale dans la reconnaissance des entités nommées. Leur modèle a été expérimenté sur différents ensembles de données, chinois et anglais, ainsi que dans des scénarios spécifiques à un domaine et généraux.

Foppiano et al. (2024a) se sont intéressés à évaluer les capacités des LLMs tels que GPT-3.5-Turbo, GPT-4 et GPT-4-Turbo dans l'IE à partir de documents scientifiques dans le domaine des sciences des matériaux. En ce qui concerne la NER, une méthodologie a été introduite afin d'analyser de manière comparative les expressions matérielles complexes, mettant particulièrement l'accent sur la standardisation des formules chimiques. Ils ont évalué les performances des LLMs dans la NER par rapport à des modèles traditionnels basés sur l'architecture BERT et des approches basées sur des règles. Leurs résultats ont montré que les LLMs, même après un *prompting* à faible échantillonnage, ne parviennent pas à surpasser la performance de la ligne de base dans la reconnaissance des entités nommées liées aux matériaux et aux propriétés physiques. Cependant, un GPT-3.5-Turbo finement réglé avec une stratégie appropriée parvient à surpasser tous les autres modèles. Ils mettent donc en évidence les réglages fins qu'il est nécessaire de mettre en place dans les domaines spécifiques tels que les sciences des matériaux.

La tâche de NER a vivement été exploitée dans le domaine biomédical, qui est un do-

maine spécifique comme le domaine juridique et pratique que nous exploitons. Notamment, dans le cadre des LLMs, DeID-GPT (Liu et al., 2023) exploite GPT-4 comme LLM dans le cadre de la désidentification, qui consiste à identifier et supprimer automatiquement les informations d'identification. Pour cela, les LLMs, tels que ChatGPT ou GPT-4, ont montré un grand potentiel dans leur capacité pour les NER, ce qui permet d'identifier des informations confidentielles. Dans l'étude de Chen et al. (2023), une évaluation complète est menée pour établir les performances de base des LLMs, tels que GPT-3.5 et GPT-4, ainsi que LLaMA, sur douze ensembles de données du domaine biomédical, couvrant six tâches extractives et génératives distinctes. Les résultats indiquent que les performances de *Zero-shot* et de *few-shot* de GPT-4 sont compétitives voire supérieures à celles des modèles BERT et BART finement ajustés, en particulier dans les tâches de raisonnement et de génération de texte. Cependant, malgré ce potentiel prometteur, des lacunes persistent dans les performances des LLMs, notamment en ce qui concerne les tâches extractives telles que la NER, où les différences peuvent atteindre jusqu'à 30% par rapport aux modèles BERT finement ajustés. De plus, les LLMs ont été identifiés comme générant des erreurs prévalentes ainsi que des données manquantes et des incohérences.

Dans le domaine biomédical, Taiyi (Luo et al., 2023) est un modèle LLM bilingue *fine-tuné* (anglais et chinois) capable de traiter diverses tâches spécifiques à la biologie. Pour préparer Taiyi, une collection exhaustive de 102 ensembles de données en anglais et 38 en chinois, couvrant plus de 10 types de tâches biomédicales, est rassemblée. Des schémas de données normalisés sont ensuite conçus et appliqués universellement pour répondre aux exigences spécifiques de chaque tâche et assurer un formatage cohérent des données. Ensuite, une stratégie de *fine-tuning* supervisé en deux étapes est proposée, ce qui permet d'optimiser considérablement les performances du modèle sur une diversité de tâches. Enfin, l'évaluation de Taiyi est menée sur 13 ensembles de tests BioNLP. Les résultats expérimentaux mettent en évidence le potentiel prometteur de Taiyi dans l'apprentissage multilingue multi-tâches, avec des performances supérieures sur des tâches spécifiques à la biologie par rapport aux LLMs généraux, ce qui appuie l'idée d'utiliser un LLM ajusté à un domaine précis.

Enfin, UniversalNER (Zhou et al., 2024) est une approche de distillation ciblée avec un *instruction tuning* axé sur des missions (*mission-focused instruction tuning*) qui est explorée pour former divers modèles, notamment pour la NER et l'OpenIE. L'*instruction tuning* élargit les exemples d'entrée-sortie utilisés dans le *fine-tuning* supervisé en y ajoutant un autre élément : les instructions. Ces instructions agissent comme des directives supplémentaires pour guider l'apprentissage des modèles, en fournissant des informations contextuelles ou des consignes spécifiques sur la tâche à accomplir. Cette étude de cas montre que ChatGPT peut être distillé pour la NER et l'OpenIE. UniversalNER a été évalué exhaustivement sur un ensemble de données qui a ensuite été mis à disposition pour la NER, comprenant 43 ensembles de données couvrant 9 domaines divers tels que la biologie, la programmation, les médias sociaux, le droit et la finance. Les auteurs ont rendu disponibles la recette de distillation, les données et les modèles UniversalNER pour faciliter les recherches futures sur la distillation ciblée[9].

Extraction des termes pertinents

Les algorithmes pour l'extraction automatique de termes pertinents peuvent être découpés en quatre catégories : (1) les approches encapsulant des informations linguistiques ou les approches linguistiques à l'aide de dictionnaires et de systèmes basés sur des règles (Wong et al., 2007) ; (2) les modèles supervisés (Witten et al., 1999; Turney, 2000; Frank et al., 1999),

9. L'ensemble des travaux de UniversalNER est accessible au lien suivant : `https://universal-ner.github.io/`

(3) les modèles non supervisés (Salton & Buckley, 1988; Rajaraman & Ullman, 2011; Velardi et al., 2001; Hasan & Ng, 2010) ; et (4) les modèles de langues (Grootendorst, 2020, 2022).

Utilisation des informations linguistiques Les premiers algorithmes d'extraction terminologique se sont orientés vers des **approches linguistiques**. Justeson & Katz (1995) ont été parmi les premiers à s'intéresser aux propriétés linguistiques des termes techniques. Ils ont examiné la terminologie utilisée dans les dictionnaires techniques de différents domaines, notamment la fibre optique ; la physique et les mathématiques ; la médecine ;et la psychologie. L'étude a conclu que les syntagmes représentaient 92,5 à 99 % des termes techniques trouvés, dont environ 70 % comportaient plus d'un mot ; dans quelques cas, ils accompagnaient des adjectifs et, dans une moindre mesure, des prépositions ; mais aucun cas n'impliquait des verbes. Les auteurs se sont davantage attachés à développer la linguistique des mots clés plutôt qu'à élaborer une solution.

Daille et al. (1994) ont obtenu, en parallèle, des conclusions équivalentes de Justeson & Katz (1995), bien qu'ils affirment que les termes techniques sont construits à partir de multi-mots (*multi-word units* - MWU) de base par le biais d'opérations de composition, et poursuivent en disant qu'un terme multi-mots complexe tel que « satellite de communication géostationnaire » est le résultat de la combinaison de deux MWU : le « satellite géostationnaire » et le « satellite de communication ». Salton (1988) est allé plus loin en suggérant d'inclure les termes discontigus, c'est-à-dire ceux composés d'éléments séparés par des mots intermédiaires, tels que « building dictionary » et « a building of a dictionary ».

Park et al. (2002) ont abordé la question sous un angle quelque peu différent. Ils se sont principalement intéressés à l'extraction de glossaires, dont l'un des principaux objectifs était de localiser et d'extraire des termes liés à un domaine spécifique. Ce qui a distingué leur travail, c'est un ensemble de patrons syntaxiques (*syntactic patterns*) qu'ils ont utilisé pour identifier les termes candidats largement répandus depuis. L'utilisation de conjonctions était pourtant jusque-là explicitement rejetée (Justeson & Katz, 1995) mais la motivation pour l'utilisation d'un ensemble particulier de modèles syntaxiques provient principalement de la nécessité de travailler pour un domaine spécifique, ce qui rend le système pertinent pour un domaine de spécialité.

Une autre idée intéressante a été émise par Barker & Cornacchia (2000), qui ont mis en avant la notion d' « extraction de mots-clés par tête » (*head-driven keyword extraction*). L'idée était de définir les mots-clés comme des groupes nominaux contenant les têtes les plus fréquentes. Les auteurs ont fait état d'une amélioration modeste par rapport aux lignes de base mais leur effort unique pour relier une théorie syntaxique (alors en vigueur) à l'extraction de mots-clés mérite d'être souligné.

Plus récemment, d'autres modèles basés sur des approches linguistiques ont permis d'identifier automatiquement les expressions poly-lexicales (combinaison d'au moins deux unités lexicales (mots)). Pour cela, plusieurs systèmes ont utilisé des analyseurs syntaxiques en dépendances par transitions (Constant & Nivre, 2016; Saied et al., 2018) ou basés sur des grammaires de substitutions d'arbre (Green et al., 2013). Lors du *workshop SemEval-2016* (Bethard et al., 2016a), la tâche 10 (Schneider et al., 2016) consistait à détecter des unités sémantiques minimales avec leur signification et la tâche 11 consistait à identifier des mots complexes (Bethard et al., 2016b). Un système de la tâche 10 s'est démarqué en étant réalisé à l'aide de dictionnaires et de systèmes basés sur des règles (Cordeiro et al., 2016). En 2018, le meilleur système de la campagne PARSEME (Ramisch et al., 2018), nommé TRAVERSAL, utilisait des arbres discriminants (Waszczuk, 2018).

Pour finir, Savary et al. (2019) soutiennent que les dictionnaires sont nécessaires pour obtenir une meilleure généralisation de l'identification d'expressions polylexicales, où la

généralisation est plus difficile que dans des tâches similaires telles que la reconnaissance d'entités nommées.

Modèles supervisés Les approches **supervisées** utilisent des documents annotés par des termes clés dits « corrects ». Ces documents permettent de former des classifieurs capables de distinguer les termes clés des « non »-termes clés d'un document. On peut donc, ici, réduire l'extraction de termes clés à un problème de classification binaire où les segments textuels candidats sont classés comme positifs (c'est-à-dire les termes-clés) ou négatifs (c'est-à-dire les non-termes-clés). Les modèles supervisés sont souvent connus comme étant des modèles basés sur des caractéristiques (*features*) (Witten et al., 1999; Marujo et al., 2012). KEA (Witten et al., 1999; Frank et al., 1999) et GenEx (Turney, 2000) utilisent des caractéristiques basées sur la fréquence et la position d'un segment textuel dans un document cible.

Bien que les approches **supervisées** soient généralement plus performantes que les approches **non supervisées** (Kim et al., 2012), la nécessité de disposer de vastes corpus annotés par l'homme pour chaque domaine d'étude a conduit à accorder une grande attention à la conception d'approches **non supervisées**.

Modèles non supervisés En second lieu, les modèles **non supervisés** sont implémentés de diverses façons. Il peut s'agir de modèles statistiques dont le plus connu consiste à utiliser le TF-IDF (Salton & Buckley, 1988; Rajaraman & Ullman, 2011; Barker & Cornacchia, 2000; Zhang et al., 2007; Liu et al., 2009) ou IDF (Velardi et al., 2001). Un modèle simple à implémenter également est *FirstPhrases*.

L'observation selon laquelle un mot-clé contient rarement un mot d'arrêt a conduit Rose et al. (2010) à développer une méthode largement utilisée, connue sous le nom de RAKE. Cette méthode extrait les mots-clés en divisant un texte en un ensemble de séquences de mots en isolant les mots vides (ou ceux définis comme *stopwords* en choisissant parmi les mots qui reviennent le plus souvent). Dans un élan similaire, le système d'extraction de phrases-clés le plus performant de SemEval 2010 revient à KP-Miner (El-Beltagy & Rafea, 2010). KP-Miner a utilisé des observations statistiques telles que la fréquence des termes pour filtrer les segments qui ont peu de chances d'être des segments clés. Plus précisément, des seuils estimés à partir des données sont appliqués pour la fréquence des segments textuels. Les segments candidats sont ensuite ordonnés par classement (*ranking*) à l'aide du modèle TF-IDF en combinaison avec un facteur d'amplification qui vise à réduire le biais en faveur des termes à mot unique.

Le et al. (2016) ont montré que l'extraction de termes-clés à partir d'un document peut bénéficier de la prise en compte de segments textuels candidats avec des balises de partie du discours autres que des noms ou des adjectifs. YAKE (Campos et al., 2020) est une technique d'extraction automatique de mots-clés qui utilise de multiples caractéristiques statistiques du texte pour attribuer des scores aux mots et aux phrases. Elle classe les candidats pour obtenir des mots-clés pour un texte donné. Adar & Datta (2015) ont extrait des termes-clés en extrayant des abréviations de la littérature scientifique et ont construit une base de données de termes-clés sémantiquement hiérarchique.

Les vecteurs d'intégration des mots ont également été utilisés pour mesurer les liens entre les mots dans les modèles basés sur les graphes (Wang et al., 2014a) que nous allons détailler ci-après (Florescu & Caragea, 2017). Il s'agit de réduire l'extraction de termes clés comme un problème de *ranking* (Hasan & Ng, 2014). Ces techniques basées sur les graphes construisent un graphe de mots à partir de chaque document cible, de sorte que les nœuds correspondent aux mots et les arêtes aux modèles d'association de mots. Les nœuds sont ensuite classés à l'aide de mesures de centralité du graphe telles que PageRank (Mihalcea &

Tarau, 2004; Liu et al., 2010) ou HITS (Litvak & Last, 2008), et les phrases les mieux classées sont renvoyées sous forme d'expressions clés. Le système dit pionnier, TextRank (Mihalcea & Tarau, 2004), est un système pour évaluer les phrases clés en appliquant PageRank à un graphe de mots construit à partir de mots adjacents dans un document.

Depuis leur introduction, de nombreuses extensions basées sur les graphes ont été proposées, qui visent à modéliser différents types d'informations. Notamment, TextRank a été étendu de nombreuses fois. SingleRank (Wan & Xiao, 2008) est une extension de TextRank en ajoutant des arêtes pondérées entre les mots qui cooccurrent dans une fenêtre de taille variable $w \geqslant 2$. Les documents voisins textuellement similaires sont inclus dans ExpandRank (Wan & Xiao, 2008) qui consiste à calculer des informations plus précises sur la cooccurrence des mots. ExpandRank, lui-même, a été étendu pour intégrer des informations provenant de réseaux de citations où les articles se citent les uns les autres (Gollapalli & Caragea, 2014).

Plusieurs approches non supervisées basées sur les graphes s'appuient sur des techniques de regroupement de mots telles que le regroupement des mots candidats en thèmes, puis l'extraction d'une phrase clé représentative de chaque thème (Liu et al., 2009; Bougouin et al., 2013). Lahiri et al. (2014) ont extrait des termes-clés de documents en utilisant diverses mesures de centralité telles que le degré de nœud, le coefficient de *clustering* et la proximité. SemGraph (Martínez-Romo et al., 2016) utilise les informations de WordNet pour enrichir les relations sémantiques entre les mots dans le graphe. PositionRank (Florescu & Caragea, 2017) vise à capturer à la fois les segments textuels très fréquents et leur position dans un document. En effet, la position relative d'un mot dans un document s'est révélée être une caractéristique très efficace dans l'extraction supervisée de termes clés (Zhang et al., 2007; Hulth, 2003) mais bien souvent, l'information sur la position n'est pas utilisée dans des méthodes non supervisées.

En résumé, PositionRank est un modèle PageRank basé sur la position qui incorpore toutes les positions des occurrences d'un mot et accorde des probabilités plus élevées aux mots trouvés au début d'un document au lieu d'utiliser une distribution uniforme sur les mots. MultiPartiteRank (Boudin, 2018) est une technique d'extraction de mots-clés basée sur un graphe. Elle utilise une structure de graphe *k-partite* (*Multipartite graph*) pour représenter les candidats et les sujets dans un seul graphe et exploite les relations pour améliorer la sélection des candidats. Elle détermine la préférence des mots-clés à l'aide d'un nouveau mécanisme de sélection. TopicalPageRank (Sterckx et al., 2015) est un extracteur de termes clés qui incorpore des informations (*topics*) provenant de modèles thématiques. Il a la particularité de ne calculer qu'un seul PageRank pour chaque texte et ce peu importe le nombre de thématiques car il est indépendant.

En hybride, des systèmes à la fois basés sur les graphes et des statistiques ont émergés. Danesh et al. (2015) ont calculé un poids initial pour chaque phrase sur la base d'une combinaison d'heuristiques statistiques telles que le score TF-IDF et la première position d'une phrase dans un document. Les phrases et leurs poids initiaux sont ensuite incorporés dans un algorithme basé sur un graphe qui produit le classement final des candidats aux termes clés.

Modèles de langues Ces **modèles** ont pris une ampleur conséquence depuis peu. De nombreux modèles se sont basés sur ces derniers afin d'extraire automatiquement des termes clés. Lors de la campagne de PARSEME, certains modèles basés sur des réseaux neuronaux ont obtenus de bonnes performances (Ramisch et al., 2018), notamment le système nommé TRAPACC (Stodden et al., 2018). Ensuite, l'édition de 2020 (Ramisch et al., 2020) a bénéficié d'avancées des modèles de langues pré-entrainés comme en témoigne le meilleur système, MTLB-struct, basé sur le modèle BERT et utilisant un entraînement multi-tâches (Taslimi-

poor et al., 2020).

Les analyseurs lexicaux ont donc cherché à se baser sur un modèle BERT prédisant les expressions poly-lexicales (Liu et al., 2021). BERTopic (Grootendorst, 2022) est un *topic model* qui étend une tâche de *clustering* en extrayant une représentation « cohérente du sujet » à travers le développement d'une variante de TF-IDF basée sur des classes. Plus précisément, BERTopic génère l'intégration de documents à l'aide de modèles linguistiques pré-entraînés basés sur des transformeurs, regroupe ces intégrations et, enfin, génère des représentations de *topics* (thèmes) à l'aide de la procédure TF-IDF basée sur les classes. Le système KeyBERT utilise les *embeddings* BERT et la similarité Cosinus simple pour trouver les sous-phrases d'un document qui sont les plus similaires au document lui-même. Tout d'abord, les *embeddings* de documents sont générés avec BERT pour obtenir une représentation au niveau du document. Ensuite, les plongements (*embeddings*) de mots sont extraits pour les mots ou N-grammes. Enfin, la similarité Cosinus est utilisée pour trouver les termes les plus similaires au document. Les termes les plus similaires peuvent alors être identifiés comme les termes qui décrivent le mieux l'ensemble du document.

La particularité de KeyBERT est de ne pas devoir être entraîné et il est directement utilisable. KeyBERT a utilisé d'autres ressources et notamment des modèles d'extraction de mots-clés (Sharma & Li, 2019). KeyBERT n'est pas unique et d'autres articles et solutions qui utilisent les BERT-embeddings existent (Sharma & Li, 2019) ainsi que KeyBERT lui-même (Issa et al., 2023). AdaptKeyBERT (Priyanshu & Vijay, 2022) est un exemple de système qui adapte KeyBERT. Il s'agit d'un pipeline pour l'entraînement d'extracteurs de mots-clés avec des bases LLMs en incorporant le concept d'attention régularisée dans une phase de pré-entraînement pour l'adaptation du domaine en aval.

LLM Parmi les modèles de langue, nous pouvons retrouver la sous-catégorie des LLMs. En effet, dans l'état de l'art, afin d'évaluer les LLMs, les tâches de l'IE et de l'extraction de termes pertinents (*Keyphrases extraction* - KPE) ont été employées. Du et al. (2023) se sont intéressés aux LLMs en les évaluant sur la tâche d'extraction de termes (ou phrases) clés (KPE). Afin d'éviter l'étiquetage humain coûteux sur des données de terrain pour des applications industrielles, ils ont utilisé un autre LLM comme source d'étiquettes. Leur évaluation s'est effectuée dans différentes langues et différents domaines. Il est intéressant de noter que d'autres recherches ont été faites sur l'utilisation de doubles LLMs : l'un pour la tâche en question et l'autre pour l'évaluation (Manakul et al., 2023; Chern et al., 2023).

Corpus et évaluation

Dans cette section, nous allons chercher un moyen efficace de pouvoir évaluer notre système d'extraction terminologique. En effet, trois aspects se sont démarqués dans l'état de l'art :

1. L'évaluation par non supervision. En effet, comme exprimé dans la sous-section précédente (3.3.1), dans plusieurs tâches du TALN les LLMs ont pu être utilisé afin d'évaluer les modèles (Du et al., 2023; Manakul et al., 2023; Chern et al., 2023). Il semble donc possible de pouvoir évaluer des données de terrain directement sans avoir recourt à un jeu de données reconnu mais qui soit dans une autre langue ou à l'étiquetage manuel.
2. L'évaluation pour le NER spécifiquement que nous allons développer dans le paragraphe associé ci-après.
3. L'évaluation pour le KPE spécifiquement que nous allons développer dans le paragraphe associé en suivant.

Corpus et évaluation pour le NER. Afin d'évaluer mais aussi d'entraîner les systèmes sur la tâche de NER, plusieurs jeux de données ont été créés et sont reconnus au sein de la communauté. L'ensemble des jeux de données reconnus au sein de la communauté sont détaillés sur le tableau 3.1. Ce tableau est divisé en quatre grandes catégories :

— Tout d'abord, la première catégorie (« Standards ») traite des jeux de données mis à disposition depuis plusieurs années qui sont installés et reconnus auprès de la communauté.
— Ensuite, la deuxième traite des domaines spécifiques du NER comme par exemple le domaine du biomédical, la science ou même les données issues de twitter.
— Ensuite, la troisième catégorie traite des jeux de données présentant des spécificités comme le fait d'être créé pour le *few-shot* ou encore le *cross-domain*.
— Enfin, depuis peu, les LLMs sont au cœur de beaucoup de sujets de recherche. Certains jeux de données ont donc été construits grâce à leur utilisation.

Catégories	Sous-catégorie	Nom	Description	Langue
Standards	News	ACE 05 (Walker et al., 2006)	compilé à partir de divers domaines de *news* (presse, forums, *etc.*) : différents types d'entités = personnes (PER), lieux (LOC), entités géopolitiques (GPE) et organisations (ORG). Aussi : Types de relations complexes : relations entre artefacts (ART), affiliations générales (GEN-AFF) et relations sociales personnelles (PER-SOC).	Multi
		CoNLL (2003 (Sang & Meulder, 2003), 2004 (Roth & tau Yih, 2004))	Corpus annoté collecté à partir de *news* et focus sur des entités générales comme People, Organizations, et Locations. Aussi des relations présentes : Work For et Live in	Anglais
		NYT (Sandhaus, 2008), NYT10 (Riedel et al., 2010), NYT11 (Hoffmann et al., 2011)	1 500 000 articles taggés manuellement par des bibliothécaires avec des étiquettes tirées d'un vocabulaire d'indexation normalisé de PERSONNE, ORGANISATION, LIEU, et DESCRIPTEUR DE SUJET. 275 000 articles étiquetés de manière algorithmique et vérifiés à la main. Produit par supervision à distance à partir d'articles du New York Times 24 relations prédéfinies.	Anglais

Catégories	Sous-catégorie	Nom	Description	Langue
		Ontonotes 5.0 (Pradhan et al., 2013b)	Corpus à grande échelle, multigenre et multilingue annoté manuellement avec des informations syntaxiques, sémantiques et discursives : L'un des ensembles de données NER les plus utilisés (en raison de sa qualité et de sa taille élevées).	Multi
		WNUT 2017 (Derczynski et al., 2017b)	Entités émergentes et rares dans des textes bruyants (de type discussion)	Anglais
	Wikipédia	MultiNERD (Tedeschi & Navigli, 2022)	Corpus annoté automatiquement couvrant 10 langues, 15 catégories de NER et 2 genres textuels. Inclut : informations de désambiguïsation pour permettre le développement de systèmes multilingues de mise en relation d'entités & des URL d'images pour encourager la création de systèmes multimodaux	Multi
		Polyglot-NER (Al-Rfou et al., 2015)	Ensemble de données d'entraînement généré automatiquement à partir de Wikipédia et de Freebase pour la tâche de reconnaissance des entités nommées. L'ensemble de données contient les données d'entraînement de base basées sur Wikipédia pour 40 langues que nous avons (avec résolution des coréférences) pour la tâche de reconnaissance des entités nommées	Multi
	Science	ScIERC (Luan et al., 2018)	Dataset pour le NER, la coreference et l'extraction de relations. Collection de 500 documents dans des termes scientifiques (résumés de papiers sur l'IA)	Anglais

Catégories	Sous-catégorie	Nom	Description	Langue
Domaines spécifiques		FabNER (Kumar & Starly, 2021)	corpus de 350 000 mots dans une collection de résumés obtenus à partir de Web of Science par le biais de revues connues dans le domaine de la recherche sur les procédés de fabrication	Anglais
	Biomédical	DIANN	Différentes entités à détecter : handicaps, négation acronymes	Anglais et espagnol
		NCBI (Dogan & Lu, 2012)	Corpus de maladies de grandes échelles composé de 6900 mentions de maladies dans 793 citations PubMed. Corpus développé par une équipe de 12 annotateurs (deux personnes par annotation) et couvre toutes les phrases d'un résumé PubMed. Catégories suivantes : maladie spécifique, classe de maladie, mention composite et modificateur.	Anglais
		E3C (Magnini et al., 2021)	Vaste corpus de documents cliniques européens avec informations temporelles et chronologiques et des informations sur les entités cliniques basées sur des taxonomies médicales, à utiliser pour le raisonnement sémantique.	Espagnol, basque, anglais, français et italien
	Twitter	Broad Twitter (Derczynski et al., 2016)	échantillonné à travers différentes régions, périodes temporelles et types d'utilisateurs de Twitter. Annotations effectuées par une combinaison d'experts en NLP et de *crowd workers*, ce qui nous permet d'exploiter le rappel de la foule tout en maintenant une qualité élevée. Plus : Entités observées comparées au fils d'actualité.	Anglais
		HarveyNER (Chen et al., 2022a)	Emplacements précis annotés dans des tweets. Mention de lieux (complexes) et descriptions informelles.	Anglais

Catégories	Sous-catégorie	Nom	Description	Langue
Spécificités	Few-Shot	Few-NERD (Ding et al., 2021)	Ensemble de données NER à grande échelle annotées manuellement. Hiérarchie de 8 types d'entités à gros grain (*coarse-grained*) et 66 types d'entités à grain fin (*fine-grained*). Composé de 188 238 phrases de Wikipedia dont 4 601 160 mots inclus et annoté en tant que contexte ou partie du type d'entité à deux niveaux.	Anglais
	Cross-domain	CrossNER (Liu et al., 2020b)	collection entièrement étiquetée de données NER : cinq domaines avec des catégories d'entités spécialisées pour différents domaines	
LLM		NuNER (Bogdanov et al., 2024)	Dataset basé sur un échantillon aléatoire de C4 (Raffel et al., 2020), reconnu pour sa diversité de domaines, comprenant des textes provenant de blogs, d'articles de presse et de messages sur les réseaux sociaux. Corpus annoté avec une grande variété d'entités et de types pour généraliser à différents problèmes de NER. Utilisation d'un LLM pour extraire et assigner librement des types à toute entité identifiée, y compris des concepts plus proches des thèmes que des types d'entités traditionnels. Dataset final comprend 1 million de phrases annotées après l'application d'un filtre éliminant les annotations considérées comme peu informatives	Anglais
		Pile-NER-type (Zhou et al., 2023a)	Dataset pour le NER multi-mots généré par GPT en utilisant la construction de données par *prompt*. Dataset collecté par *prompting* dans gpt-3.5-turbo-0301 et augmenté avec des échantillons négatifs.	Anglais

Catégories	Sous-catégorie	Nom	Description	Langue

TABLE 3.1: Liste des jeux de données regroupés dans des catégories et sous-catégories (domaine par exemple) avec leurs descriptions et langues associées.

Plusieurs autres *datasets* existent mais n'ont pas été présentés dans le tableau mais nous allons en lister quelques-uns sans les détailler : AnatEM (Pyysalo & Ananiadou, 2014), BC2GM, BC4CHEMD, bc5cdr (Wei et al., 2016), FindVehicle, GENIA, MIT Movie & MIT Restaurant (Liu et al., 2013), Polyglot-NER (Al-Rfou et al., 2015), tweetNER7 Ushio et al. (2022), WikiANN (Rahimi et al., 2019), wikineural (Tedeschi et al., 2021)

Plusieurs métriques d'évaluation ont été mises en évidence. La première métrique dite stricte (ou également appelée *span-level*) consiste à vérifier si l'entité a été parfaitement identifiée avec le bon type correspondant. Il s'agit donc de trouver le segment textuel qui s'accorde parfaitement avec le segment textuel du jeu de données et que la classe soit correctement prédite. Une autre métrique dite plus souple consiste à prédire correctement la classe associée à un segment textuel corrélé au segment textuel du jeu de données. En effet, le système peut créer de légères différences dans l'extraction de segments textuels. Il peut s'agir du bon segment mais dont la forme a été changée comme le fait de mettre le segment sous sa forme singulière, par exemple. Il peut également s'agir d'un sous ou sur ensemble du segment textuel. Dans ce cas, l'extraction du segment peut également être correcte mais le système a cherché à minimiser ou maximiser les informations dans le segment.

Corpus pour l'extraction de termes pertinents. En termes de **corpus**, certains se sont démarqués au sein de la communauté car ils intègrent les expressions poly-lexicales mais aussi les super-sens de celles-ci comme le corpus Streusle (Schneider & Smith, 2015), celui de PARSEME (Ramisch et al., 2018, 2020) déjà mentionné précédemment ainsi que DimSum, le corpus pour la tâche 10 de SemEval 2016 (Schneider et al., 2016). Certains systèmes mentionnés ci-dessus se sont confrontés à ces corpus afin de s'auto-évaluer (Liu et al., 2021).

Le Centre Aquitain des Technologies de l'Information et Electroniques (CATIE) présente plusieurs jeux de données en français pour l'extraction de termes clés (Centre Aquitain des Technologies de l'Information et Electroniques, 2023). Notamment, nous pouvons retrouver sur HuggingFace plusieurs jeux de données d'extraction de mots-clés créés à partir du jeu de données « Dataset of French Prompts (DFP) ». Il est donc pertinent de s'intéresser à ces jeux de données car ils disposent directement du *prompt* à injecter directement dans les modèles de langues et permet de les confronter directement. CATIE ont, ainsi, créé trois jeux de données issus de trois jeu de données différents. Le premier, *wikinews-fr-100_fr_prompt_keywords_extraction*, est un jeu de données dont les données originales proviennent du jeu de données wikinews-fr-100. Le deuxième s'intitule *taln-archives_fr prompt_keywords_extraction* et provient du jeu de données taln-archives. Enfin, le troisième intitulé termith-eval_fr prompt_keywords_extraction provient du jeu de données termith-eval. Il est à noter que les jeux de données d'origine sont des jeux de données pour l'extraction de termes clés qui peuvent également être utilisés.

EUROPA (Salaün et al., 2024) est un jeu de données pour la génération de phrases-clés multilingues dans le domaine juridique. Il est dérivé des arrêts de la Cour de justice de l'Union européenne (UE) et contient des instances dans les 24 langues officielles de l'UE.

Discussion et positionnement

Intégrer un système de NER en complément de l'extraction des termes pertinents offre une approche plus holistique et précise dans l'analyse de données textuelles. Alors que l'extraction des termes pertinents permet de repérer les mots-clés et les concepts importants dans un texte, le NER va plus loin en identifiant spécifiquement des entités telles que les noms de personnes, d'organisations, de lieux, de dates, etc. avec leur étiquette associée ce qui constitue un avantage pour la structuration du texte. Cette combinaison permet une compréhension plus fine du contenu en contexte. En effet, le NER offre une capacité à reconnaître les entités même si elles ne sont pas explicitement mentionnées comme des mots-clés, enrichissant ainsi la qualité de l'analyse en identifiant des informations cruciales qui pourraient être autrement manquées. Par exemple, dans une analyse de texte sur les actualités politiques, l'extraction des termes pertinents pourrait repérer des mots-clés comme « élection » ou « parti politique », mais seul un système de NER pourrait identifier les noms des candidats et des partis, fournissant ainsi une vue plus précise et complète de la situation. En combinant ces deux approches, nous maximisons la capacité à extraire des informations pertinentes et à les contextualiser, ce qui conduit à une analyse plus approfondie et des *insights* plus riches.

NER En ce qui concerne le système de NER, nous avons constaté un état de l'art conséquent avec les modèles de langues et plus précisément les SLMs qui ont marqué un véritable tournant dans le domaine. Nous nous orientons vers un modèle de langue basé sur BERT, notamment en raison de son caractère multilingue et de la multitude de modèles dérivés qu'il a inspirés. En effet, BERT et ses dérivés ont été pré-entraînés sur de vastes corpus de données provenant de différentes langues et domaines, ce qui constitue un riche panel de modèles mis à disposition en libre accès. Nous travaillons sur un corpus juridique et pratique en français, il est donc important de pouvoir adopter un système qui puisse s'adapter dans un domaine spécifique et une langue autre que l'anglais.

L'utilisation d'un modèle de langue comme CamemBERT classique s'avère judicieuse pour l'analyse de textes juridiques, compte tenu de sa pertinence dans la compréhension du langage naturel en français. Toutefois, dans le domaine spécifique du juridique, où la terminologie et la syntaxe peuvent être complexes et hautement spécialisées, un simple CamemBERT pourrait présenter des limites en termes de NER spécifiques au domaine juridique (que nous appellerons « entités juridiques »). Afin de pallier ces lacunes et d'optimiser la performance du système de NER dans le contexte juridique, il serait pertinent de créer un système d'entités juridiques supplémentaire. Ce système serait spécialement conçu pour identifier et extraire les entités juridiques telles que les lois, les articles, les décisions de justice, les parties impliquées dans un litige, etc. en s'appuyant sur des REGEX. En combinant les capacités de compréhension générale de CamemBERT avec un système d'entités juridiques spécialisé, nous pouvons garantir une reconnaissance précise et contextuelle des entités spécifiques au domaine juridique, améliorant ainsi la qualité et la pertinence des analyses dans ce domaine crucial.

Termes pertinents L'utilisation d'un LLM, notamment ChatGPT, pour l'extraction de termes pertinents dans le domaine juridique se justifie par sa capacité à traiter le langage naturel de manière contextuelle et à s'adapter aux spécificités des textes juridiques. Contrairement aux systèmes état de l'art qui peuvent manquer de flexibilité pour s'adapter efficacement aux domaines spécialisés comme le juridique, les LLMs offrent une approche plus générique et adaptable. Cependant, il est essentiel de souligner l'importance de l'*instruction tuning* pour

garantir des performances optimales dans le contexte juridique. Plusieurs revues d'IE ont en effet relevé que l'*instruction tuning* est cruciale pour améliorer la capacité des LLMs à extraire des informations pertinentes dans des domaines spécialisés comme le juridique. En prenant en compte ces recommandations, nous pouvons exploiter pleinement le potentiel des LLMs pour l'extraction de termes pertinents dans le domaine juridique, en veillant à réaliser une *instruction tuning* appropriée pour garantir des résultats précis et fiables.

Cependant, il reste à noter que nous n'avons pas trouvé de systèmes état de l'art à disposition au moment de l'implémentation du système. Nous avons donc décidé d'implémenter notre propre modèle.

Justification des choix d'évaluation Dans notre étude, nous avons fait plusieurs choix d'évaluation en raison des caractéristiques spécifiques de nos données et de notre domaine d'application.

1. **Jeux de données en anglais et pas du domaine :** Nous avons opté pour plusieurs évaluations en utilisant divers jeux de données et notamment en anglais malgré la langue de nos données en français, car il n'existait pas de corpus de spécialité dans le domaine juridique en français à notre disposition. Cette décision a été motivée par la nécessité de disposer de données riches et diversifiées pour entraîner nos modèles. Nous nous sommes donc orientés vers trois classes d'aspects d'évaluation :
   1.1 Corpus (très certainement) en anglais et potentiellement de spécialité : soit du juridique, soit d'un autre domaine.
   1.2 Corpus standard de vocabulaire général (Wikipédia par exemple) mais multilingue voir français.
   1.3 Évaluation sur un échantillon de notre corpus avec diverses méthodes d'évaluations pour une évaluation la moins biaisée possible. Cela peut s'apparenter via l'évaluation par QA à partir de l'extraction terminologique ou d'une autre tâche d'évaluation. Il peut s'agir aussi de l'utilisation des annotations déjà présentes dans les textes et de matcher la similarité.

   Les jeux de données dans le NER étant nombreux, nous avions un vaste choix à notre disposition et beaucoup de jeux de données multilingues. Cela n'a pas été forcément le cas pour l'extraction de termes pertinents.
2. **Disponibilité du dataset EUROPA :** Au moment de nos expériences, le dataset EUROPA (Salaün et al., 2024), qu'il aurait été judicieux d'utiliser, n'était pas encore disponible en libre accès. Cette limitation temporelle nous a contraints à utiliser d'autres sources de données pour nos évaluations.
3. **Évaluation par double check avec LLM :** Nous avons choisi d'évaluer nos systèmes en utilisant un double check avec un LLM. Cette approche consiste à évaluer un système contenant un LLM par un autre LLM, ce qui permet une évaluation plus rigoureuse et objective des performances.
4. **Évaluation par une autre tâche du TALN :** Nous avons choisi d'utiliser, entre autres, notre propre corpus pour certaines évaluations comme mentionné précédemment. En effet, nous avons relevé l'importance des entités juridiques mais aucun jeu de données n'existe dans ce contexte précis. Nous utilisons donc une autre tâche d'évaluation afin d'évaluer la pertinence des termes extraits. Par exemple, pour l'évaluation de l'extraction de termes pertinents, il est possible d'utiliser la tâche de QA ou de classification de documents. En ce qui concerne l'extraction de termes juridiques, Bendahman et al. (2023) s'est orientée sur une évaluation par le biais de la tâche de résumé automatique et de détection des hallucinations.

5. **Volume d'information et contexte :** Nous avons constaté qu'il n'était pas préjudiciable de détecter plus de termes que prévu dans nos évaluations. En effet, dans de nombreux cas, les *datasets* contenaient un nombre limité de termes à détecter, et une détection accrue de termes clés nous a permis d'enrichir le contexte et les connexions dans notre graphe, améliorant ainsi la compréhension globale du sujet.

6. **Évaluation de la frontière (*Boundary*) :** Nous avons adopté une approche d'évaluation de la frontière où nous vérifions que les termes clés du jeu d'évaluation que nous devrions repérer font partie de l'ensemble de tous les termes clés détectés. Bien que nous ne soyons pas préoccupés par une détection accrue de termes clés, il est crucial que tous les termes clés du corpus d'évaluation soient identifiés. Nous avons également introduit une double évaluation de la frontière, où il est acceptable d'avoir un terme clé presque équivalent avec des adjectifs supplémentaires ou des informations contextuelles, tant que le terme clé principal est correctement identifié.

3.3.2 Extraction de relations

La tâche d'extraction de relations est cruciale pour extraire des informations sémantiques à partir de textes, et il existe plusieurs approches pour y parvenir, comme explicité dans divers états de l'art récents et complets (Zhao et al., 2024; Smirnova & Cudré-Mauroux, 2019; Niklaus et al., 2018; Pawar et al., 2017). Les relations peuvent être considérées comme des relations lexico-sémantiques, comme les relations de synonymie ou d'antonymie par exemple, et il est possible d'étendre cette approche à d'autres types de relations entre les termes (hyperonymie, hyponymie, la méronymie ou partie-tout, l'holonymie : le tout par rapport à la partie, . . .).

Pour extraire des relations de synonymie, la similarité sémantique entre les termes peut être utilisée en associant chaque terme extrait à une représentation vectorielle en contexte, créée à partir de modèles basés sur les mots tels que Glove (Pennington et al., 2014a), Word2vec (Mikolov et al., 2013) et FastText (Bojanowski et al., 2017), ou de modèles de transformateurs dépendant du contexte tels que Elmo (Peters et al., 2018) et BERT (Devlin et al., 2019a). Ensuite, la similarité entre les vecteurs peut être calculée, par exemple, en utilisant la mesure du Cosinus. Si la similarité est élevée, cela indique que les termes sont synonymes. Nous pouvons considérer cette extraction de relations typées sous deux aspects : l'extraction de relations et la classification de relations.

Extraction de relations

Plusieurs travaux de recherche sur l'extraction de relations ont été proposés, deux grandes catégories d'approches existent (Granada et al., 2018) à savoir : (1) les approches à base de patrons lexico-syntaxiques (Hearst, 1992; Panchenko et al., 2013) ; et (2) les approches à base d'apprentissage automatique/apprentissage profond (Bunescu & Mooney, 2005a; Xue et al., 2018).

Les premiers travaux de recherche d'extraction de relations reviennent à Hearst (1992). Cette méthode repose sur des patrons lexico-syntaxiques et cherche à extraire des relations taxonomiques uniquement sur des textes anglophones. Plus tard, plusieurs recherches ont été faites sur des corpus en français. Panchenko et al. (2013) ont étendu les patrons afin de limiter un certain nombre de bruits sur les corpus en français. Par la suite, Panchenko et al. (2016) et Bordea et al. (2015a) ont montré que ce type d'approches produit des résultats intéressants, notamment du point de vue de la précision. Cependant, ce type d'approche ne permet pas de couvrir l'information en totalité. En effet, l'utilisation d'un nombre prédéfini

de patrons lexico-syntaxiques est problématique par la précision de certaines relations et par la gestion de l'ambiguïté de certains patrons.

Pour compenser ce problème et donc pour limiter l'utilisation des patrons lexico-syntaxiques, des approches à base d'apprentissage automatique ont été proposées que ce soit par supervision (Snow et al., 2004; Bunescu & Mooney, 2005a; Pantel & Pennacchiotti, 2006) ou par non supervision (Mintz et al., 2009; Morin & Jacquemin, 2004). L'une des premières approches de cette catégorie a été proposée par Brin (1998). Elle s'appuie sur une technique de *bootstrapping* qui consiste à sélectionner des patrons par apprentissage semi-supervisé afin de construire une base d'apprentissage. Sur la même idée, des techniques de sélection de mots par analyse distributionnelle et de sélection de traits sémantiques ont été utilisées pour identifier des relations entre des entités nommées (Etzioni et al., 2004). Des méthodes semi-supervisées ont aussi été utilisées et ont été entraînées sur des corpus annotés afin d'apprendre automatiquement des propriétés linguistiques signalant les relations visées. Plusieurs algorithmes d'apprentissage ont été employés comme les SVM (Zhao & Grishman, 2005; Bunescu & Mooney, 2005b), la régression logistique (Kambhatla, 2004), le *parsing* augmenté (Miller et al., 2000) et les CRF (Culotta et al., 2006).

Cartier (2015) a montré qu'il est possible d'hybrider les 2 types d'approches pour extraire des relations à partir d'un corpus d'un million de définitions provenant de deux ressources, à savoir : le TLFi, Trésor de la Langue Française informatisé (Dendien & Pierrel, 2003) et Wikipédia. Cette approche combine la précision des patrons lexico-syntaxiques et le rappel des méthodes statistiques par analyse distributionnelle, et est appliquée à plusieurs relations. Le but de ce travail est d'obtenir automatiquement une ressource sémantique pour le français contemporain à partir d'un corpus de textes.

Angeli et al. (2014) utilisent des techniques d'apprentissage actif (*active learning*) afin d'extraire des relations de manière supervisée à distance (*distantly supervised*). Ils appliquent l'apprentissage actif à la classe de modèles *latent-variable distantly supervised*. Ils utilisent des petits corpus d'exemples annotés afin d'améliorer la précision de bout en bout du système.

Le modèle CoRE-NEPD (Graph-based Cross-document Relation Extraction via Non-bridge Entity Enhancement and Prediction Debiasing) (Yue et al., 2024) est une approche qui améliore significativement l'extraction de relations entre des entités cibles situées dans des documents différents. Il surmonte plusieurs limitations des méthodes antérieures, notamment l'ignorance des entités non-ponts et le biais de prédiction introduit par les nombreuses instances sans relation (NA). Le modèle CoRE-NEPD est composé de trois composants : un encoder BERT, un *Graph Recurrent Network* (GRN) et un classifieur. L'encodeur BERT sert à la construction du graphe d'entités. Les entités cibles, ponts et non-ponts sont identifiées et connectées selon les arêtes de co-occurrence et sémantiquement liées. Le GRN est utilisé pour encoder le graphe d'entités, capturant ainsi les dépendances complexes entre les entités. Enfin, l'application du classifieur permet de prédire les relations en utilisant la distribution de prédiction dé-biaisée.

LLM Plusieurs tâches de l'IE ont été explorées lors de l'essor des LLMs, notamment l'extraction de relations a été abordée. ATG (Urchade et al., 2024) est une méthode pour les problèmes d'entité commune (*joint entity*) et d'extraction de relation à partir de textes non structurés. Ces problèmes ont été modélisés comme un problème de génération de séquences conditionnelles (*conditional sequence generation problem*), ce qui implique que le modèle est basé sur des segments textuels (*spans*) plutôt que des générateurs (de gauche à droite) au niveau des *tokens*. ATG génère un graphe linéarisé où les nœuds représentent des portions de texte et les arêtes des triplets de relations. Elle utilise une architecture transformeur-encodeur-décodeur avec un mécanisme de pointage sur un vocabulaire dynamique de *spans*

et de types de relations.

Zhou et al. (2023b) ont évalué la confiance des LLMs en contexte en utilisant deux aspects : le conflit de connaissances et la prédiction avec abstention. Cette confiance peut être améliorée de manière significative en utilisant des stratégies de *prompting* soigneusement conçues. Ces expériences ont notamment été menées sur la tâche d'extraction de relations et les résultats ont démontré une amélioration significative de la confiance en contexte.

Foppiano et al. (2024b) ont mené une étude pour évaluer les capacités des LLMs pour l'extraction d'informations structurées à partir de documents scientifiques dans le domaine de la science des matériaux. L'étude concerne donc, d'une part, la NER des matériaux étudiés et des propriétés physiques et, d'une autre part, l'extraction de relations entre ces entités. La performance des LLMs dans l'exécution de ces tâches est comparée aux modèles traditionnels basés sur l'architecture BERT et les approches basées sur les règles. Selon les résultats de ces études, les LLMs ne parviennent pas à surpasser dans le cadre du *Zero-Shot* mais un GPT-3.5-Turbo finement ajusté avec la stratégie appropriée pour RE surpasse tous les modèles. De plus, l'article met en évident l'efficacité de GPT-4 pour l'extraction de relations sans ajustement. Avec peu d'exemples, les *baselines* sont surpassées. Il semble donc nécessaire de s'intéresser à l'utilisation des LLMs sur nos données du fait de cette étude sur un corpus spécifique qui présente de bons résultats.

RationaleCL (Xiong et al., 2023) est une méthode qui incorpore un raisonnement généré par LLM dans la tâche du *Continual Relation Extraction* (CRE) - qui consiste à prendre en considération les relations qui émergent constamment - afin de permettre au modèle de se souvenir de ce qui a été fait dans les tâches précédentes. Le système a été expérimenté sur FewRel (Han et al., 2018) et TACRED (Zhang et al., 2017a).

De nouveau, lorsque nous nous intéressons aux domaines de spécialité, le biomédical ressort. Jimenez Gutierrez et al. (2022) ont comparé les SLMs et les LLMs pour les deux tâches de l'extraction d'information : NER et l'extraction de relations. Dans le cadre de l'extraction de relations, l'étude relève que les LLMs n'ont pas su se montrer aussi performant que les SLMs *finetuné* pour l'extraction de relations en *few-shot*. En effet, les LLMs ont du mal à gérer la catégorie des relations None-of-The-Above (NoTA). Chen et al. (2023), précédemment cités pour le NER, qui ont réalisé une étude complète de la performance des LLMs dans le domaine du biomédical pour 6 tâches du TALN notamment dans le cas de l'extraction de relations. Les résultats ont montré qu'il est préférable d'utiliser un modèle BERT affiné qui a montré de meilleures performances et ce sur les tâches extractives et classification (dont l'extraction de relations mais pas que). Luo et al. (2023) ont également réalisé une étude sur le biomédical en anglais et en chinois de la performance des LLMs *finetuné* ou non concernant plusieurs tâches de TALN. Cette étude montre que les LLMs affinés sur des données d'instruction spécifiques au domaine peuvent améliorer la performance sur les tâches en aval du domaine. Cependant, il est à noter que dans le cadre des tâches de NER et d'extraction de relations les méthodes discriminatives conventionnelles surpassent les méthodes avec LLM. Dans l'état de l'art de Zhao et al. (2024), il est intéressant de noter que l'état de l'art s'est intéressé à l'extraction de relations de domaines spécifiques dans les challenges liés à l'extraction de relations et les solutions mises en œuvre. Le domaine du légal a été mis en évidence où il est question d'extraire des relations légales entre des entités contenues dans des documents juridiques. Andrew (2018) a exploré une combinaison de techniques statistiques et basées sur des règles sans utiliser de données étiquetées pour identifier automatiquement les entités et les relations dans les documents juridiques. Parmi les méthodes d'apprentissage semi-supervisé, Thomas & Sangeetha (2021) ont proposé une méthode d'apprentissage semi-supervisée basée sur des modèles pour extraire des faits relationnels à partir de textes judiciaires. Cette méthode combine le bootstrapping

et les techniques OBIE (Ontology-Based Information Extraction) pour accélérer l'extraction des faits judiciaires. Bi-FLEET (Wang et al., 2021) est modèle développé pour l'extraction d'éléments de contrat inter-domaines. Ce modèle intègre un encodeur de relations entre les éléments des clauses avec un schéma de rétroaction bidirectionnelle, en utilisant un cadre multi-tâches pour capturer les interactions entre l'extraction d'éléments de contrat et la classification des clauses. Pour finir, ConReader (Xu et al., 2022a) est un framework qui explore les relations contextuelles à long terme, les relations de définition de termes et les relations entre des clauses similaires dans l'extraction de clauses de contrat. Cependant, les solutions actuelles pour l'extraction de relations dans le domaine juridique présentent plusieurs limites (Zhao et al., 2024). Tout d'abord, la plupart des techniques d'apprentissage profond nécessitent de grandes quantités de données étiquetées pour être efficaces. Cependant, dans le domaine juridique, il est souvent difficile de disposer de suffisamment de données étiquetées, ce qui peut limiter la performance des modèles. Ensuite, les documents juridiques varient énormément en termes de format, structure et contenu, ce qui rend difficile la généralisation des modèles d'extraction de relations à différents types de documents juridiques. De plus, les relations juridiques évoluent au fil du temps avec les changements de lois et de règlements. Les modèles doivent être continuellement mis à jour et adaptés pour rester pertinents, ce qui peut être coûteux et chronophage. Enfin, les documents juridiques contiennent souvent des relations imbriquées et des entités chevauchantes. La gestion de ces relations complexes est un défi majeur pour les modèles actuels, qui peuvent ne pas être capables de les identifier et de les extraire correctement. En résumé, bien que les techniques modernes d'extraction de relations offrent des améliorations significatives et intéressantes à explorer, elles doivent encore surmonter plusieurs défis pour être pleinement efficaces et fiables dans le domaine juridique.

Dans le cadre de l'apprentissage d'ontologies (*Ontology Learning* - OL), Babaei Giglou et al. (2023) se sont intéressés à utiliser les LLMs afin de capturer les relations linguistiques complexes. Pour cela trois tâches ont été explorées : le typage de termes, la découverte de taxonomies et l'extraction de relations non-taxonomiques. Les résultats ont mis l'accent sur deux points : les LLMs basiques ne sont pas suffisamment adaptés aux tâches explorées. Néanmoins, les auteurs soulignent que lorsqu'ils sont affinés de manière efficace (*instruction tuning*), ils performent et permettent de réduire le goulot d'étranglement de l'acquisition de connaissances, pour la construction d'ontologies. Pour nous qui cherchons à construire un graphe de connaissances dont l'idée peut se rapprocher d'une ontologie, cette étude est à prendre en considération.

Malgré le potentiel des LLMs, les SLMs ont montré régulièrement de meilleurs résultats dans le cadre de l'extraction de relations notamment à cause de deux points relevés dans les articles précédents : Le premier vient de la faible pertinence concernant l'entité et la relation dans les démonstrations récupérées pour l'apprentissage en contexte ; et le second, la forte tendance à classer à tort les exemples NoTA dans d'autres étiquettes prédéfinies. GPT-RE (Wan et al., 2023) incorpore des représentations d'entités spécifiques aux tâches dans la recherche de démonstrations pour palier au premier problème. Le second problème est abordé par leur enrichissement des démontrations avec une logique de raisonnement induit par des étiquettes de référence (*gold label-induced reasoning logic*). GPT-RE a obtenu des performances SoTA sur des jeux de données reconnus pour la tâche d'extraction de relation (SemEVAL et SciERC Luan et al. (2018)). Il a également été évalué sur TACRED (Zhang et al., 2017a) et ACE05 Walker et al. (2006) sans passé SoTA.

Zhao et al. (2024) ont mis en évidence dans leur état de l'art les techniques d'apprentissage profond existantes pour l'extraction de relations. Ils se sont intéressés aux ressources telles que les jeux de données ou les métriques pour l'évaluation de la tâche d'extraction de

relations. Ils catégorisent également les différents travaux en trois catégories : la représentation textuelle, l'encodage du contexte et la prédiction de triplets. Cet état de l'art traite de toute la palette des techniques d'apprentissage profond que ce soit les réseaux neuronaux, les transformers et SLM (pré-entraînés) ainsi que les LLMs. Ils ont d'ailleurs mis en évidence l'extraction de relations avec les LM pré-entraînés et les performances remarquables des modèles pré-entrainés et finetunés. Les modèles *few-shot* ont également présenté des perfomances remarquables. Quant aux LLMs, ils ont montré d'impressionnantes capacités en utilisant le *prompting*.

Représentation de relations

Zhao et al. (2024) ont discuté de la représentation textuelle en utilisant des techniques d'apprentissage profond dans leur état de l'art pour l'extraction de relations. Ils pointent quatre niveaux de représentations : (1) au niveau des mots (*Word-level Embeddings*), (2) au niveau des caractères, (3) au niveau de la position, et (4) au niveau syntaxique. Les représentations textuelles au niveau des mots encodent chaque mot dans un espace vectoriel à valeurs réelles. Les mots ayant des significations similaires sont représentés par des vecteurs proches dans cet espace. Les approches récentes utilisent des embeddings de mots préentraînés comme Word2Vec et GloVe, qui encodent la signification des unités d'entrée dans un espace continu. Ces représentations permettent de capturer les propriétés syntaxiques et sémantiques des mots, ce qui améliore les performances des modèles d'apprentissage profond pour l'extraction de relations. Les représentations au niveau des caractères capturent les informations morphologiques des mots. Elles sont particulièrement utiles pour gérer les mots rares ou les variantes orthographiques. Les approches de RE basées sur des caractères peuvent être combinées avec des représentations au niveau des mots pour améliorer la robustesse et la précision des modèles. Les représentations enrichies par la position incorporent des informations sur la position relative des mots dans une phrase. Cela est crucial pour déterminer les relations entre les entités qui dépendent de leur position dans le texte. Enfin, Les représentations syntaxiques utilisent des informations issues de l'analyse syntaxique, comme les étiquettes grammaticales (POS) et les dépendances syntaxiques, pour enrichir les plongements (*embeddings*) de mots. Ces représentations aident à capturer des relations plus complexes en intégrant à la fois la syntaxe et la sémantique.

En utilisant des représentations à différents niveaux (mots, caractères, positions, syntaxe), les modèles d'extraction de relations peuvent mieux comprendre et désambiguïser les relations entre les entités. De plus, des représentations textuelles plus précises permettent aux modèles d'extraction de relations d'atteindre des performances plus élevées, car elles fournissent des informations contextuelles et sémantiques cruciales pour la tâche. Et enfin, des représentations textuelles robustes permettent aux modèles d'extraction de relations de s'adapter à différents contextes et types de textes, augmentant ainsi leur applicabilité et leur efficacité dans des domaines variés.

Plusieurs travaux ont été proposés permettant l'apprentissage des vecteurs de relations pour un ensemble de paires de mots *(a, b)*, sur la base de phrases dans lesquelles ces paires de mots coexistent. Par exemple, la méthode appelée *Latent Relational Analysis* (LRA) (Turney, 2005) repose d'abord sur l'identification d'un ensemble de patrons lexicaux suffisamment fréquents. Cette méthode construit une matrice pour laquelle la fréquence des patrons lexicaux est mesurée pour chaque paire de mots. Les vecteurs de relations sont ensuite obtenus en utilisant la décomposition en valeurs singulières. Par ailleurs, plus tard, une approche inspirée du modèle GloVe (Pennington et al., 2014b), un modèle de plongements lexicaux, a été proposée (Jameel et al., 2018). L'idée est d'apprendre des vecteurs de relations en tenant compte des statistiques de cooccurrences entre une paire cible de mots (a, b) et d'autres

mots.

Une approche totalement différente de celles citées a été proposée par Washio & Kato (2018). Ils ont entraîné un réseau de neurones pour prédire les chemins de dépendances à partir d'une paire de mots. Leur approche utilise des vecteurs de mots comme entrée, par conséquent, les informations relationnelles sont codées implicitement dans les poids du réseau de neurones, plutôt que sous forme de vecteurs de relation (bien que la sortie de ce réseau de neurones, pour une paire de mots, puisse toujours être considérée comme un vecteur de relation). Le modèle Pair2Vec (Joshi et al., 2019) proposé met l'accent sur l'apprentissage des vecteurs de relations pouvant être utilisés pour les mécanismes d'attention de phrases croisées dans des tâches telles que les systèmes de questions-réponses ou l'implication textuelle. Un des problèmes liés à l'implémentation de ces approches est le temps coûteux en consommation des ressources. Par exemple, Pair2Vec aura besoin d'un temps d'entraînement entre 7 et 10 jours sur un corpus de taille représentative et une utilisation d'un matériel physique standard (Camacho-Collados et al., 2019a).

D'autres approches ont été proposées pour améliorer la représentation des vecteurs de relations en utilisant des connaissances provenant de bases de données externes. Ces approches reposent sur le fait que deux mots sont similaires (c'est-à-dire, leurs vecteurs sont proches dans l'espace de représentation) si et seulement si ils sont liés l'un à l'autre dans un graphe de connaissances donné (Faruqui et al., 2015; Speer R., 2017). L'avantage de ces approches est de travailler sur des relations typées, par exemple : l'hyperonymie (*généralisation*), l'hyponymie (*spécification*) ou la synonymie (*équivalence*), ce qui n'est pas forcément le cas dans un cadre d'apprentissage non supervisé.

Classification de type de relations

Sur la question de la classification des relations, nous pouvons diviser ce problème en deux parties en détaillant quatre points. Premièrement, il y a les méthodes traditionnelles d'extraction (Qian et al., 2008). Deuxièmement, il y a les méthodes statistiques divisées en deux points : la classification supervisée et la classification non supervisée. Différents modèles de classification binaire ont été proposés : des champs aléatoires conditionnels (CRF) (Sutton & McCallum, 2007) à la classification *Naive Bayes* (Suresh kumar & Zayaraz, 2015) en passant par l'utilisation de machine à vecteurs de support (SVM) (Uzuner et al., 2011; Minard et al., 2011). Nous pouvons classer les méthodes non supervisées en deux catégories : d'une part, les méthodes de regroupement (Gábor et al., 2016) et d'autre part, les méthodes de réseaux neuronaux.

Les méthodes basées sur les réseaux neuronaux ont été plus étudiées. En se basant sur les CNNs, Nguyen & Grishman (2015) proposent un modèle d'extraction de relations qui, à l'étape finale, classifie le type de relations. D'autres modèles CNNs ont été proposés (Zeng et al., 2014; Xu et al., 2015; Lee et al., 2017) et (dos Santos et al., 2015). Le classement par CNN peut être une technique (dos Santos et al., 2015). D'autres méthodes de classification des relations basées sur les réseaux neuronaux ont été proposées, comme l'attention multi-niveaux (Wang et al., 2016) ou les réseaux neuronaux récurrents (RNNs) basés sur des LSTMs (Zheng et al., 2017; Peng et al., 2017).

ReRe (Xie et al., 2021) décompose son extraction de relations en deux composants : la classification de relations et l'extraction d'entités. La classification de relations est vue comme un problème de classification multi-label. En ce qui concerne la classification de relations au niveau des phrases, le modèle utilise un BERT ou RoBERTa pour représenter les *tokens* de la phrase et les classifier en sortie finale en utilisant une fonction d'activation par sigmoïde. Il s'agit ici de voir le problème de l'extraction de relations comme une tâche d'apprentissage positive non étiquetée plutôt que d'utiliser la supervision à distance afin de générer

automatiquement des échantillons pour l'apprentissage dans le cadre de l'extraction qui engendre des étiquettes bruitées et des données d'entraînement déséquilibrées par la quantité d'étiquettes négatives trop grande.

Évaluation et corpus

Les jeux de données SciERC, CoNLL 2004, ACE'05 (-Rel) ou encore NYT déjà mentionnés pour le NER contiennent également des annotations concernant l'extraction de relations. NYT-HRL (NYT10-HRL basé sur NYT10 et NYT11-HRL basé sur NYT11) (Takanobu et al., 2019) est un jeu de données basé sur NYT auquel est appliqué un apprentissage par renforcement hiérarchique (*hierarchical reinforcement learning* - HRL) pour décomposer en hiérarchie la détection de relations et l'extraction des entités de sorte qu'il soit plus facile de traiter les relations qui se chevauchent.

De nombreux systèmes de classification des relations ont été évalués ces dernières années dans le cadre de campagnes d'évaluation de référence telles que SemEval-2018 (tâches 7 (Gábor et al., 2018) et 9 (Camacho-Collados et al., 2018)). La tâche 7 consistait en l'extraction et la classification de relations dans des articles scientifiques. La tâche 9, quant à elle, consistait à découvrir des termes liés par l'hyperonymie. La tâche 17 de SemEval-2015 (Bordea et al., 2015b) proposait également de trouver des relations d'hyperonymie et d'hyponymie entre des termes donnés, et SemEval-2007 (Girju et al., 2007) avec une tâche de classification des relations. La tâche 10 de SemEval-2017 s'intéressait également à l'aspect de l'extraction de relations mais était adaptée aux articles scientifiques. Il reste intéressant de se pencher sur les recherches associées.

Google-IISc Distant Supervision (GIDS) (Jat et al., 2018) est un jeu de données en anglais pour l'extraction de relations supervisée à distance (*distantly-supervised relation extraction*). La supervision à distance a été utilisée pour créer à la fois des jeux d'entraînement et de test mais introduit cependant du bruit. Pour retirer le bruit, GIDS a cherché à s'assurer que les relations labellisées étaient correctes pour chaque instance en commençant avec un échantillon de relations catégorisées humainement. Le corpus est composé de 5 types de relations binaires.

KBP37 (Zhang & Wang, 2015) est une version modifiée du jeu de données en anglais présenté dans Angeli et al. (2014). Il s'agit de collections de documents issus de la KBP de 2010 et 2013 ainsi qu'une version de juillet 2013 de corpus textuels issus de Wikipédia. Par la suite, cet ensemble de corpus de données a été annoté produisant ainsi 33811 phrases annotées. KBP37 contient 18 relations uni-directionnelles ainsi qu'une relation supplémentaire nommée « no_relation », donnant ainsi 37 classes de relations.

En ce qui concerne les domaines spécifiques, de nouveau le biomédical s'est intéressé à la tâche d'extraction de relations. ADE Corpus (Gurulingappa et al., 2012) est un jeu de données sur les effets indésirables des médicaments. Il s'agit d'un ensemble de données permettant de déterminer si une phrase est liée à un effet indésirable (Vrai) ou non (Faux) et d'extraire les relations entre l'effet indésirable et le médicament. Ce jeux de données contient donc deux composantes. D'une part « DRUG-AE.rel »et « DRUG-DOSE.rel »concernent les relations. Le premier fournit des relations entre les médicaments et les effets indésirables. Le second fournit des relations entre les médicaments et les dosages. D'une autre part, « ADE-NEG.txt » fournit toutes les phrases du corpus ADE qui ne contiennent PAS d'effets indésirables liés aux médicaments.

Dans l'état de l'art de Zhao et al. (2024) est mis en évidence le premier ensemble de données juridiques destiné à la révision de contrats (Hendrycks et al., 2021). Cet ensemble de données, nommé CUAD, permet de modéliser les relations implicites dans les documents juridiques, comme les relations criminelles dans les jugements et les relations entre

les clauses dans les contrats. Ce jeu de données est en anglais. Cependant, ce jeu de données est trop spécialisé dans un sous domaine juridique éloigné des documents que nous traitons. En effet, comme exprimé précédemment, les documents juridiques varient énormément en termes de format, structure et contenu, ce qui rend difficile la généralisation des modèles mais également des jeux de données pour différents types de documents juridiques. Cependant, ce jeu de données reste intéressant à explorer car il reste un jeu de données dans un domaine de spécialité au plus proche du nôtre.

Le jeu de données FewRel (Han et al., 2018) est un ensemble de données de classification des relations pour le *Few-Shot*, composé de 70 000 phrases sur 100 relations dérivées de Wikipédia et annotées par des *crowdworkers*. La relation de chaque phrase est d'abord reconnue par des méthodes de supervision distante, puis filtrée par des utilisateurs.

Enfin, en ce qui concerne le multilingue, MultiTACRED (Zhang et al., 2017a; Alt et al., 2020; Stoica et al., 2021) est une version multilingue de l'ensemble de données à grande échelle *TAC Relation Extraction Dataset* (TACRED) (Zhang et al., 2017b). TACRED est un ensemble de données d'extraction de relations à grande échelle comprenant 106 264 exemples construits sur des fils d'actualité et des textes web provenant du corpus utilisé dans les défis annuels de la *TAC Knowledge Base Population* (TAC KBP). TACRED couvre 41 types de relations tels qu'ils sont utilisés dans les défis TAC KBP (par exemple, *per :schools_attended* et *org :members*) ou sont étiquetés comme *no_relation* s'il n'y a pas de relation définie. Ces exemples sont créés en combinant les annotations humaines disponibles dans les défis TAC KBP et le *crowdsourcing*. MultiTACRED a traduit automatiquement les instances de l'ensemble de données TACRED original et a projeté automatiquement leurs annotations d'entités dans 12 langues dont le français. Les traductions sont validées syntaxiquement en vérifiant l'exactitude du balisage XML. Toute traduction présentant une structure de balises invalide, par exemple des paires de balises de tête ou de queue manquantes ou invalides, est rejetée. SMiLER (Seganti et al., 2021) fournit des annotations de relations pour des textes dans 14 langues avec 36 types de relations (y compris le « no relation »). Les ensembles de tests des langues permettant d'évaluer contiennent plus de 12 000 échantillons.

Afin d'évaluer l'extraction de relations, différentes métriques ont été utilisées mais s'appuient sur les mêmes que pour l'extraction d'entités. En effet, l'évaluation dite souple (ou également dite *Boundaries*) nécessite de prédire correctement les entités de manière souple avec le type de relation associé. Le triplet de relations est donc correct si le modèle prédit correctement les limites (ou *boundaries*) de l'entité sujet, de l'entité objet et de la relation d'entité. L'évaluation dite stricte nécessite en plus une prédiction précise du type d'entité.

Récemment, les LLMs ont été abondamment utilisés pour générer des ensembles de données uniquement synthétisés à partir des LLMs. Cependant, il se peut que certaines données présentent une qualité médiocre et nécessitent tout de même des données étiquetées supplémentaires. LLMaAA (Zhang et al., 2023) utilise les LLMs comme annotateurs et les place dans une boucle d'apprentissage actif pour déterminer ce qu'il faut annoter efficacement. Afin d'apprendre de manière robuste avec des pseudo-étiquettes, un processus d'annotation et d'apprentissage a été mis en place. Dans un premier temps, des exemples k-NN d'un petit pool de démonstration comme exemples en contexte sont tirés. Dans un second temps, une technique de (re-)pondération des exemples est adoptée. Elle consiste à assigner aux échantillons d'apprentissage des poids pouvant être appris. Il est intéressant de noter que LLMaAA a été expérimenté pour la tâche de NER et d'extraction de relations. Selon les résultats de leur expérience, seuls quelques centaines d'exemples annotés suffisent pour permettre le bon apprentissage des modèles utilisant LLMaAA.

Discussion

L'extraction de relations, élément clé de l'extraction d'informations, reste un domaine de recherche difficile en raison de la complexité du langage humain. Récemment, beaucoup se tournent vers la supervision à distance (*distant supervision*). La supervision à distance est le processus de spécification du concept que les mots individuels d'un passage, généralement une phrase, tentent de transmettre. Par exemple, une base de données contient la relation structurée : *concerns(NLP, this sentence)*. Notre système de supervision à distance prendrait la phrase en entrée : « *This is a sentence about NLP.* ». Sur la base de cette phrase, il reconnaîtrait les entités, (outil de reconnaissance d'entités nommées, par exemple) « *NLP* » et « *this sentence* ». Comme notre base de données indique que « *NLP* » et « *this sentence* » sont liés par le lien de *concern(s)*, il identifiera la phrase d'entrée comme exprimant la relation *concerns(NLP, this sentence)*. Il s'agit, ici, d'un exemple et le système pourra généraliser les cas d'usage. Cependant, les bases de données sont encore loin d'être totalement achevées. En théorie, en complétant les bases de données autant que possible, cela rend la supervision à distance puissante pour l'extraction de relations. D'autre part, certaines chaînes d'inférence efficaces peuvent être envisagées pour extraire davantage de relations lorsqu'aucune entité n'est mentionnée dans les bases de données.

Se tourner vers les méthodes basées sur les réseaux neuronaux semble prometteur car ils utilisent peu de caractéristiques (*features*) manuelles ou extraites par des outils de TAL au préalable. Pour compléter les réseaux neuronaux, il semblerait intéressant de s'orienter vers les mécanismes d'attention (*attention mechanism*). Nous pouvons exploiter plusieurs niveaux d'attention pour traiter le problème des multi-instances et des multi-relations. Il est cependant primordial de garder à l'esprit que l'extraction d'entités et l'extraction de relations s'influencent mutuellement. L'extraction conjointe peut être considérée comme une méthode afin d'accroître les performances.

Par rapport à cet état de l'art, nous nous positionnons dans un cadre de création de représentations vectorielles de relations. Notre approche s'appuie sur des plongements de mots/termes pour ensuite en déduire des vecteurs de relations de différents types. Nous utilisons ces représentations vectorielles pour y appliquer des méthodes d'apprentissage automatique dans un cadre de classification de relations.

3.3.3 Open Information Extraction

Han et al. (2020) soulève le problème des modèles entrainés d'extraction de relations et proposent différents axes de recherche comme l'utilisation de plus de données, l'utilisation du *FewShot Learning*, les travaux utilisant un contexte plus compliqué (document plutôt que phrases) et enfin le domaine ouvert.

Comme le montre la figure 3.3, l'extraction ouverte d'informations (*Open Information Extraction* - OpenIE) extrait des triplets de relations dites ouvertes et d'entités au sein d'un texte. Le but des systèmes ouverts consiste à ne pas utiliser de types de relations au sein du triplet prédéfini. Il s'agit donc de découvrir des types de relations inédits à partir de données non supervisées. Transformer les extractions faites en OpenIE vers des faits ontologiques est connu en tant qu' « extraction de relation ». Plusieurs revues dans le domaine de l'OpenIE ont été faites (Li et al., 2021; Zouaq et al., 2017; Glauber & Claro, 2018; Niklaus et al., 2018; Liu et al., 2022; Vo & Bagheri, 2016; Ali et al., 2019). TextRunner (Banko et al., 2007; Yates et al., 2007) a été le premier système et a introduit la notion d'OpenIE afin de réduire les efforts manuels des systèmes existants. Le but était de ne pas se limiter à un petit ensemble de relations cibles connues en avance mais de s'ouvrir afin de rester indépendant du do-

maine dans l'extraction et la découverte d'informations. TextRunner consiste à utiliser un ensemble de patrons (*patterns*) pour extraire des tuples relationnels d'une phrase et ce à la fois pour les arguments et les relations sémantiques entre les arguments.

Trois challenges majeurs ont été soulevés pour l'OpenIE : (1) l'automatisation, (2) l'hétérogénéité du corpus et (3) l'efficacité. La nécessité de l'automatisation repose sur le fait de vouloir réduire au minimum le travail manuel en développant des stratégies non supervisées sans spécifier les relations cibles à l'avance. Le système détecte automatiquement des relations possibles. L'hétérogénéité du corpus est à prendre en compte dans l'OpenIE et les outils actuels comme les analyseurs (*parsers*) syntaxiques ou de dépendances ou les NER ne sont pas adaptés car ils nécessitent bien souvent un domaine spécifique clairement identifié. Dans l'OpenIE, il est encouragé d'utiliser des méthodes d'analyse syntaxique (*shallow parsing method*) comme les systèmes de *PosTagging*. Et enfin, l'efficacité consiste à extraire rapidement les informations malgré la quantité de données (*Big data*). Le principe de l'OpenIE est de rester en surface mais de privilégier la rapidité car le fait d'avoir beaucoup de données, cela compense la précision. Les techniques de *linguistic features* comme le *PosTagging* seront donc privilégiées aux *deep features* comme le *parse tree*.

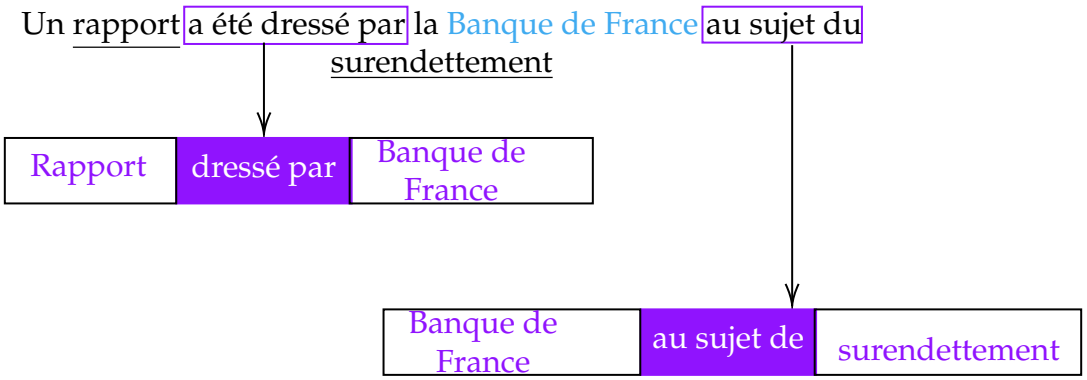

FIGURE 3.3 – Exemple d'une extraction d'information dans le cadre de l'OpenIE sur la phrase « Un rapport a été dressé par la Banque de France au sujet du surendettement ».

Ici, nous nous concentrons sur l'extraction de triplets relationnels ouverts (*Open Relational Triple*). Nous pouvons découper les systèmes de l'OpenIE en deux catégories : (1) les premiers systèmes ; et (2) les systèmes de nouvelle génération.

Systèmes de première génération. Les systèmes basés sur l'apprentissage de patrons d'extraction reposent sur une analyse de dépendance. Les systèmes d'IE ouverts comme TextRunner (Banko et al., 2007), WOEpos et WOEparse (Wu & Weld, 2010) se concentrent sur l'extraction de relations binaires de la forme (arg_1, segment textuel de relation, arg_2) à partir de textes. TextRunner est un système de type *shallow syntactic information* qui utilise un modèle *Naive Bayes* avec des fonctions POS et des blocs (*chunks*) de groupes nominaux, formées à l'aide d'exemples de *features* non lexicalisées. Leurs exemples ont été générés de manière heuristique à partir de la *Penn Treebank*. Les systèmes WOE utilisent Wikipédia comme source de données d'entraînement pour leurs extracteurs, ce qui conduit à des améliorations par rapport à TextRunner. Les systèmes WOE utilisent des *features* d'analyse de dépendances plutôt que des caractéristiques linguistiques peu profondes. Cependant, ils perdent en vitesse d'extraction.

O-CRF (Banko & Etzioni, 2008) est un système *OpenIE* qui utilise un modèle CRF. H-CRF est, quant à lui, un extracteur basé sur les ensembles (*ensemble-based extractor*) qui apprend à

combiner les output des systèmes d'extraction de relation lexicalisés et non-lexicalisés. StatSnowball (Zhu et al., 2009) utilise des champs aléatoires de Markov logiques discriminants (*discriminative Markov logic networks - discriminative* MLN) et atténue les règles strictes en apprenant leurs poids dans un sens d'estimation maximale de vraisemblance. La sélection des motifs est effectuée en résolvant une estimation de maximum de vraisemblance pénalisée par la norme l_1.

Certains systèmes sont basés sur des clauses. Il s'agit de segmenter les phrases en tuples en incorporant une étape de restructuration de phrases et en transformant les phrases complexes en un ensemble de clauses indépendantes syntaxiquement simplifiées. ClauseIE (Corro & Gemulla, 2013) est basé sur les clauses en exploitant des connaissances linguistiques sur la grammaire de la langue anglaise pour détecter d'abord les clauses dans une phrase d'entrée et pour identifier ensuite le type de chaque clause en fonction de la fonction grammaticale de ses constituants. ClausIE est basé sur l'analyse syntaxique des dépendances et un petit ensemble de lexiques indépendants du domaine.

Reverb (Fader et al., 2011) implémente deux contraintes syntaxiques et lexicales simples sur les relations binaires exprimées par des verbes, ce qui a plus que doublé l'aire sous la courbe de précision-rappel (*precision-recall curve*) par rapport aux extracteurs précédents tels que TextRunner et WOEpos. Reverb a ensuite été adapté afin d'extraire des connaissances à partir de textes français (Gotti & Langlais, 2018). Il s'agit d'un classifieur peu supervisé puisque les triplets proviennent du Wikipédia Français. R2A2 (Etzioni et al., 2011) marque un tournant dans les systèmes d'OpenIE et s'appuie sur Reverb en ajoutant un identificateur d'argument (*Arg Learner*) qui consiste à extraire les arguments d'une phrase (de relation). Ce système cherche à pallier deux problèmes de sa génération : les extractions incohérentes et les extractions non informatives.

Les informations sémantiques capturent non seulement les structures linguistiques des phrases, mais aussi le sens littéral des phrases, qui peuvent exprimer des relations plus diverses et plus adaptées par rapport aux modèles syntaxiques. Cependant, les informations sémantiques peuvent également être trop spécifiques et donc conduire au problème de canonisation (Vashishth et al., 2018; Wu et al., 2018; Galárraga et al., 2014). Certains systèmes ont donc cherché à utiliser les informations sémantiques via l'étiquetage des rôles sémantiques comme il est le cas de EXEMPLAR (de Sá Mesquita et al., 2013). L'OpenIE est étroitement liée à l'étiquetage des rôles sémantiques (*Semantic role labeling* - SRL) (Punyakanok et al., 2008; Toutanova et al., 2008) tels que agent, thème et artefact et, cela, dans la mesure où les deux tâches extraient des relations et des arguments des phrases. SRL-IE (Christensen et al., 2010) ont utilisé un analyseur SRL dans cette optique. Cependant, les systèmes SRLs s'appuient traditionnellement sur des analyseurs syntaxiques, ce qui les rend sensibles aux erreurs d'analyse et substantiellement plus lents que les systèmes OpenIE tels que Reverb. Cette différence est particulièrement importante lorsqu'il s'agit d'opérer sur un corpus de grande taille. Enfin, les SRL nécessitent des ressources sémantiques construites à la main comme Propbank (Kingsbury & Palmer, 2002) et FrameNet (Baker et al., 1998).

Certains systèmes se sont intéressés à d'autres langues et sont souvent des approches à base de règles. ArgOIE et PredPatt sont parmi les plus populaires. ArgOIE (Gamallo & Garcia, 2015) est un système multilingue basé sur des règles et qui prend en entrée des analyses de dépendances au format CoNLL-X. Il identifie les structures d'arguments dans les analyses de dépendances et extrait un ensemble de propositions de base de chaque structure d'arguments. PredPatt (White et al., 2016) consiste à extraire des prédicats et arguments. Il prend en entrée une analyse syntaxique de dépendance et retourne un ensemble de structures prédicat-arguments en appliquant des patrons syntaxiques. PredPatt fait partie d'un projet plus gros appelé « *Universal Dependency* ». Il est donc multilingue car *Universal Depen-*

dency supporte plusieurs langues.

Les premiers systèmes d'OpenIE se sont concentrés sur les relations exprimées par des verbes et oublient les relations nominales. Reprenons notre exemple de la figure 3.3. Nous constatons qu'il est primordial d'extraire de telles informations pour obtenir le triplet *(Banque de France ; au sujet de ; surendettement)*. Un des premiers systèmes à s'y être intéressé est OLLIE (Mausam et al., 2012). Le système cherche à étendre la portée syntaxique des phrases de relation pour couvrir un nombre beaucoup plus important d'expressions de relation ainsi que la représentation de l'OpenIE pour permettre des informations contextuelles supplémentaires telles que l'attribution et les modificateurs clausaux. ReNoun (Yahya et al., 2014) et RelNoun (Pal & Mausam, 2016) ont, eux aussi, proposé une approche pour extraire les relations nominales. OPENIE4 (Mausam, 2016) étend SRL-IE et RelNoun afin de produire des relations n-aires à la fois pour des noms et des verbes. Le projet autour de OPENIE-X a continué d'être amélioré et en 2020 est sorti OPENIE6 (Kolluru et al., 2020a). Il s'agit d'un système itératif basé sur une architecture d'étiquetage itératif de la grille (*Iterative Grid Labeling - IGL*) à deux dimensions.

Systèmes de nouvelle génération. Selon l'état de l'art de Vo & Bagheri (2016), les informations sur les dépendances profondes (*Deep dependency information*), qui représentent la dépendance entre les mots d'une phrase, peuvent être utilisées directement pour trouver les triplets de relations. Ces systèmes n'ont été développés que plus tard car ils prenaient trop de temps. Bien que moins populaires, certains modèles utilisent les informations de dépendances en entrée du modèle. ElSahar et al. (2017) utilisent le chemin de dépendance pour donner plus de poids aux mots entre deux entités nommées, de sorte que le modèle n'utilise les informations de dépendance que comme un complément et s'appuie davantage sur la signification sémantique pour extraire des informations.

Les informations hiérarchiques qui existent implicitement dans les langues et qui peuvent être présentées explicitement dans les bases de connaissances sont utiles à l'apprentissage de la représentation des connaissances (Wang et al., 2014b; Lin et al., 2015; Hu et al., 2015; Xie et al., 2016). En outre, les bases de connaissances contiennent des connaissances factuelles fines qui fournissent des informations de base et des structures hiérarchiques nécessaires à l'extraction de relations. Par rapport au *clustering* traditionnel, la base de données peut fournir des informations hiérarchiques qui aident à représenter et à regrouper les relations de manière plus organisée (Zhang et al., 2021) et des connaissances factuelles hiérarchiques pour l'augmentation des données (Fangchao et al., 2021).

Comme expliqué précédemment, les informations sémantiques ont besoin d'être prises en compte. Des tentatives ont été faites pour utiliser les résultats de WordNet (Fellbaum, 1998; Harabagiu et al., 1999) afin d'inclure des informations sémantiques (Liu & Yang, 2012). Cependant, récemment, avec l'arrivée des modèles de langues, les modèles d'OpenIE utilisent généralement les représentations de mots et de phrases obtenues à partir de modèles de langage pré-entraînés (Kolluru et al., 2020b; Hwang & Lee, 2020; Xinwei & Hui, 2020). Ces représentations contiennent des informations syntaxiques et sémantiques (Jawahar et al., 2019). Parallèlement, certains modèles utilisent des plongements lexicaux provenant de modèles *embeddings* tels que GloVe, ELMo et Word2Vec pour capturer les informations sémantiques (Ni et al., 2021). Multi$\hat{2}$OIE est un modèle d'étiquetage de séquence qui combine BERT (Devlin et al., 2019b) avec une attention *multi-head*. Ils appliquent le BERT multilingue. MT4CrossIE (Li et al., 2023) est un système multilingue constitué de deux étapes qui utilise un modèle pré-entrainé et *finetuné* ainsi que des données pour transférer des connaissances dans le modèle pré-entraîné.

PIVOINE (Lu et al., 2023) est un LLM pour l'IE en monde ouvert. Il s'agit d'un BLOOM

pré-entrainé sur le jeu de données InstructOpenWiki construit pour l'IE en monde ouvert enrichi avec un corpus de compréhension, d'annotations étendues et d'instructions diverses. PIVOINE a surpassé, lors d'expérimentations, les méthodes traditionnelles du monde fermé ainsi que d'autres LLMs de base. PIVOINE semble donc généralisable puisque certaines instructions et certains cas étaient non connus.

Évaluation Selon la revue de Liu et al. (2022), les jeux de données (*datasets*) sont découpés en trois groupes.

— Le premier consiste à convertir depuis d'autres jeux de données crowdsourcés, notamment ceux pour le *Question Answering* (QA). OIE2016 (Stanovsky & Dagan, 2016) est l'un des jeux de données les plus populaire parmi les *benchmarks* de OpenIE, qui s'appuie sur les annotations QA-SRL (He et al., 2015). Par la suite, AWOIE (Stanovsky et al., 2018) a étendu OIE2016. Cependant, ces deux OIE supervisés ont des problèmes d'annotation et d'extraction. D'autres jeux de données ont donc réannoté une partie du jeu de données par la suite comme Re-OIE2016 (Zhan & Zhao, 2019) et CaRB (Bhardwaj et al., 2019) ou bien en convertissant QA-SRL 2.0 (FitzGerald et al., 2018) comme LSOIE (Solawetz & Larson, 2021).
— Le deuxième groupe est directement crowdsourcé en incluant plusieurs datasets comme WiRe57 (Léchelle et al., 2018), SAOKE (Sun et al., 2018) et BenchIE (Gashteovski et al., 2021).
— Le troisième groupe représente les jeux de données construits en alignant les triplets dans la base de connaissances (*Knowledge Base* - KB) avec le texte dans le corpus comme NYT-FB (Mintz et al., 2009; Yao et al., 2011), FewRel (Han et al., 2018) ainsi que T-Rex (ElSahar et al., 2018) et les jeux de données qui en découlent T-REx SPO et T-REx DS (Hu et al., 2020).

Limites et positionnement Il est vrai que les systèmes d'IE ouverts ne nécessitent pas de données d'apprentissage spécifiques aux relations ce qui les rend utile dans le cadre d'un jeu de données peu annotées. Cependant les revues mentionnées précédemment relèvent des limites résumées en trois points.

— Un premier point mentionné concerne les jeux de données car ceux-ci sont limités. En effet, une grosse quantité d'informations est disponible sur le web mais pourtant les jeux de données se limitent principalement à Wikipédia anglophone, Newswire, NYT (Riedel et al., 2010) et Freebase. Ces *datasets* doivent donc être plus larges et ne sont disponibles qu'en anglais. Les jeux de données de l'OpenIE devraient inclure davantage de langues qui permettent de comparer plusieurs modèles.
— Le deuxième point relevé concerne l'évaluation des systèmes OpenIE qui consiste à utiliser l'extraction de relations ainsi qu'une mesure de similarité. Or, l'OpenIE étant plus verbeux, beaucoup de relations ne sont pas correctement jugées. De plus, les jeux de données contiennent bien souvent une relation par phrase alors que plusieurs relations sont souvent présentes dans les phrases et sont détectées par les systèmes d'OpenIE. Cela peut introduire du bruit dans la relation extraite. Une des façons de corriger ce problème est d'utiliser le *clustering* afin de regrouper les segments textuels pour la relation et un type de relation donné. Cependant, nous retournons sur le problème principal de l'OpenIE où il faut établir au préalable un ensemble de types de relations.
— Enfin, un troisième point relevé concerne les données de sortie. En effet, il n'existe pas de forme standard commune pour les résultats de l'OpenIE. Il est donc difficile de

comparer les modèles entre eux et la canonisation des résultats. Niklaus et al. (2018) relèvent l'importance de rendre canoniques les phrases et les entités dans l'OpenIE.

L'utilisation de l'OpenIE semble pertinente. En effet, nous choisissons de créer notre propre système OpenIE, visant à compléter les méthodes d'extraction existantes qui sont souvent limitées à un ensemble fini de types de relations prédéfinis. L'objectif principal de notre approche est de détecter de nouveaux types de relations qui pourraient ne pas être couverts par les méthodes d'extraction traditionnelles. Toutefois, nous reconnaissons également l'importance de recentrer ces nouvelles découvertes vers ce qui est déjà connu dans le graphe d'informations existant. Cette double approche vise à éviter la démultiplication d'informations tout en enrichissant la compréhension globale des relations juridiques. Combiner à la fois un système d'extraction de relations classique et un système d'OpenIE mène à trois catégories d'avantages :

1. **Flexibilité et adaptabilité** : Contrairement aux systèmes d'IE traditionnels qui se limitent à des schémas ou des types prédéfinis de relations, notre approche se veut flexible et adaptable. En utilisant des techniques d'apprentissage automatique et de traitement du langage naturel, notre système peut identifier et extraire une gamme plus large de relations, même si elles ne sont pas explicitement spécifiées à l'avance.
2. **Identification de nouveaux types de relations** : En explorant de manière proactive les données juridiques, notre système est capable de détecter des relations potentiellement nouvelles, sous-représentées voire de spécialité. Cela permet une découverte continue de connaissances et une évolution constante du système pour mieux répondre aux besoins changeants des utilisateurs.
3. **Intégration avec les connaissances existantes** : Malgré son orientation vers l'exploration de nouvelles relations, notre système reste ancré dans les connaissances existantes. En recentrant les nouvelles découvertes vers les concepts et les relations déjà établis dans le graphe d'informations, nous évitons la duplication excessive d'informations et favorisons une vue holistique et cohérente du domaine juridique.

3.4 Classification de documents

La classification de documents est un domaine de recherche vivement exploré et de plus en plus appliqué dans divers secteurs métiers. En effet, ce domaine intéresse à la fois la communauté scientifique ainsi que les entreprises. La classification de documents est une tâche de classification qui consiste à attribuer un document à une classe parmi un ensemble défini au préalable. En effet, cette classification peut prendre différentes formes comme la classification par genre, par thème, *etc.* Il faut définir préalablement les classes que nous souhaitons associer aux différents textes ainsi que les textes en eux mêmes. Souhaitons-nous classifier les documents au complet ou une brique plus petite de ces documents ? Par exemple, nous pouvons choisir d'associer à des articles d'actualité des catégories comme « Sport », « Politique », « Diplomatie », *etc.* Il peut également s'agir d'associer une polarité à des avis de films également.

Nous nous intéressons, ici, à la classification de documents selon des domaines de spécialité. En effet, comme explicité dans le chapitre 2, les textes que nous exploitons sont divisés en huit domaines spécifiques : État civil et Cimetières, Élections, Commande publique, Urbanisme, Comptabilité et Finances locales, Ressources Humaines Territoriales, Justice et Santé. Nous nous intéressons à la classification de domaine par entonnoir. En effet, en classifiant un document ou un texte, les termes clefs rattachés à celui-ci sont également classifiés permettant ainsi de rattacher à la fois un document en lui-même mais aussi l'ensemble des

termes clés qui le composent. De plus, si deux termes clés d'une relation t_1, t_2 appartiennent à un domaine d, alors la relation r appartient également à d. Ainsi, par la simple classification de documents, par transitivité, les nœuds et relations sont classifiées. Classifier les documents parmi ces huit domaines nous conduit à différents avantages.

— Le premier avantage concerne directement le graphe de connaissances. En effet, précédemment, nous avons expliqué qu'au sein du graphe de connaissances, les nœuds pouvaient être associés à des domaines. Il est intéressant voire primordial de pouvoir filtrer les nœuds et relations du graphe en fonction de ces huit domaines spécifiques.
— La seconde concerne les relations entre les nœuds ou termes clés. En effet, lors de l'étape de l'extraction de relations à partir d'un texte, les relations formées entre les différents termes clés ne sont pas forcément pertinentes dans le contexte du domaine donné. Afin de filtrer ces relations non pertinentes, une méthode possible est de les classifier selon leur domaine d'application puisque la relation ne pourra exister que si les deux termes clés appartiennent au même domaine.

La classification de documents est un sujet largement abordé dans la littérature, avec plusieurs états de l'art récents (Das et al., 2018; Dhar et al., 2020; Dwivedi & Arya, 2016; Kaur & Kumar, 2015; Al-Qurishi, 2022; da Costa et al., 2023; Gasparetto et al., 2022; Wang et al., 2023a).

Il y a également eu plusieurs défis consacrés à ce domaine comme dans le cadre de la fouille de textes cliniques lors de l'édition 2021 (Grouin et al., 2021) et la classification de documents en genre et thème de l'édition 2008 du DÉfi Fouille de Textes (DEFT) (Plantié et al., 2008; Hurault-Plantet et al., 2008) pour le français et la conférence *Text REtrieval Conference* (TREC) lors de l'édition de 2004 (Voorhees, 2004) pour l'anglais. La première tâche de DEFT'21 consistait à identifier le profil clinique d'un patient en sélectionnant les différents types de maladies correspondant aux entrées génériques des chapitres du MeSH (*Medical Subject Headings*). Celle-ci s'inscrivait dans la continuité de deux éditions précédentes (Cardon et al., 2020; Grabar et al., 2019). Les meilleurs résultats obtenus lors de cette édition reposent sur des transducteurs à états finis et des listes de mots-clés issus du MeSH et de MedDRA, une méthode qui n'utilise pas l'apprentissage (Hiot et al., 2021). D'autres participants ont utilisé des approches de classification multi-labels supervisées, complétées par des plongements lexicaux tels que CamemBERT (Gérardin et al., 2021; Bailly et al., 2021). Une autre approche consiste à utiliser le MeSH, complété de termes du corpus, pour indexer le contenu des documents (Dupont et al., 2021).

DEFT2008 consistait à identifier un article s'il appartenait à l'une des neuf classes données pour chacune des sources ce qui amenait à s'intéresser à la classification de documents. Les résultats obtenus ont révélé que les systèmes d'apprentissage supervisé, en particulier ceux basés sur les SVMs, ont montré une grande efficacité (Béchet et al., 2008). Il a également été observé que les meilleures techniques utilisées pour la classification de documents en français lors de la campagne de 2008 semblaient être pertinentes pour l'anglais, étant donné que des techniques similaires et des résultats comparables ont été présentés en 2019 par Dogan & Uysal (2019).

TREC 2004 comportait sept domaines d'intérêt appelés « *tracks* ». La *track* génomique comportait plusieurs tâches de catégorisation lors de TREC 2004, et les tâches de la *novelty track* nécessitaient l'affectation de phrases contenues dans des documents à des catégories « pertinentes » et « nouvelles ». La *track* du web comportait également une variante d'une tâche de catégorisation, bien que dans ce cas, les sujets, et non les documents, devaient être catégorisés. La *track* génomique (*Genomics Track*) de TREC 2004 comportait deux tâches (Hersh et al., 2006). La première tâche était une tâche standard d'extraction ad hoc utilisant des sujets obtenus auprès de chercheurs biomédicaux réels et des documents provenant

d'un large sous-ensemble de la base de données bibliographiques MEDLINE. La deuxième tâche était axée sur la catégorisation de documents en texte intégral, simulant la tâche des conservateurs du système *Mouse Genome Informatics* (MGI) et consistant en trois sous-tâches.

Cette deuxième tâche est décrite dans un article complémentaire (Cohen & Hersh, 2006). Le meilleur système utilise une combinaison de pondération Okapi (BM25 pour la fréquence des termes, mais avec la IDF standard), le *stemming Porter*, l'expansion des symboles par LocusLink et les enregistrements MeSH, le *blind relevance feedback* (aussi appelé *blind query expansion*), et l'utilisation des trois champs du sujet (titre, besoin, et contexte). Par la suite, le même groupe a amélioré ce système en ajoutant une technique de modélisation linguistique du Dirichlet-Prior smoothing (Zhou et al., 2006).

Dans le même esprit, l'université de Waterloo (Büttcher et al., 2004) obtient la deuxième meilleure performance en utilisant diverses approches comme la pondération Okapi, le *blind relevance feedback* et diverses formes d'expansion des requêtes spécifiques à un domaine. D'autres systèmes ont performé par l'utilisation du renforcement des poids des mots dans les documents (Li et al., 2004; Pirkola, 2004; Carpenter, 2004). Le problème avec ces résultats vient du fait qu'ils englobent l'ensemble des deux tâches et non spécifiquement la classification de documents. Un article ressort de cette conférence car il se concentre sur l'utilisation d'arbres et règles de décision pour la catégorisation de textes dans la *track* génomique (Guillén, 2004). Il est intéressant de noter que les résultats de TREC montrent que dans le cadre d'un domaine spécifique, il est important de prendre en considération le domaine de spécialité notamment en utilisant des dictionnaires spécifiques plutôt que des systèmes génériques dits « prêts à l'emploi ». Cela s'est ressenti dans les résultats obtenus par des systèmes génériques (Cohen et al., 2004; Seki et al., 2004; Nakov et al., 2004; Aronson et al., 2004).

Il existe différentes méthodes de classification de domaines, notamment la classification automatique, la classification manuelle par des experts humains et la classification hybride combinant les deux approches. L'efficacité de ces méthodes dépend de plusieurs facteurs, tels que la qualité et la quantité des données d'entraînement, la pertinence des catégories de classification et la précision des algorithmes utilisés. La classification multi-classe de documents de spécialité peut être abordée selon plusieurs approches : supervisées, non supervisées, ou semi-supervisées (ou dites hybrides). Parmi les méthodes supervisées, nous pouvons retrouver des méthodes dites classiques qui ont été utilisées récemment (Chen et al., 2016; Kalra et al., 2022; Palacharla & Vatsavayi, 2022; Li et al., 2021; Abalorio et al., 2022). En effet, certaines utilisent l'extraction de sujets pour la classification supervisée dans le cadre de textes de spécialité. D'autres utilisent une classification supervisée basée sur des poids (Tang et al., 2021).

Labani et al. (2018) ont proposé le modèle Critère de discrimination relative multivariée (*Multivariate Relative Discrimination Criterion* - MRDC) pour obtenir des caractéristiques de classification de textes, en réduisant les caractéristiques redondantes. L'évaluation de ce modèle a été réalisée sur trois corpus anglais, avec une meilleure précision obtenue en utilisant le classificateur Multinomial Naïve Bayes (MNB). Jiang et al. (2018) ont, quant à eux, introduit une approche hybride basée sur un réseau *Deep Belief Network* (DBN) et une régression softmax. La méthode RMDL (*Random Multimodel Deep Learning*) proposée par Kowsari et al. (2018) utilise des ensembles d'architectures d'apprentissage profond pour améliorer les performances de classification et déterminer les frontières entre classes.

Guo et al. (2021) exprime l'importance de la représentation textuelle pour déterminer le label dans le cadre de la classification de texte. C'est pourquoi, dans ce manuscrit, nous allons comparer différentes représentations textuelles qui ont fait leurs preuves ces dernières années puis les comparer entre elles dans le cadre de textes de spécialité (ceux de Berger-

Levrault).

Dans cet état de l'art, nous avons repéré un jeu de données pertinent pour notre cas d'usage. Il s'agit de EURLEX (57K) (Chalkidis et al., 2019) pour la classification de textes multi-labels à grande échelle. Ce jeu de données contient des documents (celui disponible sur HuggingFace[10] contient 57 000 documents) législatifs en anglais provenant d'EUR-Lex. Ce jeu de données se rapproche de notre cas d'usage puisqu'il s'agit de textes dans le domaine juridique.

Les modèles actuels, entraînés sur des corpus de données provenant du domaine général, peuvent rencontrer des difficultés lorsqu'ils sont appliqués à des métiers spécifiques caractérisés par un vocabulaire spécialisé. Ainsi, la recherche s'oriente vers le développement de méthodes de classification adaptées à ces contextes particuliers, visant à surmonter les limitations des modèles génériques et à améliorer la performance dans des domaines d'application spécifiques. Cependant, l'état de l'art pour la classification de textes étant fourni, nous nous intéressons à étudier comparativement et réemployer les méthodes de cet état de l'art.

3.5 Contributions : Résumé des choix d'orientation et d'implémentation

Pour rappel, la figure 3.2 représente les différentes briques d'acheminement du corpus jusqu'au graphe lexico-sémantique classifié parmi les huit domaines des textes de Berger-Levrault. Nous allons détailler chacun des blocs du schéma avec les positions prises par rapport à l'état de l'art effectué.

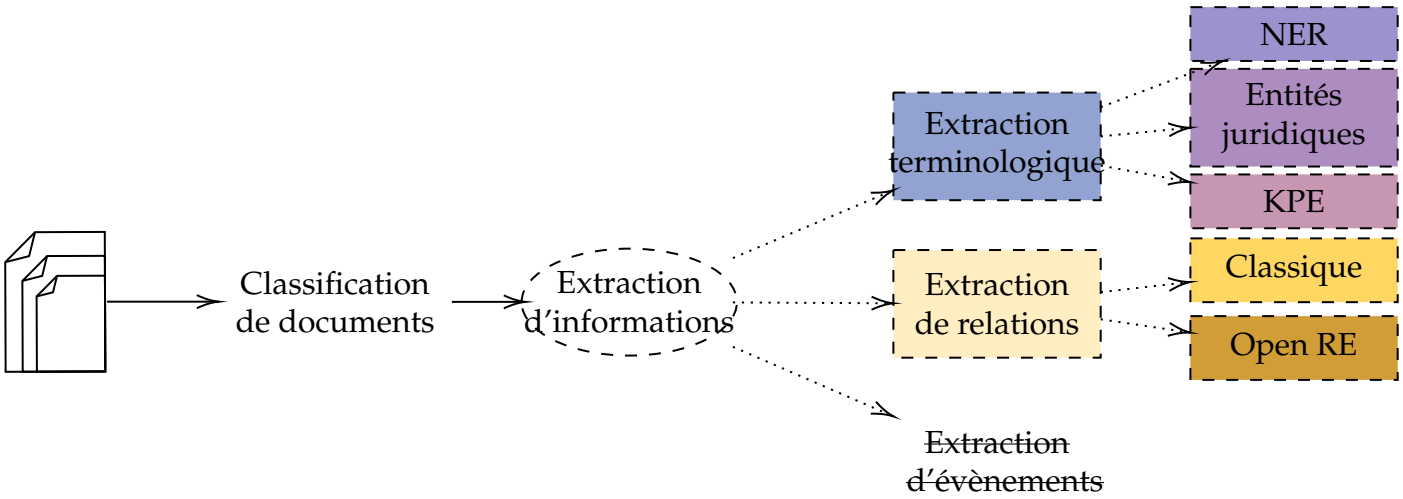

FIGURE 3.4 – Schéma des différents blocs d'orientations suite à l'état de l'art.

La figure 3.4 étaye chacun des blocs de l'état de l'art sur lesquels nous nous attardons. Les fonctionnalités, séparées par des flèches pleines, représentent les étapes. Les flèches en tirets mènent aux sous catégories qui décomposent les différentes fonctionnalités.

1. **Classification de documents :** Afin de pouvoir filtrer le graphe de connaissances en fonction de chacun des huit domaines, nous classifions les documents en fonction des

10. EURLEX dataset disponible sur HuggingFace au lien suivant : `https://huggingface.co/datasets/eurlex`

domaines de spécialité puisque la classification permet d'associer les nœuds et relations du graphe extraits de ces mêmes textes à des domaines. Dans le cadre de notre étude, nous avons pris en considération les nombreux systèmes de classification de documents présents dans l'état de l'art. Cependant, compte tenu de la nature spécifique de nos textes, qui sont principalement d'ordre juridique, il est essentiel de souligner que les systèmes existants ont déjà démontré leur efficacité dans des contextes similaires. Ainsi, plutôt que de nous concentrer uniquement sur la comparaison avec ces systèmes établis, notre approche vise à évaluer quelle méthode de classification fonctionne le mieux spécifiquement pour nos documents. Cette démarche nous permettra d'analyser plus en profondeur l'impact de la classification sur l'étape suivante de notre processus, à savoir l'extraction d'informations. En nous focalisant sur cette question, nous visons à déterminer si l'utilisation de la classification de documents améliore de manière significative la qualité et l'efficacité de l'extraction d'informations dans le contexte juridique.

2. **Extraction d'informations**

   2.1 **Extraction terminologique :** Nous nous orientons donc vers une combinaison d'un système d'extraction de termes pertinents avec un système NER pour une analyse textuelle plus complète et précise. En effet, alors que l'extraction de termes pertinents identifie les mots-clés et concepts importants, le NER va plus loin en repérant spécifiquement des entités telles que les noms de personnes, lieux, dates, etc., offrant ainsi une meilleure structuration du texte. Cette combinaison permet une compréhension plus fine du contenu, en identifiant des informations cruciales qui pourraient être manquées autrement. En combinant ces deux approches, nous maximisons la capacité à extraire et contextualiser des informations pertinentes, conduisant à une analyse plus approfondie.

      2.1.1 **NER :** Nous optons pour l'utilisation d'un modèle de langue basé sur BERT pour notre système de NER en raison de son caractère multilingue et de la diversité de ses modèles dérivés. BERT et ses variantes, ayant été pré-entraînés sur de vastes corpus de données multilingues, offrent une gamme étendue de modèles disponibles gratuitement, ce qui est particulièrement pertinent pour notre travail sur un corpus juridique en français.

      2.1.2 **Entités juridiques :** Bien que CamemBERT soit un choix judicieux pour la NER en français, sa capacité à identifier les entités juridiques spécifiques au domaine peut être limitée. Pour pallier ces lacunes, nous proposons de combiner les capacités de compréhension générale de CamemBERT avec un système d'entités juridiques spécialisé, basé sur des REGEX, pour garantir une reconnaissance précise et contextuelle des entités spécifiques au domaine juridique, améliorant ainsi la qualité et la pertinence de nos analyses.

      2.1.3 **KPE :** L'utilisation d'un LLM tel que ChatGPT pour l'extraction de termes clés dans le domaine juridique est justifiée par sa capacité à comprendre le langage naturel de manière contextuelle et à s'adapter aux spécificités des textes juridiques. Contrairement aux systèmes état de l'art qui peuvent manquer de flexibilité pour s'adapter efficacement aux domaines spécialisés comme le juridique, les LLMs offrent une approche plus générique et adaptable. Cependant, l'*instruction tuning* est essentielle pour garantir des performances optimales dans ce contexte. En prenant en compte les recommandations des revues d'IE, nous pouvons exploiter pleinement le potentiel des LLMs pour l'extraction de termes pertinents dans le domaine juridique, en veillant à réaliser une *instruction tuning* appropriée pour des résultats précis et fiables. Notons cependant

que, faute de systèmes état de l'art disponibles lors de l'implémentation, nous avons développé notre propre modèle.

2.2 **Extraction de relations :** Nous souhaitons utiliser un système d'extraction de relations combiné à un système d'OpenRE car les deux méthodes se complètent. En effet, les méthodes d'extraction existantes sont souvent limitées à des types de relations prédéfinis. Le but, ici, vise donc à détecter de nouveaux types de relations non couverts par les méthodes traditionnelles, tout en recentrant ces découvertes vers ce qui est déjà connu dans le graphe d'informations existant. Cette double approche évite la démultiplication d'informations tout en enrichissant la compréhension globale des relations juridiques. La combinaison d'un système d'extraction de relations classique avec l'OpenRE présente des avantages significatifs. Tout d'abord, notre approche se distingue par sa flexibilité et son adaptabilité. Ce système peut non seulement identifier, mais également extraire une gamme étendue de relations, dépassant ainsi les limitations imposées par les schémas prédéfinis. Deuxièmement, cette approche proactive permet la détection de nouveaux types de relations, parfois spécialisées, offrant ainsi une voie vers une découverte continue de connaissances et une amélioration constante du système. Enfin, bien que notre système soit tourné vers l'exploration de nouvelles relations, il demeure solidement ancré dans les connaissances existantes. Cette intégration évite la duplication inutile d'informations et favorise une vision globale et cohérente du domaine juridique, enrichissant ainsi la compréhension des relations établies dans le graphe d'informations.

2.2.1 **Systèmes classiques :** La supervision à distance est une approche qui spécifie les concepts que les mots individuels tentent de transmettre, en se basant sur des relations prédéfinies. Bien que cette méthode ait un potentiel considérable, les bases de connaissances nécessaires ne sont pas encore pleinement développées. Les réseaux neuronaux, en revanche, offrent une voie prometteuse, en utilisant des caractéristiques automatiquement apprises. En intégrant des mécanismes d'attention, ils peuvent traiter les défis des multi-instances et des multi-relations. Il est crucial de reconnaître l'interaction entre l'extraction d'entités et l'extraction de relations, et l'extraction conjointe peut améliorer les performances. Notre approche se concentre sur la création de représentations vectorielles de relations, en utilisant des plongements de mots pour inférer des vecteurs de relations, exploités ensuite dans des méthodes d'apprentissage automatique pour la classification.

2.2.2 ***Open Relation Extraction* (OpenRE) appartenant à l'OpenIE :** Nous souhaitons donc nous orienter sur un LLM pour l'OpenRE. Tout d'abord, les LLMs permettent de s'adapter à des jeux de données plus vastes et multilingues. Contrairement aux approches traditionnelles qui se limitent souvent à des corpus restreints comme Wikipédia anglophone, les LLMs ont exploité une variété de données linguistiques pour une meilleure généralisation. De plus, les LLMs ont la capacité d'apprendre des modèles de relations plus complexes et variés, ce qui peut aider à traiter les ambiguïtés et les bruits présents dans les données d'OpenIE. En utilisant des techniques d'apprentissage automatique telles que le *clustering*, les LLMs peuvent regrouper les segments textuels pour identifier les relations et types de relations, contribuant ainsi à une extraction plus précise et cohérente des informations. Enfin, les LLMs peuvent faciliter la canonisation des résultats de l'OpenIE en produisant des sorties plus standardisées et en fournissant des mécanismes pour la normalisation des entités

et des relations extraites, ce qui facilite la comparaison entre les modèles et la compréhension des résultats. En intégrant ces capacités dans un système similaire à PIVOINE (Lu et al., 2023), les LLMs peuvent jouer un rôle crucial dans l'amélioration de l'efficacité et de la précision de l'OpenIE, tout en permettant une meilleure généralisation et une adaptation à des environnements multilingues et variés.

CHAPITRE 4

Extraction terminologique

Après avoir établi une base théorique sur les graphes de connaissances lexico-sémantiques, ce chapitre se concentre sur l'extraction terminologique. L'objectif est d'identifier et d'extraire des termes clés à partir de textes pour alimenter notre graphe. L'extracteur terminologique, baptisé *InfoGlean KeyTerms* prend la forme de trois sous-systèmes incluant un système pour la reconnaissance des entités nommées (NER), un pour l'extraction des termes pertinents et un pour les entités juridiques. Les défis comprennent la gestion de la variabilité terminologique et l'adaptation des modèles à différents domaines. Afin de sélectionner les différents systèmes, une évaluation des différents sous-systèmes est effectuée. Enfin, nous parlerons des différents critères liés aux termes clés et méthodes introduits pour enrichir les termes extraits.

Sommaire du chapitre

4.1 Processus et Méthodes d'annotations des Experts **76**
4.2 Système InfoGlean KeyTerms pour l'Extraction Automatique de Termes . . **82**
4.3 Méthodologie et Résultats de l'évaluation des sous-systèmes d'extraction terminologique **85**
4.3.1 Structuration et Critères de l'évaluation des sous-systèmes 85
4.3.2 Présentation et Analyse des Résultats Obtenus par les Sous-Systèmes d'Extraction 94
4.3.3 Discussion critique sur les Résultats et Identification des Limites . 112
4.4 Synthèse des Contributions et Perspectives pour les Travaux Futurs **115**
4.4.1 Bilan des Contributions Principales en Extraction Terminologique . 115
4.4.2 Perspectives d'Amélioration et Suggestions pour les Recherches Futures 115

Definition 4.1 (Extraction terminologique)**.** *L'**extraction terminologique** est une tâche appartenant au domaine de l'extraction d'information. Elle consiste à extraire automatiquement les termes pertinents d'un corpus de textes donné. Cela peut prendre différentes formes comme les thèmes pertinents, segments textuels principaux voire les mots principaux (ou leur racine). Deux phases se dessinent naturellement : dans un premier temps, une liste de candidats terminologiques est créée.*

Il s'agit de déterminer les syntagmes terminologiques syntaxiquement plausibles ; dans un second temps, des modèles basés sur des méthodes statistiques et/ou symboliques permettent de filtrer ces candidats.

Comme explicité précédemment (c.f. chapitre 1), nous travaillons sur les données de Berger-Levrault à savoir des ouvrages juridiques et pratiques de secteurs métiers. Nous pouvons comparer l'ensemble du corpus à un garde-manger pour un chef cuisinier. Tout comme un chef cuisinier veillera à sélectionner les bons ingrédients pour la réalisation d'une recette d'un pays précis, l'extraction terminologique consiste à extraire les bons termes des textes pour une tâche particulière dans le cadre d'un domaine précis. Par exemple, un chef préparant un plat français classique comme un coq au vin va choisir des ingrédients spécifiques tels que du vin rouge, des champignons, et des herbes de Provence. De même, dans le domaine juridique, pour une tâche de rédaction de contrats, des termes spécifiques comme « clause de non-responsabilité », « contrat à durée déterminée », et « force majeure » seront extraits comme des ingrédients essentiels pour la tâche. En règle générale, les approches d'extraction terminologique automatique extraient des candidats terminologiques, à savoir des syntagmes terminologiques syntaxiquement plausibles, par découpage de phrases, par exemple. Les candidats sont ensuite filtrés à l'aide de méthodes statistiques (d'apprentissage automatique) et/ou symboliques. L'extraction terminologique consiste à extraire une liste de termes pertinents associés à un domaine précis. Il s'agit ici de se débarrasser du superflu en cherchant des termes un peu moins courants mais plus pertinents et ensuite de compléter cette extraction par des synonymes à l'aide d'autres sources. L'exemple 4.1 montre cette évolution : nous partons de l'ensemble des groupes nominaux par exemple et petit à petit nous filtrons en fonction du but recherché. En effet, quel est le niveau de filtre souhaité ? souhaitons nous des termes pertinents dans le texte donné ou souhaitons nous que les termes soient parfaitement adéquates au domaine cible ? En choisissant les termes clés les plus pertinents et significatifs, l'extraction terminologique permet de se concentrer sur l'essentiel de l'information. Une notion de structure est naturellement créée par l'identification de termes conceptuels et les autres termes associés au concept.

Selon certains auteurs, la dimension économique et concurrentielle du droit de l'urbanisme autant que la décentralisation et le transfert des compétences qui s'y sont développées devraient plaider pour un usage relativement important du contrat dans cette matière (Bonnefont R., « Urbanisme et contrat : les liaisons dangereuses », art. cit.). Et ce d'autant plus qu'en cette matière comme ailleurs, les personnes publiques disposent de la liberté contractuelle, constitutionnellement garantie, essentiellement sur le fondement de l'article 4 de la Déclaration des droits de l'Homme et du citoyen (Cons. const., 30 nov. 2006, no 2006-543 DC : JO, 8 déc. 2006) et reconnue par le juge administratif en tant que principe général du droit (CE, sect., 28 janv. 1998, no 138650, Société Borg-Warner : Lebon, p. 20).

ENSEMBLE DES CANDIDATS:
Certains auteurs, La dimension économique, La dimension concurrentielle, Le droit de l'urbanisme, La décentralisation, Le transfert des compétences, Un usage relativement important du contrat, Cette matière, Les personnes publiques, La liberté contractuelle, L'article 4 de la Déclaration des droits de l'Homme et du citoyen, Le juge administratif, Un principe général du droit, La Société Borg-Warner,

FILTRE NIVEAU 1:
Dimension économique, Dimension concurrentielle, Droit de l'urbanisme, Décentralisation, Transfert des compétences, Usage relativement important, Matière, Personnes publiques, Liberté contractuelle, Fondement de l'article 4, Déclaration des droits de l'Homme et du citoyen, Juge administratif, Principe général du droit, Société Borg-Warner,

FILTRE NIVEAU 2:
Droit de l'urbanisme, Décentralisation, Transfert des compétences, Liberté contractuelle, Déclaration des droits de l'Homme et du citoyen, Juge administratif, Principe général du droit,

Exemple 4.1 – Exemple d'un paragraphe avec les candidats puis les filtrages associés

Légende :

⟶ **Ensemble des candidats** : les candidats sans filtre extraits par le système (étape 1) : l'ensemble des groupes nominaux

⟶ **Filtre niveau 1** : Filtrage des candidats jugés pertinents pour donner un sens au texte

⟶ **Filtre niveau 2** : Filtrage strict des candidats : termes pertinents qui sont directement reliés au domaine cible.

4.1 Processus et Méthodes d'annotations des Experts

Les ouvrages et articles du corpus de Berger-Levrault sont partiellement annotés. Un expert d'un des 8 sous-domaines traite un document ou un article et parcourt chacun des paragraphes un à un en choisissant de l'annoter ou non par un ou plusieurs termes. En pratique, l'expert lit un paragraphe et va donner à la volée des termes qui caractérisent ce paragraphe. Il constitue une liste de termes à ce paragraphe. Il est libre de donner les termes qu'il souhaite. Il n'a pas en sa possession de guide d'annotations et peut associer une thématique ou des termes qui lui semblent pertinents et qui ne figurent pas dans le texte. Nous disposons d'annotations réalisées par les experts sur le paragraphe illustré dans l'exemple 4.2. Les annotations des experts ont été reportées par la mise en *italique* dans le texte associé par un humain de façon à faire un traitement intuitif pour faire correspondre et non pas ce qui est obtenu avec le système.

Les communes dotées d'un plan d'occupation des sols rendu public ou d'un plan local d'urbanisme approuvé peuvent, par délibération, instituer un *droit de préemption urbain* sur tout ou partie des *zones* urbaines et des *zones* d'*urbanisation future* délimitées par ce plan, dans les périmètres de protection rapprochée de prélèvement d'eau destinée à l'alimentation des collectivités humaines définis en application de l'article L. 1321-2 du Code de la santé publique, *zones* et secteurs définis par un plan de prévention des *risques technologiques* en application de l'article L. 515-16 du Code de l'environnement, dans les *zones* soumises aux servitudes prévues au II de l'article L. 211-12 du même code, ainsi que sur tout ou partie de leur territoire couvert par un plan de sauvegarde et de mise en valeur rendu public ou approuvé en application de l'article L. 313-1 lorsqu'il n'a pas été créé de *zone* d'aménagement différé ou de périmètre provisoire de *zone* d'aménagement différé sur ces territoires.

ANNOTATIONS DES EXPERTS:
Droit de préemption urbain, Préemption, Risque naturel et/ou technologique, Secteur sauvegardé, zone, zone d'aménagement concerté, d'urbanisation future

Exemple 4.2 – Exemple d'un paragraphe avec les annotations qui lui sont associées

Légende :

⟶ *Annotations des experts manuellement transformées pour faire correspondre avec le texte*

Les annotations des experts sont manuelles et les experts ne détiennent pas une ressource lexicale de référence au moment de l'annotation. Les experts n'annotent donc pas de la même façon. Les termes-clés ont été décrits avec plusieurs formes fléchies et parfois avec des informations additionnelles non pertinentes comme les déterminants (ex : « des frais »). Un terme-clé regroupe toutes les annotations par formes fléchies sous une seule forme représentative.

Nous avons récupéré l'ensemble des annotations des experts. À partir des différentes formes fléchies, nous souhaitons passer à une forme représentative appelée « terme-clé ». Pour cela, nous allons prétraiter les données afin de les unifier.

La première étape consiste à éliminer les premiers mots outils (déterminant, par exemple). Dans l'exemple réel « de participation de zones d'aménagement concerté », l'annotation est transformée en « participation de zones d'aménagement concerté ».

Ensuite, à partir de ces éléments, nous souhaitons unifier les annotations d'un même identifiant sous un même terme-clé représentatif dit référent : une forme unique pour les termes de mêmes identifiants mais de forme fléchie différente. Par exemple, « de restauration immobilière », « restauration immobilière » et « Restauration immobilière » font tous les trois référence au même terme clé. Afin de construire un identifiant pour un terme donné déjà passé par la première étape, nous lemmatisons l'ensemble des mots qui constituent ce terme. Cet identifiant sert uniquement à retrouver toutes les différentes formes fléchies.Afin de choisir le bon représentant, nous avons découpé notre problème en deux cas de figure : termes simples (mots) et complexes. Dans le cadre des termes complexes, nous avons fait le choix de prendre la forme fléchie ayant le plus grand nombre d'occurrences dans le corpus. Pour cela, une analyse statistique est effectuée sur tout le corpus afin de calculer le nombre d'occurrences de chaque annotation sur l'ensemble du corpus. Ainsi, la forme la plus fréquente représente le substitut pouvant faire référence à un terme-clé donné. Pour les termes simples, nous privilégions une forme canonique dite standard car nous souhaitons privilégier la forme singulière et générique à la forme plurielle.

Cependant, comme exprimé précédemment, le couplage (*matching*) s'est fait manuellement car il est très compliqué de le faire automatiquement puisque de gros prétraitements souvent multiples sont à faire. Par exemple, dans notre cas « Risque naturel et/ou technologique » a été couplé avec « risque technologique ». Or, bon nombre d'autres cas existent rendant difficile le pré-traitement de ces annotations. Les annotations des experts ne sont donc pas présentes dans le texte associé. Cela peut être dû au fait que le terme nécessite un prétraitement comme dans le cas précédent mais cela peut aussi être dû à un terme clé trop long qui le rend inexploitable ou il peut s'agir d'un thème associé au paragraphe. De plus, comme nous le constatons sur l'exemple 4.2, ce paragraphe manque cruellement d'informations puisque l'expert annote le paragraphe en vue de donner un bref résumé de l'entièreté du paragraphe et non en vue de donner les termes pertinents du texte. Naturellement, quand nous confrontons un paragraphe avec les termes qu'un humain pourrait sélectionner comme intéressants dans le texte pour en faire une idée générale du texte face aux annotations, nous remarquons une large différence. Cette hypothèse est vérifiable sur l'exemple 4.3, qui suit. En effet, nous avons manuellement mis en évidence en soulignant les termes du texte qu'il nous semblait judicieux d'extraire afin de pouvoir extraire de l'information du texte. Cette information est subjective mais elle représente l'intuition du problème évoquée.

Les communes dotées d'un plan d'occupation des sols rendu public ou d'un plan local d'urbanisme approuvé peuvent, par délibération, instituer un *droit de préemption urbain* sur tout ou partie des *zones* urbaines et des *zones* d'*urbanisation future* délimitées par ce plan, dans les périmètres de protection rapprochée de prélèvement d'eau destinée à l'alimentation des collectivités humaines définis en application de l'article L. 1321-2 du Code de la santé publique, *zones* et secteurs définis par un plan de prévention des *risques technologiques* en application de l'article L. 515-16 du Code de l'environnement, dans les *zones* soumises aux servitudes prévues au II de l'article L. 211-12 du même code, ainsi que sur tout ou partie de leur territoire couvert par un plan de sauvegarde et de mise en valeur rendu public ou approuvé en application de l'article L. 313-1 lorsqu'il n'a pas été créé de *zone* d'aménagement différé ou de périmètre provisoire de *zone* d'aménagement différé sur ces territoires.

Exemple 4.3 – Exemple d'un paragraphe avec des termes extraits du texte confrontés aux annotations des experts

Légende :

$\longrightarrow$ *Annotations des experts manuellement transformés pour faire correspondre avec le texte*

$\longrightarrow$ Termes annoté manuellement que nous souhaiterions extraire

Pour conforter cette idée, nous avons réalisé une série de calculs concernant les annotations des experts. Ces calculs sont présentés sur les différents tableaux qui suivent. Le tableau 4.1 présente le nombre d'occurrences des différentes annotations qu'un ouvrage peut avoir. Nous avons extrait un petit échantillon de 7 ouvrages représentatifs des différents cas. En effet, certains textes ne présentent aucune occurrence des annotations. Quant à d'autres ouvrages, beaucoup d'annotations peuvent être présentes. Nous constatons donc une certaine hétérogénéité dans les corpus des annotations.

Identifiant ouvrage	Nombre d'annotations (avec occurrences)
295452	0
295453	0
295454	594
295455	9 150
295456	1 763
...	...
350349	61
350431	389

TABLE 4.1 – Échantillon de calcul d'occurrences pour différents ouvrages. Sélection de 7 ouvrages représentatifs parmi 164 au total. Mise en évidence d'ouvrages n'ayant aucune occurrence présente.

Pareillement, le tableau 4.2 présente le nombre d'occurrences d'annotations pour un article donné et son domaine associé. Nous avons extrait 3 articles pour chacun des domaines et nous avons cherché à créer un échantillon représentatif du nombre d'occurrences des annotations. Ici encore, nous constatons une certaine hétérogénéité des annotations dans les articles. En effet, certains articles, et ce, pour chacun des domaines, ne présentent aucune occurrence des annotations et d'autres ont beaucoup d'annotations.

N° article	Nombre d'annotations (avec occurrences)	Domaine
295488_urbanisme_10	0	urbanisme
295488_urbanisme_1000	6	urbanisme
295488_urbanisme_1001	1	urbanisme
317197_ressources-humaines-territoriales_34	0	ressources-humaines-territoriales
317197_ressources-humaines-territoriales_340	4	ressources-humaines-territoriales
317197_ressources-humaines-territoriales_342	33	ressources-humaines-territoriales
295496_justice_541	24	justice
295496_justice_542	31	justice
295496_justice_543	0	justice
295542_État-civil-et-cimetières_2437	5	État-civil-et-cimetières
295542_État-civil-et-cimetières_2439	8	État-civil-et-cimetières
295542_État-civil-et-cimetières_244	0	État-civil-et-cimetières
295577_Santé_312	0	Santé
295577_Santé_314	26	Santé
295577_Santé_32	4	Santé
295530_élections_1.html	4 329	élections
295530_élections_113.html	26	élections
295530_élections_114.html	0	élections
295583_Commande-publique_119	0	Commande-publique
295583_Commande-publique_1191	53	Commande-publique
295583_Commande-publique_1194	12	Commande-publique
295585_Comptabilité-et-finances-locales_1173	0	Comptabilité-et-finances-locales
295585_Comptabilité-et-finances-locales_1177	21	Comptabilité-et-finances-locales
295585_Comptabilité-et-finances-locales_1179	3	Comptabilité-et-finances-locales

TABLE 4.2 – Échantillon de calcul d'occurrences pour différents articles et différents domaines. Sélection de 21 articles représentatifs parmi 13085 au total.

Le tableau 4.3 présente les différents calculs réalisés en globalité sur le corpus. Le tableau est séparée en deux blocs : d'une part les calculs au total et de l'autre part le nombre moyen des annotations dans des conditions spécifiques.

Différents types de calcul	**Nombre d'occurrences**
Annotations des experts au total	1 043 707
Annotations des experts au total dans les articles	134 380
Termes annotés au total	49 702
Termes annotés présents dans le corpus	34 640 (70%)
Liste unique de termes clefs à partir des annotations	41 875 (84%)
Annotations des experts par ouvrage	1 599
Annotations des experts par article	10
Annotations des experts par paragraphe	1.5
Nombre de paragraphes associés aux annotations par article	4.43
Occurrences des annotations par paragraphe	4.29

TABLE 4.3 – Calculs totaux et moyens des occurrences, nombre d'annotations et termes annotés

Le tableau 4.4 exprime différents calculs sur les annotations en fonction des domaines. Le tableau est séparé en deux parties. Tout d'abord, le tableau présente les calculs réalisés sur les articles uniquement. En effet, les articles sont associés à un domaine précis. Nous sommes donc certains du domaine associé. Par contre, ce n'est pas le cas pour les ouvrages (c.f. chapitre 5). Cependant, un système de classification estime le domaine des ouvrages. C'est pour cela que nous réalisons une estimation de calcul sur les annotations du corpus global, en second temps.

Le tableau 4.5 présente un échantillon du nombre d'annotations qu'une brique peut contenir. Une brique représente un paragraphe et est composée d'un identifiant la définissant. Nous constatons qu'un paragraphe peut être annoté ou non (dans ce cas, il a 0 annotation).

ID de la brique (paragraphe)	**Nombre d'annotations associées au paragraphe**
brique_694	2
brique_488	0
brique_1683	8
brique_1684	4
brique_1685	3
brique_1686	5
brique_1687	6
brique_485	0
brique_1688	1
brique_1689	2

TABLE 4.5 – Nombre d'annotations pour un paragraphe donné (associé à son id de brique)

Le tableau 4.6 présente des exemples d'annotations issues de notre corpus associé à son nombre d'occurrences et le nombre de paragraphes dans lesquels elles apparaissent. Nous constatons donc que certaines annotations apparaissent plusieurs fois au sein d'un même

Domaine	Moyenne par article (domaine sûr)	Somme articles	Moyenne par texte : Articles + ouvrage (estimation)	Somme texte (articles + ouvrage : estimation)	Nbre termes annotés (articles + ouvrages : estimation)
Commande publique ou Marchés publics	7,7	9 392	3,45	18 307	5 334
Comptabilité et finances locales	11	19 431	4,22	93 506	22 165
Santé	5	4 849	7,6	24 192	3 186
Justice	14,6	37 000	4,9	113 735	23 467
Ressources humaines territoriales	9,8	3 546	5	25 630	5 149
Urbanisme	16,1	21 733	6,1	64 285	10 630
Élections	31,2	6 145	4,46	17 909	4 220
État civil et cimetières	2,2	6 255	4,1	13 036	3 197
Total	12,2	108 351	5	370 600	77 348

TABLE 4.4 – Annotations des experts par domaine

paragraphe (c'est à dire au sein du texte) puisqu'il y a plus d'occurrences que de paragraphes. Le cas inverse est vrai aussi. Parfois l'annotation n'existe pas au sein du texte et dans ce cas, l'annotation sera comptabilisé par une unique occurrence. Cela signifie que l'occurrence ne représente que ce que l'expert a annoté. À partir des différents calculs réali-

Annotations	Nombre d'occurrences de l'annotation	Nombre de paragraphes
Filiation	378	66
Inceste	52	11
effet de la loi	4	1
établissement	179	36
accouchement sous X	1	1
mater semper certissima est	2	1

TABLE 4.6 – Nombre d'occurrences et nombre de paragraphes des annotations

sés sur les différents tableaux, nous constatons donc que nous avons un manque d'annotation pour l'ensemble des articles et des ouvrages. Nous pouvons donc conclure, à partir de ces calculs que les annotations des experts ne suffisent pas pour l'extraction d'information. Deux perspectives s'offraient à nous : exploiter ces exceptions parmi les annotations des experts en calculant une similarité par Cosinus entre les termes du texte pour les regrouper par exemple ou bien rajouter de l'information textuelle en développant un extracteur terminologique basé sur le corpus de Berger-Levrault. Nous nous sommes orientés vers un extracteur terminologique que nous présentons dans la section suivante afin de rajouter de l'information textuelle.

4.2 Système InfoGlean KeyTerms pour l'Extraction Automatique de Termes

Les **communes** dotées d'un **plan d'occupation des sols** rendu public ou d'un **plan local d'urbanisme** approuvé peuvent, par délibération, instituer un ***droit de préemption urbain*** sur tout ou partie des ***zones* urbaines** et des ***zones* d'*urbanisation future*** délimitées par ce plan, dans les **périmètres de protection rapprochée** de **prélèvement d'eau** destinée à l'alimentation des **collectivités humaines** définis en application de l'article L. 1321-2 du **Code de la santé publique,** *zones* et secteurs définis par un **plan de prévention des *risques technologiques*** en application de l'article L. 515-16 du **Code de l'environnement,** dans les *zones* soumises aux **servitudes prévues** au II de l'**article L. 211-12** du même code, ainsi que sur tout ou partie de leur territoire couvert par un **plan de sauvegarde et de mise en valeur** rendu public ou approuvé en application de l'article L. 313-1 lorsqu'il n'a pas été créé de ***zone* d'aménagement différé** ou de **périmètre provisoire de *zone* d'aménagement différé** sur ces territoires.

Exemple 4.4 – Exemple d'un paragraphe avec les trois différentes extractions de InfoGlean KeyTerms superposées aux annotations des experts.

Légende :

⟶ *Annotations des experts manuellement transformés pour faire correspondre avec le texte*

⟶ Termes extraits par le système de NER

⟶ **Termes extraits par l'extracteur de termes pertinents**

⟶ Termes extraits par le système d'extraction d'entités juridiques

InfoGlean KeyTerms est le nom donné à l'extracteur terminologique que nous avons développé. Cet extracteur se décompose en trois sous-systèmes : un système pour la NER, un pour l'extraction des termes pertinents et un pour les entités juridiques. L'exemple 4.4 présente les différentes extractions des trois sous-systèmes ainsi que les annotations des experts afin de les confronter. Nous utilisons des représentations différentes pour chacun des systèmes et comme précédemment, nous dénotons les annotations des experts par l'*italique*.

Le **système de NER**, nommé *Entity eXtraction And Classification Tools* (EXACT) et présenté par la coloration en bleu du souligné en pointillé sur l'exemple 4.4, est une tâche qui consiste à identifier des entités nommées spécifiques dans un texte puis de les catégoriser dans des catégories préétablies telles que des personnes, des lieux, des organisations, des dates et bien d'autres encore (Jehangir et al., 2023b). Les entités nommées sont des éléments concrets du

texte qui ont une importance particulière, s'agissant souvent de noms propres. La figure 4.1 présente un exemple étiqueté par un système NER tiré du jeu de données MUC [1], dont la 7^{ieme} édition a été utilisée par le pionnier des NER (Chinchor, 1997). Dans cet exemple, nous constatons le nom propre « Henri Nord ». En effet, sans détection de cette entité nommée le mot « Nord » a du sens et pourrait être traité différemment. Il est donc important de détecter les entités nommées à part entière afin de ne pas les tronquer en différents mots simples et ainsi être mal interprétées. Généralement, un système NER est composé de deux étapes : l'identification des entités nommées et l'étiquetage des entités dans chacune des catégories définies au préalable.



FIGURE 4.1 – Texte étiqueté par un système de reconnaissance d'entités nommées utilisé lors de la campagne d'évaluation MUC

Deux options se sont offertes à nous : soit l'utilisation de systèmes classiques étant entraînés à générer des arbres syntaxiques, soit l'exploitation de modèles de langue (SLM ou LLM). Les essais sur les systèmes classiques comme Stanza (Qi et al., 2020), Spacy (Honnibal et al., 2020) ou Flair (Akbik et al., 2018b) se sont montrés non concluants sur nos textes de spécialité car ces derniers contiennent un vocabulaire trop spécifique (et technique) ainsi que des phrases trop longues. Nous nous sommes donc orientés vers l'exploitation de modèles de langue. Et pour cela, nous utilisons un modèle CamemBERT (Martin et al., 2019) finetuné pour la tâche des NER (CamemBERT-NER [2]). De plus, le modèle choisi est « CamemBERT-NER-with-dates [3] », disponible sur *HuggingFace*, une extension du modèle avec un étiquetage supplémentaire pour les dates [4].

Après avoir fait le choix de ce système, l'essor des LLMs et les bons résultats obtenus pour les termes fréquents nous ont amené au prompt présenté sur l'exemple 4.5 injecté dans l'API d'OpenAI pour la génération de textes (ChatGPT) [5]. Nous omettons de définir un format de sortie pour la lisibilité de l'exemple mais tout format peut être défini. Par exemple, en conditions réelles, nous ajoutons aux *prompts* donnés la sortie en JSON comme nous le verrons dans la partie évaluation (c.f. section 4.3). Lai et al. (2023) pointe l'idée qu'un *prompt* en anglais serait en général de meilleure qualité mais l'article montre également que ce n'est pas le cas pour le français car il est considéré comme une langue de haute ressource (>1%). En effet, les différentes langues sont regroupées dans quatre catégories en fonction du ratio de données de cette langue du corpus CommonCrawl : Haute Ressource (H, > 1%), moyenne ressource (M, > 0.1%), faible Ressource (L, > 0.01%), et Extrêmement faible Ressource (X, < 0.01%).

L'**extracteur de termes pertinents** - ou connu sous extraction de termes clés (*keyphrase / keyword extraction*) - se nomme *EXtraction of TERM using OpenIE* (EXTERMO) et consiste à

1. *Message Understanding Conference* (MUC) : `https://aclanthology.org/venues/muc/`
2. Le modèle camembert-ner est disponible sur HuggingFace : `https://huggingface.co/Jean-Baptiste/camembert-ner`
3. Le modèle camembert-ner-with-dates est tiré de HuggingFace : `https://huggingface.co/Jean-Baptiste/camembert-ner-with-dates`.
4. La librairie de l'analyseur de date utilisé est disponible à l'adresse suivante : `https://dateparser.readthedocs.io/en/latest/`.
5. Lien vers l'API d'OpenAI ChatGPT : `https://platform.openai.com/docs/models/chatgpt`

Soit le texte suivant :
En décembre 2014, un rapport a été dressé par la Banque de France au sujet de la question préoccupante du surendettement. Justement intitulé « Étude des parcours menant au surendettement », il nous livre une analyse en amont des dépôts de dossiers auprès des commissions de surendettement.

Extraire les termes pertinents puis les entités nommées de ce texte. L'output à obtenir sont des listes dont les éléments sont séparés par des points virgules.

OUTPUT :
Termes pertinents : décembre 2014 ; rapport ; Banque de France ; surendettement ; Étude des parcours menant au surendettement ; analyse ; dépôts de dossiers ; commissions de surendettement.
Entités nommées : Banque de France ; Étude des parcours menant au surendettement.

Exemple 4.5 – Exemple de *prompt* pour l'extraction des termes pertinents et des NER.

extraire les termes (simples ou complexes) importants à partir d'un texte. Ces termes clés permettent de cerner le sujet d'un document (Mihalcea & Tarau, 2004; Beliga et al., 2015). Ils sont présentés en **gras** sur l'exemple 4.4. Deux versions ont été mises en place. La première exploite les systèmes état de l'art connus depuis quelques années comme les modèles de langue, les modèles statistiques ou encore les systèmes de *ranking* basés sur les graphes. La seconde est arrivée après l'essor des systèmes basés sur les LLMs. L'utilisation de l'API d'OpenAI pour la génération de texte (ChatGPT) a montré de bons résultats. L'exemple 4.6 présente le *prompt* utilisé dans ChatGPT pour l'extraction de termes pertinents.

Le **système d'extraction d'entités juridiques**, nommé *Juridicial entity Regular Expression eXtractor* (JUREX) et présenté par un soulignement ondulé et coloré en orange sur l'exemple 4.4, consiste à extraire les entités d'un texte comme pour le NER mais liées au domaine juridique uniquement. Nous avons observé cinq types différents d'entités juridiques : Loi, Article, Proposition, Décret et Autre. L'exemple 4.7 présente pour chacun des types d'entités juridiques des exemples réels tirés de nos textes juridiques. L'article de Bendahman et al. (2023) propose une approche consistant à extraire les entités dites d'intérêt, c'est à dire à la fois les entités juridiques et les entités nommés. Nous utilisons ce système reposant sur des règles basées sur des expressions régulières (REGEX).

Extraire les termes pertinents du texte encadré par les balises <text> et </text>. Le format d'output doit être :
- Une liste pour chacun des termes pertinents dont les éléments sont séparés par ;

<text> "Les communes dotées d'un plan d'occupation des sols rendu public ou d'un plan local d'urbanisme approuvé peuvent, par délibération, instituer un droit de préemption urbain sur tout ou partie des zones urbaines et des zones d'urbanisation future délimitées par ce plan, dans les périmètres de protection rapprochée de prélèvement d'eau destinée à l'alimentation des collectivités humaines définis en application de l'article L. 1321-2 du Code de la santé publique, zones et secteurs définis par un plan de prévention des risques technologiques en application de l'article L. 515-16 du Code de l'environnement, dans les zones soumises aux servitudes prévues au II de l'article L. 211-12 du même code, ainsi que sur tout ou partie de leur territoire couvert par un plan de sauvegarde et de mise en valeur rendu public ou approuvé en application de l'article L. 313-1 lorsqu'il n'a pas été créé de zone d'aménagement différé ou de périmètre provisoire de zone d'aménagement différé sur ces territoires." </text>

OUTPUT :
communes ; plan d'occupation des sols ; plan local d'urbanisme ; droit de préemption urbain ; zones urbaines ; zones d'urbanisation future ; périmètres de protection rapprochée de prélèvement d'eau ; alimentation des collectivités humaines ; article L. 1321-2 du Code de la santé publique ; zones ; secteurs ; plan de prévention des risques technologiques ; article L. 515-16 du Code de l'environnement ; zones soumises aux servitudes ; article L. 211-12 ; plan de sauvegarde et de mise en valeur ; article L. 313-1 ; zone d'aménagement différé ; périmètre provisoire de zone d'aménagement différé

Exemple 4.6 – Exemple de *prompt* pour l'extraction de termes pertinents.

4.3 Méthodologie et Résultats de l'évaluation des sous-systèmes d'extraction terminologique

4.3.1 Structuration et Critères de l'évaluation des sous-systèmes

Nous souhaitons évaluer notre système d'extraction terminologique en différentes phases. La première et la primordiale consiste à évaluer indépendamment chacune des composantes

— **Loi :** La loi n° 2016-162 du 19 février 2016
— **Article :** l'article 275 du Code civil (Cass., 1re civ., 15 avr. 2015, n° 14-12375) / l'article 2238 du Code civil
— **Proposition de loi :** Proposition de loi adoptée en 1ère lecture par le Sénat le 11 mars 2015 / projets de loi au Conseil des ministres du vendredi 31 juillet 2015
— **Décret :** Le décret n°2016-576 du 11 mai
— **Autre :** l'ordonnance n° 2015-1033 du 20 août 2015

Exemple 4.7 – Liste de différents exemples pour chacun des types d'entités juridiques

du système sur différents jeux de données, avec différentes techniques d'évaluation et différentes métriques. Ensuite, en seconde phase, une évaluation dite qualitative sera exécutée. Il s'agit d'une évaluation globale de notre système sur notre jeu de données.

Choix des différents sous-systèmes

Système de NER :

1. CamemBERT NER *with dates* est le modèle visé car il représente une extension du modèle de CamemBERT-NER mais avec un *tag* additionnel pour les dates. Les jeux de données, état de l'art, ne permettant pas de vérifier la pertinence des dates, nous utilisons pour l'expérimentation le modèle CamemBERT-NER classique et ajouter les informations de dates ne ferait que fausser les résultats. Pour les jeux de données standard, nous utiliserons donc le modèle CamemBERT-NER classique.
2. L'API d'OpenAI pour la génération de texte (ChatGPT) : Nous utilisons GPT-3.5-Turbo, développé par OpenAI, via l'API pour interpréter puis générer du texte. Ce modèle peut traiter des contextes de conversation jusqu'à 4096 tokens par requête. Une requête se formule sous forme de *prompt*, qui consiste à formuler des instructions ou des questions spécifiques pour obtenir des réponses précises et pertinentes du modèle. Cette méthode permet de guider le modèle vers des outputs souhaités. Un token représente généralement environ 4 caractères de texte anglais, mais pour le français, il correspond souvent à environ le 3/4 d'un mot. La tarification de l'API est fondée sur le nombre de tokens consommés, avec des coûts proportionnels au volume et à la fréquence des requêtes. GPT-3.5-Turbo peut identifier et classifier les entités nommées telles que les personnes, les organisations, les lieux, les dates, etc., en fonction du contexte fourni. GPT-3.5-Turbo peut également extraire les phrases clés en saisissant la sémantique et le contexte global du texte.
3. GliNER multilingue V_0 et $V_{2.1}$ (Zaratiana et al., 2023)[6] s'appuient sur un encodeur transformateur bidirectionnel similaire à BERT pour le NER. Il est capable d'identifier divers types d'entités et contrairement aux modèles NER conventionnels, qui sont

6. L'ensemble des modèles et papiers liés à GliNER sont disponibles sur HuggingFace : `https://huggingface.co/urchade`

souvent limités à des entités prédéfinies, GliNER offre une flexibilité en reconnaissant tout type d'entité. GliNER représente une alternative aux LLMs. GliNER multilingue est intéressant car il est capable de traiter plusieurs langues. GliNER multilingue V_0 est sous licence « Creative Commons Attribution Non Commercial 4.0 » tandis que GliNER multilingue $V_{2.1}$ est sous la licence « Apache License 2.0 ».

4. WikiNEuRal NER multilingue (Tedeschi et al., 2021) est un système qui combine des approches neuronales et basées sur la connaissance pour créer des ensembles de données multilingue de type *silver-standard* afin d'améliorer la NER multilingue. Un ensemble de données de type *silver-standard* fait référence à un ensemble de données annotées ou étiquetées avec une certaine fiabilité, mais pas avec une garantie absolue de précision. Contrairement à un *gold standard*, qui est considéré comme une référence de haute qualité avec des annotations vérifiées et validées de manière rigoureuse, un *silver-standard* peut comporter une certaine marge d'erreur ou d'incertitude dans les annotations. Dans ce système, un modèle de langage multilingue (mBERT) est affiné pendant trois époques en utilisant l'ensemble de données WikiNEuRal, qui couvre neuf langues, notamment l'allemand, l'anglais, l'espagnol, le français, l'italien, le néerlandais, le polonais, le portugais et le russe. Le modèle est entraîné conjointement sur les neuf langues pour prendre en charge la NER multilingue. Cependant, il est important de noter que WikiNEuRal est dérivé de Wikipedia, donc les performances du modèle peuvent ne pas se généraliser bien à tous les types de texte, tels que les articles de presse. Pour remédier à cette limitation et améliorer la robustesse, il est recommandé de combiner WikiNEuRal avec d'autres ensembles de données, tels que CoNLL, qui se concentre principalement sur les articles de presse. WikiNEuRal est sous la licence « Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0) » dont les conditions restreignent l'utilisation du contenu de ce référentiel à des fins de recherche non commerciale sous la licence.

Nous souhaitions nous orienter vers deux autres systèmes supplémentaires à savoir NuNER multilingue (Bogdanov et al., 2024) et Gemini (Gemini_Team, 2024). Gemini s'est avéré beaucoup trop complexe à mettre en place à cause du bruit généré dans ses réponses et parfois le manque de données dans la sortie du prompt. En effet, un post-traitement beaucoup trop conséquent aurait été nécessaire afin d'obtenir les NER. Comme nous pouvons le voir sur la figure 4.2, la sortie du prompt est modifiée au cours de la génération du résultat ce qui rend l'interprétation du résultat difficile. De plus, tous les mots à classifier du jeu de données standard n'ont pas été pris en compte et ont été ignorés dans le traitement. Nous ne pouvons que difficilement ensuite comparer les résultats. NuNER, quant à lui, n'extrait pas les NER d'un texte mais les *embeddings* les représentant. Nous avons cherché à les intégrer à une *pipeline* afin de pouvoir extraire les NER mais cela n'a pas abouti à des résultats concluants.

Système d'extraction des termes pertinents : Nous utilisons afin d'extraire les termes pertinents d'un texte l'API d'OpenAI pour la génération de texte (ChatGPT) Nous utilisons la même version que pour la NER. Nous utilisons également Gemini (Gemini_Team, 2024) étant un des concurrents direct d'OpenAI. Gemini, baptisé *Generalized Multimodal Intelligence Network*, est un LLM génératif et multimodal de type transformeur, développé par Google. Ce LLM a été entraîné sur d'énormes ensembles de données textuelles et visuelles à l'aide des *Tensor Processing Units* (TPU) de Google. Gemini est capable d'interpréter et de générer du texte, du code ainsi que des traductions dans plus de 100 langues. Il est également capable d'interpréter divers médias tels que l'audio et la vidéo et de générer des contenus

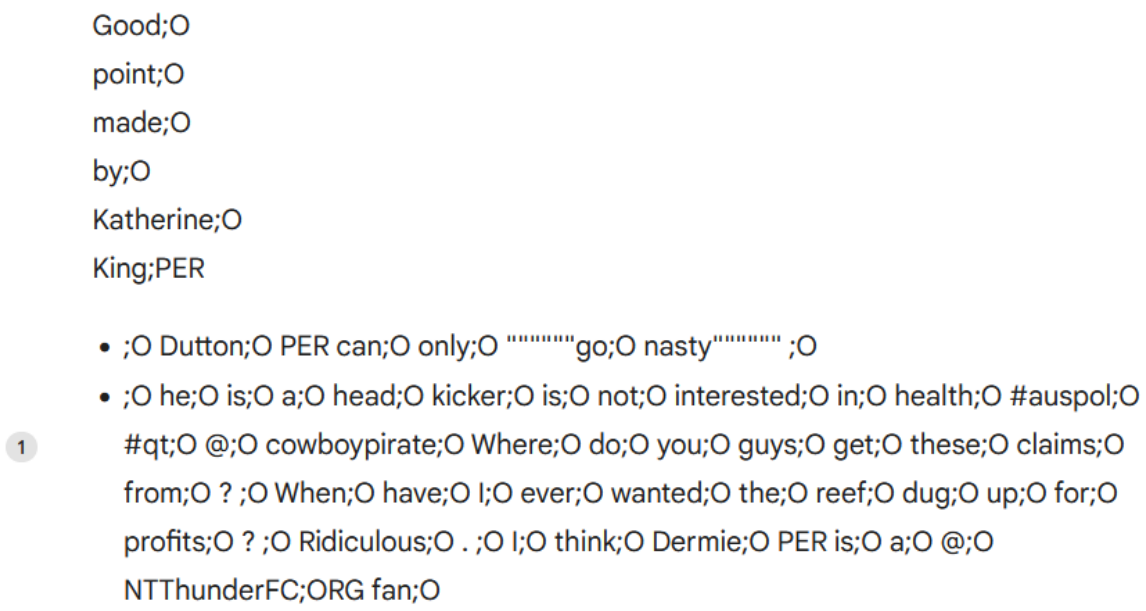

FIGURE 4.2 – Exemple de sortie de Gemini pour la tâche de NER

créatifs comme des images, des dessins, de la musique et des vidéos mais nous ne nous en servons pas pour notre cas d'usage. Nous utilisons la version de base de Gemini à savoir la $V_{1.0}$.

Leur *prompt* est donné sur l'exemple 4.3.1. Un exemple de *prompt* donné à l'API d'OpenAI pour la génération de texte (ChatGPT) en intégrant un paragraphe issu des textes de Berger-Levrault est présenté sur l'exemple 4.3.1. Si le corpus utilisé présente déjà un prompt prédéfini, nous utiliserons l'API d'OpenAI pour la génération de texte (ChatGPT) avec le prompt prédéfini directement.

Soit le texte suivant :

TEXT

Extrais moi les termes pertinents de ce texte sous forme de liste séparée par des ;

Exemple 4.8 – Prompt utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'obtenir les termes pertinents d'un texte donné.

Corpus

Nous avons donc décidé d'évaluer chacun des sous-systèmes indépendamment. Nous avons différents corpus pour chacun des sous-systèmes. Cependant, en ce qui concerne la NER et l'extraction de termes pertinents, nous pouvons regrouper ces jeux de données en trois catégories :

Soit le texte suivant :

Initiée par l'ordonnance n°2005-1527 du 8 décembre 2005, la réforme du permis de construire procède du mouvement de simplification du droit entamée depuis une quinzaine d'années. Entre objectif constitutionnel d'intelligibilité et inéluctable inflation législative, le permis de construire s'est essentiellement construit par voie d'ordonnances avec les avantages et inconvénients que cela implique. Une modernisation selon un projet de loi habilitant le gouvernement à simplifier le droit est toujours évoquée.

Extrais moi les termes pertinents de ce texte sous forme de liste séparée par des ;

OUTPUT :
ordonnance n°2005-1527 ; 8 décembre 2005 ; réforme du permis de construire ; simplification du droit ; objectif constitutionnel d'intelligibilité ; inflation législative ; permis de construire ; ordonnances ; modernisation ; projet de loi ; gouvernement ; simplifier le droit

Exemple 4.9 – Exemple de prompt utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'obtenir les termes pertinents d'un paragraphe issu du corpus de Berger-Levrault.

1. Évaluation sur un corpus standard : multilingue, français ou anglais. Il s'agit d'évaluer le système de manière générale et d'évaluer son adaptabilité dans le cas où le jeu de données évoluerait différemment.
2. Évaluation sur un corpus de spécialité mais qui n'est pas forcément du juridique. Étant donné le peu de jeux de données disponiblex, il est très fréquent qu'il soit nécessaire de s'orienter vers un jeu de données en anglais.
3. Évaluation sur notre corpus de spécialité mais de manière non supervisée. Pour cela, nous allons mettre en place des techniques d'évaluation particulières qui sont détaillées dans ce qui suit

Nous allons désormais détailler chacun de ces corpus pour chacun des sous-systèmes.

Système de NER : Parmi les jeux de données état de l'art, un bon nombre de jeux reconnus par la communauté existent. Pour chacune des catégories listées précédemment, nous allons lister nos choix tout en les justifiant.

1. **Évaluation sur un corpus standard :** Nous nous orientons vers un jeu de données issu de Wikipédia et Freebase nommé Polyglot-NER (Al-Rfou et al., 2015). Il s'agit d'un jeu de données multilingue contenant le français. Le jeu de données a été récupéré librement au format CONLL. Nous avons récupéré le jeu d'entraînement du projet.

Afin de l'évaluer, nous procédons de la manière suivante. Le jeu de données associe ligne par ligne un mot (voir un petit groupe de mots) à une classe de NER répartie dans quatre grandes catégories : « Other (autre) », « Personne », « Organisation » ou « Localisation ». Cependant, les systèmes sont des extracteurs de NER et sont donc applicables sur l'ensemble d'une phrase. Nous avons donc extrait les phrases du jeu de données en reconstruisant les phrases à partir de mots individuels et en enregistrant les indices de début et de fin de chaque phrase. Le modèle NLP est appliqué à chaque phrase pour détecter les entités nommées. Les résultats du modèle sont comparés avec les annotations dans le jeu de données. Cette comparaison permet de vérifier si les entités détectées par le modèle correspondent aux annotations manuelles. Pour chaque entité nommée détectée dans la phrase, nous recherchons l'emplacement du token dans le jeu de données. Il y a une gestion des cas où il n'y a pas de correspondance, c'est-à-dire les cas où aucune entité nommée n'est trouvée en attribuant des valeurs par défaut. Étant donné que dans le jeu de données est associé à chacun des mots voir des groupes des mots une classe et que les systèmes extraits les entités nommés existantes, nous associons à la classe « autre », tous les mots qui n'ont pas été extraits par le système. De plus, parfois le système extrait un groupe de mot découpé au sein du jeu de données. Nous redispatchons donc le groupe de mots avec chacune des étiquettes à chacun des mots individuellement qui le composent. Enfin, nous vérifions si la classe prédite est correcte avec celle du jeu de données. Puis une *accuracy* ou taux d'exactitude est calculé : quel est le ratio de bonnes prédictions sur l'ensemble du jeu de données ?

2. **Évaluation sur un corpus de spécialité :** Nous nous orientons vers le jeu de données Broad Twitter corpus dit BTC (Derczynski et al., 2016). Il s'agit d'un ensemble de *tweets* collectés à des moments, dans des lieux et pour des usages sociaux divers. L'objectif est de représenter un large éventail d'activités, afin d'obtenir un ensemble de données plus représentatif du langage utilisé dans ces formats de médias sociaux les plus difficiles à traiter. BTC se présentant sous le même format que PolyglotNER, nous appliquons le même processus d'évaluation. BTC étant en anglais, nous réalisons deux évaluations distinctes : l'une sur le corpus en anglais directement puisque certains des modèles utilisés sont multilingues et l'autre sur une traduction en français du corpus. Pour cela, après avoir regroupé les phrases complètes du jeu de données, nous appliquons le *transformer* pour la traduction Opus-mt-en-fr.

Système d'extraction de termes pertinents :

1. **Évaluation sur un corpus standard :** Nous nous orientons vers les jeux de données *wikinews-fr-100_fr_prompt_keywords_extraction* et par conséquent wikinews-fr-100[7].

   Wikinews-fr-100 est composé de 100 articles d'actualité en français provenant de wikinews. Les phrases-clés ont été annotées par des lecteurs (étudiants en informatique) dans un cadre non contrôlé (c'est-à-dire non limité à des entrées de thésaurus). Les phrases-clés de référence (attribuées par l'indexeur) sont également classées selon le schéma PRMU (*Present-Reordered-Mixed-Unseen*) (Boudin & Gallina, 2021).

2. **Évaluation sur un corpus de spécialité :** Nous nous orientons vers deux corpus de spécialités. Le premier est taln-archives_fr prompt_keywords_extraction et par conséquent TALN-archives. TALN-Archives (Boudin, 2013) est un ensemble de données composé de 1207 résumés d'articles scientifiques en français provenant des Archives

7. Retrouvez le détail des corpus de CATIE-AQ sur leur page HuggingFace : `https://huggingface.co/CATIE-AQ`

TALN. Les phrases-clés ont été annotées par les auteurs dans un cadre non contrôlé (c'est-à-dire non limité à des entrées de thésaurus). Les traductions anglaises du titre, du résumé et des phrases-clés sont également disponibles pour un sous-ensemble de documents (456 documents entièrement traduits et 719 documents partiellement traduits), ce qui permet d'expérimenter la génération de phrases-clés multilingues. Le deuxième jeu de données est Termith-eval_fr prompt_keywords_extraction et par conséquent TermITH-Eval. TermITH-Eval (Bougouin et al., 2016) est un ensemble de données composé de 400 résumés d'articles scientifiques en français provenant des bases de données FRANCIS et PASCAL de l'Institut français de l'information scientifique et technique (Inist). Les mots-clés ont été annotés par des indexeurs professionnels dans un cadre non contrôlé (c'est-à-dire, non limité à des entrées de thésaurus).

Système d'extraction des entités juridiques : Contrairement aux deux sous-systèmes précédents, ce système ne peut être évalué autrement que de manière non supervisée. Nous nous allons donc utiliser un échantillon de notre propre corpus de données. Il s'agit du même échantillon que nous allons ci-après détailler pour les trois systèmes. Nous détaillerons dans la partie suivante les techniques d'évaluation pour ce système afin de pallier le manque de données.

Évaluation sur le corpus de Berger-Levrault (échantillon) : Nous utilisons, ici, non pas tout le corpus mais un échantillon. Ce corpus ne contient que des annotations associées au paragraphe et n'est donc pas suffisamment annoté pour la tâche de d'extraction de termes clés (KPE) et n'est ni annoté pour le NER, ni pour les entités juridiques. Comme explicité précédemment, nous allons donc utiliser des techniques d'évaluation de manière non supervisée que nous détaillerons dans la partie suivante. Dans ce paragraphe, nous détaillons l'échantillon créé à partir de notre corpus de spécialité.

Chacun des systèmes sera évalué sur le même échantillon. Nous réduisons la taille du corpus car certaines techniques d'extraction demandent beaucoup de ressources, comme un coût financier par exemple en ce qui concerne l'API d'OpenAI pour la génération de texte (ChatGPT) . Nous allons donc éviter de multiplier le nombre de requêtes afin de minimiser le prix de nos expériences. De plus, certaines techniques limitant le nombre dit de *tokens*, demandent en entrée un texte restreint. Les ouvrages du corpus de spécialité étant de grosse taille, nous avons décidé de ne sélectionner que les articles du corpus et de les filtrer par domaine. Ensuite, nous avons sélectionné l'entièreté des articles du plus petit domaine. Il s'agit du domaine « élections » qui compte 250 articles. Pour chacun des autres domaines, nous sélectionnons 250 articles. Nous tirons aléatoirement un plus grand nombre d'articles (ici, 300 articles) et nous retirons les articles les plus petits afin d'avoir le plus d'information. Nous obtenons donc 2 000 articles que nous avons également découpés en paragraphe. Nous obtenons 3 280 paragraphes et cet ensemble a été filtré si les paragraphes étaient vides (en cas de saut de page par exemple) réduisant cet ensemble à 3 262 paragraphes. Un article est composé en moyenne de 1 782 mots tandis qu'un paragraphe est lui composé de 39 mots, en moyenne. Le tableau 4.7 répertorie différents calculs par domaine : le nombre moyen des mots présents dans les articles et les paragraphes ainsi que le nombre de paragraphes.

Techniques d'évaluation et métriques

Taux d'exactitude Sur les corpus standards annotés, nous pouvons directement calculer le taux d'exactitude. Nous calculons le nombre d'éléments prédits correctement face à l'ensemble des éléments du jeu de données. Cependant, il est à noter que les systèmes ne vont

Domaine	Nbre mots articles	Nbre mots paragraphes	Nbre paragraphes
Commande publique	1 833	63	276
Comptabilité et finances locales	1 002	20	338
Justice	2 174	37	463
Ressources humaines territoriales	990	49	321
Santé	1 036	28	313
Urbanisme	2 780	39	612
Élections	3 767	46	637
État civil et cimetières	705	22	302

TABLE 4.7 – Nombre moyen de mots par domaine

Légende :

$\longrightarrow$ **Calculs** :

Nbre mots articles : Nombre de mots moyen pour les articles d'un domaine

Nbre mots paragraphes : Nombre de mots moyen pour les paragraphes d'un domaine

Nbre paragraphes : Nombre de paragraphes associés à un domaine (après avoir découpé les 250 articles en paragraphes et avoir filtré les paragraphes vides)

pas classifier chaque mot du jeu de données mais comme exprimé précédemment, ils vont extraire directement les entités nommées à partir des phrases. Nous rajoutons donc une couche de « souplesse ». Cela signifie que pour chaque entité extraite, nous recherchons l'ensemble des mots correspondants dans les jeux de données et nous les associons à l'étiquette. Nous estimons que le fait d'associer l'ensemble des mots qui constituent le groupe de mots à cette même classe donnera un bon aperçu de la qualité du système. En effet, cela permet d'une part d'estimer la capacité d'un système à extraire les entités nommées et d'une autre part, nous nous évaluons sur un corpus standard où la classification d'entités nommées est privilégiée afin de faciliter les traitements.

Statistiques sur les mots Dans le cadre de l'évaluation de différents systèmes pour la NER ou la KPE, plusieurs calculs peuvent fournir des informations pertinentes sur les performances et les caractéristiques des systèmes.

1. **Comparaison du nombre de termes multi-mots vs mots simples** : Cette comparaison permet de mesurer la capacité du système à identifier des expressions complexes par rapport à des termes uniques. Les termes multi-mots, comme « Communautés de communes », « Pacte civil de solidarité (PACS) » ou « Tribunal administratif », sont souvent plus informatifs et spécifiques que des termes simples comme « communautés », « pacte » ou « tribunal ». Un système efficace doit donc être capable de reconnaître et extraire ces expressions composées, car elles fournissent un contexte plus riche et pertinent. La proportion de termes multi-mots par rapport aux termes simples peut indiquer la finesse et la précision du système dans la capture des concepts clés.
2. **Nombre moyen de mots**

   2.1 **Nombre moyen de mots au sein d'un terme** : Calculer la longueur moyenne des termes extraits peut offrir une perspective sur la nature des termes identifiés. Un nombre moyen de mots élevé pourrait indiquer que le système tend à extraire

des phrases ou des expressions complexes, ce qui pourrait être bénéfique dans les domaines spécifiques où les concepts sont souvent exprimés en plusieurs mots, comme dans notre cas pour « Pacte civil de solidarité (PACS) » par exemple. Cependant, il faut être vigilant à être équilibré sans tomber dans l'autre extrême et ainsi avoir un risque de bruit. En effet, les termes très longs sont plus susceptibles de contenir des mots non pertinents ou des erreurs de segmentation, ce qui peut introduire du bruit dans les résultats d'extraction ou de reconnaissance. Cette présence de bruit peut compromettre la qualité globale de l'analyse.

2.2 **Nombre moyen si les termes qui ont au moins deux mots sont filtrés** : Les termes simples (un seul mot) peuvent biaiser la moyenne générale vers des valeurs plus faibles. En excluant ces termes simples, on obtient une mesure plus équilibrée qui reflète mieux la capacité du système à identifier des expressions composées. Cela permet de comparer plus justement les performances des systèmes sur des termes multi-mots, sans l'influence disproportionnée des termes simples. En se concentrant sur les termes multi-mots, on peut identifier si un système rencontre des difficultés spécifiques à extraire des expressions complexes ou, au contraire, s'il excelle dans cette tâche. Cela peut aider à cibler des améliorations précises dans le développement du système. Si le système a un nombre moyen de mots beaucoup plus conséquent, cela signifie qu'il extrait beaucoup de mots simples. Étant donné que la moyenne est abaissée, connaître la moyenne en filtrant les termes pour ne considérer que ceux composés d'au moins deux mots permettent de vérifier qu'un trop gros nombre de termes n'est pas extrait pour les multi-mots pour les mêmes raisons explicitées précédemment.

Prompts utilisés Le tableau 4.8 résume les divers prompts utilisés pour différentes évaluations, notamment la NER et la KPE sur des corpus standards. Pour le NER, le prompt spécifie la classification de chaque mot en tant qu'entité nommée ou non, avec une catégorisation spécifique pour les entités nommées telles que les personnes, les organisations et les lieux. Pour le KPE, les prompts sont intégrés directement dans les jeux de données, mais pour l'exemple du corpus de Berger-Levrault, le prompt indique d'extraire à la fois les termes pertinents et les entités nommées à partir du texte encadré par des balises spécifiques.

Evaluation par LLM Le dernier prompt du tableau 4.8 sert à l'évaluation pour le NER et le KPE. En effet, étant donné le manque de labels pour l'échantillon du corpus de Berger-Levrault, les différents modèles sont évalués par système de vote auprès de LLMs. Pour évaluer la pertinence des entités nommées ou des termes pertinents extraits d'un domaine donné, nous utilisons un mécanisme de vote entre deux LLMs distincts. Le prompt fourni aux LLMs permet d'attribuer une note entre 1 et 5 ainsi qu'un score de confiance flottant compris entre 0 et 1 (pourcentage %) pour chaque entité, en fonction de leur pertinence par rapport au domaine spécifié. Après exécution du prompt sur les deux LLMs, nous calculons une moyenne des notes et des scores de confiance pour obtenir une évaluation consolidée. Cette méthode de vote permet de lisser les biais individuels des modèles et d'obtenir une estimation plus fiable de la pertinence des entités extraites. Le résultat final est présenté au format JSON, incluant la moyenne des scores et des pourcentages de confiance. En évaluant ainsi la pertinence par domaine, nous nous assurons que les entités extraites correspondent bien aux attentes et aux exigences spécifiques du domaine souhaité.

Évaluation	Prompt utilisé
NER sur les corpus standards	Soit le fichier CoNLL suivant : [TEXT] Pour chacune des lignes (chaque mot), dire s'il s'agit ou non d'une entité nommée. S'il s'agit d'une entité nommée classifier dans l'une de ces classes suivantes : [O,PER, ORG,LOC] Sinon classifier dans la classe 'O' Output : mot ;classeAssociée ;
KPE corpus standards	*prompts directement dans les jeux de données*
NER & KPE sur l'échantillon du corpus de Berger-levrault	Extraire les termes pertinents puis les entités nommées du texte encadré par les balises <text> et </text>. Le format d'output doit être : - Une liste pour chacun dont les éléments sont séparés par ; - Pour les entités nommées un couple (Classe, Entité) <text> [TEXT] </text>
Évaluation pour le NER et KPE	Soit les entités nommées (ou termes pertinents) extraites pour le domaine "NOMDOMAINE" séparées par des points virgules données entre les balises <NE> et </NE> <NE> LISTE NE OU KP </NE> Donne une note entre 1 et 5 ainsi qu'un score flottant de confiance compris entre 0 et 1 pour chacune des entités afin d'évaluer la pertinence par rapport au domaine Donne le résultat au format JSON

TABLE 4.8 – Différents prompts utilisés pour les différentes évaluations

4.3.2 Présentation et Analyse des Résultats Obtenus par les Sous-Systèmes d'Extraction

Résultats pour le système de NER

Modèle	PolyglotNER	BTC	
		EN	FR
CamemBERT NER with dates	86 %	80%	81,5 %
L'API d'OpenAI (ChatGPT)	**99,9** %	**99,8** %	**99,8** %
Wikineural NER multilingue	87,5 %	81,1 %	82 %
GliNER multilingue V_0	86,5 %	82,1 %	*82,2* %
GliNER multilingue $V_{2.1}$	*89,1* %	*84,6%*	82%

TABLE 4.9 – Résultats des différents modèles sur les corpus standard de NER PolyglotNER et Broad Twitter Corpus (BTC)

Évaluation sur les jeux de données standard Le tableau 4.9 présente les résultats de l'évaluation pour la tâche de NER des performances de différents modèles sur deux corpus standard : PolyglotNER et Broad Twitter Corpus (BTC), dans les langues anglaise (EN) et française (FR). L'API d'OpenAI pour la génération de texte (ChatGPT) est le modèle de NER le plus performant selon cette évaluation, suivi par GliNER multilingue $V_{2.1}$. Les autres modèles, bien qu'efficaces, montrent des marges de progression par rapport à ces modèles. Nous pouvons classer l'ensemble des résultats en trois catégories : les modèles ayant obtenu de très bonnes performances, ceux ayant obtenu une bonne performance et enfin ceux ayant obtenu des performances assez bonnes voire passable.

— **Très bonne performance** : L'API d'OpenAI (ChatGPT) se démarque largement des autres modèles, avec des scores quasi parfaits sur tous les tests, indiquant une supériorité claire en termes de précision et de généralisation pour la tâche de NER. En effet, l'API d'OpenAI (ChatGPT) affiche des performances presque parfaites, avec des scores très élevés et constants sur tous les corpus et langues. Cela suggère une très grande efficacité et fiabilité pour la tâche de NER.

— **Bonne Performance** : GliNER multilingue $V_{2.1}$ se classe comme le deuxième meilleur modèle, montrant des améliorations significatives par rapport à sa version antérieure et à d'autres modèles comme CamemBERT et Wikineural. En effet, GliNER multilingue $V_{2.1}$ présente une amélioration notable par rapport à la version précédente (V_0), particulièrement sur PolyglotNER et BTC en anglais. Les performances en français restent constantes par rapport à V_0.

— **Assez bonne performance voire passable** : CamemBERT NER with dates, Wikineural NER multilingue, et GliNER multilingue V_0 affichent des performances respectables mais ne peuvent rivaliser avec l'API d'OpenAI et GliNER multilingue $V_{2.1}$. GliNER multilingue V_0 montre des performances légèrement meilleures que CamemBERT sur PolyglotNER et BTC en anglais. Ses performances en français sont similaires à celles de Wikineural NER multilingue. Wikineural NER multilingue a des performances comparables à CamemBERT, mais avec une légère amélioration sur PolyglotNER. Les performances sont également assez équilibrées entre l'anglais et le français sur le BTC. CamemBERT NER with dates montre des performances solides mais pas exceptionnelles. Il est légèrement plus performant en français (FR) qu'en anglais (EN) sur le corpus BTC.

Évaluation sur l'échantillon du corpus de spécialité de Berger-Levrault Le tableau 4.10 présente les résultats des calculs effectués sur les entités nommées (NE) extraites par différents modèles. Voici une interprétation détaillée des résultats :

— **Pourcentage de multimots :** Ce pourcentage indique la proportion de termes extraits par chaque système qui sont composés de plusieurs mots. Les modèles GliNER multilingue V_0 et $V_{2.1}$, ainsi que l'API d'OpenAI, ont les pourcentages les plus élevés de termes multi-mots, suggérant qu'ils détectent plus de termes composés. Le graphique 4.3 présente visuellement ces résultats pour chacun des systèmes. Les résultats sont séparés en deux couleurs distinctes : bleue pour les calculs réalisés sur la liste unique des termes extraits et orange pour les calculs réalisés sur l'ensemble des termes pertinents.

— **Nombre moyen de mots par terme :** Cela représente le nombre moyen de mots par terme extrait, incluant les termes simples et multimots. L'API d'OpenAI a le nombre moyen de mots par terme le plus élevé, indiquant qu'elle extrait des termes plus longs en moyenne. Le graphique 4.4 présente visuellement ces résultats pour chacun des systèmes. Les résultats sont séparés en deux couleurs distinctes : bleue pour

les calculs réalisés sur la liste unique des termes extraits et orange pour les calculs réalisés sur l'ensemble des termes pertinents.

— **Nombre moyen de mots si multimots :** Ce calcul est similaire au précédent mais ne considère que les termes composés de plusieurs mots. L'API d'OpenAI se distingue encore une fois avec une moyenne de 3 mots pour les termes multimots, contre 2 mots pour les autres systèmes. Le graphique 4.5 présente visuellement ces résultats pour chacun des systèmes. Les résultats sont séparés en deux couleurs distinctes : bleue pour les calculs réalisés sur la liste unique des termes extraits et orange pour les calculs réalisés sur l'ensemble des termes pertinents.

— **Nombre moyen de termes par texte :** Ce chiffre indique combien de termes sont extraits en moyenne par texte pour chaque système. L'API d'OpenAI extrait le plus grand nombre de termes par texte, suivi par CamemBERT et GliNER V_0. Le graphique 4.6 présente visuellement ces résultats pour chacun des systèmes. Les résultats sont séparés en deux couleurs distinctes : bleue pour les calculs réalisés sur la liste unique des termes extraits et orange pour les calculs réalisés sur l'ensemble des termes pertinents.

L'API d'OpenAI se distingue, donc, par un pourcentage élevé de multimots, un nombre moyen élevé de mots par terme (incluant et excluant les multimots) et le plus grand nombre moyen de termes extraits par texte. Cela peut indiquer une tendance à identifier des termes plus complexes et une extraction plus exhaustive des entités. Les modèles GliNER (V_0 et $V_{2.1}$) et CamemBERT ont des performances similaires avec des pourcentages de multimots élevés et des nombres moyens de mots par terme modérés. Wikineural NER multilingue semble être plus conservateur avec un pourcentage de multimots plus faible et un nombre moyen de termes par texte plus bas.

L'analyse des résultats obtenus dans la figure 4.7 montre les scores moyens attribués aux différents modèles par les LLMs pour les différents domaines spécifiques liés à la NER. La figure 4.8, quant à elle, présente les scores de confiance (c'est à dire le pourcentage) moyens attribués aux différents modèles par les LLMs pour les différents domaines spécifiques liés à la NER. Nous allons observer chacun des résultats pour chaque domaine.

1. **Justice :** En ce qui concerne les scores compris entre 0 et 5, GliNER V_0 semble le plus performant pour ce domaine, tandis que Wikineural a des performances inférieures. En effet, les trois systèmes ayant les meilleurs scores sont : GliNER V_0 avec un score de 4.24, Gemini avec 3.84 et GliNER $V_{2.1}$ avec 3.83. En ce qui concerne le pourcentage de confiance, nous voyons une cohérence avec les scores obtenus puisque GliNER V_0 montre une performance particulièrement élevée (84,9%) tandis que Wikineural a une confiance significativement plus faible (55,5%).

2. **État civil et cimetières :** GliNER V_0 et l'APi d'OpenAI (ChatGPT) obtiennent de bons résultats, mais Wikineural est moins performant. En effet, les trois systèmes ayant les meilleurs scores sont : GliNER V_0 avec un score de 3.95, l'APi d'OpenAI (ChatGPT) avec 3.05 et GliNER $V_{2.1}$ avec 3.02. GliNER V_0 obtient une bonne confiance ici également (78,8%), tandis que Wikineural affiche une faible confiance (44,7%).

3. **Comptabilité et finances locales :** Gemini et GliNER V2.1 montrent de bonnes performances, alors que Wikineural est à nouveau moins performant. En effet, les trois systèmes ayant les meilleurs scores sont : Gemini avec un score de 3.90, l'APi d'OpenAI (ChatGPT) avec 3.73 et GliNER $V_{2.1}$ avec 3.90. Gemini (82,8%) et GliNER $V_{2.1}$ (79%) montrent une forte confiance, mais il est intéressant de noter que l'APi d'OpenAI (ChatGPT) a une confiance beaucoup plus faible. En effet, elle obtient un score de 3.73 pour une confiance de 50,5%.

Système utilisé	Pourcentage multimots		Nbre mots moyen		Nbre mots moyen si multimots		Nbre moyen termes x texte	
CamemBERT NER with dates	71,8%	56%	3	2	3	3	7	16
L'API d'OpenAI (ChatGPT)	77,9%	78,54%	4	3	4	4	10	13
Wikineural NER multilingue	63,3%	52,4 %	3	3	4	4	3	7
GliNER multilingue V_0	78,9%	59,7 %	3	2	3	3	6	17
GliNER multilingue $V_{2.1}$	79,9 %	66,6 %	3	2	3	3	3	6

TABLE 4.10 – Tableau récapitulatifs des différents calculs effectués sur les NER extraits avec les différents modèles donnés.

Légende :

⟶ **Calculs** :

Pourcentage multimots : pourcentage de multimots dans l'ensemble des termes détectés par chaque système

Nbre mots moyen : nombre moyen de mots par terme clef pour chaque système

Nbre mots moyen si multimots : nombre moyen de mots par terme clef pour en ne prenant que les multimots

Nbre moyen termes x texte : nombre moyen de termes par texte pour chaque système.

⟶ **Couleurs** :

Liste unique : Les résultats colorés représentent les résultats obtenus avec la liste unique des termes extraits.

4. **Commande publique :** Gemini se distingue par un score élevé, mais GliNER V2.1 est moins performant dans ce domaine. En effet, les trois systèmes ayant les meilleurs scores sont : Gemini avec un score de 4.19, l'APi d'OpenAI (ChatGPT) avec 3.30 et CamemBERT avec 3.38. Gemini(74,4%), CamemBERT (74,3%) et GliNER (70,9%) sont proches en termes de confiance, avec Wikineural légèrement en retrait 64,3%. Encore une fois les résultats obtenus en termes de score ne reflètent pas la confiance attribuée.

5. **Urbanisme :** Gemini obtient le meilleur score, tandis que Wikineural affiche une performance faible. En effet, les trois systèmes ayant les meilleurs scores sont : Gemini avec un score de 3.87, CamemBERT avec 3.72 et GliNER $V_{2.1}$ avec 3.35. CamemBERT(79,6%) ainsi que Gemini (78%) se démarquent avec une forte confiance, alors que GliNER V_0 montre une confiance relativement faible (61,1%).

6. **Santé :** CamemBERT, l'APi d'OpenAI (ChatGPT) et Gemini montrent de bonnes performances, mais GliNER $V_{2.1}$ est notablement inférieur. En effet, les trois systèmes ayant les meilleurs scores sont : CamemBERT avec un score de 3.56, l'APi d'OpenAI (ChatGPT) avec 3.24 et Gemini avec 3.51. Les résultats restent globalement bas. Cela peut s'expliquer par le fait que les LLMs s'attendent à avoir une terminologie spécialisée avec un domaine propre à la santé employant des termes médicaux techniques. Les LLMs s'attendent à avoir des entités allant du nom de maladies, de médicaments, des procédures médicales, organes du corps, à des termes généraux de soins de santé. En ce qui concerne le pourcentage de confiance, CamemBERT (76,2%) et Gemini (70,2%) affichent une confiance relativement forte, tandis que GliNER $V_{2.1}$

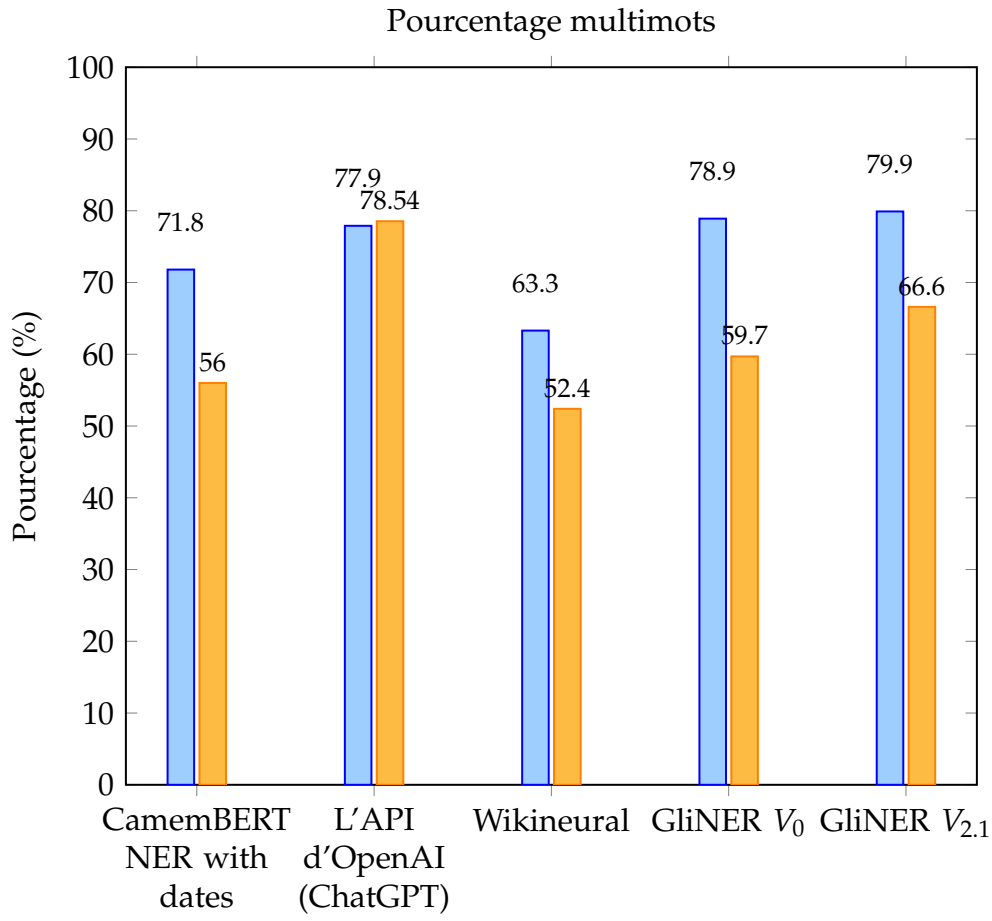

FIGURE 4.3 – Pourcentage de multimots parmi l'ensemble des termes pertinents extraits : l'ensemble ainsi que la liste unique.

Légende :

⟶ **Pourcentage multimots :** pourcentage de multimots dans l'ensemble des termes détectés par chaque système

⟶ **Couleurs** :

Liste unique : Les résultats colorés représentent les résultats obtenus avec la liste unique des termes extraits.

Ensemble des termes pertinents : Les résultats colorés en orangé représentent les résultats obtenus pour l'ensemble des termes pertinents extraits.

(24%5) montre une très faible confiance. Cela va en accord avec les scores obtenus.

7. **Ressources humaines Territoriales :** Tous les modèles, sauf Wikineural, montrent des performances élevées, avec Gemini en tête avec un score de 4.56. Gemini (88,2%)et les deux versions de GliNER (V_0 : 84,8% $V_{2.1}$: 84,5%) montrent une très forte confiance, tandis que Wikineural (56%) reste à un niveau inférieur.
8. **Élections :** GliNER $V_{2.1}$ et V_0 sont en tête, tandis que Wikineural reste le moins performant. l'APi d'OpenAI (ChatGPT) (76,8%) et GliNER V_0 (76,8%) montrent une bonne confiance, tandis que Wikineural (50,3%) affiche à nouveau une faible confiance. Les scores obtenus ainsi que les taux de confiance sont donc en accord.

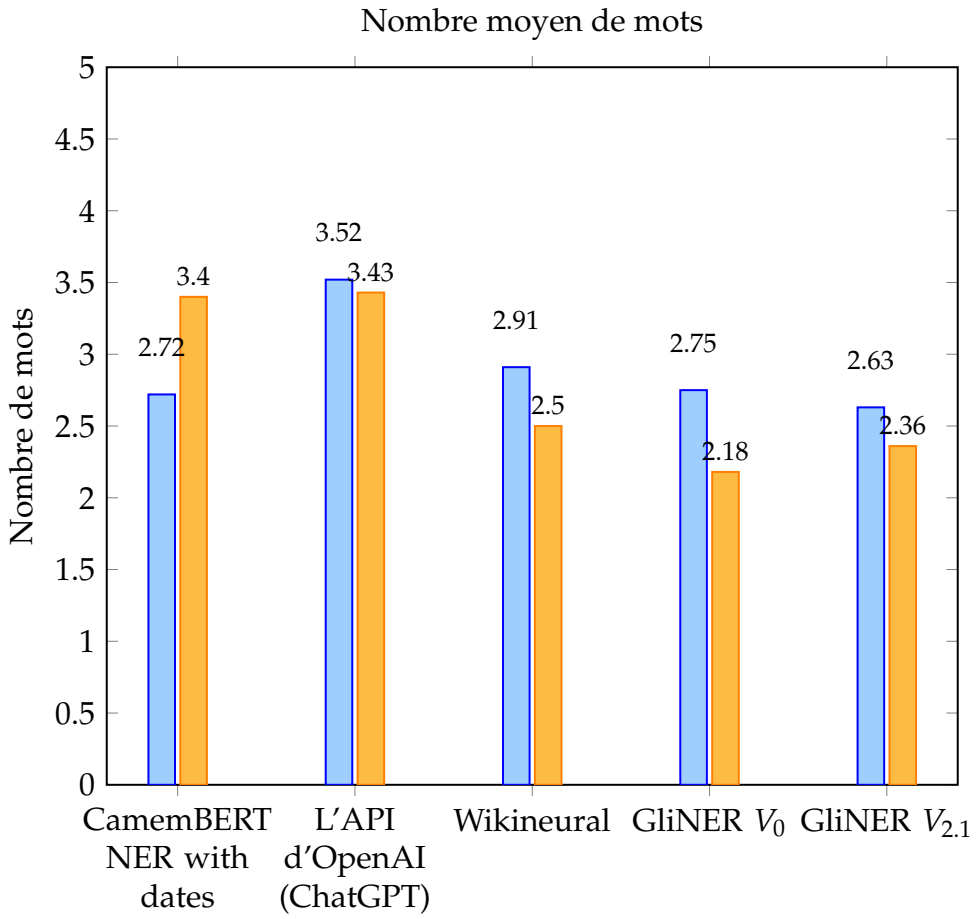

FIGURE 4.4 – Nombre moyen de mots contenus dans les termes extraits : calcul moyen pour la liste unique des termes extraits ainsi que l'ensemble des termes extraits.

Légende :

⟶ **Nbre mots moyen :** nombre moyen de mots par terme clef pour chaque système

⟶ **Couleurs** :

Liste unique : Les résultats colorés représentent les résultats obtenus avec la liste unique des termes extraits.

Ensemble des termes pertinents : Les résultats colorés en orangé représentent les résultats obtenus pour l'ensemble des termes pertinents extraits.

L'analyse montre que certains modèles comme GliNER V_0 et $V_{2.1}$, ainsi que Gemini, sont les plus performants dans la majorité des domaines étudiés. En revanche, Wikineural semble systématiquement avoir des scores plus bas, indiquant qu'il est moins adapté ou performant pour ces tâches spécifiques de reconnaissance des entités nommées. En effet, GliNER V_0 et $V_{2.1}$ ont montré de très bonnes performances dans plusieurs domaines, en particulier les ressources humaines territoriales, la justice et les élections. Gemini a été performant dans presque tous les domaines, particulièrement en ressources humaines territoriales, commande publique et justice. En ce qui concerne CamemBERT et l'APi d'OpenAI

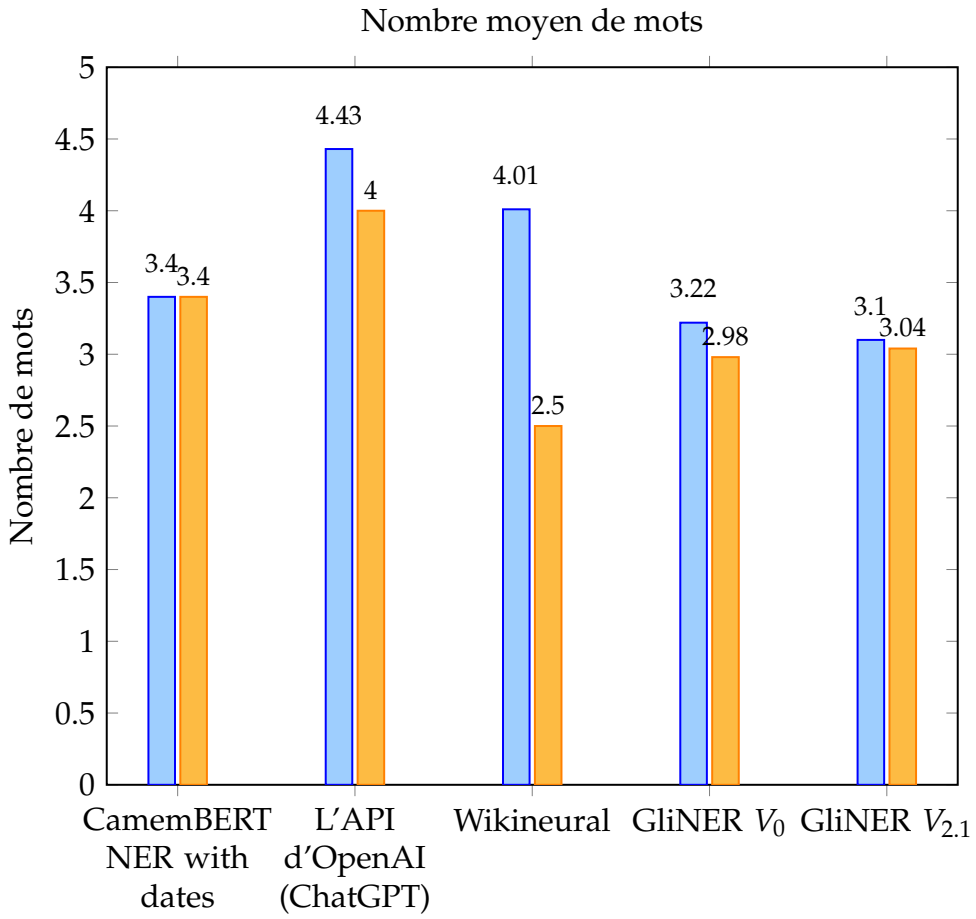

FIGURE 4.5 – Nombre moyen de mots contenus dans les termes extraits : calcul moyen pour la liste des multimots unique ainsi que l'ensemble des multimots.

Légende :

⟶ **Nbre mots moyen si multimots :** nombre moyen de mots par terme clef pour en ne prenant que les multimots

⟶ **Couleurs** :

Liste unique : Les résultats colorés représentent les résultats obtenus avec la liste unique des termes extraits.

Ensemble des termes pertinents : Les résultats colorés en orangé représentent les résultats obtenus pour l'ensemble des termes pertinents extraits.

(ChatGPT), ils affichent des performances stables et relativement élevées, surtout en justice et santé. Pour finir, Wikineural est, quant à lui, globalement moins performant dans tous les domaines comparés aux autres modèles. L'analyse des scores de confiance, quant à eux, montre que certains modèles, en particulier Gemini et les versions de GliNER, ont une forte confiance dans la reconnaissance des entités nommées pour la plupart des domaines étudiés. CamemBERT est également performant dans plusieurs domaines spécifiques. En revanche, Wikineural a systématiquement des scores de confiance plus bas, ce qui suggère qu'il pourrait être moins fiable pour ces tâches. Pour améliorer les performances dans des domaines spécifiques comme la santé et l'urbanisme, il serait utile de continuer à affiner les

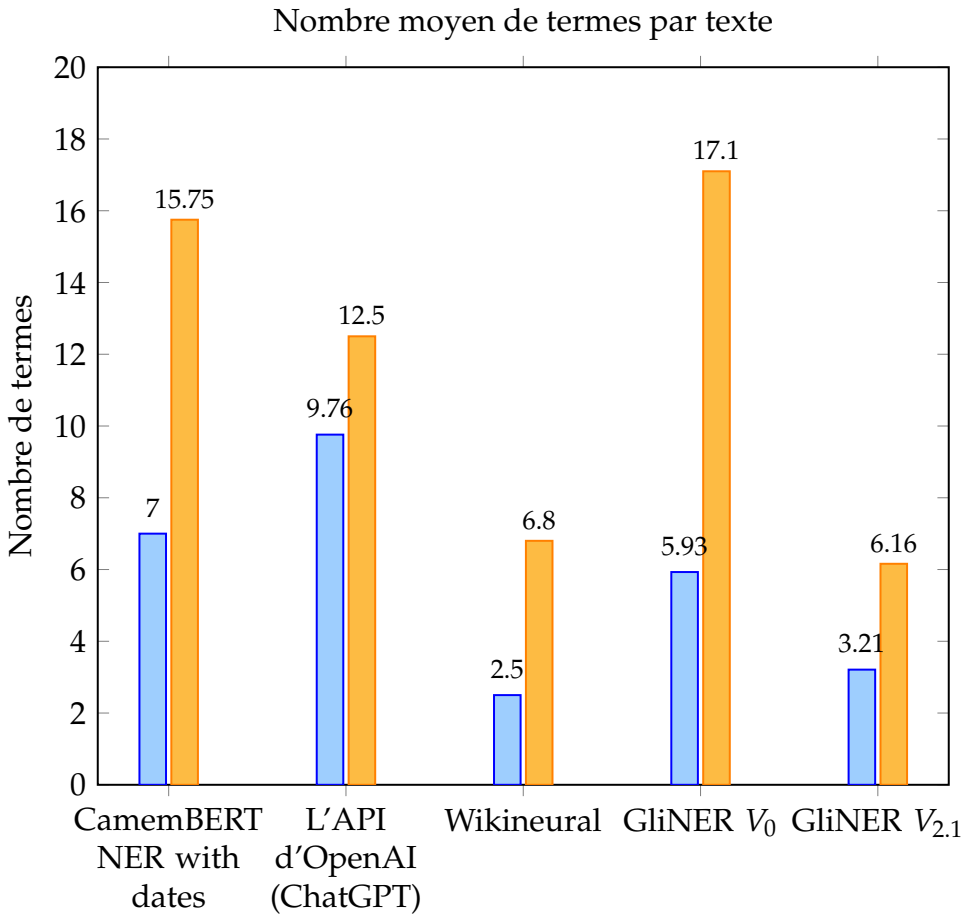

FIGURE 4.6 – Nombre moyen de termes par texte

Légende :

⟶ **Nbre moyen termes x texte :** nombre moyen de termes par texte pour chaque système.

⟶ **Couleurs** :

Liste unique : Les résultats colorés représentent les résultats obtenus avec la liste unique des termes extraits.

Ensemble des termes pertinents : Les résultats colorés en orangé représentent les résultats obtenus pour l'ensemble des termes pertinents extraits.

modèles avec des corpus spécialisés et de haute qualité. Nous constatons donc une homogénéité dans les résultats obtenus entre les scores et les scores de confiance. En effet, GliNER V_0 et $V_{2.1}$ montrent une forte confiance dans plusieurs domaines, particulièrement en ressources humaines territoriales et justice. Gemini, quant à lui, est performant dans presque tous les domaines, avec des scores de confiance élevés, particulièrement en ressources humaines territoriales. CamemBERT affiche une confiance relativement élevée, surtout en urbanisme et santé. l'APi d'OpenAI (ChatGPT) a obtenu une variabilité de confiance avec des scores élevés dans certains domaines (élections), mais également faible comme en comptabilité et finances locales. Pour finir, Wikineural est globalement moins performant en termes de confiance dans tous les domaines comparés aux autres modèles.

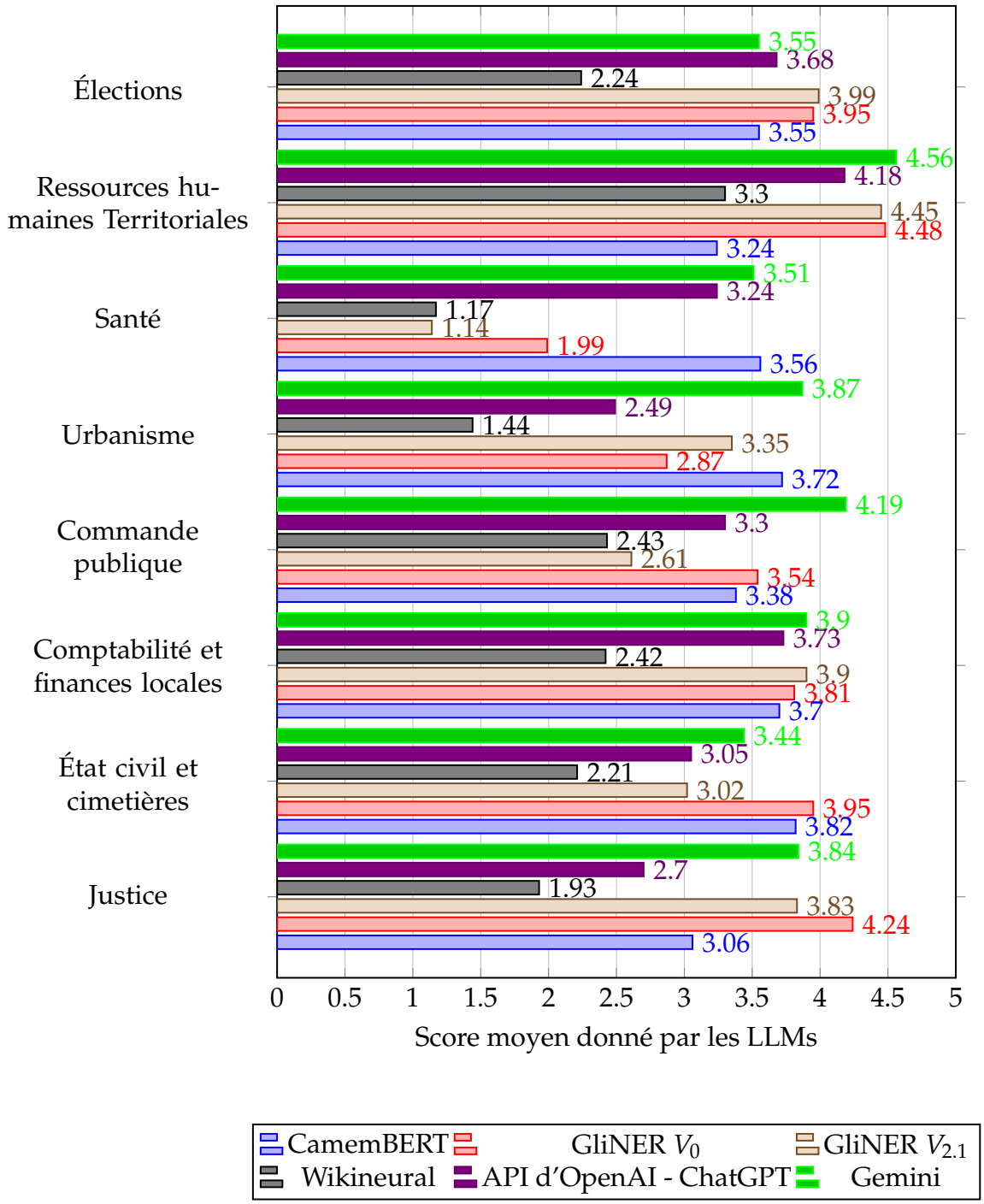

FIGURE 4.7 – Score moyen attribué par les LLMs pour chacun des domaines pour le NER

Résultats pour le système de KPE

Évaluation sur les jeux de données standard Les résultats présentés dans le tableau 4.11 montrent les performances de deux modèles, l'API d'OpenAI pour la génération de texte (ChatGPT) et Gemini, pour le KPE à partir de trois sources différentes : Wikinews-fr-100,

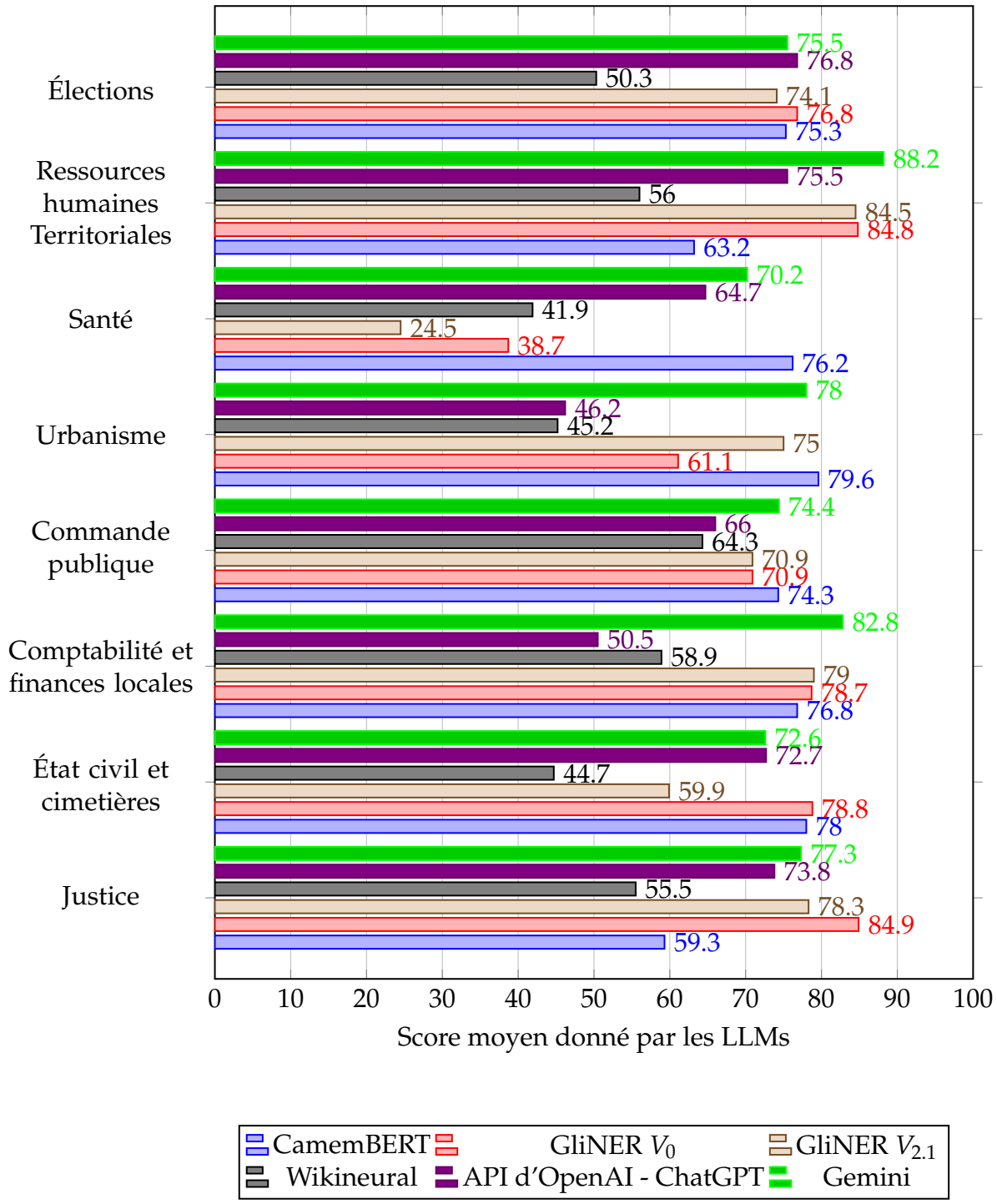

FIGURE 4.8 – Score de confiance moyen attribué par les LLMs pour chacun des domaines pour le NER

Taln-archives-fr et Termith-eval-fr.

Pour le modèle l'API d'OpenAI (ChatGPT), les performances varient selon la source de données. Pour l'extraction de termes pertinents à partir de Wikinews-fr, le modèle a atteint un score de 92,7%. Pour le modèle Gemini, les performances sont également différentes selon la source de données. Pour l'extraction de termes pertinents son meilleur résultat a été obtenu sur Termith-eval-fr avec un score de 81%. Plus généralement, nous constatons que le modèle de l'API d'OpenAI (ChatGPT) a des performances globalement supérieures à celles du modèle Gemini, quelle que soit la source de données. De plus, les performances varient

Modèle utilisé	Wikinews-fr-100 prompt keywords extraction	Taln-archives-fr prompt keywords extraction	Termith-eval-fr prompt keywords extraction
API d'OpenAI (ChatGPT)	92,7 %	72,5 %	85 %
Gemini	53,6 %	76,1 %	81 %

TABLE 4.11 – Résultats de différents systèmes pour l'extraction de termes pertinents (KPE).

significativement en fonction de la source de données utilisée pour l'extraction des termes pertinents.

Les performances de l'API d'OpenAI (ChatGPT) pour le KPE à partir de l'ensemble de données TALN-archives peuvent être relativement faibles en raison de la nature spécifique et complexe des articles scientifiques qui le composent. En effet, les articles scientifiques sont souvent rédigés dans un langage technique et spécialisé, ce qui peut poser des défis d'interprétation pour le modèle. De plus, le modèle pourrait manquer de données d'entraînement spécifiquement adaptées à ce domaine, ce qui limite sa capacité à extraire des termes pertinents de manière précise. Par conséquent, les performances peuvent être affectées par des difficultés à comprendre le contexte et la structure des articles, ainsi que par un niveau de bruit plus élevé dans les données. Une adaptation spécifique du modèle ou un entraînement supplémentaire sur des données scientifiques pourraient contribuer à améliorer ses performances dans ce contexte particulier. Il est intéressant de noter ce point étant donné que les données de Berger-Levrault sont spécifiques. Bien que les résultats de Gemini pour le KPE soient légèrement supérieurs à ceux de l'API d'OpenAI (ChatGPT) pour TALN-archives, il est important de considérer le compromis entre la performance brute et la complexité du post-traitement. En effet, il ne s'agit là que d'un jeu de données parmi les trois évalués bien que ce soit le jeu de données dit de spécialité, TermITH-Eval l'est également et l'API d'OpenAI (ChatGPT) l'a surpassé. De plus, Gemini a nécessité un lourd post-traitement trop complexe et qui peut entraîner plusieurs problèmes. La mise en œuvre et l'optimisation d'un post-traitement sophistiqué peuvent nécessiter des ressources supplémentaires en termes de temps et d'efforts de développement. Par conséquent, malgré ses performances légèrement supérieures pour TALN-archives, choisir Gemini n'est pas être judicieux en confrontant le post-traitement complexe face au peu de pourcentage gagné.

Évaluation sur l'échantillon du corpus de spécialité de Berger-Levrault Le graphique 4.9 compare le nombre moyen de mots pour différents types de termes. Il permet de montrer que les deux systèmes, à savoir l'API OpenAI (ChatGPT) et Gemini, extraient des termes avec une longueur moyenne similaires. Il n'y a pas de différence significative entre l'API OpenAI (ChatGPT) et Gemini en termes de longueur moyenne des termes extraits et ce pour chaque catégorie. Le graphique 4.10 compare les systèmes par rapport aux pourcentages de multimots et leur nombre moyen de termes par texte. Nous observons cette fois-ci, une différence notable entre les deux systèmes. En effet, en ce qui concerne le pourcentage de multi-mots, Gemini montre un pourcentage plus élevé de multi-mots par rapport à l'API OpenAI (ChatGPT), indiquant une capacité potentiellement supérieure à extraire des expressions complexes ou des phrases ou bien extrait « trop »d'informations dans les termes, c'est-à-dire du bruit. Cette information n'est pas vérifiable avec ces pourcentages. En ce qui concerne le nombre moyen de termes par texte, Gemini a tendance à extraire un nombre moyen plus faible de termes par texte que l'API OpenAI (ChatGPT), suggérant une extrac-

tion plus sélective ou une capacité différente à identifier les termes pertinents. Dans notre cas d'usage, il semble préférable d'extraire un plus grand nombre d'informations du fait de la richesse d'un corpus de spécialité.

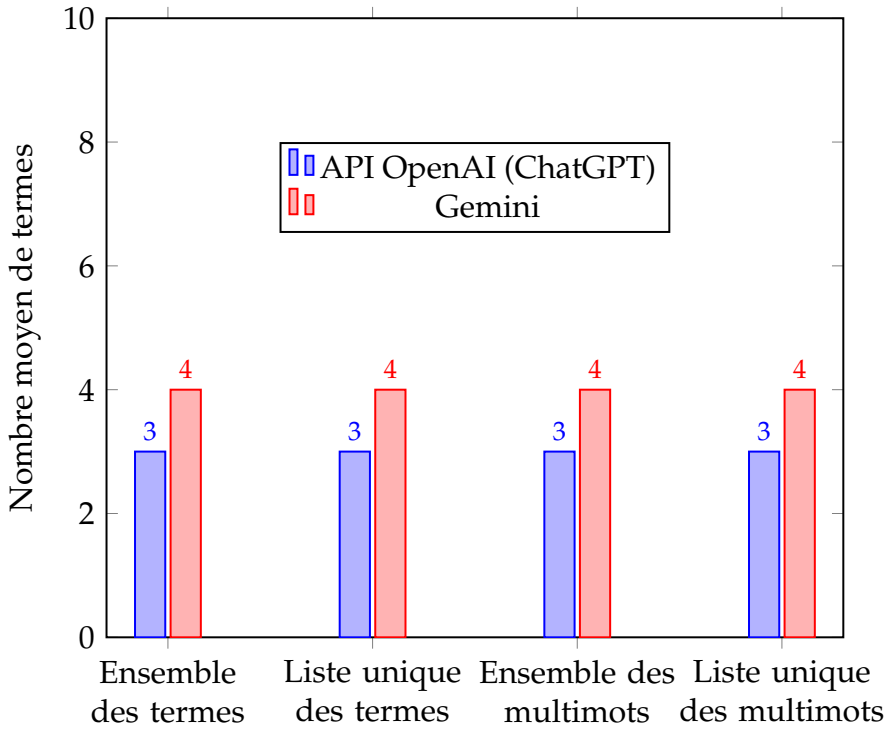

FIGURE 4.9 – Calcul du nombre moyen de mots des termes pertinents extraits par l'API d'OpenAI et Gemini.

Légende :

⟶ **Calculs** :

Nbre mots moyen : nombre moyen de mots par terme clef pour chaque système

Nbre mots moyen si multimots : nombre moyen de mots par terme clef en ne prenant que les multimots

Calculs réalisés sur l'ensemble des termes pertinents extraits puis sur la liste unique des termes pertinents.

Au vu des résultats sur les jeux de données standards, il est donc préférable d'utiliser l'API OpenAI (ChatGPT) car il semble plus pertinent du fait des résultats et fournit donc un nombre plus important de termes avec un nombre de mots moyen équivalent à Gemini.

L'analyse des scores moyens attribués par les LLMs pour chaque domaine dans la figure 4.11 permet de tirer plusieurs conclusions sur les performances des modèles de l'API d'OpenAI (ChatGPT) et Gemini dans la tâche de KPE. Les scores de l'API d'OpenAI (ChatGPT) varient de 3.31 (Justice) à 4.99 (Élections). Les scores de Gemini, eux, varient de 3.62 (État civil et cimetières) à 4.88 (Ressources humaines Territoriales). Gemini a des scores plus élevés que l'API d'OpenAI (ChatGPT) dans 5 des 8 domaines (Justice, Urbanisme, Santé, Ressources humaines Territoriales, et Élections). L'API d'OpenAI (ChatGPT) surpasse Gemini dans 3 domaines (État civil et cimetières, Comptabilité et finances locales, et Commande

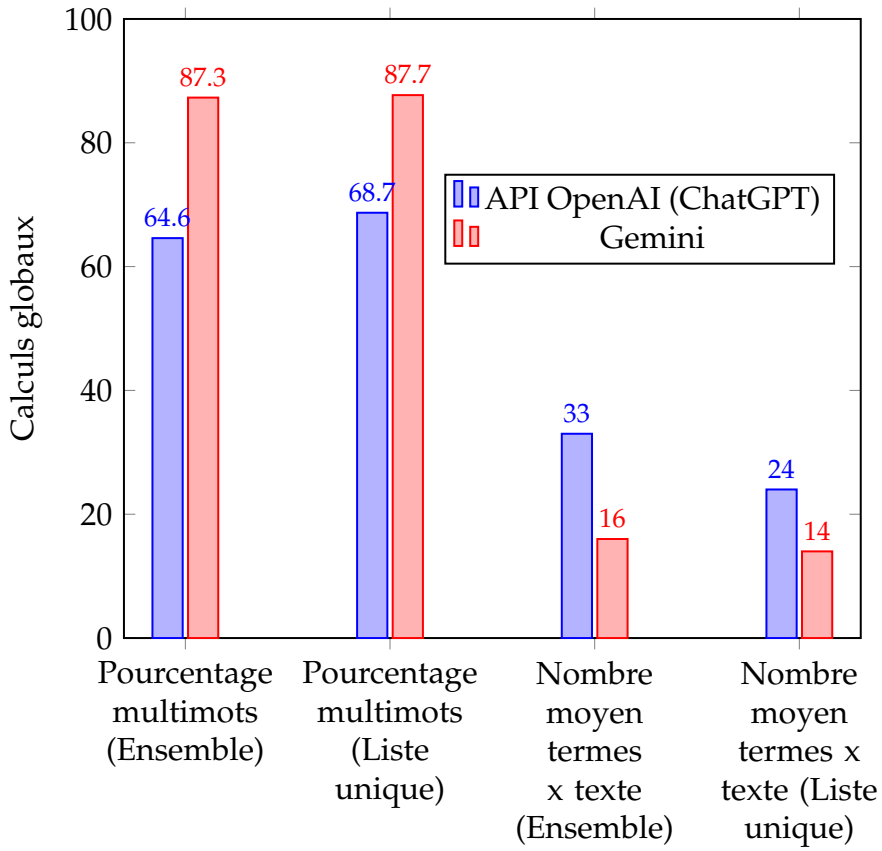

FIGURE 4.10 – Statistiques sur les termes pertinents extraits par l'API d'OpenAI et Gemini.

Légende :

⟶ **Calculs** :

Pourcentage multimots : pourcentage de multimots dans l'ensemble des termes détectés par chaque système

Nbre moyen termes x texte : nombre moyen de termes par texte pour chaque système.

Calculs réalisés sur l'ensemble des termes pertinents extraits puis sur la liste unique des termes pertinents.

publique). L'analyse des scores de confiance moyens attribués par les LLMs à l'API d'OpenAI (ChatGPT) et Gemini pour chacun des domaines dans la tâche de KPE révèle plusieurs points intéressants comme nous pouvons le constater sur la figure 4.12. Les scores de l'API d'OpenAI (ChatGPT) varient de 74,4 % (Urbanisme) à 95 % (Élections). Les scores de Gemini varient de 77,7 % (État civil et cimetières) à 88,8 % (Ressources humaines Territoriales). L'API d'OpenAI (ChatGPT) a une confiance globalement plus élevée dans les domaines de la justice, état civil et cimetières, comptabilité et finances locales, commande publique, santé, et surtout les élections. Gemini montre une meilleure performance dans les domaines de l'urbanisme et des ressources humaines territoriales. Nous allons détailler les scores et les scores de confiances par domaine.

1. **Justice :** Gemini présente une meilleure performance dans ce domaine avec des scores

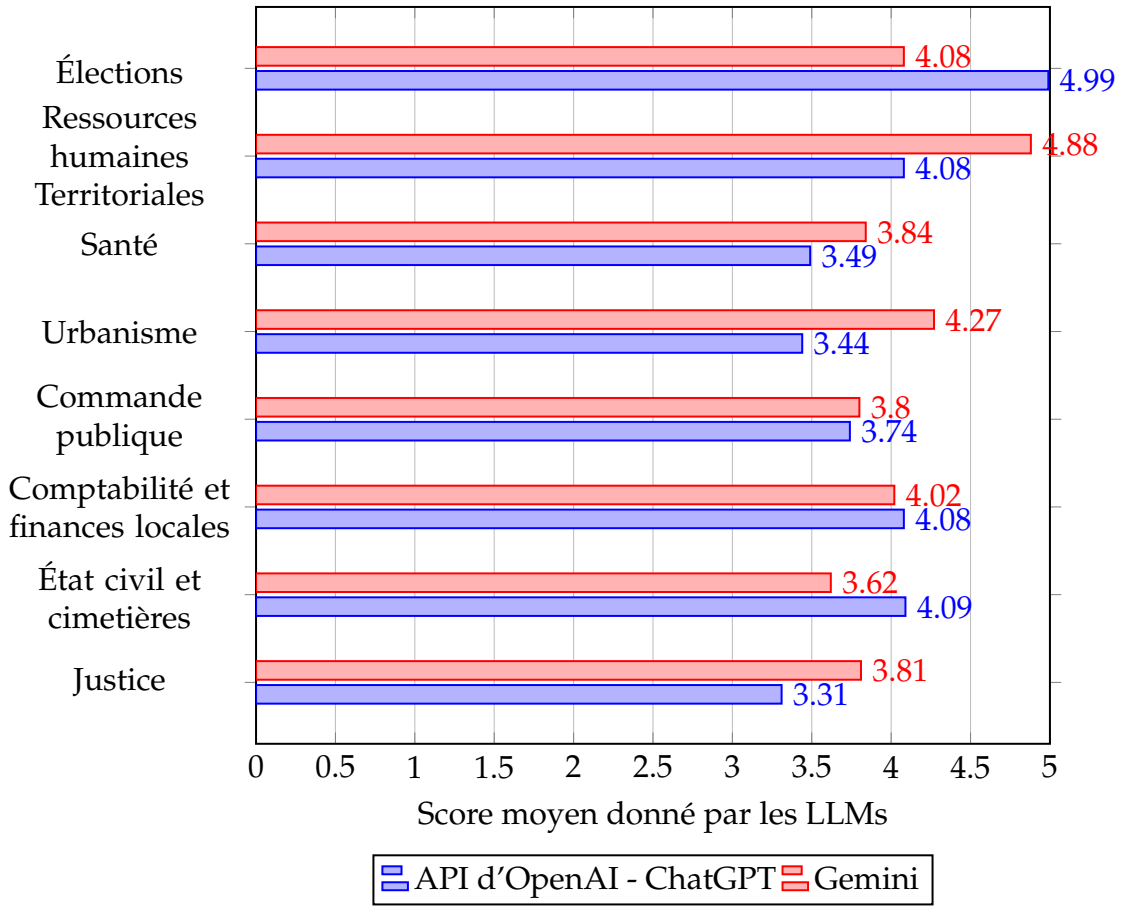

FIGURE 4.11 – Score de pertinence moyen attribué par les LLMs pour chacun des domaines pour le KPE. Il s'agit d'associer un score entre 1 et 5 à un terme en fonction de sa pertinence dans un domaine donné.

moyens (3.81) et de confiance élevés (78,7%) même si elle reste plus faible que l'API d'OpenAI (ChatGPT) (83,1%).

2. **État civil et cimetières :** l'API d'OpenAI (ChatGPT) a des scores élevés pour les deux métriques (4.09 et 86,1%), suggérant une fiabilité et une performance supérieure dans ce domaine.
3. **Comptabilité et finances locales :** Les deux modèles ont des scores élevés et proches, mais l'API d'OpenAI (ChatGPT) montre une confiance légèrement plus élevée, indiquant une performance légèrement meilleure.
4. **Commande publique :** Les scores sont proches, avec Gemini ayant un score moyen légèrement supérieur et l'API d'OpenAI (ChatGPT) un score de confiance plus élevé, indiquant des performances comparables.
5. **Urbanisme :** Gemini présente des scores moyens (4.27) et de confiance (85%3) significativement plus élevés, suggérant une meilleure extraction des termes clés dans ce domaine.
6. **Santé :** Les scores moyens et de confiance sont comparables, avec Gemini légèrement en tête pour le score moyen (3.84) et l'API d'OpenAI (ChatGPT) pour le score de

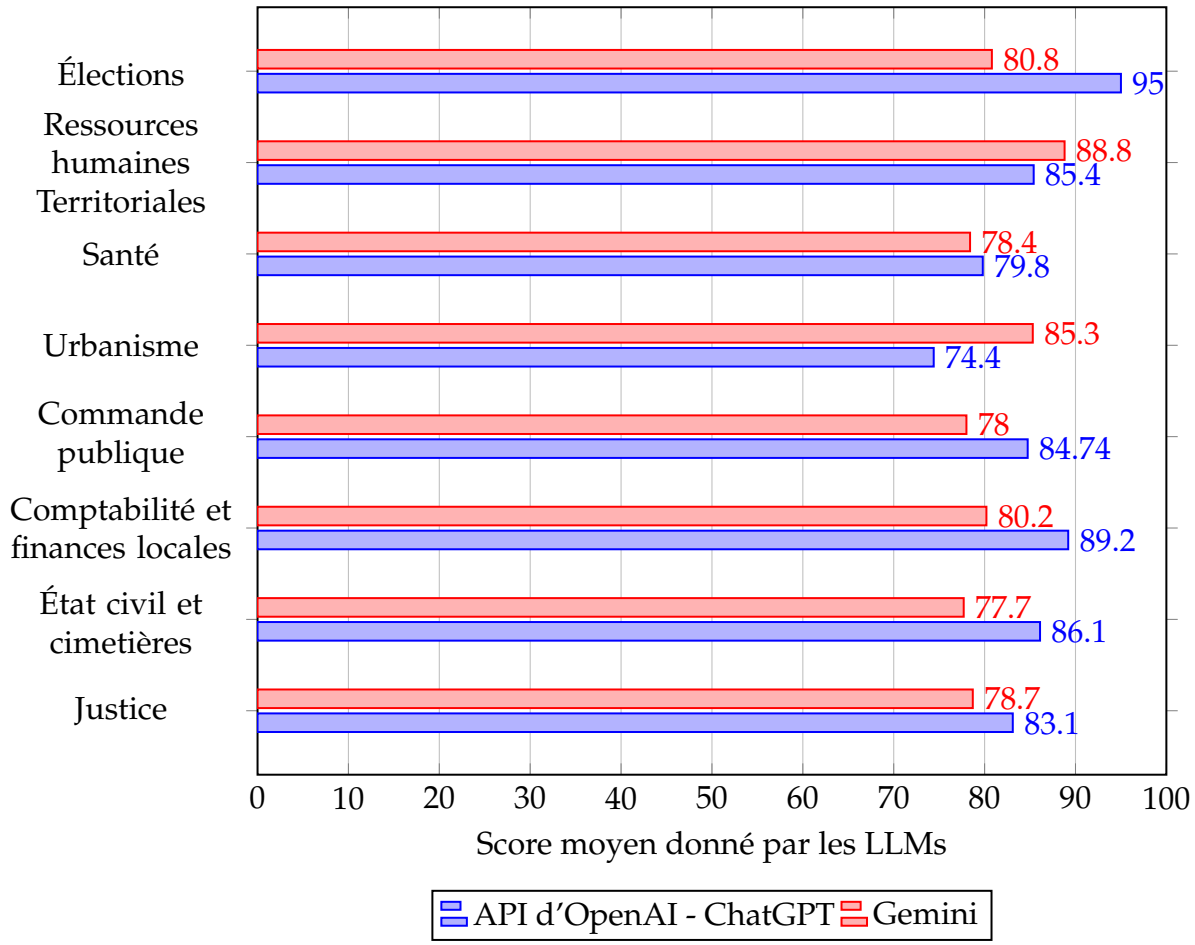

FIGURE 4.12 – Score de confiance moyen attribué par les LLMs pour chacun des domaines pour le KPE

confiance (79,8%), indiquant des performances similaires.

7. **Ressources humaines Territoriales :** Gemini a des scores moyens (4.88) et de confiance (88,8%) plus élevés, indiquant une forte performance et fiabilité dans ce domaine.
8. **Élections :** l'API d'OpenAI (ChatGPT) montre des scores très élevés pour les deux métriques (4.99 et 95%), suggérant une très forte fiabilité et performance dans l'extraction des phrases clés pour les élections.

Les résultats obtenus dans les 8 domaines montrent une cohérence entre les scores moyens et les scores de confiance, indiquant des domaines où les modèles sont à la fois précis et fiables. Les deux modèles se sont distingués de part et d'autre. C'est pourquoi, les scores élevés dans des domaines spécifiques indiquent que chaque modèle a des points forts distincts, ce qui pourrait être exploité pour des applications spécifiques où un modèle peut être préféré à l'autre.

Résultats pour le système d'extraction d'entités juridiques

Comme expliqué précédemment, l'évaluation ne peut être réalisée que sur l'échantillon du corpus de spécialité de Berger-Levrault car les REGEX utilisées ont été créées à la main à partir d'exemple de textes du corpus. De plus, il s'agit d'une tâche spécifique et il n'existe pas de corpus annoté pour l'extraction d'entités juridiques en français.

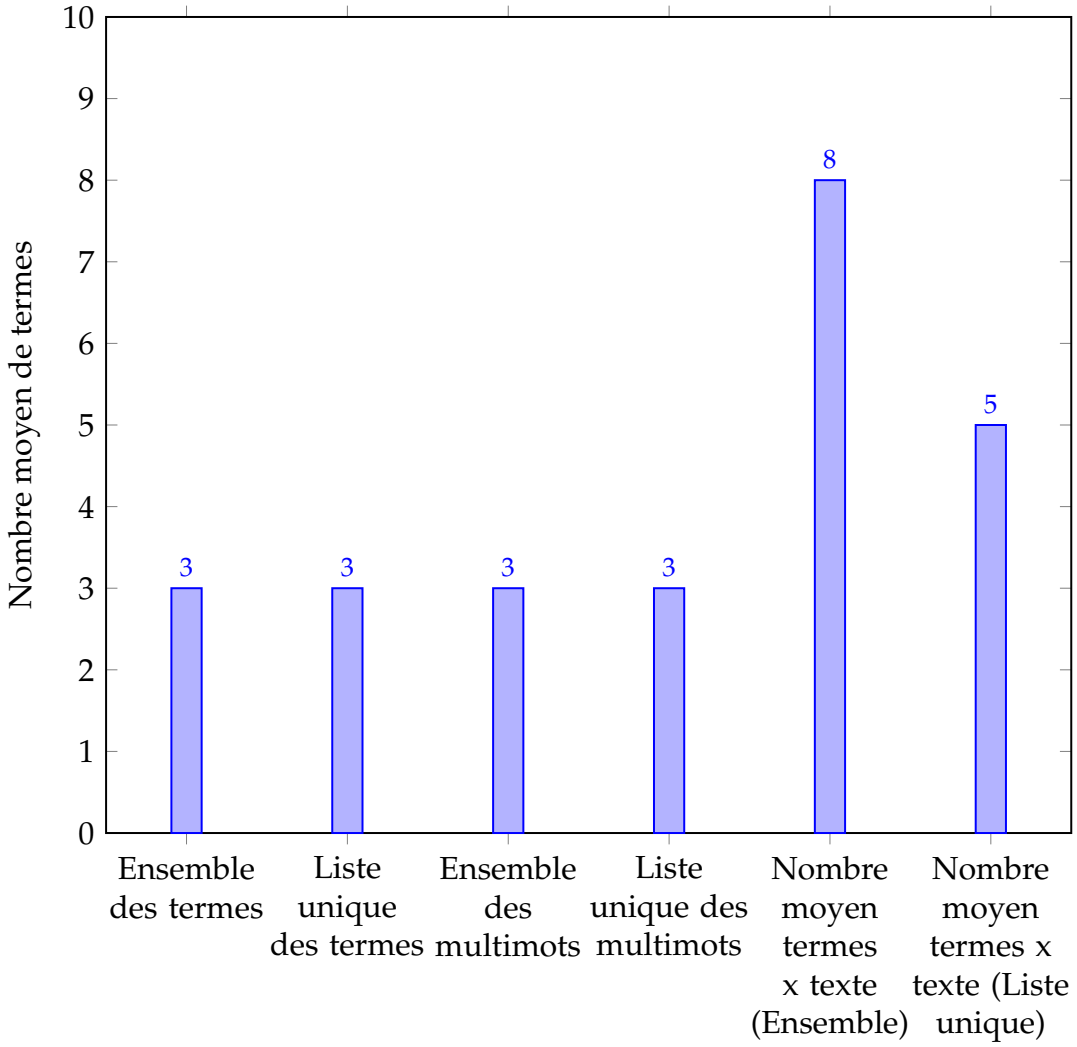

FIGURE 4.13 – Calcul du nombre moyen de mots des entités juridiques extraites par REGEX.

Légende :

⟶ **Calculs** :

Nbre mots moyen : nombre moyen de mots par terme clef pour chaque système

Nbre mots moyen si multimots : nombre moyen de mots par terme clef pour en ne prenant que les multimots

Nbre moyen termes x texte : nombre moyen de termes par texte pour chaque système.

Calculs réalisés sur l'ensemble des termes pertinents extraits puis sur la liste unique des termes pertinents.

Statistiques sur les entités extraites Le graphique présenté sur la figure 4.13 utilise un format de barres horizontales pour présenter le nombre moyen de termes extraits par les REGEX dans différents contextes : pour l'ensemble des termes, pour la liste unique des en-

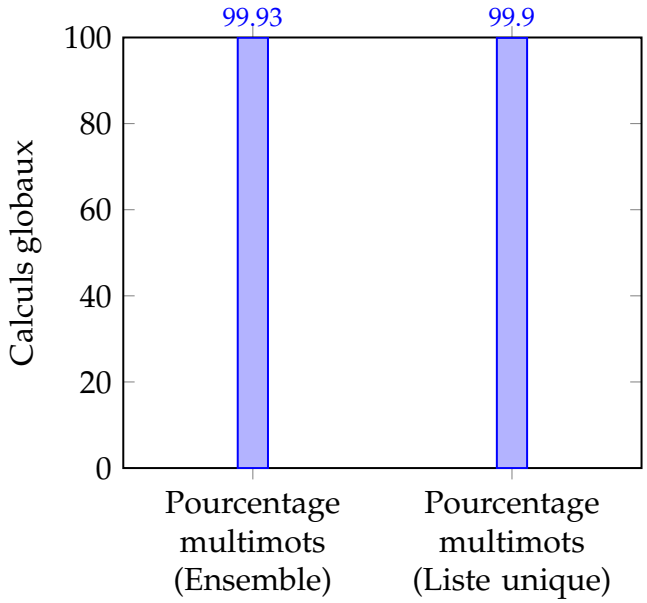

FIGURE 4.14 – Statistiques sur les entités juridiques extraites par REGEX.
Légende :
⟶ **Calculs** :
Pourcentage multimots : pourcentage de multimots dans l'ensemble des termes détectés par chaque système
Calculs réalisés sur l'ensemble des termes pertinents extraits puis sur la liste unique des termes pertinents.

tités extraites, pour l'ensemble des multimots extraits et la liste unique des multimots. Les REGEX semblent extraire en moyenne 3 termes ou multimots, que ce soit dans l'ensemble total ou dans la liste unique des entités juridiques. Cela suggère une cohérence dans la capacité des REGEX à identifier des entités juridiques à partir de textes variés, en maintenant un niveau de détail. Le graphique présenté sur la figure 4.14 utilise également des barres horizontales pour présenter différents calculs globaux réalisés sur les entités juridiques extraites par les REGEX : le pourcentage de multimots pour l'ensemble des entités extraites, le pourcentage de multimots pour la liste unique des entités extraites, le nombre moyen de termes par texte pour l'ensemble des entités extraites ainsi que le nombre moyen de termes par texte pour la liste unique des entités extraites. Lorsque nous nous intéressons au pourcentage de multimots, les REGEX identifient principalement des multimots dans les entités juridiques extraites, avec un pourcentage élevé dépassant les 99%. Cela indique une capacité robuste à détecter des entités composées de plusieurs mots, ce qui est souvent crucial dans le domaine juridique pour identifier des concepts complexes comme « décret n°2019-423 », par exemple. Lorsque nous nous intéressons au nombre moyen de termes par texte, le système extrait en moyenne 8 termes lorsque tous les termes pertinents sont considérés, et 5 termes lorsqu'on ne considère que les termes uniques. Cela suggère une capacité à extraire un nombre substantiel de termes par document ou texte, lorsque nous comparons avec le nombre moyen de termes extraits pour le NER et le KPE précédemment explicité.

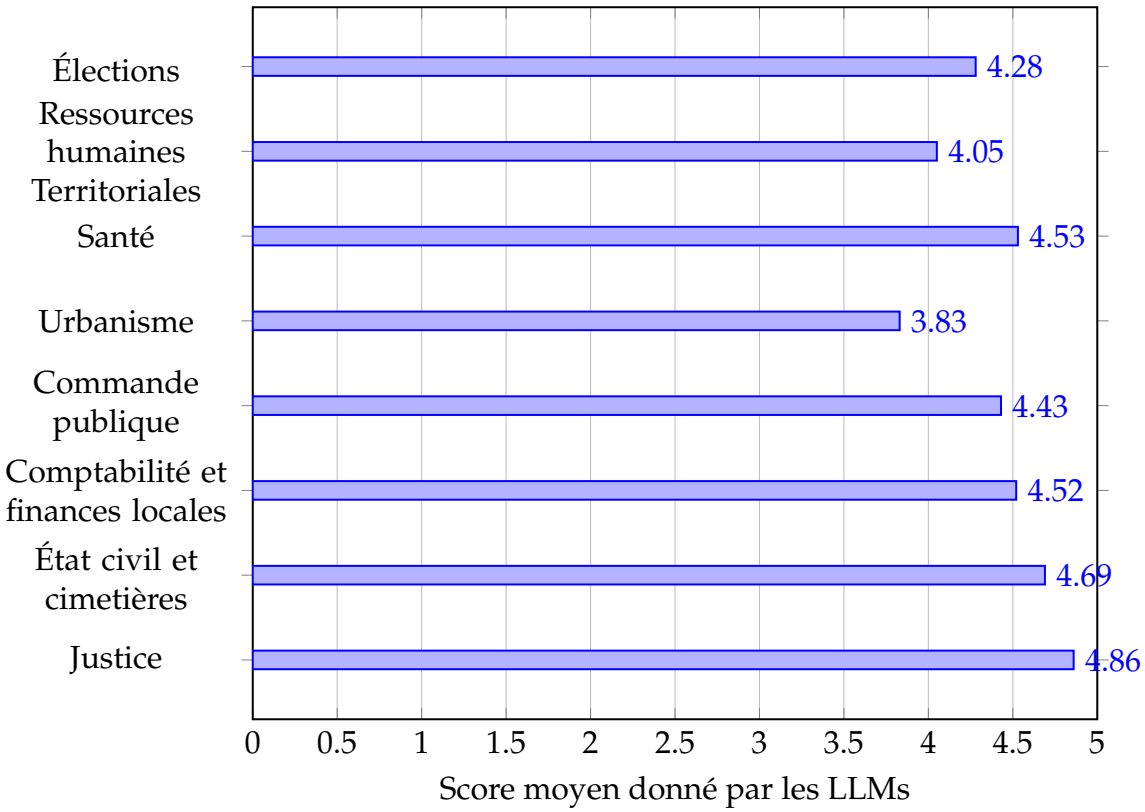

FIGURE 4.15 – Score moyen attribué par les LLMs pour chacun des domaines pour le système des entités juridiques

Évaluation par LLMs Le graphique présenté sur la figure 4.15 montre le score moyen attribué par les LLMs à notre système d'extraction d'entités juridiques appliqué sur différents domaines spécifiques. Le système basé sur l'utilisation de REGEX montre une performance généralement élevée dans l'extraction d'entités juridiques à travers une variété de domaines. Les scores varient entre environ 3.83 et 4.86, reflétant une grande précision et une bonne fiabilité globale. Les domaines tels que la justice, l'état civil, la comptabilité et la santé se distinguent avec des scores particulièrement élevés. Bien que tous les domaines montrent des scores élevés, il y a des variations qui peuvent être dues à la complexité des contextes spécifiques et à la diversité des entités extraites. En conclusion, ce graphique montre que les REGEX sont efficaces et fiables dans l'extraction d'entités juridiques à travers l'ensemble des domaines spécifiques du juridique.

Le graphique présenté sur la figure 4.16 est un graphique à barres horizontales qui montre le score moyen de confiance en pourcentage attribué par les LLMs à notre système d'extraction d'entités juridiques appliqué sur différents domaines spécifiques. Globalement, le système basé sur l'utilisation de REGEX semble performant dans l'extraction d'entités juridiques pour la plupart des domaines évalués. Les scores varient entre environ 78,8 % et 97,6 %, indiquant des niveaux de confiance divers mais généralement élevés. Les domaines tels que la *justice*, *l'état civil*, la *santé* et les *élections* montrent des scores particulièrement élevés. En conclusion, ce graphique met en lumière la capacité des REGEX à extraire des entités juridiques avec des niveaux de confiance élevés dans une gamme variée de domaines, ce qui est crucial pour notre application.

En conclusion, les REGEX se révèlent être une méthode efficace et pertinente pour l'extraction d'entités juridiques, offrant une précision stable et une capacité robuste à traiter des documents du document juridique et à extraire les termes composés de plusieurs mots.

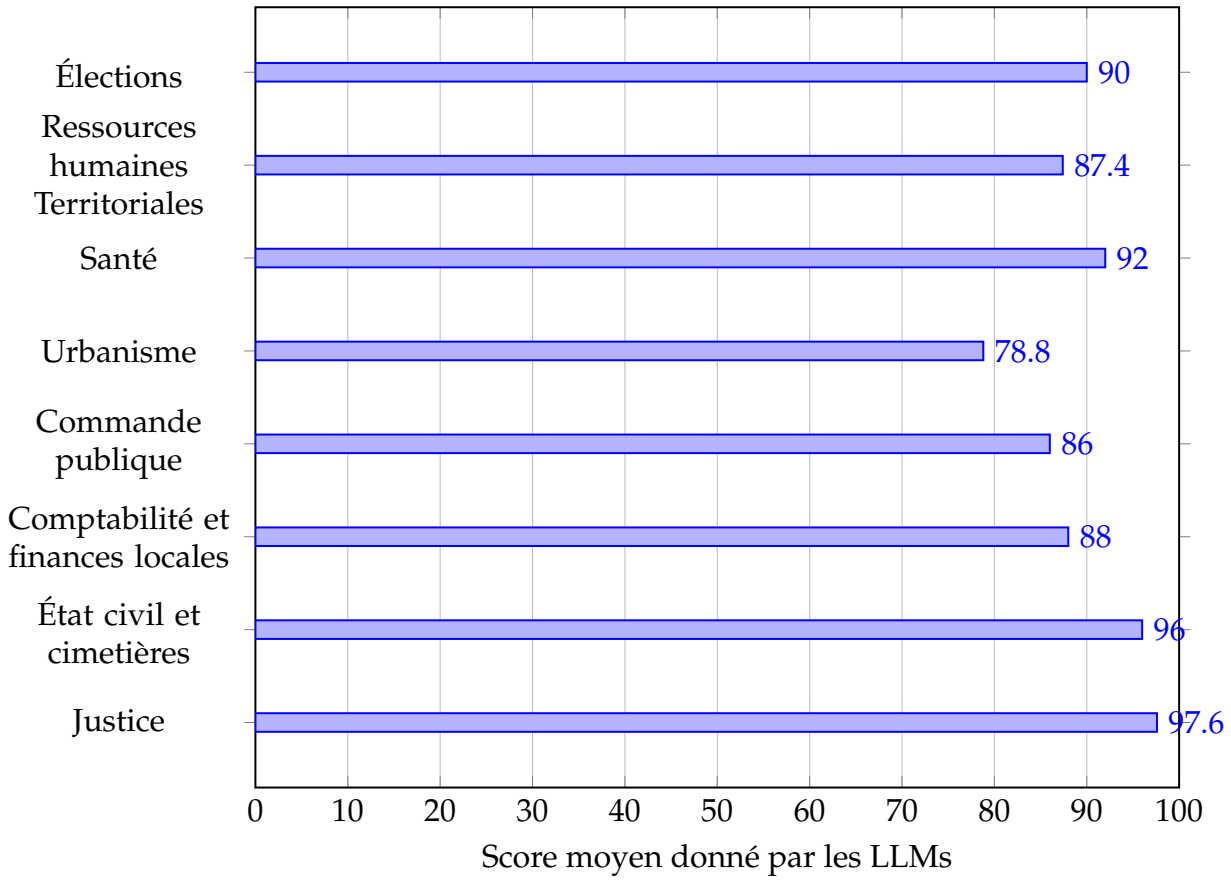

FIGURE 4.16 – Score de confiance moyen attribué par les LLMs pour chacun des domaines pour le système d'extraction d'entités juridiques

4.3.3 Discussion critique sur les Résultats et Identification des Limites

Dans cette section, nous nous sommes intéressé uniquement à *InfoGlean KeyTerms* puisque nous avons vu que les annotations effectuées par des experts sont variables en termes de qualité et de consistance. En effet, notamment l'absence de guide standardisé d'annotations peut introduire des incohérences qui affectent l'évaluation des performances des modèles. Cependant, avant de discuter autour de notre système, nous pouvons envisager d'introduire des guides d'annotations standardisés pour les experts afin d'assurer une meilleure consistance et qualité des annotations. Cela améliorerait la qualité des données d'entraînement et, par conséquent, la performance des modèles.

InfoGlean KeyTerms

Les résultats obtenus à partir de l'extraction terminologique ont été évalués selon plusieurs critères et à l'aide de différents modèles. Voici les points clés des résultats observés :

— **Pourcentage de Multimots** : Les termes extraits ont été évalués en fonction de leur composition en termes de nombre de mots. Les modèles GliNER multilingue V_0 et $V_{2.1}$ ainsi que l'API d'OpenAI ont montré les pourcentages les plus élevés de termes multimots, suggérant une capacité supérieure à identifier des termes composés de plusieurs mots.

— **Nombre Moyen de Mots par Terme** : Ce critère mesure la longueur des termes extraits. L'API d'OpenAI s'est distinguée en extrayant des termes plus longs en moyenne que les autres systèmes, ce qui pourrait indiquer une meilleure capture de la com-

plexité terminologique.

— **Nombre Moyen de Mots si Multimots** : En se concentrant uniquement sur les termes composés de plusieurs mots, l'API d'OpenAI a également montré une longueur moyenne plus élevée de 3 mots par terme, contre 2 mots pour les autres systèmes. Toutefois, nous constatons donc des résultats proches et similaires.
— **Nombre Moyen de Termes par Texte** : Cette métrique évalue la densité de termes extraits par texte. Les résultats ont montré une variation importante entre les systèmes, avec certains extrayant plus de termes par texte que d'autres.
— **Performance par Domaine** : Les résultats ont été également segmentés par domaines spécifiques, tels que l'urbanisme et la santé. Des différences notables ont été observées dans les performances des modèles selon les domaines, avec certains modèles étant plus efficaces dans certains contextes que d'autres.

Les performances des différents systèmes varient significativement selon les domaines. Par exemple, les systèmes comme Gemini et CamemBERT ont montré de bons résultats en urbanisme mais moins performants en santé, ce qui indique une spécialisation et une adaptation limitée des modèles à des contextes variés. Nous pouvons donc envisager de développer un système différent en fonction des domaines en minimisant le nombre de modèles différents afin de diminuer le coût de développement et d'application des modèles.

Il est toutefois important de noter que, bien que l'extraction de termes plus longs puisse être indicatrice de précision, elle peut également introduire de la complexité inutile, surtout si les termes extraits ne sont pas couramment utilisés ou reconnus dans le domaine d'application. Nous pourrions mettre en place une évaluation qualitative ou chercher à filtrer les termes trop longs en mettant en place des mécanismes de filtrage plus avancés pour sélectionner les termes les plus pertinents et éliminer ceux qui sont redondants ou peu utiles. Cela inclut l'utilisation de méthodes de ranking basées sur les graphes ou l'intégration de LLMs ou de SLMs.

Une évaluation intéressante à mettre en place concerne la densité des termes. En effet, une extraction de trop nombreux termes par texte peut diluer la pertinence et rendre difficile la compréhension et l'utilisation pratique des résultats. Il est crucial de trouver un équilibre entre la quantité et la qualité des termes extraits. Une analyse qualitative de la pertinence des termes extraits serait utile à mettre en place bien que difficile.

Limites des jeux de données standards

Le tableau 4.12 présente une comparaison entre les termes attendus avec leur classe associée et les termes prédits avec leurs classes associées. Ce tableau met en lumière les cas où les prédictions sont correctes malgré les erreurs présentes dans le jeu de données initial. Le jeu de données est utilisé pour identifier mot par mot s'il s'agit d'une entité nommée et pour lui attribuer une étiquette correspondante. Cependant, comme nous le constatons dans le tableau 4.12, qui illustre quelques exemples d'erreurs provenant de plusieurs de nos modèles, le jeu de données comporte plusieurs anomalies. Par exemple, le jeu de données omet parfois de considérer une expression complète, n'attribuant ainsi aucune classe alors que le groupe de mots forme une entité. De même, il arrive qu'une classe incorrecte soit assignée parce que le jeu de données n'a étiqueté qu'un mot comme étant une localisation, alors qu'un groupe de mots pourrait former une organisation. Il aurait été pertinent de pouvoir associer plusieurs étiquettes. En outre, bien que certains mots soient effectivement des entités, ils ne sont parfois pas reconnus comme tels dans le jeu de données.

Ces observations soulignent les limites du jeu de données, indiquant qu'il n'est pas toujours parfaitement fiable pour une annotation précise. Cependant, il est important de rappeler que les jeux de données constituent un repère précieux pour évaluer les performances des

systèmes de NER et comparer les différents modèles. Malgré ses quelques imperfections, un jeu de données bien conçu reste un outil essentiel pour le développement et l'amélioration des modèles NER.

Terme Attendu	Classe Attendue	Terme Prédit	Classe Prédite
Compagnie	O	Compagnie Thor	ORG
Thor	O	Compagnie Thor	ORG
Thierry	O	Thierry Smits	PER
Smits	O	Thierry Smits	PER
Madeleine	O	Madeleine de Maupin	PER
de	O	Madeleine de Maupin	PER
Maupin	O	Madeleine de Maupin	PER
Pierre	O	Pierre Bossan	PER
Bossan	O	Pierre Bossan	PER
Chronique	O	Chronique anglo - saxonne	MISC
anglo	O	Chronique anglo - saxonne	MISC
-	O	Chronique anglo - saxonne	MISC
saxonne	O	Chronique anglo - saxonne	MISC
éditions	O	éditions Dupuis	ORG
Dupuis	O	éditions Dupuis	ORG
Mouloudia	O	Mouloudia Club d'Oran	ORG
Club	O	Mouloudia Club d'Oran	ORG
d'Oran	O	Mouloudia Club d'Oran	ORG
Hauts	O	Hauts - de - Seine	LOC
-	O	Hauts - de - Seine	LOC
de	O	Hauts - de - Seine	LOC
-	O	Hauts - de - Seine	LOC
Seine	O	Hauts - de - Seine	LOC
Paris	O	Paris	LOC
Chine	O	Chine	LOC
Philippe	O	Philippe	PER
Jean	O	Jean	PER
Emmanuelle	O	Emmanuelle Baillet	PER
Baillet	O	Emmanuelle Baillet	PER
Louis	O	Louis - Gabriel Monnier	PER
Gabriel	O	Louis - Gabriel Monnier	PER
Monnier	O	Louis - Gabriel Monnier	PER
New	O	New York	LOC
York	O	New York	LOC
Iron	O	Iron Man	PER
Man	O	Iron Man	PER
Spider	O	Spider Man	PER
Man	O	Spider Man	PER
Madagascar	O	Madagascar	LOC
Jacques	O	Jacques	PER
Drevet	O	Drevet	PER
Marcel	O	Marcel	PER
Olivier	O	Olivier	PER
Marie	O	Marie	PER

Terme Attendu	Classe Attendue	Terme Prédit	Classe Prédite
Simpson	O	Simpson	PER
Philippe	O	Philippe	PER
Peythieu	O	Peythieu	PER
Lou	O	Lou	PER
Abraham	O	Abraham	PER
Pierre	O	Pierre	PER
France	O	France	LOC
d'Aubervilliers	O	d'Aubervilliers	LOC

TABLE 4.12: Dataset face aux prédictions réalisées

4.4 Synthèse des Contributions et Perspectives pour les Travaux Futurs

4.4.1 Bilan des Contributions Principales en Extraction Terminologique

Dans ce chapitre, l'accent est mis sur l'extraction terminologique à partir des corpus textuels. Un premier système basé sur les annotations des experts a été mis en place. Or, les annotations des experts varient en qualité et en consistance, ce qui a affecté la performance des modèles. Les experts annotent de manière subjective sans guide d'annotation standardisé, ce qui a introduit des incohérences. Un nouveau système baptisé InfoGlean KeyTerms a donc été développé, composé de trois sous-systèmes principaux : la reconnaissance des entités nommées (NER), l'extraction des termes pertinents, et l'extraction des entités juridiques. Ce chapitre décrit le processus de sélection et d'évaluation de ces sous-systèmes, en s'appuyant sur les données juridiques et pratiques fournies par Berger-Levrault.

Les principales étapes du processus d'extraction terminologique incluent la génération d'une liste de candidats terminologiques par le découpage syntaxique des phrases, suivie de l'application de modèles statistiques et d'apprentissage automatique pour filtrer ces candidats et extraire les termes pertinents. L'extraction terminologique vise ainsi à identifier et structurer les termes clés associés à un domaine précis, permettant de se concentrer sur l'essentiel de l'information et de créer une structure conceptuelle des termes.

Les résultats obtenus montrent que l'approche utilisée permet de détecter efficacement les termes pertinents et de structurer les connaissances textuelles, facilitant ainsi l'accès et l'utilisation des informations pour diverses applications industrielles et scientifiques.

4.4.2 Perspectives d'Amélioration et Suggestions pour les Recherches Futures

Tout d'abord, il a été noté que les annotations des experts étaient limitées car elles n'avaient pas de guide d'annotations. Afin de standardiser ces annotations, nous souhaitons donc développer un guide d'annotation standardisé pour les experts afin d'assurer une meilleure consistance et qualité des annotations. L'utilisation d'outils d'annotation assistée pourrait également améliorer la précision.

Ensuite, les termes et les connaissances évoluent rapidement, surtout dans les domaines techniques et juridiques. Le système doit donc être régulièrement mis à jour pour maintenir sa pertinence et son efficacité. Nous devons donc chercher à évaluer la capacité du système

à intégrer de nouvelles entités ou créer un nouveau module traitant les nouvelles informations arrivants. Nous devons donc développer des pipelines automatisés pour la mise à jour des modèles et des données afin de maintenir la pertinence et l'efficacité du système face aux évolutions rapides des connaissances. En effet, il est crucial qu'une information dite « vraie » par le passé qui soit devenue obsolète puisse être annotée dans le graphe de connaissances. De plus, de nouveaux domaines peuvent faire leur apparition, nous devons donc adapter et tester le système sur d'autres domaines et vérifier leur généralisation. Nous devons donc gérer cet aspect de généralisation à d'autres domaines qui pourraient nécessiter des ajustements des modèles.

En terme d'évolution, nous aimerions également enrichir les termes extraits par des synonymes provenant de sources diverses externes dans le domaine juridique et pratique comme légifrance[8] par exemple.

Enfin, nous avions cherché à créer une application d'évaluation continue au sein de cette thèse qui n'a pas abouti, nous souhaiterions donc mettre en place un système d'évaluation continue pour mesurer la performance du système sur des jeux de données diversifiés et en temps réel. Ceci pourrait inclure des retours d'utilisateurs et des analyses de performance post-déploiement.

En résumé, bien que les résultats du système d'extraction terminologique soient prometteurs, notamment en termes de détection des termes composés dans les 8 domaines spécifiques, il présente des marges de progression notables en termes de généralisation des modèles, et de mise à jour continue des données et des systèmes.

8. Lien du site de légifrance : `https://www.legifrance.gouv.fr/`

CHAPITRE 5

Extraction de relations à l'aide du réseaux lexical JeuxDeMots

Le chapitre précédent a traité de l'extraction terminologique, fournissant une base de termes clés pour ce chapitre. Ici, nous explorons l'extraction de relations en utilisant le réseau lexical JeuxDeMots. Les défis incluent la gestion de la rigidité des listes de relations prédéfinies et l'amélioration de la spécificité des relations extraites. Le chapitre précédent a présenté un système d'extraction terminologique. L'ensemble de termes clés extrait précédemment constitue une base d'entrée pour le système d'extraction de relations. En effet, partir d'une base de termes clés permet d'orienter un système d'extraction de relations en connaissant au préalable l'ensemble des termes sources et cibles qu'une relation peut composer. Les termes clés étant jugés comme pertinents pour un domaine donné, les relations extraites auront plus de chance d'être pertinentes.

Sommaire du chapitre

5.1 Algorithme d'extraction par classification de relations **117**
5.1.1 *Relational Embeddings Model* (REM) : Description du système 118
5.1.2 Évaluation 121
5.2 Classification de domaine **123**
5.2.1 Analyse des données pour la classification de domaine 125
5.2.2 *JURIdical DOCument Classification* (JURID'OCC) : Description du système 125
5.2.3 Évaluation 129
5.3 Discussion **133**
5.3.1 Évaluation 133
5.3.2 Relations dans le graphe de connaissances lexico-sémantique . . . 134
5.3.3 Relations en non-contextuel 134

5.1 Algorithme d'extraction par classification de relations

Différents types de relations peuvent être extraits du texte, en fonction du niveau de spécificité, de la granularité et du domaine. Les relations d'entités nommées sont entre des en-

tités nommées, telles que des personnes ou des lieux. Par exemple (Joe Biden, président_de, États-Unis). Les relations sémantiques de rôle font référence aux relations entre un prédicat et ses arguments, comme (agent, manager, John) ou (manger, patient, pomme). Les relations ontologiques sont des liens conceptuels qui structurent et organisent les connaissances en fonction de leurs relations fondamentales dans une ontologie. Les relations ontologiques jouent un rôle crucial dans la structuration et l'organisation des connaissances en établissant des liens conceptuels fondamentaux. Parmi ces relations, les relations de taxonomie, telles que l'hyperonymie et l'hyponymie, permettent de relier des termes généraux à des termes plus spécifiques, comme « animal » et « chien » reliés par la relation (chien, sous-classe_de, animal). Les relations partitives, incluant l'holonymie et la méronymie, décrivent les liens entre un tout et ses parties, par exemple, entre « voiture » et « roue », nous avons les relations (voiture, est_holonyme, roue) et (roue, est_méronyme, voiture). De plus, les relations de substances typiques relient une substance à ses manifestations, comme « eau » avec « pluie » ou « glace ». Les lieux typiques établissent des connexions entre des lieux et les activités qui y sont généralement associées, tel que « école » et « apprentissage ». Enfin, les relations de temps et événements typiques lient des périodes de temps à des événements couramment observés, comme « été » et « vacances ». Ensemble, ces relations forment une base ontologique solide qui facilite la compréhension et l'organisation des informations. Tous ces types de relations fournissent des informations précieuses sur le texte. Il s'agit ici de types de relations bien définis qui sont fréquemment réemployés comme la synonymie, l'équivalence, la hiérarchie ou encore l'association. Cependant, cela nécessite de définir chaque type au préalable et s'il y en a beaucoup à déterminer cela nécessite un coût manuel. Une autre forme d'extraction de relations existe. Il s'agit des relations ouvertes qui consistent à extraire l'information relationnelle depuis le texte comme les verbes par exemple permettant une plus grande liberté. Cependant cette technique nécessite de gérer un grand nombre de possibilités.

Afin de pallier les problèmes à la fois d'un système trop complexe que requiert une extraction de relations ouvertes et afin de limiter le nombre de types de relations, nous avons créé un système d'extraction de relations progressif et de manière itérative. Nous avons donc réalisé notre système en plusieurs étapes :

1. Réduction du problème d'extraction de relations à une extraction de relations de taxonomie
2. Extension à d'autres types de relations essentielles comme la synonymie, la méronymie ou d'autres.
3. Classification de domaines pour améliorer les résultats
4. Extraction de relations ouvertes

5.1.1 *Relational Embeddings Model* (REM) : Description du système

La figure 5.1 représente le système d'extraction de relations mis en place, nommé *Relational Embeddings Model* (REM). Il est valable à la fois lorsque nous avons réduit le problème à une extraction de relations taxonomique (REM V_0) ainsi que lors de l'extension à d'autres types de relations (REM V_1). Chaque étape ou bloc du système représente un **module** ou « sous-programme » du système.

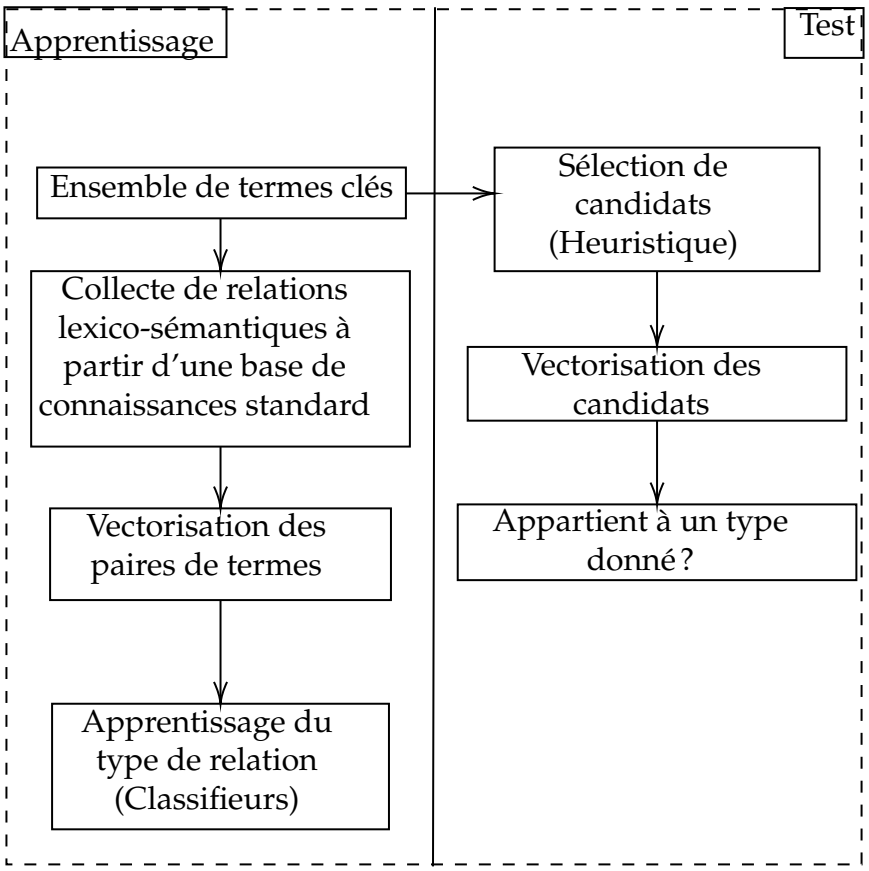

FIGURE 5.1 – Schéma (général) du système d'extraction de relations à l'aide du réseaux lexical JeuxDeMots

Réduction du problème d'extraction de relations à une extraction de relations de taxonomie

Une relation de taxonomie est une hiérarchie de concepts organisée selon des critères d'inclusion, permettant de structurer des connaissances de manière ordonnée. Cette hiérarchie s'appuie sur des relations d'hyperonymie et d'hyponymie. L'hyperonymie désigne la relation entre un terme général, appelé hyperonyme, et des termes plus spécifiques, appelés hyponymes, qui héritent des propriétés de l'hyperonyme. Par exemple, « animal » est un hyperonyme pour « chien », « chat » et « oiseau », qui sont ses hyponymes. Ainsi, l'hyponymie est l'inverse de l'hyperonymie et illustre comment des concepts spécifiques s'inscrivent dans des catégories plus larges. Cette structuration permet de comprendre et d'organiser les concepts en fonction de leurs niveaux de généralité et de spécificité, facilitant ainsi la classification et la recherche d'informations.

Étape 1 (Apprentissage) : Extraction de relations hyperonymiques et hyponymiques à partir de JeuxDeMots : En exploitant rezoJDM, nous avons extrait un ensemble de relations de type hyperonyme, dont l'identifiant unique dans rezoJDM est *r_isa*, et hyponyme, identifié par *r_hypo*, se limitant à ces deux types de relations sémantiques. Nous extrayons les relations de poids positif dans rezoJDM à partir des termes clés issus de l'extraction terminologique précédemment effectuée. Pour ce faire, nous requêtons rezoJDM en injectant la liste des termes clés. rezoJDM propose un filtrage par type de relations. Nous spécifions donc le type de relations visé : ici *r_hypo* puis *r_isa*. Nous effectuons ensuite un filtrage des relations extraites en commençant par supprimer les relations de poids négatif ou nul car nous nous

intéressons uniquement aux relations considérées comme « vraies ».

Parmi les instances de relations restantes, soit (x, R, y) un triplet représentant une relation de la forme (terme source, type de relation, terme cible). Nous gardons cette relation si y appartient aux termes clés. Les paires de termes ainsi obtenues forment la base de notre jeu de données pour l'apprentissage et l'évaluation. Après avoir récupéré à la fois les instances de relations d'hyperonymie et d'hyponymie, nous avons inversé chacune des instances de relations afin de constituer un plus large nombre de relations dans chacune des classes. En effet, lorsque nous inversons l'instance de relation d'hyperonymie, c'est-à-dire que nous intervertissons le terme cible avec le terme source, cette instance devient alors une hyponymie. Nous pouvons faire de même pour transformer une instance de relation d'hyponymie en instance de relation d'hyperonymie.

Étape 2 (Apprentissage) : Entraînement de modèles Word2Vec pour les plongements lexicaux : Nous avons formé des représentations vectorielles continues des termes-clés en entraînant un modèle Word2Vec (Mikolov et al., 2013) sur notre corpus. Pour cela, chaque terme clé a été remplacé par son référent afin de réduire le nombre de formes fléchies. En utilisant l'architecture CBOW via la bibliothèque Gensim[1], nous avons traité des n-grammes comme des unités complètes. En effet, le système est capable d'apprendre des vecteurs de mots composés et non de les déduire comme il se fait généralement. Par exemple, *fiche_de_paie* au lieu de 'fiche de paie' qui serait obtenu à partir du vecteur moyen des mots qui le composent, à savoir : fiche et paie. Cela permet de capturer leur signification entière et pour cela, nous utilisons le module phrases de Gensim[2].

Étape 3 (Apprentissage) : Construction de vecteurs de relations typées à partir des embeddings : À partir des vecteurs plongements lexicaux (*embeddings*) obtenus, nous avons dérivé des représentations vectorielles pour les relations en utilisant la soustraction vectorielle normée. Par exemple, si V_1 est le vecteur pour marché public et V_2 pour marché, alors (V_1 – V_2) représente la relation (marché public, *r_isa*, marché). Cette méthode permet de respecter la non-symétrie des relations *r_isa* et *r_hypo*, fournissant une base pour la classification.

Étape 4 (Apprentissage) : Développement de classifieurs binaires pour la prédiction de relations sémantiques : En utilisant la bibliothèque *Scikit-learn*[3], nous avons développé plusieurs classifieurs binaires, incluant des arbres de décision, des méthodes ensemblistes et des machines à vecteur de support (SVM). Ces classifieurs, alimentés par les vecteurs de relations, permettent de déterminer si une paire de termes correspond à une relation entre un terme spécifique et un terme générique, ou vice versa. Nous avons travaillé avec un ensemble équilibré de paires de relations réparties entre l'hyperonymie (*r_isa*) et l'hyponymie (*r_hypo*).

Étape 5 (Test) : Objectif et impact de l'approche sur l'enrichissement des réseaux lexico-sémantiques : L'objectif de notre approche est d'apprendre et d'évaluer les relations sémantiques extraites de rezoJDM, afin d'utiliser les classifieurs binaires pour prédire de nouvelles relations non encore reconnues par JeuxDeMots. Bien que rezoJDM soit l'un des plus grands

1. Lien vers la bibliothèque Gensim : `https://radimrehurek.com/gensim/models/word2vec.html`
2. Lien vers le modèle de Phrases implémenté dans Gensim : `https://radimrehurek.com/gensim/models/phrases.html`
3. Lien modèles supervisés Scikit-Learn : `https://scikit-learn.org/stable/supervised_learning.html`

réseaux lexico-sémantiques pour le français, il reste incomplet pour les domaines spécialisés. Par conséquent, l'utilisation de ces classifieurs pour la prédiction de nouvelles relations représente un axe important pour l'enrichissement de telles ressources. Pour cette étape, nous sélectionnons des candidats parmi les termes clés et effectuons une sélection aléatoire de potentielles paires. L'objectif est de ne pas surcharger le nombre de candidats sélectionnés ; mais à terme, nous visons à obtenir une sélection heuristique. Nous appliquons une stratégie de tirage aléatoire afin d'extraire les termes clés n'ayant pas de relations contenues dans rezoJDM. Les candidats sont ensuite vectorisés comme dans la phase d'apprentissage, puis injectés dans le classifieur. Ces candidats appartiennent-ils à un type donné ? Si oui, nous vérifions la confiance du classifieur face à cette prédiction. Si le classifieur est suffisamment confiant (score de confiance du classifieur > seuil, ici 95%), l'instance de relation est alors prédite comme valide.

Extension à d'autres types de relations

Dans le cadre de l'extension de l'extraction de relations de taxonomie, nous envisageons d'appliquer ce processus à un ensemble varié de types de relations, incluant notamment la synonymie, l'antonymie, les relations « meronymie »(*has part*) et « holonymie »(*part of*), ainsi que d'autres types disponibles dans la base de connaissances rezoJDM. Le processus d'extraction des relations hyperonymiques et hyponymiques reste identique, à la différence près que nous changeons le type de relation d'hyperonymie par un type de relation spécifique. Le principe reste inchangé par rapport au schéma initial donné sur la figure 5.1. Cependant des modifications d'adaptation ont été réalisées et vont être détaillées ci-après.

Modification de l'étape 1 (Apprentissage) : Lors de la construction des vecteurs de relations typées à partir des *embeddings*, nous procédons à une extension prenant en compte à la fois les relations symétriques et non symétriques, contrairement aux relations taxonomiques qui étaient uniquement non symétriques. Pour différencier ces relations, nous appliquons une opération arithmétique distincte : pour les relations asymétriques, nous effectuons une **soustraction** entre les vecteurs des deux termes clés d'une relation afin de maintenir le sens directionnel de la relation. En revanche, pour les relations symétriques, nous prenons la **valeur absolue** de cette différence, car le sens est bidirectionnel.

Modification de l'étape 4 (Apprentissage) : Nous développons des classifieurs binaires pour la prédiction des relations sémantiques. Pour chaque type de relation, nous utilisons le type spécifié en tant que classe positive et un « type contraire » en tant que classe négative pour l'apprentissage. Chaque type de relation dispose de son propre classifieur binaire. Pour la création d'une nouvelle classification binaire, nous utilisons également un seuil lors du test pour étendre les classifieurs binaires. L'objectif est de déterminer si une relation donnée appartient ou non à un type de relation spécifique. Pour ce faire, nous apprenons le type recherché en tant que classe positive, tandis que d'autres types de relations aléatoires servent de classe négative, en maintenant la symétrie ou l'asymétrie en fonction du type de la classe positive. Ensuite, nous appliquons un seuil : si la prédiction pour notre type de relation donné atteint 95% (seuil fixé), alors la relation est affectée à la classe positive ; sinon, elle est affectée à la classe dite négative (« sans classe »).

5.1.2 Évaluation

Comparaison à un système état de l'art

Évaluation de la représentation vectorielle des instances de relation : Corpus, système état de l'art et détails Afin d'évaluer notre système d'extraction de relations, nous nous intéressons à valider notre système de représentation vectorielle des relations. Nous confrontons, ensuite, notre système à un autre système état de l'art sur une tâche de **classification de types de relations**. Nous avons à notre disposition un jeu de données afin de valider notre représentation de paires de termes clés dans une relation. Il s'agit de l'ensemble de paires de termes clés liés par une relation donnée extraite depuis rezoJDM. Nous avons utilisé notre propre corpus métier afin de vérifier cette représentation vectorielle sur un corpus de spécialité directement. Cependant, nous devons prendre en considération le problème d'équilibre de notre jeu de données. Il est, en effet, préférable d'obtenir un corpus équilibré, à savoir un même nombre de paires de termes pour chaque type de relation pour la phase de classification. Cela permettra de valider ou d'invalider les différentes représentations vectorielles créées à partir des termes clés que ce soit notre système ou un autre système état de l'art. Pour cela, nous avons décidé de niveler par le bas le nombre de paires associées à un type de relation. Les autres ensembles de paires de relations sont donc en plus grand nombre. Nous piochons donc aléatoirement ce nombre fixé d'éléments dans chaque corpus. Nous répétons l'opération cinq fois afin d'obtenir cinq corpus représentatifs de notre classe. Lorsque nous nous comparons à un autre système, nous récupérons les représentations des mêmes paires de termes extraites. Si un système a récupéré plus de paires qu'un autre, nous retirons celles en surplus. Les paires qui constituent l'évaluation doivent être exactement les mêmes pour les deux systèmes.

Nous avons souhaité nous comparer à un système état-de-l'art de représentation de vecteurs de relations. Nous avons fait le choix de nous comparer à RELATIVE (*RELations as LATent dIscourse VEctors*), une méthode récente datant de 2019 d'apprentissage de vecteurs de relations basée sur des principes qui se sont avérés efficace (Camacho-Collados et al., 2019b). Cette méthode, à partir d'un corpus textuel et d'un modèle de plongements lexicaux de mots, apprend des plongements de paires de termes qui sont reliés sémantiquement. Cette approche est différente de la nôtre car, en effet, nous apprenons des plongements lexicaux et en déduisons par une opération arithmétique simple, une représentation vectorielle des paires de termes extraits depuis rezoJDM. RELATIVE, quant à lui, apprend directement des représentations vectorielles de relations. Nous avons fait le choix de RELATIVE car il a obtenu de très bons résultats lors de son expérimentation sur des jeux de données représentatifs dans le domaine de la classification de documents comme *20News* (Lang, 1995) ou encore *Reuters* (Lewis et al., 2004). Nous nous sommes tournés vers ce système car il donne une expérimentation très complète. Nous avons été nous même intrigués de pouvoir le réutiliser afin de représenter nos relations.

Résultats Nous avons appliqué une série de classifieurs binaires sur le jeu de données équilibré comme expliqué ci-dessus. Du fait du nivellement par le bas, nous lançons nos opérations sur chacun des corpus, un à la fois pour obtenir un résultat au plus juste en moyennant les cinq jets. Pour cette étude, nous n'avons pas approfondi le travail de sélection des meilleurs paramètres des classificateurs mais avons simplement identifié les meilleures représentations de relations étant donné un même paramétrage des classificateurs. Pour chacun des cinq tirages aléatoires et, afin d'obtenir des résultats fiables, nous utilisons le principe de la validation croisée (*Cross-validation*). Une fonction est disponible dans *Scikit-learn*[4].

4. Lien cross_val_score Scikit-learn : `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html`

Afin d'évaluer la qualité de nos représentations vectorielles de relation ainsi que celles de « Relative », nous utilisons deux mesures d'évaluation, à savoir : (1) le taux d'exactitude (*Accuracy rate*) et (2) la F-mesure. Ces deux mesures ont été indiquées directement dans un paramètre d'entrée à la fonction de validation croisée.

Le tableau 5.1 présente les résultats obtenus pour les couples d'instances de relation suivantes :

— Synonymie et antonymie
— Hyperonymie et hyponymie
— Meronymie et holonymie

Nous nous intéressons à ces trois cas de figures pour chacun des systèmes (renseignés dans le tableau). Le taux d'exactitude et la F-mesure sont abrégés respectivement par « Acc » et « F1 ». L'ensemble des détails sont regroupés dans la légende associée à ce tableau. Nous constatons des résultats pour notre système supérieur en taux d'exactitude et en F-mesure par rapport à Relative. Nous observons de bons résultats pour les SVM avec un taux d'exactitude supérieur à 73% pour chacun des couples de relations. Nous constatons des résultats pertinents et équilibrés pour l'ensemble des couples et des classifieurs. L'ensemble des bons résultats peuvent s'expliquer par notre corpus qui est de spécialité. Notre système est modelé dans ce but. Ce qui n'est pas le cas de RELATIVE. Néanmoins, nous constatons une sous-performance pour le classifieur DT dans le cadre du couple (« Fait partie de », « est une partie de ») ainsi que pour k-NN pour les relations symétriques lorsque $k \geqslant 3$. De plus, les résultats en globalité n'atteignent pas de très hauts pourcentages car RELATIVE a trouvé des relations entre les termes-clés différentes que celles de notre système. Nous avons donc dû niveler par le bas pour obtenir 1 206 paires de termes-clés à cause de l'antonymie, notamment. L'antonymie représente un terme et son contraire, il a été normal de trouver peu de termes clés concordant et cela a contribué à la chute du nombre de paires de termes-clés. En effet, RELATIVE extrait des représentations de paires de relations en contexte, ce qui n'est pas le cas de notre système qui à l'inverse reste générique.

5.2 Classification de domaine

L'extraction de relations à partir de textes est un domaine de recherche actif dans le traitement automatique des langues. Les relations existent entre différents termes clefs présents dans les textes. Cependant, le langage étant ambigu, certaines relations ne sont pas pertinentes car les termes qui le composent se trouvent dans des contextes différents. Une méthode possible pour filtrer les relations qui ne nous intéressent pas est de classifier ces relations en fonction de leur domaine d'application. Ainsi, si deux termes clefs d'une relation t_1, t_2 appartiennent à un domaine d, alors la relation r appartient également à d. De plus, en classifiant les documents, les termes clés rattachés au document sont également classifiés, permettant ainsi de rattacher un document à l'un des domaines. Il existe différentes méthodes de classification de domaines, notamment (1) la classification automatique, (2) la classification manuelle par des experts humains et (3) la classification hybride combinant les deux approches. L'efficacité de ces méthodes dépend de plusieurs facteurs, tels que la qualité et la quantité des données d'entraînement, la pertinence des catégories de classification et la précision des algorithmes utilisés.

Ici, nous nous intéressons à la question de la classification de documents selon des domaines de spécialité. En effet, nous nous demandons si la classification de documents peut aider l'extraction de relations. L'objectif final est de pouvoir filtrer les relations d'un graphe de connaissances lexico-sémantique à partir de textes en fonction de huit domaines spé-

Méthode	(Synonymie, Antonymie)				(Hyperonymie, Hyponymie)				(Fait partie de, est une partie de)			
	REM		RELATIVE		REM		RELATIVE		REM		RELATIVE	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SVM	0,73	0,73	0,55	**0,51**	0,75	0,75	0,57	0,55	**0,79**	**0,79**	**0,54**	0,54
DT	**0,72**	0,71	0,52	0,52	0,71	0,71	0,53	0,53	0,66	0,66	0,50	0,50
RF	**0,78**	**0,78**	0,55	0,55	**0,78**	**0,78**	0,57	0,57	0,74	0,73	**0,50**	**0,50**
***k*-PPV** ($k = 1$)	0,74	0,74	0,54	0,53	**0,80**	**0,80**	0,53	0,53	0,7	0,69	**0,52**	**0,52**
***k*-PPV** ($k = 3$)	0,61	0,61	**0,54**	**0,53**	**0,79**	**0,79**	0,58	0,55	0,73	0,73	**0,54**	**0,53**
***k*-PPV** ($k = 5$)	0,63	0,63	0,54	**0,53**	**0,79**	**0,78**	0,56	0,56	0,76	0,75	**0,53**	**0,53**

TABLE 5.1 – Tableau comparatif des performances des modèles lors de la classification de types de relations avec Relative puis notre système REM.

Légende :

⟶ **Systèmes** :

REM : *Relational Embeddings Model* : Notre modèle d'extraction de relations

RELATIVE : *RELations as LATent dIscourse VEctors*

⟶ **Métriques** :

Acc : *Accuracy* à savoir le taux d'exactitude

F1 : F-mesure

⟶ **Classifieurs** :

SVM : Machines à Vecteurs de Support (*support-vector machine*), un algorithme d'apprentissage supervisé utilisé pour les tâches de classification.

DT : *Decision Tree*, un modèle de décision qui utilise un arbre de décision pour prendre des décisions basées sur les observations des données.

RF : Forêt aléatoire (*Random Forest*), un ensemble d'arbres de décision pour améliorer la précision de la classification.

k-PPV ou **k-NN** : k plus proches voisins (*K-Nearest Neighbors*) avec $k = (1, 3, 5)$, un algorithme de classification basé sur la proximité des points de données dans l'espace des caractéristiques.

⟶ **Couleurs** :

Meilleurs résultats (pour Acc et pour F1) pour un classifieur donné

Résultats les plus faibles (pour Acc et pour F1) pour un classifieur donné

cifiques : *État civil et Cimetières, Élections, Commande publique, Urbanisme, Comptabilité et Finances locales, Ressources Humaines Territoriales, Justice* et *Santé*. Nous avons à notre disposition des articles déjà classifiés dans un domaine précis, mais aussi des ouvrages dont il n'est pas possible de récupérer leur(s) domaine(s). Pour résoudre ce problème, nous avons élaboré trois stratégies possibles : la classification de documents, la classification directe des termes clés des experts et la classification directe des relations. Nous avons effectué une analyse des données et un développement expérimental comparatif des performances des différentes stratégies pour évaluer leur efficacité.

5.2.1 Analyse des données pour la classification de domaine

Dans cette sous-section, nous allons analyser les annotations des experts en fonction des domaines. Nous nous intéressons ici à la liste unique unifiée des termes clés représentatifs des experts, qui comprend un total de 41 875 termes-clés répartis entre les ouvrages et les articles.

Domaine	**Nombre de termes mono-domaines**	**Nombre de termes multi-domaines**
Justice	1 299	20
Comptabilité et Finances locales	893	47
Commande publique	692	99
État civil et Cimetières	587	194
Santé	438	236
Urbanisme	303	239
RH territoriales	187	307
Élections	118	334

TABLE 5.2 – Répartition des termes annotés des experts en termes mono-domaines et en termes multi-domaines

Les termes-clés utilisés ne sont pas toujours exclusifs à un seul domaine. En effet, lorsque nous analysons les termes-clés des articles, 4 517 d'entre eux sont associés à un seul domaine, tandis que 650 sont associés à plusieurs domaines. Cela peut poser un problème, car cela souligne le phénomène de domaines croisés, qui indique que certains domaines sont étroitement liés. Le tableau 5.2 présente la répartition des termes clés qui sont exclusifs à un seul domaine et ceux qui sont associés à plusieurs domaines. Pour chaque domaine, le nombre d'éléments qui sont exclusifs à ce domaine et le nombre d'éléments qui appartiennent également à d'autres domaines sont indiqués. De ce tableau, nous pouvons constater que le sous-corpus « *justice* » est très annoté. Nous constatons aussi un lexique conséquent de termes clés mono-domaines. Cela constitue assez d'informations à notre disposition pour constituer les lexiques et ainsi aider à la classification de domaines en comparant la similarité entre les termes clés.

Afin de vérifier la possible intersection entre les différents domaines, nous présentons une matrice de corrélation des différents domaines sur la figure 5.2. Cette matrice indique que si un terme clé est associé à un domaine donné, il est souvent également associé à un autre domaine si le coefficient de corrélation est élevé. Nous avons constaté que le domaine de la *santé* et celui des *ressources humaines territoriales* présentent la plus forte corrélation, atteignant 54 %. Nous avons également remarqué une corrélation légèrement inférieure à 30 % entre « *commande publique* » et « *comptabilité et finances locales* », ainsi qu'entre « *justice* » et « *état civil et cimetières* ».

5.2.2 *JURIdical DOCument Classification* (JURID'OCC) : Description du système

Nous avons mis en place un module, intitulé *JURIdical DOCument Classification* (JURID'OCC), qui vient s'ajouter aux autres modules présentés dans la section 5.1. Le but étant de vérifier si ce module de classification de documents permet d'améliorer l'extraction de

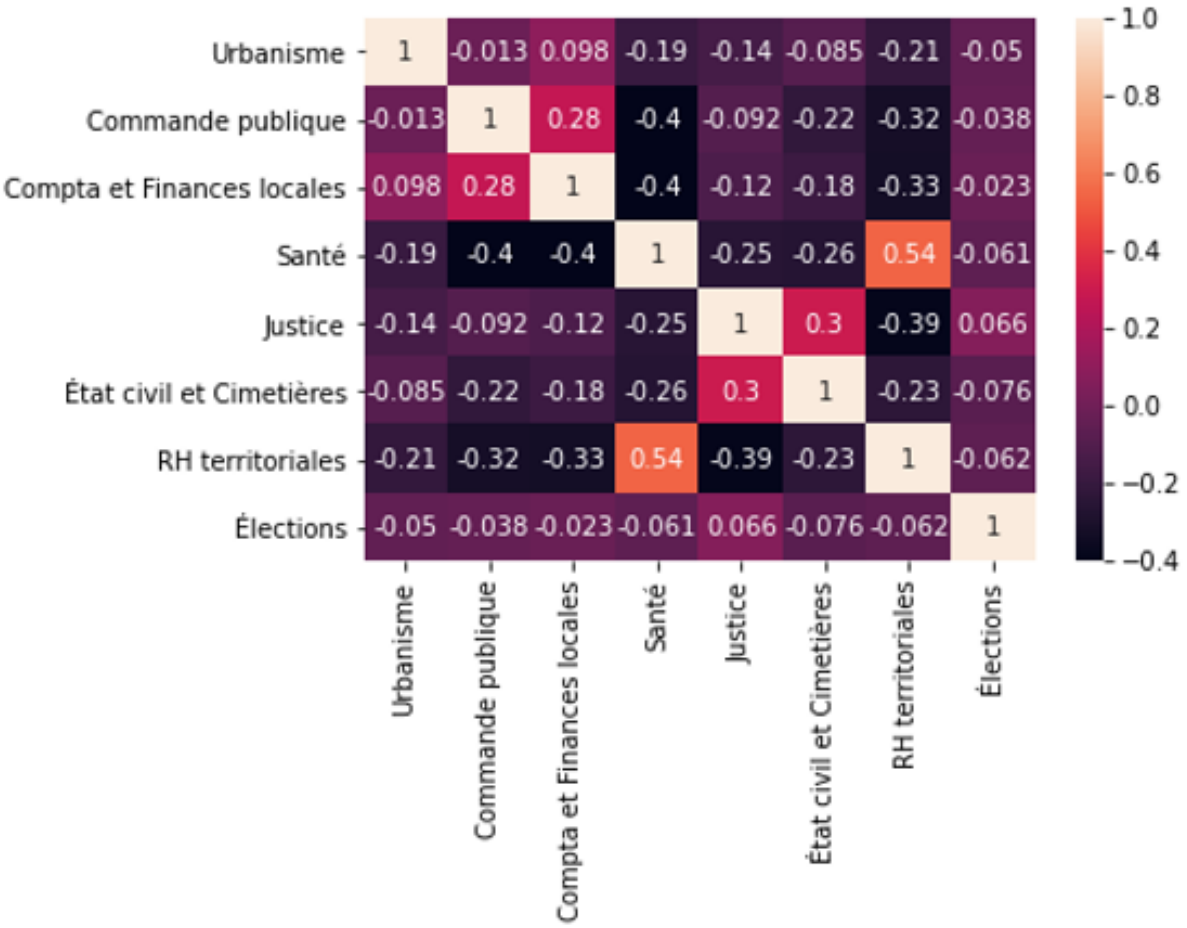

FIGURE 5.2 – Matrice de corrélation de *Pearson* entre les différents domaines pour les annotations unifiées des experts

relations entre les termes extraits. L'architecture du système complet est illustrée sur la figure 5.3. À partir de documents textuels bruts, le but est de pouvoir extraire des relations entre les idées clés et de pouvoir les filtrer en fonction du domaine.

L'architecture du système de classification de documents qui vient se greffer en début de processus est détaillée sur la figure 5.4. Ce module consiste à attribuer à chacun des paragraphes de nos textes (articles ou ouvrages) un des 8 domaines après les avoir prétraités. Le système prend en entrée d'une part les articles qui sont déjà associés à un domaine spécifique et d'une autre part les ouvrages dont le domaine est à déterminer. Dans le cas des ouvrages, nous utilisons un système de classification qui prend en entrée une représentation du texte plutôt que le texte en lui-même qui peut être un sac de mots ou une représentation vectorielle. Le fonctionnement est expliqué plus en détail dans la partie traitant de la classification de document ci après. Cette classification est un module supplémentaire et les étapes qui suivent (hors des pointillés, c.f. figure 5.3) peuvent être effectuées indépendemment. Cette architecture est générique : elle est composée d'un extracteur terminologique et de relations qui peuvent être n'importe lesquels et peuvent ainsi être substitués en cas de modification des systèmes. Nous utilisons le système d'extraction terminologique détaillé dans le chapitre précédent (c.f. chapitre 4). Il en est de même pour le module d'extraction de relations. Les expérimentations réalisées dans ce chapitre utilisent le système d'extraction de relations détaillé dans la section précédente (c.f. section 5.1). En cas d'améliorations pour n'importe quel module du schéma 5.3, les systèmes pourront être substitués et chacune des étapes restera pertinente.

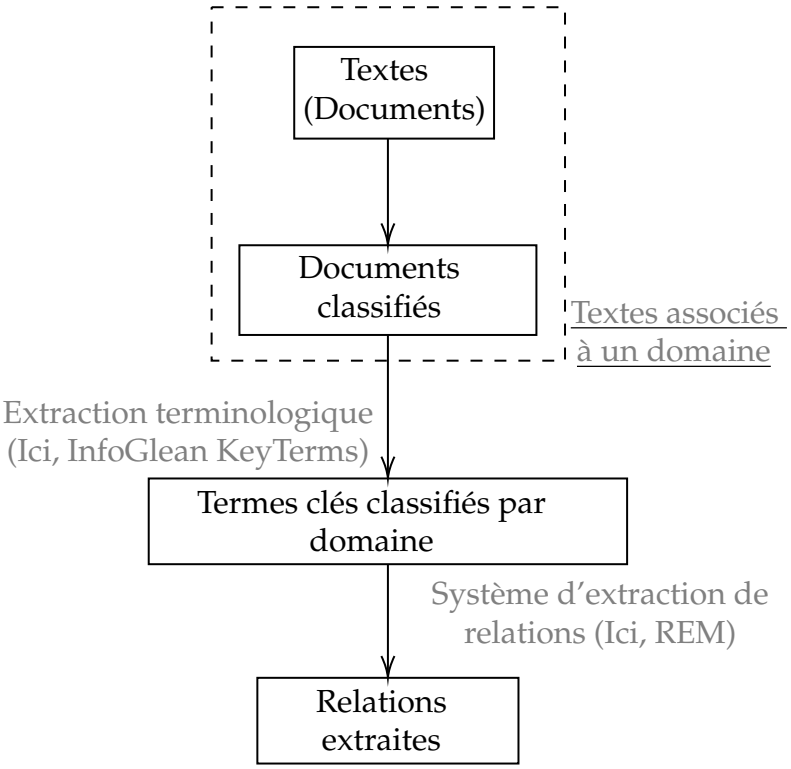

FIGURE 5.3 – Schéma (général) du système d'extraction de relations en utilisant la classification de documents

Classification de documents

Nous utilisons la classification automatique et supervisée pour attribuer un document ou texte (paragraphe par exemple) à un domaine spécifique. Cela implique de donner une catégorie à chaque texte que nous voulons classer. Toutefois, il est difficile d'apprendre directement à partir de données textuelles brutes. Par conséquent, nous devons utiliser une représentation des textes pour convertir chaque texte en une donnée mathématique plus facilement interprétable par le système d'apprentissage. Nous choisissons de découper nos documents en paragraphe pour deux raisons. La première vient de la limitation de la mémoire de certaines méthodes de représentation des documents. En effet, certaines méthodes ne peuvent emmagasiner qu'un nombre limité de *tokens* (par exemple, BERT). La deuxième vient du fait qu'un ouvrage peut traiter différents sujets de domaines différents. En effet, de manière générale, les articles traitent un thème bien spécifique tandis que les ouvrages traitent des thèmes plus généraux. Il existe différentes méthodes pour représenter les documents mais cette représentation de documents est détaillée ci-après. À partir des représentations de documents, nous réalisons une classification. Pour ce faire, nous avons utilisé plusieurs modèles de classification multi-classes tels que *MultinomialNB*, *LinearSVC* et *SGDClassifier*. Ces modèles nous permettent d'assigner une des 8 classes apprises. Cependant, dans le cadre de la classification binaire, nous ne pouvons choisir qu'entre deux classes distinctes, qui sont généralement considérées comme « oui » ou « non ».

Pré-traitement des textes

Dans le cadre de ce travail de recherche, les textes ont été prétraités selon les méthodes classiques. Ce pré-traitement n'est employé que pour la classification de documents. Pour ce faire, nous avons utilisé les bibliothèques couramment employées pour ce type de tâche.

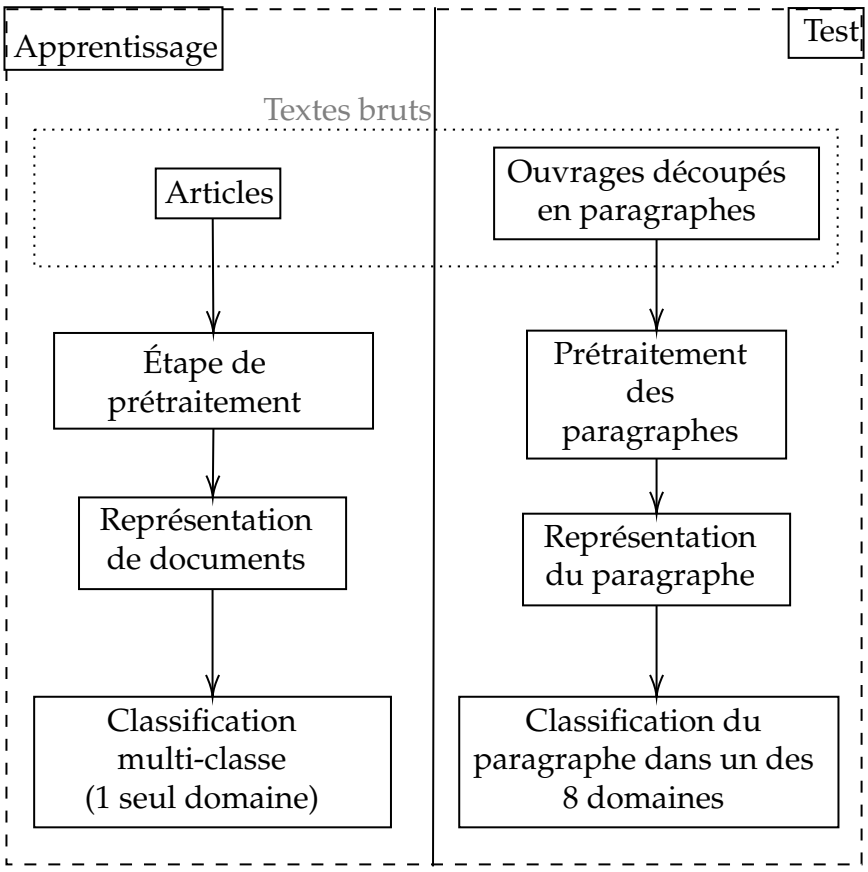

FIGURE 5.4 – Schéma de l'architecture du système de classification de documents (JURI-D'OCC)

Tout d'abord, nous avons converti chaque article en minuscules. Ensuite, nous avons réalisé une **tokenisation** des phrases, qui consiste à découper les phrases en mots distincts. Des mots courants, non significatifs de la langue française ont donc été enlevés (*stopwords*). Nous pouvons, après cette étape, également enlever les signes de ponctuation. Finalement, nous avons effectué une **lemmatisation** pour déterminer les racines des mots. Pour cette étape, nous avons utilisé la bibliothèque Stanza (Qi et al., 2020).

Représentation de documents

Nous utilisons trois méthodes courantes afin de représenter les documents : les méthodes classiques, qui consistent à représenter les documents sous forme de sacs contenant à la fois les mots et leur occurrence, les *embeddings*, qui traduisent les documents en vecteurs, et les modèles de langage, qui sont des modèles statistiques permettant de modéliser la distribution de séquences de mots dans une langue donnée. La première façon de représenter les documents consiste à les représenter par sac de mots (ou *Bag of Words* - BoW, en anglais). Nous nous intéressons à deux façons de le décrire : *Classic BoW* et *TF-IDF*. Ensuite, en ce qui concerne les *embeddings*, deux méthodes sont possibles : utilisation des plongements lexicaux (*word embeddings*) avec le modèle *word2vec* ou utilisation de *doc2vec* qui consiste à créer un vecteur du document directement. Nous utilisons Gensim. Concernant les modèles de langue, nous nous intéressons aux modèles CamemBERT et FlauBERT. Nous utilisons la méthode de représentation phrase par phrase grâce à Sentence-BERT pour créer une représentation de chacune des phrases, puis nous appliquons la moyenne. Pour notre étude, nous faisons appel à Sentence-CamemBERT et Sentence-FlauBERT, que nous récupérons depuis

HuggingFace.

5.2.3 Évaluation

Corpus de test

Afin d'évaluer différents modèles de classification de documents, nous utilisons les articles de notre corpus de spécialité qui sont déjà catégorisés entre les différents domaines. Cela permet d'avoir un corpus déjà labélisé et de spécialité. Nous avons constaté (voir figure 5.5) que nous avions une composition très déséquilibrée. C'est pourquoi, nous avons équilibré notre corpus afin de créer un corpus de test équilibré. Pour cela, nous avons réduit le nombre d'échantillons au nombre d'articles de la plus petite classe à savoir "élections". Nous nous retrouvons donc avec 250 articles dans chacune des classes après le rééquilibrage. Cela peut paraître peu, mais cela permet d'avoir un premier aperçu de résultats fiables et une conclusion définitive peut être prise sur un autre jeu de données plus conséquent et représentatif par la suite.

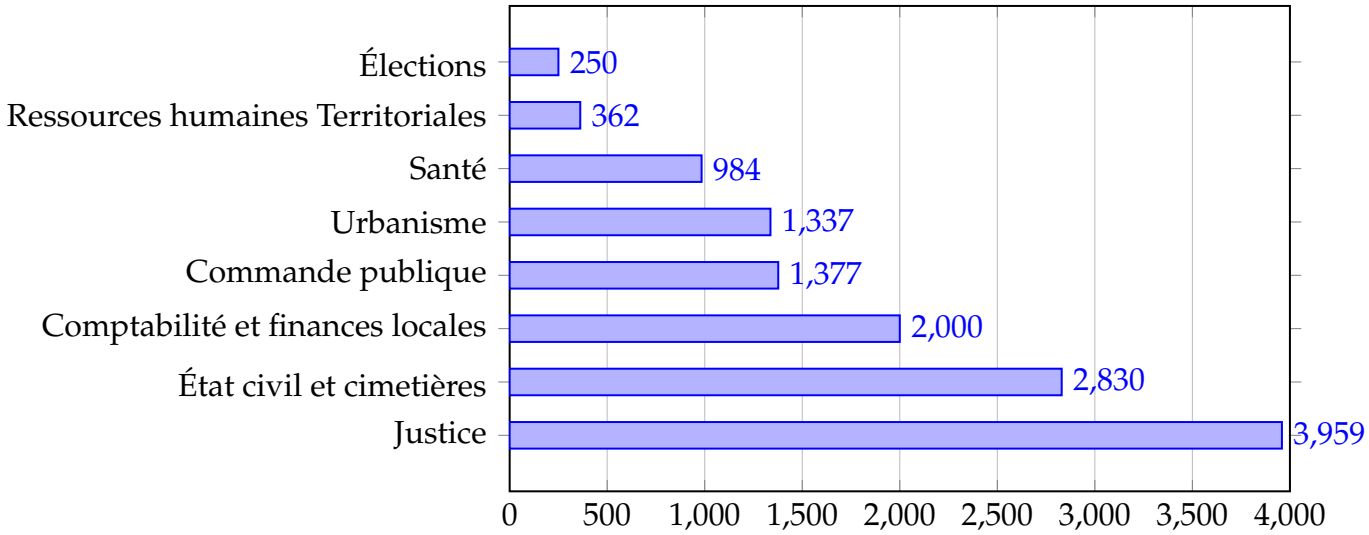

FIGURE 5.5 – Répartition des documents en fonction des différents domaines dans le jeu de données complet

Classification de documents

Nous allons tout d'abord décrire les scores obtenus pour l'attribution des textes aux 8 domaines. Cela permettra de sélectionner les résultats les plus performants pour la classification afin de vérifier leur pertinence pour l'extraction de relations. Nous comparerons ainsi les résultats obtenus avec et sans classification de documents. Les résultats sont évalués selon une mesure unique : la F-mesure. Les résultats de la classification de documents sont présentés dans le tableau 5.3. Ce tableau présente, parmi tous les tests réalisés, uniquement les meilleurs scores obtenus pour chacune des catégories de représentation de documents pour éviter de surcharger la visualisation des résultats. Des détails complémentaires à ces résultats sont disponibles en annexes (c.f. A). La légende associée au tableau présente le code afin d'identifier les meilleurs résultats pour chacun des 8 domaines. Nous constatons de bons résultats principalement en utilisant les méthodes classiques. Nous choisissons de

partir sur ces méthodes malgré les bons résultats obtenus avec les autres modèles présentés dans le tableau car les modèles de langues et Doc2vec nécessitent un modèle spécifique pour une langue donnée. En effet, *classic BoW* et TF-IDF ne nécessitent pas un modèle entraîné pour une langue en particulier. En cas d'ouvrages nouvellement rédigés, si nous rencontrons une autre langue, nous pourrons tout de même utiliser notre modèle sans devoir chercher un modèle approprié pour une langue donnée, ce qui est le cas des autres méthodes. Les résultats obtenus peuvent paraître surprenant car les méthodes dites classiques semblent plus simples d'utilisation. Nous pouvons donc penser que ce non-entrecroisement permet de distinguer les différents domaines et par conséquent que les méthodes de sacs de mots sont suffisantes pour cette tâche. Toutefois, il est à noter que les *tokenizers* utilisés en amont tels que Stanza ou Spacy sont tout de même entraînés sur une langue. Ce détail est donc à relever mais ne doit pas primer sur le choix des modèles.

Catégories	**Représentation de documents**	**Prédiction**	**Commande publique**	**Comptabilité et finances locales**	**Justice**	**RH territoriales**	**Santé**	**Urbanisme**	**Élections**	**État civil & cimetières**	**Moyenne**
Plongements lexicaux	Doc2Vec	*most similar*	0,93	0,89	0,91	0,77	0,78	0,92	0,98	0,92	0,89
	Word2Vec	*Linear SVC*	0,48	0,32	0,02	0,22	0,38	0,07	0,40	0,57	0,31
Méthodes classiques	Classic BoW	*Linear SVC*	**0,99**	**0,95**	0,88	**0,96**	0,94	**0,96**	**0,99**	0,91	**0,95**
	TF-IDF	*Linear SVC*	0,94	0,91	0,93	0,95	**0,97**	0,92	**0,99**	0,94	0,94
		SGD Classifier	0,94	0,89	**0,94**	0,95	0,95	0,93	**0,99**	**0,95**	0,94
Modèles de langues	Sentence CamemBERT	*Linear SVC*	0,94	0,89	0,90	0,90	0,92	0,92	0.94	0,97	0,92

TABLE 5.3 – Résultats de classification de documents pour chacun des domaines puis la moyenne de l'ensemble des domaines.

Légende :

⟶ **Code couleur** :

Meilleurs résultats pour un domaine donné

⟶ **Métriques** :

F1-Score

Classification de documents dans le cadre de l'extraction de relations

L'ensemble des résultats pour l'extraction de relations à la fois avec et sans classification se trouvent dans le tableau 5.4. Nous avons explicité les résultats pour chacun des domaines classifiés mais nous avons mis les moyennes des résultats dans les deux dernières colonnes puisque le modèle d'extraction de relations est par défaut moyenné car le domaine n'est pas connu. Pour ces résultats nous constatons deux choses : la classification de domaine amène de meilleurs résultats dans le cadre des relations asymétriques (*hyperonymie vs hyponymie* et fait partie de vs est une partie de) mais aussi dans le cadre des relations symétrique (ici, *synonymie vs antonymie* uniquement) pour 2 classifieurs sur les quatre évalués (SVM et *Decision Tree*). Les relations asymétriques ont une directionnalité claire, ce qui peut rendre les distinctions plus marquées et, par conséquent, plus faciles à modéliser par des algorithmes comme RF et k-NN. Par exemple, dans les relations hyperonymie vs hyponymie, il est plus évident de catégoriser un terme comme « animal » en tant qu'hyperonyme et « chien » en tant qu'hyponyme. La classification de domaine aide à renforcer ces distinctions en fournissant un contexte supplémentaire spécifique au domaine, ce qui améliore les performances des classificateurs. En revanche, les relations symétriques, telles que synonymie et antonymie, sont intrinsèquement plus complexes et subtiles. Les synonymes partagent des significations très proches, et les antonymes sont opposés mais dans un même contexte de signification. Ces nuances fines peuvent être plus difficiles à capter pour des classificateurs comme RF et k-NN, surtout lorsque les domaines spécifiques introduisent une variabilité supplémentaire. La classification de domaine pourrait complexifier la tâche en introduisant des subtilités contextuelles qui perturbent la capacité des algorithmes à généraliser correctement les relations symétriques.

En résumé, les relations asymétriques bénéficient de la classification de domaine parce que celle-ci renforce des distinctions directionnelles claires, facilitant la tâche des classificateurs RF et k-NN. En revanche, les relations symétriques, avec leurs nuances plus subtiles et contextuelles, peuvent souffrir de cette même classification de domaine, rendant la tâche plus difficile pour ces algorithmes. Nous pouvons donc en conclure qu'il y a un intérêt à utiliser préalablement une classification de domaine avant d'extraire les relations asymétriques. Hypothétiquement, nous supposons que cela améliore aussi la qualité des relations extraites mais cela reste à confirmer à l'aide d'une analyse qualitative que nous envisageons de faire en perspectives.

Statistiques sur les instances de relations extraites

Le tableau 5.5 présente des statistiques sur des instances de relations extraites, en tenant compte de la place des termes dans la relation (source ou cible) et de leur domaine (même domaine ou domaine différent). La constance du nombre moyen de mots à 1 suggère que les termes extraits sont principalement des mots simples, indépendamment de leur position dans la relation ou de leur domaine.Les termes sources ont une proportion plus élevée de multimots (environ 8 %) par rapport aux termes cibles (moins de 1%). Par contre, la différence entre les domaines n'affecte que légèrement le pourcentage de multimots, suggérant que la présence de multimots est relativement indépendante du domaine des termes.

Ce tableau met en évident plusieurs limites dans notre approche :

1. **Simplification excessive des termes :** Le nombre moyen de mots étant constant à 1 pour les termes sources et cibles indique une forte simplification des termes. Cela pourrait négliger la complexité des termes composés ou des expressions multi-mots qui pourraient être pertinentes pour certaines relations.

Type de relation	Classifieur	Commande publique	Comptabilité et finances locales	Justice	RH territoriales	Santé	Urbanisme	Elections	Etat civil & cimetières	Moyenne x domaine	Moyenne sans la classification
Hyperonymie VS Hyponymie	SVM	0,78	0,80	0,79	0,86	0,83	0,84	0,89	0,83	**0,83**	0,75
	DT	0,57	0,68	0,71	0,77	0,79	0,72	0,78	0,71	**0,72**	0,71
	RF	0,75	0,74	0,76	0,81	0,83	0,78	0,85	0,80	**0,79**	0,78
	k-NN (k=5)	0,65	0,81	0,80	0,84	0,78	0,82	0,87	0,77	**0,79**	0,78
Synonymie VS Antonymie	SVC	0,85	0,74	0,74	0,79	0,85	0,78	0,84	0,80	**0,80**	0,73
	DT	0,76	0,76	0,76	0,76	0,76	0,76	0,76	0,76	**0,75**	0,72
	RF	0,79	0,61	0,64	0,69	0,75	0,65	0,76	0,73	0,70	**0,78**
	k-NN (k=5)	0,65	0,61	0,64	0,66	0,66	0,63	0,65	0,60	0,64	**0,74**
Fait partie de VS est une partie de	SVC	0,75	0,84	0,83	0,84	0,91	0,80	0,66	0,84	**0,82**	0,79
	DT	0,76	0,72	0,75	0,72	0,81	0,61	0,54	0,61	**0,68**	0,65
	RF	0,79	0,81	0,76	0,86	0,89	0,79	0,63	0,86	**0,80**	0,73
	k-NN (k=5)	0,28	0,80	0.78	0,81	0,81	0,72	0,55	0,79	**0,75**	**0,75**

TABLE 5.4 – Tableau comparatif des performances des classificateurs / Résultat de l'extraction de relations à la fois lors de la classification de domaine et sans classification de domaine

Légende :

⟶ **Système utilisé** : REM : *Relational Embeddings Model* : Notre modèle d'extraction de relations

⟶ **Métrique** : F-mesure

⟶ **Classifieurs** :

SVM : Machines à Vecteurs de Support (*support-vector machine*), un algorithme d'apprentissage supervisé utilisé pour les tâches de classification.

DT : *Decision Tree*, un modèle de décision qui utilise un arbre de décision pour prendre des décisions basées sur les observations des données.

RF : Forêt aléatoire (*Random Forest*), un ensemble d'arbres de décision pour améliorer la précision de la classification.

k-PPV ou **k-NN** : k plus proches voisins (*K-Nearest Neighbors*) avec $k = 5$ (valeur par défaut), un algorithme de classification basé sur la proximité des points de données dans l'espace des caractéristiques.

⟶ **Code couleur** : **Meilleurs résultats pour un classifieur donné (ne concerne que les moyennes)**

2. **Manque de variation dans les termes cibles :** Le faible pourcentage de multimots pour les termes cibles (moins de 1 %) suggère une possible sous-représentation des termes composés ou complexes en position de cible. Cela pourrait limiter la richesse

Calcul effectué	Place dans la relation	Même domaine	Domaine différent	Total
Nombre de mots moyen	Source	1	1	1
	Cible	1	1	1
Pourcentage multimots	Source	8,26%	8,12%	8,21%
	Cible	0,86 %	0,95 %	0,89 %

TABLE 5.5 – Statistiques effectuées sur les termes des instances de relations extraites

Légende :

⟶ **Spécificité** :

Place dans la relation : Une relation du type (mot_source, type_relation, mot_cible) est composée d'un mot « source » et d'un mot « cible ».

Différenciation de domaine : le terme source et le terme cible ont ils au moins un domaine commun ? Si oui, ils appartiennent au même domaine, sinon ils sont rangés dans « domaine différent ».

⟶ **Calculs** :

Pourcentage multimots : pourcentage de multimots dans l'ensemble des termes soit source soit cible des instances détectées

Nombre mots moyen : nombre moyen de mots par terme clef soit source soit cible des instances de relations

Calculs réalisés sur l'ensemble des paires de termes reliés par un type de relation.

des relations extraites.

3. **Influence limitée du domaine :** Le faible impact du domaine sur les pourcentages de multimots suggère que la différenciation par domaine n'est pas pleinement exploitée. Cela implique une trop grande généricité dans les paires de relations qui est notamment dûe au fait d'avoir des mots simples uniquement. De plus, le contexte des termes et leur désambiguïsation ne sont pas pris en compte. Des termes polysémiques peuvent mener à des relations incorrectes si leur sens précis n'est pas identifié.

5.3 Discussion

5.3.1 Évaluation

L'utilisation d'un corpus standard pour valider les instances de relations est cruciale pour assurer l'objectivité et la crédibilité des résultats. En effet, valider les relations extraites en se basant uniquement sur notre propre corpus peut introduire un biais significatif vis-à-vis des données. Lorsqu'on utilise un corpus interne pour à la fois l'apprentissage et la validation, on risque de surévaluer la performance de notre modèle car ce dernier a déjà été exposé aux particularités de ces données. Cela est particulièrement problématique si l'apprentissage se fait sur une partie des relations tirées de JeuxDeMots et que la validation se fait sur une autre partie du même corpus. Cette méthode peut orienter le système d'évaluation pour obtenir de bons résultats artificiellement élevés, car le modèle a déjà une connaissance préalable des structures et des relations présentes dans le corpus.

En revanche, un système qui n'a jamais vu les données avant l'évaluation offre une mesure beaucoup plus rigoureuse et impartiale de la performance. Un tel système évalue la

capacité du modèle à généraliser et à appliquer correctement les règles d'extraction de relations sans dépendre des spécificités des données d'entraînement. Comparativement, valider notre modèle sur un corpus standard, externe et reconnu, élimine ces biais et permet de tester la robustesse et l'efficacité du modèle dans des conditions variées et non anticipées. Cela assure que les résultats obtenus ne sont pas uniquement le produit d'une familiarité avec les données, mais qu'ils reflètent réellement la capacité du modèle à extraire des relations pertinentes dans divers contextes.

En perpsective d'analyse statistique sur les instances de relations, nous pourrions nous intéresser aux informations sur les types spécifiques de relations extraites. En effet, nous pourrions voir comment se comporte les instances de relations en fonction des types de relations ainsi que de chercher à déterminer la pertinence des relations en analysant les termes impliqués. Nous pourrions fournir une analyse qualitative des termes et des relations. Une évaluation qualitative pourrait révéler des insights supplémentaires sur la pertinence et l'exactitude des relations extraites.

5.3.2 Relations dans le graphe de connaissances lexico-sémantique

En perspective, nous souhaitons intégrer de nouveaux types spécifiques pour le graphe de connaissances lexico-sémantiques qui viendraient compléter les informations déjà extraites :

1. L'intégration des relations entre les termes clés annotés par des experts et les termes pertinents extraits du texte peut grandement améliorer la précision et la pertinence de l'extraction de relations. Les annotations d'experts apportent une validation qualitative et une profondeur sémantique qui enrichissent les données textuelles brutes. En reliant les termes clés, identifiés par leur importance et leur contexte spécifique, aux termes pertinents trouvés dans le texte, on crée un réseau de relations plus cohérent et significatif. Cela permet non seulement de vérifier la validité des termes extraits, mais aussi d'améliorer la compréhension globale des relations sémantiques présentes dans le corpus, rendant ainsi les résultats plus fiables et utilisables pour des analyses approfondies. Nous pourrions voir les annotations des experts comme des concepts clés ayant une importance conséquente pour le graphe de connaissance ce qui pourrait en découler une certaines importances des termes associés à ce concept.
2. L'ajout de regroupements des termes pertinents à travers le clustering des différentes formes fléchies et des synonymes, qu'ils proviennent de sources externes ou du corpus de Berger-Levrault, est essentiel pour enrichir la modélisation des relations. En utilisant des sources externes, on peut incorporer des connaissances lexicales et sémantiques plus larges et variées, qui complètent et renforcent les données internes. Ce processus de clustering permet de reconnaître et de regrouper des termes ayant des formes (fléchies) différentes mais partageant des significations similaires, améliorant ainsi la précision et la couverture des relations. Modéliser ces regroupements par une relation spécifique permet de structurer explicitement les liens entre ces termes, facilitant leur utilisation dans des tâches ultérieures d'extraction et d'analyse de relations. Cela garantit également une meilleure gestion de la variation lexicale et une compréhension plus profonde des concepts présents dans le texte.

5.3.3 Relations en non-contextuel

Le tableau 5.6 présente des statistiques sur les relations extraites à partir d'un ensemble de termes clés. Les 1 277 termes clés initiaux constituent le point de départ pour l'extrac-

tion des relations entre ces termes. Parmi les 87 650 instances de relations extraites, 25 704 (29,3%) concernent des relations où les termes appartiennent au même domaine. Cela montre qu'environ un tiers des relations établies sont entre termes ayant une certaine cohérence thématique mais que donc deux tiers n'ont pas de cohérence et sont soit incohérentes soit d'ordre générique. Il y a 14 498 relations (16,5%) entre termes qui apparaissent dans le même document. Ce pourcentage relativement faible suggère que de nombreuses relations se forment entre termes provenant de documents différents, ce qui peut diminuer la pertinence contextuelle des relations. Cela se confirme d'autant plus lorsque nous nous forcalisons sur les instances de relations d'un même paragraphe. En effet, seulement 1 660 relations (1,9%) concernent des termes qui apparaissent dans le même paragraphe. Ce chiffre est très faible comparé aux autres catégories. Le fait que seulement 1,9% des relations extraites proviennent du même paragraphe présente plusieurs problèmes potentiels :

— **Pertinence contextuelle** : Les relations extraites de termes dans le même paragraphe sont généralement plus pertinentes et précises, car les termes sont plus susceptibles d'être liés de manière significative dans un contexte plus restreint. Le faible pourcentage indique que la majorité des relations peuvent manquer dans ce contexte précis, rendant certaines connexions moins significatives.
— **Généricité des relations** : Avec une si faible proportion de relations provenant du même paragraphe, il y a un risque accru que les relations extraites soient trop génériques. Les termes liés peuvent être associés de manière vague ou abstraite, sans une ancre contextuelle forte qui les lie de manière pertinente et spécifique. Les relations génériques peuvent manquer de précision et ne pas refléter les nuances et les détails nécessaires pour des applications spécifiques, comme la compréhension fine du contenu ou les inférences sémantiques précises.
— **Qualité des relations extraites** : La qualité des relations est cruciale pour la construction de graphes de connaissances (et même la recherche d'informations et l'analyse sémantique). Un faible pourcentage de relations provenant du même paragraphe peut indiquer que beaucoup de relations extraites sont basées sur des associations faibles ou fortuites, ce qui peut dégrader la qualité globale des données relationnelles.

Le faible pourcentage (1,9%) de relations provenant du même paragraphe soulève des préoccupations quant à la pertinence et à la spécificité des relations extraites. Pour améliorer la pertinence, il pourrait être nécessaire d'affiner les méthodes d'extraction pour mieux capturer les relations contextuelles étroites, ce qui aiderait à éviter des relations trop génériques et à renforcer la qualité des connexions sémantiques extraites.

Détail nombre	**Nombre total (pourcentage)**
Termes clés initiaux	1 277
Instances de relations / paires de termes	87 650
Relations du même domaine	25 704 (29,3%)
Relations d'un même document	14 498 (16,5%)
Relations d'un même document et même domaine (un document a plusieurs domaine)	14 240 (16,2 %)
Relations d'un même paragraphe ($\Longrightarrow$ même document et même domaine)	1 660 (1,9 %)

TABLE 5.6 – Statistiques réalisées sur les relations extraites (Echantillon pour les 4 types de relations : *syn, isa, hypo, has_part*)

Le tableau 5.7 présente des exemples d'instances de relations extraites à partir des termes

clés d'un même paragraphe, dit brique, chacune accompagnée de son type de relation et d'un poids associé. Par exemple, « vote » et « blanc » sont liés par une relation hyperonymique (*r_hypo*) avec un poids de 101, tandis que « infirmier » et « ambulancier » sont considérés synonymes (*r_syn*) avec un poids de 56. Bien que ces relations soient correctes et pertinentes dans leur contexte juridique, elles manquent de profondeur et de spécificité nécessaire pour générer une compréhension approfondie du texte. En effet, des relations telles que « communication » et « note » ou « examen » et « concours » peuvent être valides, mais elles n'apportent pas assez de détails ou d'information pour créer une véritable connaissance contextuelle. Elles restent trop génériques, ne permettant pas de capturer les nuances spécifiques du domaine juridique. Cette limitation peut rendre difficile l'utilisation de ces relations pour la construction de graphes de connaissances liés aux domaines juridiques et pratiques car elles n'offrent pas une granularité suffisante pour distinguer des concepts clés et leurs interrelations de manière significative.

Terme1	**Terme2**	**Type de relation**	**poids**
vote	blanc	r_hypo	101
infirmier	ambulancier	r_syn	56
communication	note	r_syn	220
territoire	municipalité	r_hypo	25
inscription	mention	r_hypo	25
protection	inscription	r_hypo	25
examen	concours	r_syn	35
zone	plage	r_hypo	35

TABLE 5.7 – Détail des instances de relations de même brique

Pour résumé, l'analyse des statistiques sur les relations extraites à partir des termes clés révèle plusieurs défis significatifs liés à la pertinence contextuelle et à la spécificité des relations. Bien que 29,3% des relations concernent des termes appartenant au même domaine, et 16,5% des termes apparaissent dans le même document, seulement 1,9% des relations extraites proviennent du même paragraphe. Ce faible pourcentage suggère que la majorité des relations peuvent manquer du contexte précis nécessaire pour une pertinence et une précision accrues. Les relations extraites de termes situés dans le même paragraphe sont souvent plus pertinentes et spécifiques, car elles bénéficient d'un contexte restreint et significatif. La faible proportion de telles relations indique que de nombreuses connexions établies sont probablement trop génériques et manquent de l'ancrage contextuel fort, crucial pour des applications spécifiques comme la compréhension fine du contenu ou les inférences sémantiques précises. Un faible pourcentage de relations issues du même paragraphe peut indiquer une prévalence d'associations faibles ou fortuites, dégradant ainsi la qualité globale des données relationnelles. Pour surmonter ces limitations et améliorer la pertinence des relations extraites, nous nous orientons vers l'extraction de relations ouvertes. Cette approche vise à capturer des relations plus contextuelles, basées directement sur les éléments du texte, afin de renforcer la spécificité et la qualité des connexions sémantiques extraites. En affinant les méthodes d'extraction pour mieux saisir les relations contextuelles étroites, nous espérons éviter les relations trop génériques et améliorer significativement la pertinence et l'utilité des relations extraites pour des applications concrètes dans le domaine juridique et au-delà.

CHAPITRE 6

Extractions de relations ouvertes à l'aide des LLMs

Le chapitre précédent a exposé un système d'extraction de relations utilisant le réseau lexical rezoJDM, lequel présente plusieurs limitations. D'abord, il dépend d'une liste fixe de types de relations, nécessitant la création de nouveaux classifieurs et apprentissages pour chaque nouveau type, ce qui manque de flexibilité. Ensuite, son niveau d'extraction est trop générique, se concentrant sur des termes extraits plutôt que sur les textes eux-mêmes, manquant ainsi les verbes d'action spécifiques au corpus de spécialité. Enfin, même au sein de sa propre structure, il reste trop général, comme discuté précédemment. Ce chapitre propose une nouvelle technique d'extraction de relations, visant à surmonter ces limitations en utilisant des approches dynamiques telles que l'OpenIE et en focalisant l'extraction sur les verbes d'action spécifiques aux textes, offrant ainsi une extraction plus précise et adaptée aux besoins spécifiques des corpus étudiés. Le chapitre précédent a exposé les limitations des approches basées sur des listes fixes de types de relations. Ce chapitre propose des techniques d'extraction de relations ouvertes, notamment à l'aide de grands modèles de langage (LLMs). Les défis consistent à améliorer la flexibilité et l'adaptabilité des systèmes d'extraction de relations pour mieux répondre aux besoins spécifiques des corpus de spécialité étudiés.

Sommaire du chapitre

6.1 Extraction de relations ouvertes par LLM **139**
6.2 Clustering d'instances similaires et découverte de relations **142**
6.3 Évaluation **145**
6.3.1 Extraction de relations ouvertes par LLM 146
6.3.2 Clustering de triplets de relations 153
6.3.3 Bilan 158
6.4 Discussion **159**
6.4.1 Bilan 159
6.4.2 Discussion 159

Le chapitre précédent a présenté un système d'extraction de relations basé sur le réseau lexical rezoJDM. Bien que ce système offre une certaine robustesse et une couverture

linguistique intéressante, il présente plusieurs limitations qui nécessitent l'exploration de techniques alternatives pour une extraction de relations plus fine et adaptée à des besoins spécifiques.

1. **Limitation de la liste fixe des types de relations :** Le système actuel repose sur une liste prédéfinie de types de relations. Cela signifie que pour chaque nouveau type de relation à détecter, il est impératif de créer un nouveau classifieur et de procéder à un nouvel apprentissage. Cette rigidité limite la flexibilité et l'adaptabilité du système, surtout dans des domaines où les relations entre les concepts évoluent rapidement. Pour pallier cette limitation, nous nous tournons vers des approches comme l'*Open Information Extraction* (OpenIE), qui permettent d'extraire des relations sans dépendre d'une liste fixe et préalablement définie. L'OpenIE offre une solution plus dynamique et adaptable, capable de s'ajuster à de nouveaux types de relations émergentes sans nécessiter une refonte complète du système.
2. **Niveau d'extraction trop générique :** Une autre limitation majeure du système actuel est son niveau d'extraction trop générique. Le système ne se limite pas à extraire des relations à partir des textes, mais se concentre plutôt sur les termes extraits. Cette approche globale ne permet pas de capturer les nuances et les spécificités des actions décrites dans les textes. Pour une extraction plus précise et utile, surtout dans un corpus de spécialité, il est essentiel de se concentrer sur les verbes d'action spécifiques aux textes analysés. Cela implique le besoin de techniques d'extraction basées directement sur les textes pour identifier et extraire les verbes d'action pertinents, qui reflètent mieux les relations spécifiques du domaine étudié.
3. **Généralité persistante au sein de la même brique :** Même au sein de la même brique (ou paragraphe) du système, une trop grande généralité persiste. Les limitations abordées dans le chapitre précédent montrent clairement que l'approche actuelle ne parvient pas à capturer les relations de manière suffisamment fine et détaillée. Il est donc crucial de repenser l'architecture du système pour permettre une extraction plus granulaire et contextuellement pertinente.

Ce chapitre présente une nouvelle technique d'extraction de relations visant à surmonter les limitations identifiées. En intégrant des approches plus dynamiques comme l'OpenIE et en focalisant l'extraction sur des éléments textuels spécifiques comme les verbes d'action, nous visons à améliorer la précision et la pertinence des relations extraites. Cette nouvelle technique permettra non seulement de répondre aux besoins actuels de flexibilité et d'adaptabilité, mais également de fournir des relations plus spécifiques et contextuellement adaptées aux corpus de spécialité étudiés.

Dans ce chapitre, nous explorons une nouvelle technique d'extraction de relations, appelée *Open Information Extraction* (OpenIE). Contrairement aux systèmes traditionnels d'extraction de relations, qui reposent sur des schémas relationnels prédéfinis et ne peuvent pas fonctionner efficacement dans des scénarios ouverts, l'OpenIE offre une flexibilité accrue en permettant la découverte de types de relations inédites à partir de données non supervisées.

Definition 6.1. *L'Open Information Extraction (OpenIE) est une approche de l'extraction d'informations qui vise à identifier et extraire automatiquement des relations et faits à partir de textes non structurés, sans avoir besoin de schémas prédéfinis ou de modèles spécifiques. Contrairement aux méthodes traditionnelles d'extraction d'informations qui nécessitent des ontologies ou des structures de données définies à l'avance, OpenIE utilise des techniques de traitement du langage naturel pour générer des triplets (sujet, relation, objet) à partir de données textuelles brutes. Ces triplets capturent les relations sémantiques implicites dans les phrases, facilitant ainsi l'analyse et l'organisation des*

connaissances extraites pour diverses applications comme la compréhension du langage, la recherche d'information, et la construction de graphes de connaissances.

Cette méthode n'est pas limitée à un ensemble restreint de relations cibles connues à l'avance et est indépendante du domaine fixé initialement, rendant ainsi l'extraction de relations plus adaptable et universelle. L'utilisation d'un LLM, comme celui fourni par l'API d'OpenAI pour la génération de texte (ChatGPT) , s'avère particulièrement pertinente pour l'extraction de relations en monde ouvert. Grâce à sa capacité de zero-shot et few-shot learning, le LLM peut générer des relations pertinentes à partir d'un corpus non annoté en relations. Cela signifie qu'il peut identifier et extraire des relations sans nécessiter de grands volumes de données annotées au préalable, ce qui est essentiel dans des contextes où les relations ne sont pas préalablement définies. Nous avons choisi d'utiliser le LLM de l'API d'OpenAI pour la génération de texte (ChatGPT) pour notre système d'extraction de relations. Ce modèle permet une extraction d'informations entre les termes d'un texte de manière flexible et dynamique. Par exemple, il peut identifier des verbes liant deux termes, ce qui ouvre de nombreuses possibilités pour découvrir des relations spécifiques et contextuelles.

6.1 Extraction de relations ouvertes par LLM

Nous utilisons l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'extraire des triplets de relations à partir de la liste terminologique d'un texte. Le système décrit est un processus d'extraction de relations sémantiques en monde ouvert à partir de textes, basé sur l'utilisation de termes clés pertinents. L'objectif est d'identifier des triplets de relations (Sujet, Verbe, Objet) dans un texte donné, où les termes clés peuvent apparaître comme sujet, verbe ou objet dans ces triplets. Il s'agit de l'ensemble des termes clés extraits par notre système d'extraction terminologique explicité au chapitre 4. Une fois les termes pertinents identifiés, le processus consiste à injecter un texte donné ainsi que la liste de termes clés donnée pour ce texte afin d'extraire des triplets (Sujet, Verbe, Objet) qui incluent ces termes. Les verbes dans les triplets sont à leur forme infinitive. Le prompt est donné sur l'exemple 6.1. Le principe de ce prompt consiste à pour identifier dans le texte tous les verbes et leurs relations associées. Ensuite, il vérifie si les termes pertinents apparaissent dans les triplets (en tant que Sujet, Verbe ou Objet). Et enfin, les verbes doivent être convertis en leur forme infinitive. Les triplets extraits sont formatés en JSON.

Soit le texte entre les balises <text> et </text>

<text> TEXT </text>

Soit la liste des termes pertinents extraits de ce texte : LISTE TERMES PERTINENTS

A partir des termes pertinents de ce texte, pour chaque verbe de ce texte qui leur est associé, donne moi les triplets qui contiennent les termes pertinents de ce texte. Ces triplets sont de type (Sujet, Verbe, Objet) où les termes pertinents peuvent se retrouver n'importe ou dans le triplet et ce, même en partie. Les verbes doivent être dans forme infinitive. Donne moi les résultats au format JSON.

Exemple 6.1 – Prompt générique utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'obtenir des triplets de relations à partir des termes pertinents d'un texte.

Légende :

⟶ **Input** :

Texte : texte fourni entre les balises <text> et </text>

Liste des termes clés : liste de termes pertinents donnée pour le texte. Il s'agit de l'ensemble des termes clés extraits par notre système d'extraction terminologique explicité au chapitre 4

⟶ **Extraction des Triplets :**

But : identifier dans le texte tous les verbes et leurs relations associées.

Vérification du fait que les termes pertinents apparaissent dans les triplets (en tant que Sujet, Verbe ou Objet).

Les verbes doivent être convertis dans leur forme infinitive.

⟶ **Output :**

JSON : Les triplets extraits sont formatés en JSON.

Un exemple de *prompt* donné à l'API d'OpenAI pour la génération de texte (ChatGPT) en intégrant un paragraphe issu des textes de Berger-Levrault associé à ses termes clés est présenté sur l'exemple 6.2.

Soit le texte entre les balises <text> et </text>

<text> Initiée par l'ordonnance n°2005-1527 du 8 décembre 2005, la réforme du permis de construire procède du mouvement de simplification du droit entamée depuis une quinzaine d'années. Entre objectif constitutionnel d'intelligibilité et inéluctable inflation législative, le permis de construire s'est essentiellement construit par voie d'ordonnances avec les avantages et inconvénients que cela implique. Une modernisation selon un projet de loi habilitant le gouvernement à simplifier le droit est toujours évoquée. </text>

Soit la liste des termes pertinents extraits de ce texte : Ordonnance n°2005-1527; Réforme du permis de construire; Simplification du droit; Objectif constitutionnel d'intelligibilité; Inflation législative; Permis de construire; Voie d'ordonnances; Avantages et inconvénients; Modernisation Projet de loi habilitant le gouvernement

A partir des termes pertinents de ce texte, pour chaque verbe de ce texte qui leur est associé, donne moi les triplets qui contiennent les termes pertinents de ce texte. Ces triplets sont de type (Sujet, Verbe, Objet) où les termes pertinents peuvent se retrouver n'importe ou dans le triplet et ce, même en partie. Les verbes doivent être dans forme infinitive. Donne moi les résultats au format JSON.

- -

OUTPUT:

— "Sujet" : "Ordonnance n°2005-1527", "Verbe" : "initier", "Objet" : "réforme du permis de construire"
— "Sujet" : "Réforme du permis de construire", "Verbe" : "procéder", "Objet" : "simplification du droit"
— "Sujet" : "Permis de construire", "Verbe" : "se construire", "Objet" : "voie d'ordonnances"
— "Sujet" : "Voie d'ordonnances", "Verbe" : "impliquer", "Objet" : "avantages et inconvénients"
— "Sujet" : "Modernisation", "Verbe" : "évoquer", "Objet" : "projet de loi habilitant le gouvernement"

Exemple 6.2 – Exemple de prompt utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'obtenir les triplets de relations à partir des termes pertinents d'un paragraphe issu du corpus de Berger-Levrault. La sortie présentée n'est pas en JSON afin de ne pas surcharger le visuel.

6.2 Clustering d'instances similaires et découverte de relations

Pour structurer les relations découvertes, nous utilisons un processus de clustering. Les termes similaires reliés par un type de relation sont regroupés en clusters. Par exemple, un cluster de mots peut lier un cluster de termes, facilitant ainsi l'organisation des relations découvertes. Lorsqu'un nouveau lien est trouvé, il est comparé avec les clusters existants : (1) Si le lien est très proche d'un cluster existant, il est ajouté à ce cluster. (2) Si aucun cluster proche n'existe, un nouveau cluster est créé pour ce lien. Différemment, le clustering par prompting se distingue comme une technique innovante pour structurer les relations découvertes. Contrairement à la méthode traditionnelle de comparaison des nouveaux liens avec des clusters existants, le clustering par prompting utilise des prompts spécifiques pour guider le regroupement des triplets relationnels. Cette méthode repose sur l'analyse des triplets extraits et leur regroupement en clusters selon des types de relations généralisés.

Le prompt donné dans l'exemple 6.3 consiste à traiter un ensemble de triplets relationnels extraits d'un domaine donné. Ces triplets sont initialement présentés avec des verbes d'action spécifiques comme types de relations. L'objectif est de regrouper ces triplets en clusters, selon des types de relations plus généraux, tout en préservant le sens global de chaque triplet. Ce prompt requiert trois fondamentaux à respecter. Tout d'abord, en ce qui concerne le *Clustering de triplets* en lui même, le système doit analyser ces triplets, identifier les relations similaires ou associées, et les regrouper en clusters. Ensuite, un fondamental à respecter concerne la **généralisation des types de relations**. En effet, les verbes d'action spécifiques dans chaque triplet doivent être remplacés par des verbes plus généraux qui maintiennent le sens global des relations dans chaque cluster. Par exemple, des verbes comme « construire » et « édifier » pourraient être regroupés sous un verbe général comme « bâtir ». Enfin, le dernier fondamental concerne **l'évaluation de la confiance**. Pour chaque triplet et chaque type de relation attribué, un score de confiance est accordé, indiquant la pertinence du triplet pour le type de relation donné. Le résultat final, présenté au format JSON, montre les clusters de triplets avec leurs nouveaux types de relations généralisés et les scores de confiance associés.

Soit le fichier JSON suivant :

TRIPLETS AU FORMAT JSON

Il s'agit de triplets de relations extraits pour le domaine de NOM DOMAINE. Fais un cluster de ces triplets de relations en regroupant les types de relations que l'on pourrait associer en leur attribuant un type de relation commun. Pour le moment, les types de relations sont les verbes d'action de chaque triplet mais le but est de trouver un verbe qui maintiendrait le sens global des triplets d'un même cluster. Accorde un score de confiance à chacun des triplets pour un type de relations donné.
Donne le résultat au format JSON

Exemple 6.3 – Prompt générique utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin de regrouper les triplets de relations en cluster à partir d'un ensemble de triplets lié à un domaine

Légende :

⟶ **Input** :

Fichier JSON (triplets de relations) : Un triplet de relations typique dans ce contexte est de la forme (sujet, verbe d'action, objet), où le verbe d'action représente la relation entre le sujet et l'objet.

⟶ **Clustering de triplets :**

Le système doit analyser ces triplets et les regrouper en clusters. Chaque cluster doit regrouper des triplets qui ont des relations similaires ou associées.

Généralisation des Types de Relations : les verbes d'action spécifiques de chaque triplet doivent être remplacés par des verbes plus généraux, comme regrouper « construire » et « édifier » sous « bâtir ».

Évaluation de la Confiance : Pour chaque triplet et pour chaque type de relation attribué, le système doit accorder un score de confiance. Ce score indique à quel point le triplet est pertinent pour le type de relation donné.

⟶ **Output :**

JSON : Le résultat final doit être présenté au format JSON, montrant les clusters de triplets avec leurs nouveaux types de relations généralisés et les scores de confiance associés.

Le tableau 6.1 présente des exemples de triplets de relations regroupés en cluster pour le domaine de l'urbanisme. Les relations similaires ont été regroupée sous un verbe d'action commun. Chaque triplet a un score de confiance basé sur la pertinence de son verbe d'origine avec le verbe regroupant les relations du même cluster. Chaque triplet a donc été attribué à un cluster en fonction de la similitude de ses relations avec d'autres triplets.

Cluster	Verbe	Relations	Confiance
Déclaration	déclarer	(Conseil constitutionnel, déclarer, conforme à la Constitution)	1.0
Déclaration	déclarer	(Conseil constitutionnel, déclarer, exclusion par l'article L. 321-1 du préjudice moral des préjudices réparés par l'indemnité d'expropriation)	1.0
Déclaration	déclarer	(Conseil constitutionnel, déclarer, fixation de l'indemnité en cas de mutation récente)	1.0
Déclaration	déclarer	(déclaration de cession, faire l'objet, d'une déclaration de cession conforme au formulaire Cerfa no 13644*02)	0.9
Déclaration	déclarer	(DC, constituer, le guichet unique pour le dépôt des déclarations de cession)	0.8
Déclaration	déclarer	(DC, être établie, en quatre exemplaires adressés par LRAR ou déposés contre récépissé à la mairie de la commune dont dépend le fonds ou l'immeuble objet du bail)	0.8
Vérification	vérifier	(juge administratif, vérifier, expropriation n'est pas inutile)	1.0
Vérification	vérifier	(juge administratif, vérifier, autres possibilités qui s'offrent à la collectivité expropriante ou compte tenu de projets similaires existant dans la commune)	1.0
Vérification	vérifier	(juge administratif, s'assurer, nécessité spatiale de l'expropriation)	0.9
Vérification	vérifier	(juge administratif, s'assurer, nécessité spatiale de l'expropriation)	0.9
Vérification	vérifier	(Conseil d'État, s'assurer, nécessité de recourir à la procédure d'expropriation)	0.9
Reconnaissance	reconnaître	(Conseil constitutionnel, reconnaître, constitutionnalité de la procédure d'expropriation d'urgence)	1.0
Reconnaissance	reconnaître	(Convention européenne des droits de l'Homme, reconnaître, droit pour les pouvoirs publics de recourir à la procédure d'expropriation)	1.0

Cluster	Verbe	Relations	Confiance
Reconnaissance	reconnaître	(Cour européenne des droits de l'Homme, reconnaître, existence d'un droit à une indemnité raisonnable)	1.0
Imposition	imposer	(Convention européenne des droits de l'Homme, imposer, respect des biens et de la propriété)	1.0
Imposition	imposer	(Cour européenne des droits de l'Homme, exiger, indemnité soit versée dans des délais qui ne soient pas anormalement longs)	0.9
Indemnisation	indemniser	(préjudice subi par les propriétaires et exploitant, indemniser, préjudice subi)	1.0
Indemnisation	indemniser	(préjudice subi par les propriétaires et exploitant, indemniser, préjudice subi)	1.0
Pouvoir	pouvoir	(autorité administrative, pouvoir, prescrire des normes spéciales)	1.0
Pouvoir	pouvoir	(maire, pouvoir, prescrire les mesures d'urgence)	1.0
Pouvoir	pouvoir	(autorité compétente, pouvoir, déléguer ce pouvoir)	1.0
Action	ouvrir	(action, ouvrir, à la commune une action en nullité de la cession)	1.0

TABLE 6.1: Exemple de clustering réalisé à partir de quelques relations d'un fichier JSON du domaine de l'urbanisme

Légende :

⟶ **Cluster** : Nom du cluster

⟶ **Verbe** : Verbe d'action représentatif du cluster

⟶ **Relations** :Triplets de relations

⟶ **Confiance** : Pour chaque triplet et pour chaque type de relation attribué, le système a accordé un score de confiance. Ce score indique à quel point le triplet est pertinent pour le type de relation donné.

6.3 Évaluation

L'évaluation de cette nouvelle technique ne peut pas être réalisée de manière classique en raison de la dépendance des relations au texte et aux verbes spécifiques utilisés. Par conséquent, il est impossible d'utiliser un corpus standard pour cette évaluation. Nous proposons d'utiliser un autre LLM pour évaluer les résultats obtenus afin de comparer leur pertinence

par rapport au système présenté dans le chapitre précédent (c.f. 5). Cette approche nous permet d'obtenir un aperçu qualitatif de l'efficacité de notre nouvelle méthode d'extraction de relations en monde ouvert.

6.3.1 Extraction de relations ouvertes par LLM

Description de l'évaluation

Dans cette sous-section, nous nous concentrons sur l'évaluation de l'extraction de relations ouvertes. Nous nous intéressons à la totalité des relations qui n'ont pas été regroupées sous un même type de relation ou verbe d'action. Le prompt donné dans l'exemple 6.4 sert à l'évaluation des triplets de relations extraits. En effet, étant donné le manque de labels pour l'échantillon du corpus de Berger-Levrault, le système est évalué auprès du LLMs de l'API d'OpenAI pour la génération de texte (ChatGPT) . Dans le cadre de l'analyse et de l'évaluation de triplets JSON relatifs à un domaine donné, nous avons procédé à une évaluation en attribuant une note de pertinence et un score de confiance à chaque triplet. Cette démarche vise à déterminer la pertinence des informations fournies par rapport à un domaine donné. Afin d'évaluer la pertinence ainsi que la confiance des triplets extraits pour un domaine donné, nous fournissons le prompt à l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'attribuer une note de pertinence entre 1 et 5 ainsi qu'un score de confiance flottant compris entre 0 et 1 (pourcentage %) pour chaque triplet extrait, en fonction de leur pertinence par rapport au domaine spécifié. Une note élevée (proche de 5) indique une grande pertinence et un score de confiance proche de 1 indique une haute confiance. Le résultat final est présenté au format JSON, incluant la moyenne des scores et des pourcentages de confiance. En évaluant ainsi la pertinence par domaine, nous nous assurons que les triplets extraits correspondent bien aux attentes et aux exigences spécifiques du domaine souhaité.

Soit le fichier JSON entre les balises <text> et </text>

<text> TEXT </text>

Donne une note entre 1 et 5 ainsi qu'un score flottant de confiance compris entre 0 et 1 pour chacun des triplets afin d'évaluer la pertinence par rapport au domaine juridique et pratique pour l'urbanisme
Donne le résultat au format JSON

Exemple 6.4 – Prompt générique utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'évaluer les triplets de relations extraits.

Le tableau 6.2 présente des exemples de triplets extraits associés aux scores donnés ainsi qu'une justification du score. Il répertorie plusieurs cas d'évaluation. Dans cet exemple, nous avons évalué la pertinence de chaque triplet en fonction de sa contribution potentielle à la compréhension et à l'application des règles et pratiques dans le domaine de l'urbanisme.

Triplets extraits	Scores attribués	Justification
(Établissement public de coopération intercommunale, se prononcer sur, le retrait)	Note : 3 Confiance : 0.7	Le triplet est pertinent pour comprendre le rôle des établissements publics de coopération intercommunale (EPCI) dans le processus de retrait, mais manque de détails spécifiques.
(Conseil municipal, se prononcer sur, le retrait)	Note : 3 Confiance : 0.7	Pertinent pour les délibérations locales, il est crucial dans le cadre de décisions municipales mais reste générique.
(Commune, se retirer de, la communauté urbaine)	Note : 4 Confiance : 0.8	Très pertinent car il aborde la capacité des communes à quitter des structures intercommunales, une question fréquente en urbanisme.
(Commune, se retirer de, la communauté d'agglomérations)	Note : 4 Confiance : 0.8	Similaire au cas précédent, avec une application spécifique aux communautés d'agglomérations, donc très pertinent.
(Commune, se retirer de, la communauté de communes)	Note : 4 Confiance : 0.8	Encore une fois, pertinent pour la compréhension des relations intercommunales, avec une application directe.
(Établissement public de coopération intercommunale, se retirer de, la communauté de communes)	Note : 5 Confiance : 0.95	Important pour les EPCI et leurs relations avec les communautés de communes, reflétant des dynamiques clés en urbanisme.
(Commune, appartenir au, syndicat mixte)	Note : 4 Confiance : 0.8	Pertinent bien que l'appartenance à un syndicat mixte est une réalité administrative courante (pertinent pour le graphe de connaissances)

TABLE 6.2 – Exemple de scores de pertinence et de confiance moyen attribués par les LLMs pour un triplet donné en fonction du domaine d'urbanisme avec la justification donnée par le LLM associée.

Légende :

$\longrightarrow$ **Scores attribués** :

Score de pertinence : Note accordée par un LLM entre 1 et 5 pour chacun des triplets afin d'évaluer la pertinence par rapport à un domaine donné.

Score de confiance : Score de confiance en pourcentage accordé par un LLM pour chacun des triplets par rapport à un domaine donné.

Calculs réalisés sur des exemples de triplets extraits pour chacun le domaine de l'urbanisme.

Domaine	Score de pertinence /5	Score de confiance
Commande publique	3,9	77,86 %
Comptabilité et finances locales	4,1	79,14 %
Élections	4,4	83,35 %
État civil et cimetières	3,9	82,5 %
Justice	4,6	79,42 %
RH territoriales	3,7	68,93 %
Santé	3,7	74,74 %
Urbanisme	3,7	75,38 %

TABLE 6.3 – Score de pertinence et de confiance moyen attribué aux différents triplets relationnels par les LLMs pour chacun des domaines

Légende :

⟶ **Calculs** :

Score de pertinence : Note accordée par un LLM entre 1 et 5 pour chacun des triplets afin d'évaluer la pertinence par rapport à un domaine donné.

Score de confiance : Score de confiance en pourcentage accordé par un LLM pour chacun des triplets par rapport à un domaine donné.

Calculs réalisés sur l'ensemble des triplets extraits pour chacun des domaines.

Résultats

Le tableau 6.3 présente les résultats de l'extraction de relations en OpenIE, évaluant la pertinence et la confiance des triplets extraits par rapport à divers domaines. La pertinence est notée sur une échelle de 1 à 5, avec des scores plus élevés indiquant une meilleure adéquation des triplets avec le domaine. Les domaines avec les scores de pertinence les plus élevés sont la *Justice* (4,6) et les *Élections* (4,4), ce qui suggère que les triplets extraits dans ces domaines sont perçus comme particulièrement pertinents. Les domaines avec les scores de pertinence les plus bas sont les *RH territoriales*, la *Santé* et l'*Urbanisme* (3,7 chacun), indiquant une pertinence perçue moindre. Le score de confiance est donné en pourcentage, représentant la confiance accordée aux triplets extraits. Les domaines avec les scores de confiance les plus élevés sont les *Élections* (83,35%) et l'*État civil et cimetières* (82,5%), montrant une grande confiance dans la pertinence des triplets. Les *RH territoriales* affichent le score de confiance le plus bas (68,93%), suggérant une moindre certitude dans ce domaine.

En général, il semble y avoir une corrélation positive entre la pertinence et la confiance. Par exemple, les *Élections* et la *Justice* ont des scores élevés dans les deux catégories. Cependant, il y a des exceptions comme la *Santé* et l'*Urbanisme*, qui ont des scores de pertinence moyens mais des scores de confiance relativement plus élevés. Les domaines avec une haute pertinence et confiance, tels que *Justice*, *Élections*, *État civil et cimetières* sont des domaines où les triplets extraits sont considérés à la fois très pertinents et fiables. Les domaines avec une pertinence et une confiance plus faibles tels que *RH territoriales*, *Santé*, *Urbanisme* peuvent nécessiter des améliorations dans l'extraction ou l'analyse des triplets pour améliorer la qualité perçue des résultats.

En résumé, les résultats montrent une bonne performance générale dans l'extraction de relations pour les domaines clés comme la *Justice* et les *Élections*. Pour les domaines avec

des scores plus bas, des ajustements dans les techniques d'extraction ou dans le *prompt* du modèle utilisé peuvent être nécessaires pour améliorer la pertinence et la confiance des résultats. Cela indique que, bien que l'extraction de relations soit globalement efficace, il y a encore des marges de progression dans certains domaines spécifiques pour atteindre une qualité uniforme.

Domaine	Nbre relations par article	Place dans la relation	Nbre de mots moyen	% multimots
Urbanisme	30	Source	2,63	59,54 %
		Cible	6,89	80,92 %
		Verbe action relation	1,22	18,24 %
Santé	17	Source	2,35	48,86 %
		Cible	6,1	81,11 %
		Verbe action relation	1,54	39,1 %
RH territoriales	30,6	Source	2	41,36 %
		Cible	4,22	61,92 %
		Verbe action relation	1,1	5,37 %
Justice	27	Source	1,52	29,63 %
		Cible	3,73	65,84 %
		Verbe action relation	1,34	27,16 %
État civil et cimetières	14,7	Source	5,02	69,9 %
		Cible	8,87	83,98 %
		Verbe action relation	1,27	16,51 %
Élections	26,4	Source	6,05	70,81 %
		Cible	7,17	97,84 %
		Verbe action relation	2,31	61,1 %
Comptabilité et finances locales	12,86	Source	2,23	55,56 %
		Cible	4,64	100 %
		Verbe action relation	1,33	28,89 %
Commande publique	28,43	Source	1,96	59,1 %
		Cible	3,91	43,47 %
		Verbe action relation	1,16	12,81 %

TABLE 6.4 – Statistiques effectuées sur les triplets extraits en fonction du domaine

Légende :

⟶ **Spécificité** :

Place dans la relation : Une relation du type (mot_source, Verbe_action_relation, mot_cible) est composée d'un mot « source », d'une relation représentée par un verbe d'action et d'un mot « cible ».

⟶ **Calculs** :

Nbre relations par article : Nombre moyen de triplets extraits par article pour le domaine donné.

Pourcentage multimots : pourcentage de multimots dans l'ensemble des termes soit source soit cible soit verbe d'action de la relation des triplets détectés

Nombre mots moyen : nombre moyen de mots par terme clef soit source soit cible soit verbe d'action de la relation des triplets détectés

Calculs réalisés sur l'ensemble des triplets extraits.

L'analyse des résultats obtenus dans le tableau 6.4 peut être structurée autour des différents aspects mentionnés : le nombre de relations par article, la place dans la relation, le nombre de mots moyen et le pourcentage de multimots.

1. **Nombre de relations par article :** Les domaines ayant un nombre élevé de relations par article (comme Urbanisme, RH territoriales et Commande publique) montrent une richesse en termes de relations extraites. Cependant, les domaines avec un nombre relativement faible de relations par article (comme Comptabilité et finances locales et État civil et cimetières) peuvent signaler une moindre densité d'information dans les articles ou une difficulté du système à extraire des relations pertinentes.
2. **Place dans la relation :** Les domaines comme Urbanisme et Élections montrent un nombre moyen de mots élevé pour les sources, ce qui peut indiquer des entités ou des concepts plus complexes en tant que sources. Les cibles sont souvent plus longues que les sources dans presque tous les domaines, avec des domaines comme État civil et cimetières et Élections ayant des cibles particulièrement longues. Cela peut être dû à la nature descriptive et détaillée des entités cibles dans ces domaines. Les verbes d'action relation sont généralement courts, mais des domaines comme Élections montrent des verbes plus longs, peut-être en raison de la spécificité et de la précision requises dans les relations décrites. Etant donné qu'une relation a été extraite par cette forme de « verbe d'action », il vaut mieux que le nombre moyen de mots ne dépasse pas 3 mots.
3. **Nombre de mots moyen :** Les domaines avec un nombre moyen de mots plus élevé, comme État civil et cimetières et Élections, peuvent impliquer des descriptions plus détaillées et spécifiques. En revanche, des domaines comme Justice et Commande publique, avec un nombre moyen de mots plus faible, pourraient impliquer des termes plus concis et directes.
4. **Pourcentage de multimots :** Un pourcentage élevé de multimots, comme dans les domaines des Élections et de la Santé, suggère l'utilisation de termes complexes et composés, ce qui pourrait refléter la complexité des concepts traités dans ces domaines. Cependant, un pourcentage faible de multimots, comme dans RH territoriales et Commande publique, pourrait indiquer une tendance à utiliser des termes simples et directs, mais cela pourrait aussi limiter la précision et la richesse des relations extraites.

Les légères différences peuvent être impliquées par la nature des données, la complexité conceptuelle ainsi que le formalisme et la rigueur des textes. En effet, les différences peuvent être attribuées à la nature intrinsèque des domaines. Par exemple, l'Urbanisme et les Élections peuvent nécessiter des descriptions plus détaillées et des relations plus complexes, d'où un nombre de mots moyen et un pourcentage de multimots plus élevés. Les domaines traitant de sujets plus techniques ou spécialisés (comme État civil et cimetières) ont tendance à utiliser plus de multimots et des termes plus longs pour décrire précisément les entités et leurs relations. Enfin, des domaines comme la Justice peuvent utiliser un langage plus formel et précis, ce qui se reflète dans des termes plus courts et moins de multimots. En somme, les résultats montrent une diversité en termes de richesse et complexité des relations extraites selon les domaines. Les domaines plus techniques ou nécessitant des descriptions détaillées tendent à avoir un nombre moyen de mots plus élevé et un pourcentage de multimots plus important, tandis que des domaines plus simples ou directs présentent des termes plus courts et moins de multimots. Comprendre ces nuances peut aider à adapter les méthodes d'extraction de relations pour mieux capturer les spécificités de chaque domaine.

Nous allons désormais comparer les résultats avec ceux obtenus dans le tableau 5.5, en

mettant en évidence les différences et les similarités. Pour faciliter la comparaison, nous allons analyser chaque métrique en détail.

1. **Nombre de mots moyen :** Le système actuel a obtenu des nombres de mots moyens pour les sources et les cibles qui varient considérablement selon les domaines, avec des valeurs allant de 1,52 à 6,05 pour les sources et de 3,73 à 8,87 pour les cibles. Le système précédent a quant à lui comptabilisé un nombre de mots moyen fixe : 1 pour les sources et 1 pour les cibles, indépendamment du domaine. Nous avions réalisé, cependant, les statistiques non pas sur chacun des domaines mais sur l'appartenance à un même domaine ou non. Dans le cas de l'OpenIE, le système extrait des relations locales et donc ces dernières sont du même domaine à coup sûr. Le système actuel montre une plus grande variabilité et complexité dans les termes extraits, ce qui reflète la spécificité des relations en fonction du contexte local.
2. **Pourcentage de multimots :** Les pourcentages de multimots du système d'OpenIE pour les sources varient de 29,63 % à 70,81 %, et pour les cibles de 43,47 % à 100 % selon les domaines. Le système précédent a obtenu un pourcentage de multimots pour les sources de 8,21 % au total, et pour les cibles de 0,89 % au total. Le système actuel montre des pourcentages de multimots beaucoup plus élevés, ce qui indique une plus grande utilisation de termes complexes ou composés. Cela peut être dû à une extraction de relations plus sophistiquée ou à une analyse de domaines intrinsèquement plus détaillés. La table précédente présente des pourcentages de multimots très bas, suggérant une tendance à utiliser des termes simples et directs. Nous constatons donc, de nouveau, une différence significative en terme de spécificité des relations et ainsi une amélioration dans la méthodologie d'extraction des relations.

La comparaison entre les deux système met en évidence une nette différence dans la richesse et la complexité des relations extraites. Le tableau 6.4 montre une variabilité et une sophistication accrues, reflétant probablement une analyse plus détaillée des domaines et une extraction de relations plus précise. En revanche, la tableau précédent (5.5) présente des valeurs fixes et des pourcentages de multimots très bas, indiquant une approche potentiellement plus simplifiée. Les domaines tels que Urbanisme, État civil et cimetières, et Élections montrent une richesse et une complexité accrues dans les termes et les relations, reflétant des informations plus détaillées et spécifiques dans l'extraction du système actuel. De plus, les domaines avec des termes plus longs et des pourcentages élevés de multimots montrent une précision accrues dans la description des entités et des relations.

Nous allons désormais confronter nos résultats de pertinence et de confiance des relations extraites (c.f. tableau 6.3) avec le nombre de multimots obtenus (c.f. tableau 6.4). Cela permet d'analyser notamment l'impact des multimots sur la pertinence et la confiance des relations extraites. Notre analyse se divise en trois croisements corrélés :

1. **Corrélation entre la proportion de multimots et la pertinence/confiance :** Les domaines comme les *Élections* et l'*État civil et cimetières* ont des scores élevés en pertinence et en confiance, ainsi qu'une proportion élevée de multimots (surtout pour les termes « cible »). À l'inverse, les *RH territoriales* montrent des scores plus bas en pertinence et en confiance, avec une proportion plus faible de multimots.
2. **Complexité des termes :** Les domaines avec des scores de pertinence et de confiance élevés (par exemple, *Justice*, *Élections*) tendent à avoir des termes plus complexes (multimots) dans les cibles, ce qui peut indiquer que des relations plus détaillées et spécifiques contribuent à une meilleure évaluation. C'est en effet ce qui est souhaité car, pour rappel, c'est ce que nous avions reproché au système précédent (c.f chapitre 5)

3. **Variabilité des verbes d'action :** Les domaines où les verbes d'action sont principalement des mots simples (par exemple, *RH territoriales*) ont des scores de pertinence et de confiance plus bas, ce qui pourrait signifier que des verbes plus spécifiques et complexes (multimots) pourraient améliorer la qualité des relations extraites. Cette corrélation sera à creuser dans les perspectives.

L'analyse suggère que la proportion et la complexité des multimots dans les triplets extraits influencent positivement les scores de pertinence et de confiance. Des relations plus détaillées et spécifiques, en particulier dans les termes cibles, semblent corrélées à une meilleure évaluation globale. Pour les domaines avec des scores plus faibles, augmenter la complexité et la spécificité des termes extraits pourrait améliorer les résultats.

6.3.2 Clustering de triplets de relations

Description de l'évaluation

Le prompt donné dans l'exemple 6.5 sert à l'évaluation des différents clusters regroupant les triplets de relations. En effet, étant donné le manque de labels pour l'échantillon du corpus de Berger-Levrault, le système est évalué auprès du LLMs de l'API d'OpenAI pour la génération de texte (ChatGPT) . Ce prompt consiste à évaluer les clusters de relations JSON relatifs à un domaine donné avec une note attribuée entre 1 et 5 ainsi qu'un score de confiance flottant compris entre 0 et 1 (pourcentage %) pour chaque cluster. Un score de pertinence élevé (proche de 5) indique une grande pertinence et un score de confiance proche de 1 indique une haute confiance. En évaluant ainsi la pertinence par domaine, nous nous assurons que les clusters extraits correspondent bien aux attentes et aux exigences spécifiques du domaine souhaité.

Soit le fichier JSON suivant :

CLUSTER DES TRIPLETS AU FORMAT JSON

Il s'agit de triplets de relations extraits pour le domaine NOM DOMAINE regroupés en clusters de même type de relation. Donne une note entre 1 et 5 ainsi qu'un score flottant de confiance compris entre 0 et 1 pour chacun des clusters afin d'évaluer sa pertinence par rapport au domaine juridique et pratique pour NOM DOMAINE
Donne le résultat au format JSON

Exemple 6.5 – Prompt générique utilisé dans l'API d'OpenAI pour la génération de texte (ChatGPT) afin d'évaluer les clusters des triplets de relations d'un domaine donné.

Le tableau 6.5 présente des exemples de cluster sous leur type de relation associés aux scores donnés ainsi qu'à une justification des scores. Il répertorie plusieurs cas d'évaluation.

Dans cet exemple, les notes et scores de confiance visent à évaluer la pertinence des clusters dans le contexte du domaine juridique et pratique de la comptabilité et des finances locales.

Cluster (Type de relation)	**Scores attribués**	**Justification du LLM**
Destiner	Note : 4 Confiance : 0.95	Le cluster traite de la destination des subventions vers diverses actions, pertinent dans le contexte de la comptabilité et des finances locales.
Soutenir	Note : 5 Confiance : 1	Les relations concernent le soutien aux actions et aux politiques locales, ce qui est crucial dans le domaine financier et comptable des collectivités locales.
Être	Note : 3 Confiance : 0.65	Les relations dans ce cluster sont liées à des modifications législatives et aux responsabilités de contrôle, mais avec un score de confiance moyen, indiquant une pertinence limitée dans ce contexte spécifique.
Prévoir	Note : 3 Confiance : 0.75	Ce cluster aborde les prévisions de dépenses et les mesures d'assainissement par les autorités locales, mais aussi avec un score de confiance moyen, ce qui diminue sa pertinence dans le domaine spécifique de la comptabilité et des finances locales.

TABLE 6.5 – Exemple de scores de pertinence et de confiance moyen attribués par les LLMs pour un cluster de relations donné en fonction du domaine *comptabilité et des finances locales* avec la justification associée.

Légende :

⟶ **Scores attribués** :

Score de pertinence : Note accordée par un LLM entre 1 et 5 pour chacun des clusters afin d'évaluer la pertinence par rapport à un domaine donné.

Score de confiance : Score de confiance en pourcentage accordé par un LLM pour chacun des clusters de relations par rapport à un domaine donné.

⟶ **Scores attribués** : Chaque type de relation regroupe des triplets ayant des verbes d'action similaires

⟶ **Justification** : Explication des notes attribuées par le LLM

Calculs réalisés sur des exemples de clusters de relations pour le domaine de *comptabilité et des finances locales*.

Résultats

Les résultats obtenus sur l'efficacité du clustering des triplets relationnels dans différents domaines sont analysés à partir des scores de pertinence et de confiance attribués par

les LLMs dans le tableau 6.6. Le score de pertinence évalue la qualité du regroupement des triplets dans chaque cluster par rapport au domaine donné. Les scores varient entre 3 (pour *RH territoriales*) et 4 (pour *Commande publique* et *Urbanisme*), ce qui suggère une bonne à très bonne pertinence des regroupements dans la plupart des domaines. Le score de confiance exprimé en pourcentage mesure la fiabilité ou la certitude des LLMs concernant les clusters formés pour chaque domaine. Des scores élevés comme 90,5% à 96% (pour Commande publique et Urbanisme) indiquent une confiance élevée dans les regroupements effectués. *Commande publique* et *Urbanisme* obtiennent les scores les plus élevés à la fois en pertinence (4) et en confiance (90,5% et 96% respectivement). Cela suggère que les clusters formés pour ces domaines sont à la fois très pertinents et très fiables. *Élections, Justice*, et *Santé* montrent également de bons résultats avec des scores de pertinence autour de 3,5 et des scores de confiance supérieurs à 88%. Ces domaines bénéficient donc de clusters qui sont jugés comme appropriés et sûrs par les LLMs. *État civil et cimetières* et *RH territoriales* montrent des scores légèrement inférieurs en pertinence (autour de 3) mais compensent avec des scores de confiance élevés (87% et 90,1% respectivement). *Comptabilité et finances locales* présente des scores légèrement inférieurs en pertinence et en confiance par rapport aux autres domaines, bien que toujours respectables (3 et 80%).

Globalement, les résultats montrent une efficacité raisonnable dans le clustering des triplets relationnels pour chaque domaine étudié. Les scores élevés de pertinence et de confiance dans plusieurs domaines indiquent que les clusters formés sont généralement bien adaptés aux types de relations spécifiques étudiés. Pour améliorer l'efficacité du clustering, il pourrait être intéressant d'approfondir l'analyse des clusters moins performants (comme *Comptabilité et finances locales*) pour identifier des pistes d'optimisation. Il pourrait aussi être intéressant d'employer d'autres techniques de clustering (ex : K plus proches voisins).

Domaine	**Score de pertinence**	**Score de confiance**
Commande publique	4	90,5%
Comptabilité et finances locales	3	80 %
Élections	3,8	88,4 %
État civil et cimetières	3,14	87 %
Justice	3,46	88,73%
RH territoriales	3	90,1 %
Santé	3,34	89,1 %
Urbanisme	4	96 %

TABLE 6.6 – Scores de pertinence et de confiance moyen attribué aux différents clusters par les LLMs pour chacun des domaines.

Légende :

⟶ **Calculs** :

Score de pertinence : Note accordée par un LLM entre 1 et 5 pour chacun des triplets afin d'évaluer la pertinence par rapport à un domaine donné.

Score de confiance : Score de confiance en pourcentage accordé par un LLM pour chacun des triplets par rapport à un domaine donné.

Calculs réalisés sur l'ensemble des triplets extraits pour chacun des domaines.

Les résultats du tableau 6.7 fournissent des statistiques sur les clusters construits à partir

des triplets pour différents domaines. La colonne des domaines liste les différents domaines analysés. Le nombre moyen de triplets par clusters indique la moyenne du nombre de triplets par clusters pour chaque domaine. Il semble que *Commande publique* ait le score le plus élevé (12,25), ce qui pourrait indiquer une forte densité de triplets ou de clusters dans ce domaine. En revanche, *RH territoriales* a le score le plus bas (1,14), suggérant une faible densité.

Le Pourcentage de clusters avec un seul triplet montre la proportion de clusters qui ne contiennent qu'un seul triplet. Les domaines comme *Justice*, *RH territoriales* et *Santé* ont des pourcentages très élevés (76,19 %, 85,71 % et 85,97 % respectivement), indiquant que la plupart des clusters dans ces domaines sont constitués d'un seul triplet. Cela pourrait indiquer une faible capacité de clustering ou une forte variabilité des triplets dans ces domaines. En revanche, des domaines comme *Urbanisme* et *Comptabilité et finances locales* ont des pourcentages très bas (1 %), ce qui signifie que presque tous les clusters dans ces domaines contiennent plusieurs triplets.

Le nombre maximum de triplets dans un cluster représente le nombre maximum de triplets trouvés dans un seul cluster pour chaque domaine. *Commande publique* a le plus grand nombre (32), suggérant une grande variabilité dans la taille des clusters dans ce domaine. D'autres domaines comme *RH territoriales* et *Comptabilité et finances locales* ont des nombres beaucoup plus faibles (2 et 3 respectivement), indiquant une taille de cluster plus uniforme et généralement plus petite.

En résumé, les domaines *Commande publique*, *Urbanisme* et *Comptabilité et finances locales* montrent une meilleure performance de clustering par rapport aux autres domaines, avec une plus grande capacité à regrouper les triplets similaires. Les domaines *Justice*, *RH territoriales* et *Santé* montrent une performance de clustering plus faible, avec une majorité de clusters contenant un seul triplet.

Domaine	Nombre de triplets x clusters	Pourcentage cluster avec un seul triplet	Nombre max triplet dans 1 cluster
Commande publique	12,25	25%	32
Comptabilité et finances locales	2,25	1 %	3
Élections	6,4	20 %	20
État civil et cimetières	2,43	50 %	6
Justice	1,67	76,19 %	11
RH territoriales	1,14	85,71 %	2
Santé	1,47	85,97 %	15
Urbanisme	4	1 %	8

TABLE 6.7 – Statistiques sur les différents clusters pour chacun des domaines.

Légende :

$\longrightarrow$ **Calculs** :

Nombre de triplets x clusters : moyenne du nombre de triplets par cluster pour chaque domaine.

Pourcentage cluster avec un seul triplet : proportion de clusters qui ne contiennent qu'un seul triplet.

Nombre max triplet dans 1 cluster : nombre maximum de triplets trouvés dans un seul cluster pour chaque domaine.

Calculs réalisés sur l'ensemble des clusters de triplets pour chacun des domaines.

Pour analyser la performance de clustering du système en comparant les deux tableaux (c.f. 6.6 et 6.7) et, nous allons considérer les scores de pertinence et de confiance par domaine et les mettre en relation avec les statistiques réalisées sur les différents clusters. Nous allons comparer les deux tableaux en trois catégories :

1. **Commande publique et Urbanisme :** Ces domaines montrent une forte densité de triplets par clusters, des pourcentages faibles de clusters avec un seul triplet et des scores élevés en pertinence et confiance, suggérant une bonne performance de clustering. La capacité du système à regrouper des triplets similaires est efficace.
2. **Comptabilité et finances locales, Élections, État civil et cimetières :** Bien que la densité des triplets par clusters varie, ils maintiennent des scores de pertinence et de confiance modérément élevés. La faible proportion de clusters avec un seul triplet pour « Comptabilité et finances locales » et « Urbanisme » indique une bonne performance de clustering.
3. **Justice, RH territoriales et Santé :** Ces domaines montrent des densités faibles de triplets par clusters et des pourcentages élevés de clusters avec un seul triplet, mais malgré cela, ils obtiennent des scores de confiance relativement élevés. Cela pourrait suggérer que bien que le clustering ne soit pas optimal, les triplets individuels sont toujours jugés pertinents et fiables.

En résumé, le système montre une meilleure performance de clustering dans les domaines où la densité de triplets par clusters est élevée et où le pourcentage de clusters avec un seul triplet est faible. Pour les domaines avec de faibles densités et de hauts pourcentages de clusters avec un seul triplet, la pertinence et la confiance restent élevées, indiquant que le

système est capable de fournir des informations fiables, même si le regroupement pourrait être amélioré.

Discussion

Le clustering des triplets de relations présente plusieurs limitations importantes. Premièrement, la gestion d'un grand nombre de triplets pose un défi majeur. En effet, il est nécessaire de traiter continuellement le flux de nouveaux triplets et de clusters. Une approche possible serait de regrouper les clusters existants, mais cela ne résout pas fondamentalement le problème de la gestion du flux. Alternativement, on pourrait vérifier les types de relations et regrouper tous les triplets partageant le même type de relation. Actuellement, la méthode utilisée consiste à concaténer tous les types de relations, ce qui peut entraîner la duplication des clusters. Cette méthode nécessite donc des améliorations pour éviter ces redondances et optimiser le regroupement des triplets.

Dans l'analyse des résultats, il est apparu que la technique de clustering par prompting, bien que puissante et flexible, présente certaines limites en termes de complexité et de gestion des relations extraites. La gestion d'un grand nombre de triplets et la nécessité de traiter continuellement le flux de nouveaux triplets et clusters posent des défis majeurs. Une méthode alternative serait de revenir à un clustering plus traditionnel, où chaque nouveau lien est immédiatement comparé aux clusters existants : (1) si le lien est très proche d'un cluster existant, il est ajouté à ce cluster ; (2) si aucun cluster proche n'existe, un nouveau cluster est créé pour ce lien. Cette approche permet une structuration plus directe et souvent plus intuitive des relations, réduisant le besoin d'interventions complexes et de raffinements continus des prompts. En adoptant cette méthode, on peut tirer parti de la robustesse et de la clarté offertes par des techniques de clustering éprouvées, tout en maintenant une cohérence élevée dans l'organisation des relations extraites. Tester cette approche pourrait offrir une gestion plus dynamique et continue des triplets, améliorant ainsi la qualité et la cohérence des clusters formés.

6.3.3 Bilan

L'évaluation de l'extraction de relations ouvertes par les LLMs montre une performance prometteuse en termes de diversité et de précision des relations extraites. Les modèles de langue, grâce à leur capacité à comprendre et générer du texte en langage naturel, ont démontré une aptitude à identifier des relations complexes et variées à partir de vastes corpus textuels. Cependant, des défis subsistent, notamment en ce qui concerne la gestion des ambiguïtés et la nécessité d'affiner les modèles pour des domaines spécifiques afin d'augmenter la pertinence des relations extraites. La comparaison avec des techniques traditionnelles a également révélé que les LLMs peuvent offrir des gains significatifs en termes de couverture et de flexibilité, bien que leur mise en œuvre nécessite des ressources computationnelles importantes.

L'évaluation du clustering des triplets de relations a mis en évidence les forces et les faiblesses de l'approche par prompting. Cette technique, qui consiste à regrouper les relations similaires en clusters cohérents, a montré des résultats satisfaisants en termes de précision et de cohérence des clusters. Cependant, la complexité de l'algorithme et la nécessité d'une configuration précise des prompts ont été identifiées comme des points nécessitant une attention particulière. La comparaison avec une méthode de clustering plus traditionnelle, où chaque nouveau lien est comparé aux clusters existants pour décider de son ajout ou de la création d'un nouveau cluster, a souligné la robustesse et la simplicité de cette dernière. Ainsi, bien que le prompting offre une flexibilité accrue, la méthode traditionnelle pourrait

s'avérer plus efficace dans des scénarios nécessitant une gestion simple et claire des relations.

6.4 Discussion

6.4.1 Bilan

Ce chapitre s'est concentré sur l'extraction de relations ouvertes à l'aide des LLMs. L'objectif principal était de démontrer comment les LLMs peuvent être exploités pour identifier et extraire des relations entre différentes entités à partir de vastes ensembles de données textuelles. En utilisant des techniques avancées telles que le clustering d'instances similaires et la découverte de relations, ce chapitre a permis de mettre en évidence la flexibilité et l'efficacité des LLMs dans ce domaine.

Les principales contributions de ce chapitre incluent :

1. **Extraction de relations ouvertes par LLM :** Cette section a montré comment les LLMs peuvent être utilisés pour extraire des relations sans se limiter à des schémas prédéfinis, permettant ainsi une plus grande adaptabilité et couverture des données textuelles.
2. **Clustering de triplets de relations :** En regroupant les triplets de relations similaires, il a été possible de découvrir des motifs récurrents et de structurer l'information extraite de manière plus cohérente et exploitable. Cette méthode permet de ne pas surcharger le nombre (infini) de type de relations puisque le type de relation est identifié en fonction du texte.
3. **Évaluation des performances :** Des méthodes rigoureuses ont été utilisées pour évaluer la précision et la robustesse des relations extraites, montrant ainsi la viabilité de l'approche basée sur les LLMs pour des tâches complexes d'extraction de relations.

6.4.2 Discussion

L'utilisation des LLMs pour l'extraction de relations ouvertes présente plusieurs avantages significatifs. Premièrement, leur capacité à comprendre et générer du texte dans des contextes variés permet une extraction plus précise et contextuellement appropriée des relations. Contrairement aux méthodes traditionnelles qui reposent sur des règles strictes et des schémas prédéfinis, les LLMs peuvent s'adapter à de nouvelles informations et à des domaines diversifiés sans nécessiter de réentraînement intensif.

Cependant, plusieurs défis demeurent. La qualité des relations extraites et leur pertinence peuvent varier considérablement en fonction des domaines et des prompts utilisés pour le clustering. De plus, les LLMs peuvent parfois générer des relations incorrectes ou incohérentes, particulièrement lorsque les données d'entrée sont ambiguës ou peu claires. Il est donc crucial de continuer à améliorer les techniques de prétraitement des données et d'affiner les modèles pour minimiser ces erreurs.

Une autre limitation concerne le coût computationnel associé à l'utilisation des LLMs. Leur entraînement et leur déploiement nécessitent des ressources considérables en termes de puissance de calcul et de mémoire, ce qui peut représenter un obstacle pour leur adoption à grande échelle. Les recherches futures devraient donc se concentrer sur l'optimisation des modèles pour les rendre plus accessibles et plus efficaces.

Enfin, l'intégration de techniques traditionnelles de comparaison de liens et de création de nouveaux clusters avec des approches basées sur les LLMs nécessite une gestion sophistiquée des données pour éviter les redondances et garantir la cohérence des clusters formés. Cette dualité entre méthodes traditionnelles et avancées offre une opportunité unique d'exploiter le meilleur des deux mondes, en utilisant les LLMs pour découvrir des relations nouvelles et inattendues tout en s'appuyant sur des techniques éprouvées pour valider et structurer ces découvertes.

En conclusion, bien que les LLMs ouvrent de nouvelles perspectives prometteuses pour l'extraction de relations ouvertes, il est essentiel de poursuivre les efforts de recherche pour surmonter les défis actuels et maximiser leur potentiel. Les travaux futurs devraient s'orienter vers l'amélioration de la précision, la réduction des coûts computationnels et l'implémentation d'une nouvelle technique de clustering, afin de rendre cette technologie pleinement opérationnelle et bénéfique pour un large éventail d'applications.

CHAPITRE 7

Conclusion et perspectives

Ce chapitre consiste à synthétiser les contributions de la thèse et discuter des directions futures. Les avancées réalisées dans cette thèse ouvrent de nouvelles perspectives pour l'extraction terminologiques et de relations dans le cadre de la construction de graphes de connaissances à partir de textes de spécialité.

Sommaire du chapitre

7.1 Bilan général **161**
7.2 Perspectives **163**
7.2.1 Extraction terminologique 163
7.2.2 Extraction de relations à l'aide du réseau lexical issu de JeuxDeMots 163
7.2.3 Extractions de relations ouvertes à l'aide des LLMs 164
7.2.4 Graphes de connaissances lexico-sémantiques 165

Cette thèse a exploré en profondeur la structuration des connaissances en graphes à partir de textes de spécialité du domaine public. Chaque chapitre a contribué de manière significative à l'atteinte de cet objectif global, en introduisant des systèmes innovants et en ouvrant des perspectives prometteuses pour les recherches futures. Cette conclusion récapitule les points marquants de chaque chapitre, en mettant l'accent sur les systèmes développés et les perspectives proposées.

7.1 Bilan général

En conclusion, ce manuscrit de thèse a exploré en profondeur les graphes de connaissances lexico-sémantiques comme une structure cruciale pour la représentation et l'exploitation de la connaissance textuelle. Après avoir posé les bases en introduisant les concepts fondamentaux et en posant le contexte nécessaire à notre exploration, nous avons abordé les défis significatifs liés à la définition et à l'implémentation de cette structure complexe, ainsi que notre approche pour enrichir ce graphe en utilisant des ressources existantes. Nous avons démontré comment les graphes de connaissances lexico-sémantiques peuvent servir de fondement robuste à diverses applications, notamment dans le domaine de l'analyse sémantique de texte. Cette thèse contribue à l'avancement des méthodes de représentation et

d'exploitation de la connaissance dans le domaine du traitement automatique du langage naturel (TALN).

Nous avons dressé un tableau détaillé de l'état de l'art en matière de construction automatique de bases de connaissances à partir de ressources textuelles spécialisées du secteur public en français. Après avoir introduit les concepts fondamentaux et le cadre théorique de cette thèse, nous avons examiné en profondeur les méthodes et approches existantes, évaluant leurs avantages et limitations spécifiques. Notre exploration a inclus une analyse approfondie de l'extraction d'information qui implique extraire la terminologie puis les relations. L'extraction terminologique a été explorée à la fois pour la reconnaissance d'entités nommées (NER) et l'extraction de termes pertinents (KPE). Nous avons identifié les lacunes des approches existantes et justifié l'intégration des larges modèles de langues (LLM) et des réseaux lexicaux comme JeuxDeMots pour améliorer l'extraction de relations.

Nous avons approfondi la thématique de l'extraction terminologique dans le contexte des graphes de connaissances lexico-sémantiques. Après avoir posé une base théorique solide, notre attention s'est portée sur le développement de l'extracteur terminologique *Info-Glean KeyTerms*, structuré en trois sous-systèmes distincts : un système pour la reconnaissance des entités nommées (NER), un autre pour l'extraction des termes pertinents (KPE), et un troisième dédié aux entités juridiques. Ces sous-systèmes ont été conçus pour gérer efficacement la variabilité terminologique et pour s'adapter aux spécificités des différents domaines, particulièrement ceux du secteur public en français. Une évaluation rigoureuse des performances de chaque sous-système a été réalisée afin de sélectionner les meilleurs modèles et techniques pour enrichir notre graphe de connaissances. En conclusion, ce chapitre a introduit et détaillé les critères essentiels ainsi que les méthodes utilisées pour l'enrichissement continu des termes clés extraits, renforçant ainsi notre approche dans la construction et l'exploitation de graphes de connaissances lexico-sémantiques pour le traitement avancé du langage naturel.

Ensuite, ce manuscrit de thèse a exploré de manière approfondie l'extraction de relations à travers l'utilisation du réseau lexical JeuxDeMots, capitalisant sur la base de termes clés préalablement extraits. En établissant un lien direct entre l'extraction terminologique et l'extraction de relations, nous avons pu orienter efficacement notre système vers la détection et la spécification des relations pertinentes au sein du domaine étudié. Les défis relevés ont inclut la nécessité de gérer la rigidité inhérente aux listes de relations prédéfinies, ainsi que l'amélioration continue de la spécificité des relations extraites. En intégrant ces efforts, cette recherche a visé à enrichir notre compréhension et notre capacité à structurer la connaissance lexico-sémantique, contribuant ainsi aux avancées dans le domaine du traitement automatique du langage naturel appliqué à des textes spécialisés du secteur public en français.

Enfin, ce travail de thèse a abordé de front les défis associés à l'extraction de relations dans les textes spécialisés, en particulier en confrontant les limitations des approches traditionnelles utilisant des listes fixes de types de relations. Nous avons mis en lumière les restrictions du réseau lexical rezoJDM, soulignant sa dépendance à une liste préétablie de types de relations et sa tendance à générer des résultats trop généraux, qui manquent de précision dans les contextes spécialisés. Pour surmonter ces défis, cette recherche a introduit une nouvelle approche basée sur l'extraction de relations ouvertes, exploitant des techniques telles que l'Open Information Extraction (OpenIE) et l'utilisation de grands modèles de langage (LLMs). Ces avancées visent à améliorer la flexibilité et l'adaptabilité des systèmes d'extraction de relations, permettant une identification plus précise des verbes d'action spécifiques aux corpus spécialisés étudiés. En intégrant ces innovations, cette thèse contribue à enrichir les méthodologies en traitement automatique du langage naturel, offrant des perspectives prometteuses pour l'analyse avancée de textes dans divers domaines du secteur public en

français.

7.2 Perspectives

7.2.1 Extraction terminologique

Les perspectives finales pour l'extraction terminologique sont les suivantes :

1. **Standardisation des annotations :** Pour assurer une meilleure consistance et qualité des annotations, il est primordial de développer un guide d'annotation standardisé pour les experts. Ce guide facilitera l'application cohérente des annotations. L'intégration d'outils d'annotation assistée est également recommandée pour améliorer la précision et l'efficacité du processus.
2. **Adaptation aux évolutions des connaissances :** Étant donné que les termes et les connaissances évoluent rapidement, particulièrement dans les domaines techniques et juridiques, le système doit être régulièrement mis à jour. Il est nécessaire de développer des pipelines automatisés pour la mise à jour des modèles et des données. Cela permettra de maintenir la pertinence et l'efficacité du système face aux nouvelles informations et aux évolutions des connaissances.
3. **Enrichissement par l'intégration de synonymes :** En vue d'améliorer la richesse terminologique, il est proposé d'enrichir les termes extraits en intégrant des synonymes provenant de sources externes pertinentes, telles que légifrance dans le domaine juridique. Cela contribuera à élargir la couverture terminologique et à améliorer la précision des résultats.
4. **Évaluation continue de la performance :** Afin de mesurer et d'améliorer la performance du système, il est recommandé de mettre en place un système d'évaluation continue. Celui-ci devrait inclure l'évaluation sur des jeux de données diversifiés en temps réel, ainsi que la collecte de retours d'utilisateurs et des analyses de performance post-déploiement. Cette démarche permettra de garantir l'efficacité et l'adaptabilité du système dans des environnements réels.

En conclusion, ces perspectives visent à renforcer la robustesse et l'efficacité de l'extraction terminologique pour la construction de graphes de connaissances à partir de texte, en répondant aux défis liés à la standardisation, à l'évolution des connaissances, à l'enrichissement terminologique et à l'évaluation continue de la performance.

7.2.2 Extraction de relations à l'aide du réseau lexical issu de JeuxDeMots

Les perspectives finales pour l'extraction de relations à l'aide du réseau lexical issu de JeuxDeMots sont les suivantes :

1. **Utilisation d'un corpus standard pour l'évaluation :** Il est crucial d'adopter un corpus standard pour valider les instances de relations afin d'assurer l'objectivité et la crédibilité des résultats. L'utilisation d'un corpus externe élimine les biais potentiels introduits par l'utilisation d'un corpus interne pour à la fois l'apprentissage et la validation. Cela permet une évaluation plus rigoureuse et impartiale de la performance du modèle, en testant sa capacité à généraliser au-delà des données spécifiques sur lesquelles il a été formé.

2. **Analyse statistique et qualitative des relations extraites :** Une analyse approfondie des instances de relations extraites devrait inclure à la fois des analyses statistiques et qualitatives. La première permettrait de comprendre comment les types spécifiques de relations se comportent dans différents contextes, tandis que la seconde fournirait des insights sur la pertinence et l'exactitude des relations en examinant les termes impliqués. Cette approche combinée enrichit la compréhension des structures relationnelles dans le corpus, facilitant des analyses plus nuancées et informatives.
3. **Intégration de nouveaux types spécifiques de relations :** Pour enrichir le graphe de connaissances lexico-sémantiques, il est proposé d'intégrer de nouveaux types spécifiques de relations. Cela inclut la création de relations entre les termes clés annotés par des experts et les termes pertinents extraits du texte. Ces annotations enrichissent les données textuelles en devenant les concepts clés d'un réseau de relations plus cohérent. De plus, l'ajout de regroupements de termes pertinents à travers le clustering de formes fléchies et de synonymes provenant de sources externes comme le corpus de Berger-Levrault permet d'améliorer la précision et la couverture des relations extraites.
4. **Défis et perspectives futures :** Il est essentiel de reconnaître les défis liés à la pertinence contextuelle et à la spécificité des relations extraites. Par exemple, l'analyse montre que la majorité des relations ne proviennent pas du même paragraphe, indiquant un potentiel d'amélioration dans la capture de relations contextuelles étroites. Pour surmonter ces limitations, une orientation vers l'extraction de relations ouvertes peut être envisagée, visant à capturer des relations plus spécifiques et contextuelles directement à partir du texte.

En résumé, ces perspectives visent à renforcer la robustesse et l'efficacité de l'extraction de relations pour la construction de graphes à partir de textes, en s'attaquant aux défis de validation, d'analyse et d'intégration de nouvelles connaissances. Cela contribuera à améliorer la qualité et l'utilité des résultats obtenus pour des applications pratiques notamment dans le domaine juridique et au-delà.

7.2.3 Extractions de relations ouvertes à l'aide des LLMs

Les LLMs représentent une avancée significative dans l'extraction de relations en permettant une adaptation à de nouveaux domaines et informations sans nécessiter un réentraînement intensif. Leur capacité à générer du texte dans des contextes variés contribue à une extraction plus précise et contextuellement appropriée des relations, dépassant les limitations des méthodes traditionnelles basées sur des règles strictes et des schémas prédéfinis. Cependant, trois points sont à relever dans nos perspectives pour l'extraction de relations ouvertes à l'aide des LLMs :

1. **Défis actuels à surmonter :** Malgré leurs avantages, les LLMs présentent plusieurs défis importants. La qualité et la pertinence des relations extraites peuvent varier selon les domaines et les prompts utilisés pour le clustering. Les LLMs peuvent également générer des relations incorrectes ou incohérentes lorsque les données d'entrée sont ambiguës. Il est crucial de continuer à améliorer les techniques de prétraitement des données et d'affiner les modèles pour minimiser ces erreurs. De plus, le coût computationnel élevé associé à leur utilisation nécessite des efforts d'optimisation pour rendre cette technologie plus accessible et efficace à grande échelle.
2. **Intégration de techniques hybrides :** Pour maximiser l'efficacité de l'extraction de relations, une approche hybride intégrant à la fois des techniques traditionnelles de

comparaison de liens et des méthodes basées sur les LLMs est recommandée. Cette approche permet de bénéficier des forces de chaque méthode : les LLMs pour découvrir des relations nouvelles et inattendues, et les techniques traditionnelles pour valider et structurer ces découvertes. La gestion sophistiquée des données est essentielle pour éviter les redondances et maintenir la cohérence des clusters formés.

3. **Perspectives futures :** Les recherches futures devraient se concentrer sur plusieurs axes d'amélioration :
   — Amélioration de la précision des LLMs dans l'extraction de relations complexes et spécifiques.
   — Réduction du coût computationnel par l'optimisation des modèles et des algorithmes.
   — Développement de nouvelles techniques de clustering pour une gestion efficace et continue des relations extraites.
   — Exploration de méthodes pour minimiser les erreurs de génération de relations incohérentes ou incorrectes.

En conclusion, bien que les LLMs représentent une avancée prometteuse pour l'extraction de relations ouvertes, leur adoption à grande échelle nécessite de surmonter plusieurs défis techniques et computationnels. En combinant les approches modernes avec les méthodes éprouvées, il est possible d'exploiter pleinement le potentiel des LLMs pour l'extraction de relations ouvertes, notamment dans le domaine juridique et au-delà.

7.2.4 Graphes de connaissances lexico-sémantiques

En perspectives, nous envisageons de créer un graphe de connaissances lexico-sémantique à partir de l'ensemble des données extraites dans notre étude. Ce graphe sera construit en intégrant la terminologie spécifique extraite, les relations typées entre les concepts, les relations ouvertes pour capturer les liens non spécifiés à priori, ainsi que les domaines associés à chaque terme. Notre approche consistera à modéliser ces éléments dans une structure de graphe afin de représenter de manière claire et cohérente les connaissances sémantiques et lexicales découvertes. Nous prévoyons d'utiliser des méthodes d'intégration de données avancées pour aligner et fusionner les informations extraites, garantissant ainsi la pertinence et la précision du graphe résultant. Cette ressource sera non seulement un outil essentiel pour visualiser et explorer les relations entre les concepts de manière interactive, mais aussi une base solide pour des analyses plus approfondies et des applications dans divers domaines, notamment la recherche d'information et la recommandation de contenu.

Voici les étapes à prévoir pour la construction de graphes de connaissances lexico-sémantiques :

— **Construction du graphe de connaissances :** La première étape consiste à utiliser les informations extraites, telles que la terminologie, les relations typées, les relations ouvertes et les domaines associés aux termes, pour construire le graphe de connaissances lexico-sémantiques. Cela implique de définir une structure pour représenter les entités (termes) et les relations entre elles, en utilisant des techniques appropriées de modélisation de connaissances.
— **Modélisation des entités et des relations :** Il est essentiel de définir des schémas qui spécifient comment les entités et les relations extraites seront représentées dans le graphe. Cela peut inclure l'utilisation de vocabulaires contrôlés, de taxonomies ou de schémas spécifiques au domaine pour assurer la cohérence et la compréhension sémantique des données intégrées.
— **Intégration des données :** Les informations extraites doivent être intégrées de manière cohérente dans le graphe de connaissances. Cela nécessite un processus d'ali-

gnement et de fusion des données, en tenant compte des correspondances et des similitudes entre les termes et les relations extraites. Les technologies de fusion de graphes et d'intégration de données peuvent être utilisées pour cette étape.

— **Validation et qualité des données :** Avant de finaliser le graphe de connaissances, il est crucial de procéder à une validation approfondie pour assurer la qualité des données. Cela comprend la vérification de l'exactitude des relations extraites, la cohérence des associations de domaines et la conformité aux standards définis. Des techniques de validation automatisées et manuelles peuvent être employées pour cette vérification.

— **Utilisation et évolution du graphe de connaissances :** Une fois construit, le graphe de connaissances peut être utilisé pour diverses applications telles que la recherche d'informations, l'analyse sémantique, et la recommandation de contenu. Il est également important de prévoir des mécanismes pour maintenir et mettre à jour le graphe au fil du temps, en intégrant de nouvelles informations et en ajustant les structures pour refléter les évolutions dans le domaine d'étude.

— **Perspectives d'amélioration continue :** Les perspectives futures devraient inclure l'amélioration continue du graphe de connaissances en tenant compte des retours d'utilisateurs, des avancées technologiques et des nouvelles données disponibles. Cela peut impliquer l'exploration de techniques avancées d'apprentissage automatique pour enrichir et étendre automatiquement le graphe, ainsi que l'évaluation régulière de sa pertinence et de son efficacité pour les utilisateurs finaux.

En résumé, la construction d'un graphe de connaissances lexico-sémantiques à partir des données extraites de la thèse représente une opportunité significative pour organiser et exploiter les connaissances contenues dans le texte. En suivant ces perspectives, nous pourrons créer une ressource précieuse pour la recherche et l'analyse dans le domaine juridique et pratique.

Liste des contributions

Articles publiés dans le cadre de la thèse

— Gosset, C., Billami, M. B., Lafourcade, M., Bortolaso, C., & Derras, M. (2021). Extraction automatique de relations sémantiques d'hyperonymie et d'hyponymie dans un corpus métier (Automatic extraction of hypernym and hyponym relations in a professional corpus). In Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale (pp. 162–170). Lille, France: ATALA

— Gosset, C., Billami, M. B., Lafourcade, M., Bortolaso, C., & Derras, M. (2022b). Relation-Level Vector Representation for Relation Extraction and Classification on Specialized Data. In International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2022)

— Gosset, C., Billami, M. B., Lafourcade, M., Bortolaso, C., & Derras, M. (2022a). Création et validation de représentations vectorielles pour des relations lexico-sémantiques : application à l'identification/classification de relations à partir d'un corpus métier. In Conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC2022)

— Gosset, C. (2023). Classification de documents métiers pour l'aide à l'extraction et la classification de relations lexico-sémantiques typées et pondérées. Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA), (pp. 37–40)

Articles publiés annexes à la thèse

— Billami, M. B., Kandi, M., Nicolaieff, L., Ducharlet, K., Gosset, C., Rey, S., Bortolaso, C., & Derras, M. (2021a). Vers une étude comparative de différentes approches de classification automatique de textes provenant des secteurs métiers. In Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (PIA-APIA) Bordeaux, France

— Billami, M. B., Nicolaieff, L., Gosset, C., & Bortolaso, C. (2021b). Participation de Berger-Levrault (BL.Research) à DEFT 2021 : de l'apprentissage des seuils de validation à la classification multi-labels de documents (Berger-Levrault (BL). In Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi

Fouille de Textes (DEFT) (pp. 82–94). Lille, France: ATALA

— Dufour, S., Mehdi Kandi, M., Boutamine, K., Gosset, C., Billami, M. B., Bortolaso, C., & Miloudi, Y. (2022). BL.Research at SemEval-2022 Task 8: Using various Semantic Information to evaluate document-level Semantic Textual Similarity. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 1221–1228). Seattle, United States: Association for Computational Linguistics

Bibliographie

Abalorio, C. C., Sison, A. M., Medina, R. P., & Dalaorao, G. A. (2022). Extended Max-Occurrence with Normalized Non-Occurrence as MONO Term Weighting Modification to Improve Text Classification. International Journal of Advanced Computer Science and Applications, 13(4). Number : 4. (Cité en page 67.)

Adar, E. & Datta, S. (2015). Building a scientific concept hierarchy database (schbase). In Annual Meeting of the Association for Computational Linguistics. (Cité en page 42.)

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. A. (2022). Large language models are few-shot clinical information extractors. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 38.)

Akbik, A., Blythe, D., & Vollgraf, R. (2018a). Contextual string embeddings for sequence labeling. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics (pp. 1638–1649). Santa Fe, New Mexico, USA : Association for Computational Linguistics. (Cité en page 34.)

Akbik, A., Blythe, D., & Vollgraf, R. (2018b). Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics (pp. 1638–1649). (Cité en page 83.)

Al-Qurishi, M. (2022). Recent advances in long documents classification using deep-learning. In International Conference on Natural Language and Speech Processing. (Cité en page 66.)

Al-Rfou, R., Kulkarni, V., Perozzi, B., & Skiena, S. (2015). Polyglot-NER : Massive multilingual named entity recognition. (Cité en pages 46, 49 et 89.)

Ali, S., Mousa, H. M., & Hussien, M. (2019). A review of open information extraction techniques. IJCI. International Journal of Computers and Information. (Cité en page 60.)

Alt, C., Gabryszak, A., & Hennig, L. (2020). TACRED revisited : A thorough evaluation of the TACRED relation extraction task. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1558–1569). Online : Association for Computational Linguistics. (Cité en page 59.)

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J. I., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning : concepts, cnn architectures, challenges, applications, future directions. Journal of Big Data, 8. (Cité en page 35.)

Andrew, J. J. (2018). Automatic extraction of entities and relation from legal documents. In NEWS@ACL. (Cité en page 54.)

Angeli, G., Tibshirani, J., Wu, J., & Manning, C. D. (2014). Combining distant and partial supervision for relation extraction. In Conference on Empirical Methods in Natural Language Processing. (Cité en pages 53 et 58.)

Antoniou, G. & Harmelen, F. V. (2004). Web ontology language : Owl. In Handbook on Ontologies. (Cité en page 20.)

Aronson, A. R., Humphrey, S. M., Ide, N. C., Kim, W., Loane, R. F., Mork, J. G., Smith, L. H., Tanabe, L. K., Wilbur, W. J., Xie, N., Demner-Fushman, D., & Liu, H. (2004). Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics medline citations. In Text Retrieval Conference. (Cité en page 67.)

Ashok, D. & Lipton, Z. C. (2023). Promptner : Prompting for named entity recognition. ArXiv, abs/2305.15444. (Cité en page 38.)

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. Database, 2018. (Cité en page 23.)

Babaei Giglou, H., D'Souza, J., & Auer, S. (2023). Llms4ol : Large language models for ontology learning. In International Semantic Web Conference : Springer International Publishing. (Cité en page 55.)

Bailly, A., Blanc, C., & Guillotin, T. (2021). Classification multi-label de cas cliniques avec camembert (multi-label classification of clinical cases with camembert). In JEPTALNRECITAL. (Cité en page 66.)

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 62.)

Balachandran, K. & Ranathunga, S. (2016). Domain-Specific Term Extraction for Concept Identification in Ontology Construction. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 34–41). Omaha, NE, USA : IEEE. (Cité en page 25.)

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. In CACM. (Cité en pages 60 et 61.)

Banko, M. & Etzioni, O. (2008). The tradeoffs between open and traditional relation extraction. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 61.)

Barker, K. & Cornacchia, N. (2000). Using noun phrase heads to extract document keyphrases. In Canadian Conference on AI. (Cité en pages 41 et 42.)

Bartolini, I., Moscato, V., Postiglione, M., Sperlì, G., & Vignali, A. (2022). COSINER : COntext SImilarity data augmentation for Named Entity Recognition. In T. Skopal, F. Falchi, J. Lokoč, M. L. Sapino, I. Bartolini, & M. Patella (Eds.), Similarity Search and Applications (pp. 11–24). Cham : Springer International Publishing. (Cité en page 32.)

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An Overview of Graph-Based Keyword Extraction Methods and Approaches. Journal of information and organizational sciences, (pp. 1–20). (Cité en page 84.)

Bendahman, N., Pinel-Sauvagnat, K., Hubert, G., & Billami, M. B. (2023). Quelles évolutions sur cette loi ? entre abstraction et hallucination dans le domaine du résumé de textes juridiques. 18e Conférence en Recherche d'Information et Applications (CORIA-TALN), (pp. 18–36). (Cité en pages 51 et 84.)

Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., & Zesch, T., Eds. (2016a). Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California. Association for Computational Linguistics. (Cité en page 41.)

Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., & Zesch, T., Eds. (2016b). Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego, California : Association for Computational Linguistics. (Cité en page 41.)

Bhardwaj, S., Aggarwal, S., & Mausam, M. (2019). Carb : A crowdsourced benchmark for open ie. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 64.)

Billami, M. B., Kandi, M., Nicolaieff, L., Ducharlet, K., Gosset, C., Rey, S., Bortolaso, C., & Derras, M. (2021a). Vers une étude comparative de différentes approches de classification automatique de textes provenant des secteurs métiers. In Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (PIA-APIA) Bordeaux, France. (Non cité.)

Billami, M. B., Nicolaieff, L., Gosset, C., & Bortolaso, C. (2021b). Participation de Berger-Levrault (BL.Research) à DEFT 2021 : de l'apprentissage des seuils de validation à la classification multi-labels de documents (Berger-Levrault (BL). In Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT) (pp. 82–94). Lille, France : ATALA. (Non cité.)

Black, W. J., Rinaldi, F., & Mowatt, D. (1998). Facile : Description of the ne system used for muc-7. In Message Understanding Conference. (Cité en page 27.)

Bogdanov, S., Constantin, A., Bernard, T., Crabbé, B., & Bernard, E. (2024). Nuner : Entity recognition encoder pre-training via llm-annotated data. (Cité en pages 37, 48 et 87.)

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics (ACL), 5, 135–146. (Cité en page 52.)

Bordea, G., Buitelaar, P., Faralli, S., & Navigli, R. (2015a). SemEval-2015 Task 17 : Taxonomy Extraction Evaluation (TExEval). In SemEval@NAACL-HLT. (Cité en page 52.)

Bordea, G., Buitelaar, P., Faralli, S., & Navigli, R. (2015b). SemEval-2015 Task 17 : Taxonomy Extraction Evaluation (TExEval). Proceedings of the 9th International Workshop on Semantic Evaluation, (pp. 902–910). (Cité en page 58.)

Borst, W. (1997). Construction of Engineering Ontologies for Knowledge Sharing and Reuse. Phd thesis - research ut, graduation ut, University of Twente, Netherlands. (Cité en page 20.)

Boudin, F. (2013). TALN archives : a digital archive of French research articles in natural language processing (TALN archives : une archive numérique francophone des articles de recherche en traitement automatique de la langue) [in French]. In E. Morin & Y. Estève (Eds.), Proceedings of TALN 2013 (Volume 2 : Short Papers) (pp. 507–514). Les Sables d'Olonne, France : ATALA. (Cité en page 90.)

Boudin, F. (2018). Unsupervised Keyphrase Extraction with Multipartite Graphs. arXiv :1803.08721 [cs]. (Cité en page 43.)

Boudin, F. & Gallina, Y. (2021). Redefining absent keyphrases and their effect on retrieval effectiveness. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (pp. 4185–4193). Online : Association for Computational Linguistics. (Cité en page 90.)

Bougouin, A., Barreaux, S., Romary, L., Boudin, F., & Daille, B. (2016). TermITH-eval : a French standard-based resource for keyphrase extraction evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1924–1927). Portorož, Slovenia : European Language Resources Association (ELRA). (Cité en page 91.)

Bougouin, A., Boudin, F., & Daille, B. (2013). Topicrank : Graph-based topic ranking for keyphrase extraction. In International Joint Conference on Natural Language Processing. (Cité en page 43.)

Brin, S. (1998). Extracting Patterns and Relations from the World Wide Web. In International Workshop on The World Wide Web and Databases (WebDB), Springer (pp. 172–183). (Cité en page 53.)

Buitelaar, P., Cimiano, P., & Magnini, B. (2003). Ontology Learning from Text : An Overview. IOS Press. (Cité en page 23.)

Bunescu, R. C. & Mooney, R. J. (2005a). A Shortest Path Dependency Kernel for Relation Extraction. In Proceedings of the conference on human language technology and empirical methods in natural lan (pp. 724–731). (Cité en pages 52 et 53.)

Bunescu, R. C. & Mooney, R. J. (2005b). Subsequence Kernels for Relation Extraction. In Advances in Neural Information Processing Systems 18 (NIPS 2005). (Cité en page 53.)

Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2004). Domain-specific synonym expansion and validation for biomedical information retrieval (multitext experiments for trec 2004). In Text Retrieval Conference. (Cité en page 67.)

Béchet, F., El-Bèze, M., & Torres-Moreno, J.-M. (2008). En finir avec la confusion des genres pour mieux séparer les thèmes. In Actes de l'atelier de clôture de la 4ème édition du Défi Fouille de Texte (DEFT). (Cité en page 66.)

Camacho-Collados, J., Anke, L. E., & Schockaert, S. (2019a). Relational Word Embeddings. In Association for Computational Linguistics (ACL). (Cité en page 57.)

Camacho-Collados, J., Delli Bovi, C., Espinosa Anke, L., Oramas, S., Pasini, T., Santus, E., Shwartz, V., Navigli, R., & Saggion, H. (2018). SemEval-2018 Task 9 : Hypernym Discovery. In Proceedings of The 12th International Workshop on Semantic Evaluation (pp. 712–724). New Orleans, Louisiana : Association for Computational Linguistics. (Cité en page 58.)

Camacho-Collados, J., Espinosa-Anke, L., Shoaib, J., & Schockaert, S. (2019b). A Latent Variable Model for Learning Distributional Relation Vectors. In Proceedings of IJCAI. event-place : Macau, China. (Cité en page 122.)

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! keyword extraction from single documents using multiple local features. Information Sciences, 509, 257–289. (Cité en page 42.)

Cao, Y., Hu, Z., Chua, T.-s., Liu, Z., & Ji, H. (2019). Low-resource name tagging learned with weakly labeled data. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 261–270). Hong Kong, China : Association for Computational Linguistics. (Cité en page 32.)

Cardon, R., Grabar, N., Grouin, C., & Hamon, T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine (DEFT'2020). (Cité en page 66.)

Carpenter, B. (2004). Phrasal queries with lingpipe and lucene : Ad hoc genomics text retrieval. In Text Retrieval Conference. (Cité en page 67.)

Cartier, E. (2015). Extraction automatique de relations sémantiques dans les définitions : approche hybride, construction d'un corpus de relations sémantiques pour le français. In Actes de la 22e conférence sur le TALN (pp. 131–145). (Cité en page 53.)

Castano, S., Falduti, M., Ferrara, A., & Montanelli, S. (2022). A knowledge-centered framework for exploration and retrieval of legal documents. Information Systems, 106, 101842. (Cité en page 22.)

Centre Aquitain des Technologies de l'Information et Electroniques (2023). Dfp (revision 1d24c09). (Cité en page 49.)

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., & Androutsopoulos, I. (2019). Large-scale multi-label text classification on EU legislation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6314–6322). Florence, Italy : Association for Computational Linguistics. (Cité en page 68.)

Chang, J. & Han, X. (2023). Multi-level context features extraction for named entity recognition. Computer Speech & Language, 77, 101412. (Cité en page 36.)

Chen, C. & Zhang, Y. (2018). Learning how to self-learn : Enhancing self-training using neural reinforcement learning. 2018 International Conference on Asian Language Processing (IALP), (pp. 25–30). (Cité en page 34.)

Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst. Appl., 66, 245–260. (Cité en page 67.)

Chen, P., Xu, H., Zhang, C., & Huang, R. (2022a). Crossroads, buildings and neighborhoods : A dataset for fine-grained location recognition. In North American Chapter of the Association for Computational Linguistics. (Cité en page 47.)

Chen, Q., Du, J., Hu, Y., Keloth, V. K., Peng, X., Raja, K., Zhang, R., Lu, Z., & Xu, H. (2023). Large language models in biomedical natural language processing : benchmarks, baselines, and recommendations. ArXiv, abs/2305.16326. (Cité en pages 40 et 54.)

Chen, S., Aguilar, G., Neves, L., & Solorio, T. (2021). Data augmentation for cross-domain named entity recognition. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 5346–5356). Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. (Cité en page 32.)

Chen, X., Li, L., Deng, S., Tan, C., Xu, C., Huang, F., Si, L., Chen, H., & Zhang, N. (2022b). Lightner : A lightweight tuning paradigm for low-resource ner via pluggable prompting. (Cité en page 31.)

Chen, Y., Zheng, Y., & Yang, Z. (2022c). Prompt-based metric learning for few-shot ner. ArXiv, abs/2211.04337. (Cité en page 37.)

Chern, I.-C., Chern, S., Chen, S., Yuan, W., Feng, K., Zhou, C., He, J., Neubig, G., Liu, P., et al. (2023). Factool : Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. arXiv preprint arXiv :2307.13528. (Cité en page 44.)

Chinchor, N. A. (1997). Appendix e : Muc-7 named entity task definition (version 3.5). In Message Understanding Conference. (Cité en page 83.)

Chinchor, N. A. (1998). Overview of muc-7. In Message Understanding Conference. (Cité en page 27.)

Cho, M., Ha, J., Park, C., & Park, S. (2020). Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. Journal of Biomedical Informatics, 103, 103381. (Cité en page 35.)

Choudhry, A., Khatri, I., Gupta, P., Gupta, A., Nicol, M., Meurs, M.-J., & Vishwakarma, D. K. (2023). Adversarial Adaptation for French Named Entity Recognition. arXiv :2301.05220 [cs]. (Cité en page 38.)

Christensen, J., Mausam, Soderland, S., & Etzioni, O. (2010). Semantic role labeling for open information extraction. In HLT-NAACL 2010. (Cité en page 62.)

Cimiano, P., Hotho, A., & Staab, S. (2005). Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. jair, 24, 305–339. (Cité en page 23.)

Cohen, A. M., Bhupatiraju, R. T., & Hersh, W. R. (2004). Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In Text Retrieval Conference. (Cité en page 67.)

Cohen, A. M. & Hersh, W. R. (2006). The trec 2004 genomics track categorization task : classifying full text biomedical documents. Journal of Biomedical Discovery and Collaboration, 1, 4 – 4. (Cité en page 67.)

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing : deep neural networks with multitask learning. In International Conference on Machine Learning. (Cité en page 34.)

Constant, M. & Nivre, J. (2016). A transition-based system for joint lexical and syntactic analysis. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 41.)

Cordeiro, S., Ramisch, C., & Villavicencio, A. (2016). UFRGS&LIF at SemEval-2016 task 10 : Rule-based MWE identification and predominant-supersense tagging. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 910–917). San Diego, California : Association for Computational Linguistics. (Cité en page 41.)

Corro, L. D. & Gemulla, R. (2013). Clausie : clause-based open information extraction. Proceedings of the 22nd international conference on World Wide Web. (Cité en page 62.)

Cotterell, R. & Duh, K. (2017). Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In G. Kondrak & T. Watanabe (Eds.), Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2 : Short Papers) (pp. 91–96). Taipei, Taiwan : Asian Federation of Natural Language Processing. (Cité en page 30.)

Cui, L., Wu, Y., Liu, J., Yang, S., & Zhang, Y. (2021). Template-based named entity recognition using bart. In Findings. (Cité en page 31.)

Culotta, A. & McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. In AAAI Conference on Artificial Intelligence. (Cité en page 33.)

Culotta, A., McCallum, A., & Betz, J. (2006). Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text. In Proceedings of the Human Language Technology Conference of the North American Chapter of the (pp. 296–303). (Cité en page 53.)

da Costa, L. S., Oliveira, I. L., & Fileto, R. (2023). Text classification using embeddings : a survey. Knowledge and Information Systems, 65, 2761–2803. (Cité en page 66.)

Dai, X. & Adel, H. (2020). An analysis of simple data augmentation for named entity recognition. In D. Scott, N. Bel, & C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics (pp. 3861–3867). Barcelona, Spain (Online) : International Committee on Computational Linguistics. (Cité en page 32.)

Daille, B., Gaussier, É., & Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In International Conference on Computational Linguistics. (Cité en page 41.)

Dandapat, S. & Way, A. (2016). Improved Named Entity Recognition using Machine Translation-based Cross-lingual Information. Computación y Sistemas, 20(3), 495–504. (Cité en page 30.)

Danesh, S., Sumner, T. R., & Martin, J. H. (2015). Sgrank : Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction. In International Workshop on Semantic Evaluation. (Cité en page 43.)

Das, A. P., Nayak, K., & Nayak, M. (2018). A Survey on Machine Learning Based Text Categorization. IOSR Journal of Computer Engineering (IOSR-JCE), 20, 51–56. (Cité en page 66.)

Das, S. S. S., Katiyar, A., Passonneau, R., & Zhang, R. (2021). Container : Few-shot named entity recognition via contrastive learning. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 38.)

de Sá Mesquita, F., Schmidek, J., & Barbosa, D. (2013). Effectiveness and efficiency of open relation extraction. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 62.)

Delestre, C. & Amar, A. (2022). DistilCamemBERT : une distillation du modèle français CamemBERT. In CAp (Conférence sur l'Apprentissage automatique) Vannes, France. (Cité en page 206.)

Dendien, J. & Pierrel, J. (2003). Le trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence. In Traitement Automatique des Langues, HERMÈS. (Cité en page 53.)

Derczynski, L., Bontcheva, K., & Roberts, I. (2016). Broad twitter corpus : A diverse named entity recognition resource. In International Conference on Computational Linguistics. (Cité en pages 47 et 90.)

Derczynski, L., Nichols, E., Van Erp, M., & Limsopatham, N. (2017a). Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In Proceedings of the 3rd Workshop on Noisy User-generated Text (pp. 140–147). Copenhagen, Denmark : Association for Computational Linguistics. (Cité en page 36.)

Derczynski, L., Nichols, E., van Erp, M., & Limsopatham, N. (2017b). Results of the WNUT2017 shared task on novel and emerging entity recognition. In Proceedings of the 3rd Workshop on Noisy User-generated Text (pp. 140–147). Copenhagen, Denmark : Association for Computational Linguistics. (Cité en page 46.)

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019a). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In North American Chapter of the Association for Computational Linguistics. (Cité en pages 27, 37, 38 et 52.)

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019b). Bert : Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics. (Cité en page 63.)

Dhar, A., Mukherjee, H., Dash, N. S., & Roy, K. (2020). Text categorization : past and present. Artificial Intelligence Review. (Cité en page 66.)

Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., Zheng, H., & Liu, Z. (2021). Few-nerd : A few-shot named entity recognition dataset. ArXiv, abs/2105.07464. (Cité en page 48.)

Dogan, R. I. & Lu, Z. (2012). An improved corpus of disease mentions in pubmed citations. In BioNLP@HLT-NAACL. (Cité en page 47.)

Dogan, T. & Uysal, A. K. (2019). On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification. Arab J Sci Eng, 44(11), 9545–9560. Number : 11. (Cité en page 66.)

dos Santos, C., Xiang, B., & Zhou, B. (2015). Classifying Relations by Ranking with Convolutional Neural Networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers) (pp. 626–634). Beijing, China : Association for Computational Linguistics. (Cité en page 57.)

Du, W., Advani, L., Gambhir, Y., Perry, D. J., Shiralkar, P., Xing, Z., & Colak, A. (2023). Effective proxy for human labeling : Ensemble disagreement scores in large language models for industrial nlp. ArXiv, abs/2309.05619. (Cité en page 44.)

Dufour, S., Mehdi Kandi, M., Boutamine, K., Gosset, C., Billami, M. B., Bortolaso, C., & Miloudi, Y. (2022). BL.Research at SemEval-2022 Task 8 : Using various Semantic Information to evaluate document-level Semantic Textual Similarity. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022) (pp. 1221–1228). Seattle, United States : Association for Computational Linguistics. (Non cité.)

Dupont, Y., González-Gallardo, C.-E., Lejeune, G., Millour, A., & Tanguy, J.-B. (2021). QUEER@DEFT2021 : Identification du Profil Clinique de patients et Notations Automatique de Copies d'Étudiants. In JEPTALNRECITAL. (Cité en page 66.)

Dutta, H. & Gupta, A. (2022). PNRank : Unsupervised ranking of person name entities from noisy OCR text. Decision Support Systems, 152, 113662. (Cité en page 29.)

Dwivedi, S. K. & Arya, C. (2016). Automatic text classification in information retrieval : A survey. ICoICT 2016. (Cité en page 66.)

Eftimov, T., Koroušić Seljak, B., & Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PLOS ONE, 12(6). (Cité en page 27.)

El-Beltagy, S. R. & Rafea, A. A. (2010). KP-miner : Participation in SemEval-2. In International Workshop on Semantic Evaluation. (Cité en page 42.)

El Ghosh, M., Naja, H., Abdulrab, H., & Khalil, M. (2017). Ontology Learning Process as a Bottom-up Strategy for Building Domain-specific Ontology from Legal Texts :. In Proceedings of the 9th International Conference on Agents and Artificial Intelligence (pp. 473–480). Porto, Portugal : SCITEPRESS - Science and Technology Publications. (Cité en page 22.)

Elhadad, M. K., Badran, K. M., & Salama, G. I. M. (2017). A novel approach for ontology-based feature vector generation for web text document classification. Int. J. Softw. Innov., 6, 1–10. (Cité en page 22.)

ElSahar, H., Demidova, E., Gottschalk, S., Gravier, C., & Laforest, F. (2017). Unsupervised open relation extraction. ArXiv, abs/1801.07174. (Cité en page 63.)

ElSahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J. S., Laforest, F., & Simperl, E. P. B. (2018). T-rex : A large scale alignment of natural language with knowledge base triples. In International Conference on Language Resources and Evaluation. (Cité en page 64.)

Emani, C. K. & Haralambous, Y. (2017). Un système de questions-réponses automatiques dans le domaine légal : le cas des réglementations maritimes [a question answering system in the legal domain : the case of maritime regulations]. In ICON. (Cité en page 22.)

Emani, C. K. & Haralambous, Y. (2018). Un système de questions-réponses dans le domaine légal : le cas des réglementations maritimes. Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), (pp. 47–72). (Cité en page 22.)

Etzioni, O., Cafarella, M. J., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2004). Methods for Domain-Independent Information Extraction from the Web : An Experimental Comparison. In American Association for Artificial Intelligence (AAAI) (pp. 391–398). (Cité en page 53.)

Etzioni, O., Fader, A., Christensen, J., Soderland, S., & Mausam (2011). Open information extraction : The second generation. In International Joint Conference on Artificial Intelligence. (Cité en page 62.)

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 62.)

Fangchao, L., Yan, L., Lin, H., Han, X., & Sun, L. (2021). Element intervention for open relation extraction. ArXiv, abs/2106.09558. (Cité en page 63.)

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E. H., & Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In Proceedings of NAACL (pp. 1606–1615). (Cité en page 57.)

Fellbaum, C. (1998). WordNet : an electronic lexical database. Language, speech, and communication. Cambridge, Mass : MIT Press. (Cité en pages 4 et 63.)

Feng, J., Huang, M., Zhao, L., Yang, Y., & Zhu, X. (2018). Reinforcement learning for relation classification from noisy data. In AAAI Conference on Artificial Intelligence. (Cité en page 32.)

FitzGerald, N., Michael, J., He, L., & Zettlemoyer, L. (2018). Large-scale qa-srl parsing. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 64.)

Florescu, C. & Caragea, C. (2017). PositionRank : An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers) (pp. 1105–1115). Vancouver, Canada : Association for Computational Linguistics. (Cité en pages 42 et 43.)

Foppiano, L., Lambard, G., Amagasa, T., & Ishii, M. (2024a). Mining experimental data from materials science literature with large language models. ArXiv, abs/2401.11052. (Cité en page 39.)

Foppiano, L., Lambard, G., Amagasa, T., & Ishii, M. (2024b). Mining experimental data from materials science literature with large language models. (Cité en page 54.)

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-manning, C. G. (1999). Domain specific keyphrase extraction. In Proceedings of the 16th International Joint Conference on AI (pp. 668–673). (Cité en pages 40 et 42.)

Fries, J. A., Wu, S., Ratner, A. J., & Ré, C. (2017). Swellshark : A generative model for biomedical named entity recognition without labeled data. ArXiv, abs/1704.06360. (Cité en pages 32 et 33.)

Fritzler, A., Logacheva, V., & Kretov, M. (2019). Few-shot classification in named entity recognition task. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC '19 (pp. 993–1000). New York, NY, USA : Association for Computing Machinery. (Cité en page 31.)

Fu, C., Chen, T., Qu, M., Jin, W., & Ren, X. (2019). Collaborative policy learning for open knowledge graph reasoning. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 25.)

G, V., Gupta, D., Anil, A., & S, A. (2019). An Ontology Driven Question Answering System for Legal Documents. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), volume 1 (pp. 947–951). (Cité en page 22.)

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep Bayesian active learning with image data. In D. Precup & Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research (pp. 1183–1192). : PMLR. (Cité en page 33.)

Galárraga, L., Heitz, G., Murphy, K. P., & Suchanek, F. M. (2014). Canonicalizing open knowledge bases. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. (Cité en page 62.)

Gamallo, P. & Garcia, M. (2015). Multilingual open information extraction. In Portuguese Conference on Artificial Intelligence. (Cité en page 62.)

Gashteovski, K., Yu, M., Kotnis, B., Lawrence, C. V., Glavas, G., & Niepert, M. (2021). Benchie : Open information extraction evaluation based on facts, not tokens. ArXiv, abs/2109.06850. (Cité en page 64.)

Gasmi, H., Bouras, A., & Laval, J. (2018). Lstm recurrent neural networks for cybersecurity named entity recognition. ICSEA, 11, 2018. (Cité en page 36.)

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms : From text to predictions. Inf., 13, 83. (Cité en page 66.)

Gemini_Team (2024). Gemini : A family of highly capable multimodal models. (Cité en page 87.)

Gérardin, C., Vaillant, P., Wajsbürt, P., Gilavert, C., Bellamine, A., Kempf, E., & Tannier, X. (2021). Classification multilabel de concepts médicaux pour l'identification du profil clinique du patient (multilabel classification of medical concepts for patient's clinical profile identification). In JEPTALNRECITAL. (Cité en page 66.)

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2007). SemEval-2007 task 04 : classification of semantic relations between nominals. In Proceedings of the 4th International Workshop on Semantic Evaluations - SemEval '07 (pp. 13–18). Prague, Czech Republic : Association for Computational Linguistics. (Cité en page 58.)

Glauber, R. & Claro, D. B. (2018). A systematic mapping study on open information extraction. Expert Syst. Appl., 112, 372–387. (Cité en page 60.)

Gollapalli, S. D. & Caragea, C. (2014). Extracting keyphrases from research papers using citation networks. In AAAI Conference on Artificial Intelligence. (Cité en page 43.)

Gosset, C. (2023). Classification de documents métiers pour l'aide à l'extraction et la classification de relations lexico-sémantiques typées et pondérées. Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA), (pp. 37–40). (Non cité.)

Gosset, C., Billami, M. B., Lafourcade, M., Bortolaso, C., & Derras, M. (2021). Extraction automatique de relations sémantiques d'hyperonymie et d'hyponymie dans un corpus métier (Automatic extraction of hypernym and hyponym relations in a professional corpus). In Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale (pp. 162–170). Lille, France : ATALA. (Non cité.)

Gosset, C., Billami, M. B., Lafourcade, M., Bortolaso, C., & Derras, M. (2022a). Création et validation de représentations vectorielles pour des relations lexico-sémantiques : application à l'identification/classification de relations à partir d'un corpus métier. In Conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC2022). (Non cité.)

Gosset, C., Billami, M. B., Lafourcade, M., Bortolaso, C., & Derras, M. (2022b). Relation-Level Vector Representation for Relation Extraction and Classification on Specialized Data. In International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2022). (Non cité.)

Gotti, F. & Langlais, P. (2018). From french wikipedia to erudit : A test case for cross-domain open information extraction. Computational Intelligence, 34, 420 – 439. (Cité en page 62.)

Grabar, N., Grouin, C., Hamon, T., & Claveau, V. (2019). Recherche et extraction d'information dans des cas cliniques. Présentation de la campagne d'évaluation DEFT 2019. In JEPTALNRECITAL. (Cité en page 66.)

Granada, R. L., Vieira, R., Trojahn, C., & Aussenac-Gilles, N. (2018). Evaluating the complementarity of taxonomic relation extraction methods across different languages. ArXiv, abs/1811.03245. (Cité en page 52.)

Green, S., de Marneffe, M.-C., & Manning, C. D. (2013). Parsing Models for Identifying Multiword Expressions. Computational Linguistics, 39(1), 195–227. (Cité en page 41.)

Grinberg, D., Lafferty, J. D., & Sleator, D. D. (1995). A robust parsing algorithm for link grammars. In International Workshop/Conference on Parsing Technologies. (Cité en page 28.)

Grootendorst, M. (2020). KeyBERT : Minimal keyword extraction with BERT. (Cité en page 41.)

Grootendorst, M. R. (2022). Bertopic : Neural topic modeling with a class-based tf-idf procedure. ArXiv, abs/2203.05794. (Cité en pages 41 et 44.)

Grouin, C., Grabar, N., & Illouz, G. (2021). Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021. In JEPTALNRECITAL. (Cité en page 66.)

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing ? International Journal of Human-Computer Studies, 43(5), 907–928. (Cité en page 20.)

Guarino, N., Oberle, D., & Staab, S. (2009). Handbook on Ontologies, chapter What Is an Ontology ?, (pp. 1–17). Steffen Staab & Rudi Studer : Berlin, Heidelberg, springer berlin heidelberg edition. (Cité en page 19.)

Guillén, R. (2004). Categorization of genomics text based on decision rules. In Text Retrieval Conference. (Cité en page 67.)

Guo, H., Li, X., Zhang, L., Liu, J., & Chen, W. (2021). Label-Aware Text Representation for Multi-Label Text Classification. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7728–7732). Toronto, ON, Canada : IEEE. (Cité en page 67.)

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics, 45(5), 885 – 892. Text Mining and Natural Language Processing in Pharmacogenomics. (Cité en page 58.)

Gábor, K., Buscaldi, D., Schumann, A.-K., QasemiZadeh, B., Zargayouna, H., & Charnois, T. (2018). SemEval-2018 Task 7 : Semantic Relation Extraction and Classification in Scientific Papers. Number 679 in Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018). Association for Computational Linguistics (ACL). (Cité en page 58.)

Gábor, K., Tellier, I., Charnois, T., Zargayouna, H., & Buscaldi, D. (2016). Détection et classification non supervisées de relations sémantiques dans des articles scientifiques. JEP-TALN- RECITAL 2016, (pp.1̃5). (Cité en page 57.)

Han, X., Gao, T., Lin, Y., Peng, H., Yang, Y., Xiao, C., Liu, Z., Li, P., Sun, M., & Zhou, J. (2020). More data, more relations, more context and more openness : A review and outlook for relation extraction. In AACL. (Cité en page 60.)

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., & Sun, M. (2018). Fewrel : A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In Conference on Empirical Methods in Natural Language Processing. (Cité en pages 54, 59 et 64.)

Harabagiu, S. M., Miller, G. A., & Moldovan, D. I. (1999). WordNet 2 - A Morphologically and Semantically Enhanced Resource. Workshop SIGLEX'99 : Standardizing Lexical Resources, (pp. 1–8). (Cité en pages 4 et 63.)

Harris, D. & Harris, S. L. (2022). Digital Design and Computer Architecture, chapter Sequential Logic Design, (pp. 106–169). Morgan Kaufmann : Boston. (Cité en page 26.)

Hasan, K. S. & Ng, V. (2010). Conundrums in unsupervised keyphrase extraction : Making sense of the state-of-the-art. In International Conference on Computational Linguistics. (Cité en page 41.)

Hasan, K. S. & Ng, V. (2014). Automatic keyphrase extraction : A survey of the state of the art. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 42.)

He, L., Lewis, M., & Zettlemoyer, L. (2015). Question-answer driven semantic role labeling : Using natural language to annotate natural language. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 64.)

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th Conference on Computational Linguistics, 2 (COLING'92) (pp. 539—-545). Stroudsburg, PA, USA. (Cité en page 52.)

Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). Cuad : An expert-annotated nlp dataset for legal contract review. ArXiv, abs/2103.06268. (Cité en page 58.)

Hersh, W. R., Bhupatiraju, R. T., Ross, L., Roberts, P. M., Cohen, A. M., & Kraemer, D. (2006). Enhancing access to the bibliome : the trec 2004 genomics track. Journal of Biomedical Discovery and Collaboration, 1, 3 – 3. (Cité en page 66.)

Hiot, N., Minard, A.-L., & Badin, F. (2021). DOING@DEFT : utilisation de lexiques pour une classification efficace de cas cliniques. In DEFT - JEPTALNRECITAL. (Cité en page 66.)

Hobbs, J. R., Appelt, D. E., Bear, J., Israel, D. J., Kameyalna, M., & Tyson, M. (1993). Fastus : A system for extracting information from text. In Human Language Technology - The Baltic Perspectiv. (Cité en page 27.)

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9, 1735–1780. (Cité en pages 27 et 36.)

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., & Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 45.)

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2022). Knowledge Graphs. ACM Comput. Surv., 54(4), 1–37. (Cité en page 22.)

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy : Industrial-strength Natural Language Processing in Python. (Cité en page 83.)

Hou, Y., Che, W., Lai, Y., Zhou, Z., Liu, Y., Liu, H., & Liu, T. (2020). Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1381–1393). Online : Association for Computational Linguistics. (Cité en page 31.)

Hou, Y., Zhou, Z., Liu, Y., Wang, N., Che, W., Liu, H., & Liu, T. (2019). Few-shot sequence labeling with label dependency transfer. ArXiv, abs/1906.08711. (Cité en page 30.)

Hsieh, J.-T., Li, C., & Liu, W. (2017). Effective word representation for named entity recognition. (Cité en page 26.)

Hu, X., Wen, L., Xu, Y., Zhang, C., & Yu, P. S. (2020). Selfore : Self-supervised relational feature learning for open relation extraction. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 64.)

Hu, Z., Huang, P.-Y. B., Deng, Y., Gao, Y., & Xing, E. P. (2015). Entity hierarchy embedding. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 63.)

Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., Peng, B., Gao, J., & Han, J. (2020). Few-shot named entity recognition : A comprehensive study. ArXiv, abs/2012.14978. (Cité en page 29.)

Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., Peng, B., Gao, J., & Han, J. (2021). Few-shot named entity recognition : An empirical baseline study. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 10408–10423). Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. (Cité en page 34.)

Huang, Y., He, K., Wang, Y., Zhang, X., Gong, T., Mao, R., & Li, C. (2022). Copner : Contrastive learning with prompt guiding for few-shot named entity recognition. In International Conference on Computational Linguistics. (Cité en page 37.)

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. ArXiv, abs/1508.01991. (Cité en page 36.)

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 216–223). (Cité en page 43.)

Humphreys, K., Gaizauskas, R. J., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., & Wilks, Y. (1998). University of sheffield : Description of the lasie-ii system as used for muc-7. In Message Understanding Conference. (Cité en page 27.)

Hurault-Plantet, M., Berthelin, J.-B., El Ayari, S., Grouin, C., Loiseau, S., & Paroubek, P. (2008). Résultats de l'édition 2008 du DÉfi Fouille de Textes. In Actes de JEP-TALN-RECITAL Avignon, France : TALN. (Cité en page 66.)

Hwang, H. & Lee, C. (2020). Bert-based korean open information extraction. KIISE Transactions on Computing Practices. (Cité en page 63.)

Issa, B., Jasser, M. B., Chua, H. N., & Hamzah, M. (2023). A comparative study on embedding models for keyword extraction using keybert method. 2023 IEEE 13th International Conference on System Engineering and Technology (ICSET), (pp. 40–45). (Cité en page 44.)

Jameel, S., Bouraoui, Z., & Schockaert, S. (2018). Unsupervised learning of distributional relation vectors. In Proceedings of ACL, volume 1 : Long Paper (pp. 23—-33). Australia. (Cité en page 56.)

Jat, S., Khandelwal, S., & Talukdar, P. P. (2018). Improving distantly supervised relation extraction using word and entity based attention. CoRR, abs/1804.06987. (Cité en page 58.)

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does bert learn about the structure of language ? In Annual Meeting of the Association for Computational Linguistics. (Cité en page 63.)

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023a). A survey on Named Entity Recognition — datasets, tools, and methodologies. Natural Language Processing Journal, 3, 100017. (Cité en page 27.)

Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023b). A survey on named entity recognition — datasets, tools, and methodologies. Natural Language Processing Journal, 3, 100017. (Cité en page 82.)

Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2018). Text classification based on deep belief network and softmax regression. Neural Computing and Applications, (pp. 61–70). (Cité en page 67.)

Jimenez Gutierrez, B., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., & Su, Y. (2022). Thinking about GPT-3 in-context learning for biomedical IE? think again. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), Findings of the Association for Computational Linguistics : EMNLP 2022 (pp. 4497–4512). Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. (Cité en page 54.)

Jin, Y., Xie, J., Guo, W., Luo, C., Wu, D., & Wang, R. (2019). Lstm-crf neural network with gated self attention for chinese ner. IEEE Access, 7, 136694–136703. (Cité en page 36.)

Joshi, M., Choi, E., Levy, O., Weld, D. S., & Zettlemoyer, L. (2019). pair2vec : Compositional word-pair embeddings for cross-sentence inference. In Proceedings of NAACL. (Cité en page 57.)

Justeson, J. S. & Katz, S. M. (1995). Technical terminology : some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1, 9 – 27. (Cité en page 41.)

Kalra, V., Kashyap, I., & Kaur, H. (2022). Classification based topic extraction using domain-specific vocabulary : a supervised approach. Indonesian Journal of Electrical Engineering and Computer Science, 26(1), 442. Number : 1. (Cité en page 67.)

Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the ACL Morristown, NJ, USA. (Cité en page 53.)

Kaur, M. & Kumar, V. (2015). OPTIMIZATION OF TEXT CLASSIFICATION USING SUPERVISED AND UNSUPERVISED LEARNING APPROACH. International Journal of Computer Science and Information Technologies (IJCSIT), 6, 3385–3387. (Cité en page 66.)

Khorashadizadeh, H., Amara, F. Z., Ezzabady, M., Ieng, F., Tiwari, S., Mihindukulasooriya, N., Groppe, J., Sahri, S., Benamara, F., & Groppe, S. (2024). Research trends for the interplay between large language models and knowledge graphs. (Cité en page 23.)

Kim, S. N., Medelyan, O., Kan, M.-Y., & Baldwin, T. (2012). Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation, 47, 723 – 742. (Cité en page 42.)

Kingsbury, P. R. & Palmer, M. (2002). From treebank to propbank. In International Conference on Language Resources and Evaluation. (Cité en page 62.)

Kolluru, K., Adlakha, V., Aggarwal, S., Mausam, M., & Chakrabarti, S. (2020a). Constrained iterative labeling for open information extraction. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 63.)

Kolluru, K., Aggarwal, S., Rathore, V., Mausam, & Chakrabarti, S. (2020b). Imojie : Iterative memory-based joint open information extraction. ArXiv, abs/2005.08178. (Cité en page 63.)

Korkontzelos, I., Piliouras, D., Dowsey, A. W., & Ananiadou, S. (2015). Boosting drug named entity recognition using an aggregate classifier. Artificial Intelligence in Medicine, 65(2), 145–153. (Cité en page 27.)

Kowsari, K., Heidarysafa, M., Brown, D. E., Meimandi, K. J., & Barnes, L. E. (2018). RMDL : Random Multimodel Deep Learning for Classification. Proceedings of the International Conference on Information System and Data Mining. (Cité en page 67.)

Krone, J., Zhang, Y., & Diab, M. (2020). Learning to classify intents and slot labels given a handful of examples. In T.-H. Wen, A. Celikyilmaz, Z. Yu, A. Papangelis, M. Eric, A. Kumar, I. Casanueva, & R. Shah (Eds.), Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI (pp. 96–108). Online : Association for Computational Linguistics. (Cité en page 31.)

Kumar, A. & Starly, B. (2021). "fabner" : information extraction from manufacturing process science domain literature Journal of Intelligent Manufacturing, 33, 2393 – 2407. (Cité en page 47.)

Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. Eng. Appl. Artif. Intell., 70, 25–37. (Cité en page 67.)

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In SNLP'07 : 7th International Symposium on NLP (pp.8̃). (Cité en pages 9 et 17.)

Lafourcade, M. (2011). Lexique et analyse sémantique de textes - structures, acquisitions, calculs, et jeux de mots. PhD thesis, Université Montpellier 2. (Cité en pages 4, 9 et 17.)

Lafourcade, M. & Brun, N. (2020). Jeuxdemots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences. Lexique. (Cité en page 14.)

Lahiri, S., Choudhury, S. R., & Caragea, C. (2014). Keyword and keyphrase extraction using centrality measures on collocation networks. ArXiv, abs/1401.6571. (Cité en page 43.)

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT Beyond English : Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. arXiv :2304.05613 [cs]. (Cité en page 83.)

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In North American Chapter of the Association for Computational Linguistics. (Cité en page 36.)

Lang, K. (1995). Newsweeder : Learning to filter netnews. In A. Prieditis & S. Russell (Eds.), Machine Learning Proceedings 1995 (pp. 331–339). San Francisco (CA). (Cité en page 122.)

Le, T. T. N., Nguyen, M. L., & Shimazu, A. (2016). Unsupervised keyphrase extraction : Introducing new kinds of words to keyphrases. In Australasian Conference on Artificial Intelligence. (Cité en page 42.)

Léchelle, W., Gotti, F., & Langlais, P. (2018). Wire57 : A fine-grained benchmark for open information extraction. ArXiv, abs/1809.08962. (Cité en page 64.)

Lee, J. Y., Dernoncourt, F., & Szolovits, P. (2017). MIT at SemEval-2017 Task 10 : Relation Extraction with Convolutional Neural Networks. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 978–984). Vancouver, Canada : Association for Computational Linguistics. (Cité en page 57.)

Lee, S. & Ko, Y. (2020). Named-entity recognition using automatic construction of training data from social media messaging apps. IEEE Access, 8, 222724–222732. (Cité en page 26.)

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1 : A new benchmark collection for text categorization research. J. Mach. Learn. Res., 5, 361–397. (Cité en page 122.)

Li, C., Li, W., Tang, Z., Li, S., & Xiang, H. (2021). An Improved Term Weighting Method Based on Relevance Frequency for Text Classification. preprint, In Review. (Cité en pages 60 et 67.)

Li, D., Hu, B., & Chen, Q. (2022). Prompt-based text entailment for low-resource named entity recognition. ArXiv, abs/2211.03039. (Cité en page 38.)

Li, J., Shang, S., & Shao, L. (2020). Metaner : Named entity recognition with meta-learning. In Proceedings of The Web Conference 2020, WWW '20 (pp. 429–440). New York, NY, USA : Association for Computing Machinery. (Cité en page 31.)

Li, J., Zhang, X., Zhang, M., & Zhu, X. (2004). Thuir at trec 2004 : Genomics track. In Text Retrieval Conference. (Cité en page 67.)

Li, M. & Zhang, R. (2023). How far is Language Model from 100% Few-shot Named Entity Recognition in Medical Domain. arXiv :2307.00186 [cs]. (Cité en page 39.)

Li, T., Wang, Z., Chai, L., Yang, J., Bai, J., Yin, Y., Liu, J., Guo, H., Yang, L., el abidine, H. Z., & Li, Z. (2023). MT4CrossOIE : Multi-stage Tuning for Cross-lingual Open Information Extraction. ArXiv. (Cité en page 63.)

Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., & Zhang, C. (2020). Bond : Bert-assisted open-domain named entity recognition with distant supervision. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20 (pp. 1054–1064). New York, NY, USA : Association for Computing Machinery. (Cité en page 34.)

Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In AAAI Conference on Artificial Intelligence. (Cité en page 63.)

Litvak, M. & Last, M. (2008). Graph-based keyword extraction for single-document summarization. In International Conference on Computational Linguistics. (Cité en page 43.)

Liu, A., Du, J., & Stoyanov, V. (2019). Knowledge-augmented language model and its application to unsupervised named-entity recognition. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 1142–1150). Minneapolis, Minnesota : Association for Computational Linguistics. (Cité en page 32.)

Liu, J., Pasupat, P., Cyphers, D. S., & Glass, J. R. (2013). Asgard : A portable architecture for multilingual dialogue systems. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, (pp. 8386–8390). (Cité en page 49.)

Liu, L., Shang, J., Xu, F. F., Ren, X., Gui, H., Peng, J., & Han, J. (2017). Empower sequence labeling with task-aware neural language model. In AAAI Conference on Artificial Intelligence. (Cité en page 33.)

Liu, N. F., Hershcovich, D., Kranzlein, M., & Schneider, N. (2021). Lexical semantic recognition. In P. Cook, J. Mitrović, C. P. Escartín, A. Vaidya, P. Osenova, S. Taslimipoor, & C. Ramisch (Eds.), Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021) (pp. 49–56). Online : Association for Computational Linguistics. (Cité en pages 44 et 49.)

Liu, P., Gao, W., Dong, W., Huang, S., & Zhang, Y. (2022). Open information extraction from 2007 to 2022 - a survey. ArXiv, abs/2208.08690. (Cité en pages 60 et 64.)

Liu, R., Xu, G., Jia, C., Ma, W., Wang, L., & Vosoughi, S. (2020a). Data boost : Text data augmentation through reinforcement learning guided conditional generation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9031–9041). Online : Association for Computational Linguistics. (Cité en page 31.)

Liu, X. & Zhou, M. (2013). Two-stage NER for tweets with clustering. Information Processing & Management, 49(1), 264–273. (Cité en page 29.)

Liu, Y. & Yang, B. (2012). Joint inference : a statistical approach for open information extraction. Applied Mathematics & Information Sciences, 2, 627S–633S. (Cité en page 63.)

Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 43.)

Liu, Z., Huang, Y., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Li, Y., Shu, P., Zeng, F., Sun, L., Liu, W., Shen, D., Li, Q., Liu, T., Zhu, D., & Li, X. (2023). Deid-gpt : Zero-shot medical text de-identification by gpt-4. (Cité en page 40.)

Liu, Z., Li, P., Zheng, Y., & Sun, M. (2009). Clustering to find exemplar terms for keyphrase extraction. In Conference on Empirical Methods in Natural Language Processing. (Cité en pages 42 et 43.)

Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., & Fung, P. (2020b). Crossner : Evaluating cross-domain named entity recognition. In AAAI Conference on Artificial Intelligence. (Cité en page 48.)

Lou, Y., Qian, T., Li, F., & Ji, D. (2020). A graph attention model for dictionary-guided named entity recognition. IEEE Access, 8, 71584–71592. (Cité en page 33.)

Lu, K., Pan, X., Song, K., Zhang, H., Yu, D., & Chen, J. (2023). Pivoine : Instruction tuning for open-world information extraction. ArXiv, abs/2305.14898. (Cité en pages 63 et 71.)

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific In Conference on Empirical Methods in Natural Language Processing. (Cité en pages 46 et 55.)

Luo, L., Ning, J., Zhao, Y., Wang, Z., Ding, Z., Chen, P., Fu, W., Han, Q., Xu, G., Qiu, Y., Pan, D., Li, J., Li, H., Feng, W., Tu, S., Liu, Y., Yang, Z., Wang, J., Sun, Y., & Lin, H. (2023). Taiyi : A bilingual fine-tuned large language model for diverse biomedical tasks. Journal of the American Medical Informatics Association : JAMIA. (Cité en pages 40 et 54.)

Ma, R., Zhou, X., Gui, T., Tan, Y. C., Zhang, Q., & Huang, X. (2021). Template-free prompt tuning for few-shot ner. ArXiv, abs/2109.13532. (Cité en page 37.)

Ma, T., Jiang, H., Wu, Q., Zhao, T., & Lin, C.-Y. (2022). Decomposed meta-learning for few-shot named entity recognition. In Findings of the Association for Computational Linguistics : ACL 2022 (pp. 1584–1596). Dublin, Ireland : Association for Computational Linguistics. (Cité en page 38.)

Ma, Y., Zhang, P., & Ma, J. (2018). An Ontology Driven Knowledge Block Summarization Approach for Chinese Judgment Document Classification. IEEE Access, 6, 71327–71338. Conference Name : IEEE Access. (Cité en page 22.)

Magnini, B., Altuna, B., Lavelli, A., Speranza, M., & Zanoli, R. (2021). The e3c project : European clinical case corpus. In Annual Conference of the Spanish Society for Natural Language Processing. (Cité en page 47.)

Manakul, P., Liusie, A., & Gales, M. J. (2023). Selfcheckgpt : Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv :2303.08896. (Cité en page 44.)

Mannai, M., Karâa, W. B. A., & Ghezala, H. H. B. (2018). Information Extraction Approaches : A Survey. In D. K. Mishra, A. T. Azar, & A. Joshi (Eds.), Information and Communication Technology, Advances in Intelligent Systems and Computing (pp. 289–297). Singapore : Springer. (Cité en page 22.)

Martin, L., Muller, B., Suarez, P. O., Dupont, Y., Romary, L., de la Clergerie, E. V., Seddah, D., & Sagot, B. (2019). Camembert : a tasty french language model. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 83.)

Martínez-Romo, J., Araujo, L., & Fernandez, A. D. (2016). Semgraph : Extracting keyphrases following a novel semantic graph-based approach. Journal of the Association for Information Science and Technology, 67. (Cité en page 43.)

Marujo, L., Gershman, A., Carbonell, J., Frederking, R., & ao P. Neto, J. (2012). Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and coreference normalization. In Proceedings of LREC 2012 : ELRA. (Cité en page 42.)

Mausam, M. (2016). Open information extraction systems and downstream applications. In International Joint Conference on Artificial Intelligence. (Cité en page 63.)

Mausam, M., Schmitz, M., Soderland, S., Bart, R., & Etzioni, O. (2012). Open language learning for information extraction. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 63.)

Mihalcea, R. & Tarau, P. (2004). TextRank : Bringing Order into Texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) Barcelona, Spain. (Cité en pages 42, 43 et 84.)

Mikheev, A., Grover, C., & Moens, M. (1998). Description of the ltg system used for muc-7. In Message Understanding Conference. (Cité en page 27.)

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Proceedings of the International Conference on Learning Representations (ICLR) (pp. 1–12). (Cité en pages 26, 52 et 120.)

Miller, S., Fox, H., Ramshaw, L., & Weischedel, R. (2000). A Novel Use of Statistical Parsing to Extract Information from Text. In "1st Meeting of the NAACL". (Cité en page 53.)

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations : What makes in-context learning work? ArXiv, abs/2202.12837. (Cité en page 38.)

Minard, A.-L., Ligozat, A.-L., & Grau, B. (2011). Multi-class SVM for Relation Extraction from Clinical Reports. Proceedings of Recent Advances in Natural Language Processing, (pp.6̃). (Cité en page 57.)

Mintz, M. D., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In Annual Meeting of the Association for Computational Linguistics. (Cité en pages 32, 53 et 64.)

Morin, E. & Jacquemin, C. (2004). Automatic Acquisition and Expansion of Hypernym Links. Computers and the Humanities, 38, 363–396. (Cité en page 53.)

Moscato, V., Postiglione, M., & Sperlí, G. (2023). Few-shot named entity recognition : Definition, taxonomy and research directions. ACM Transactions on Intelligent Systems and Technology, 14, 1 – 46. (Cité en page 30.)

Na, S.-H., Kim, H., Min, J., & Kim, K. (2019). Improving LSTM CRFs using character-based compositions for Korean named entity recognition. Computer Speech & Language, 54, 106–121. (Cité en page 36.)

Nakov, P., Schwartz, A. S., Stoica, E., & Hearst, M. A. (2004). Biotext team experiments for the trec 2004 genomics track. In Text Retrieval Conference. (Cité en page 67.)

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T. E., Sastry, G., Krueger, G., Schnurr, D., Such, F. P., Hsu, K., Thompson, M., Khan, T., Sherbakov, T., Jang, J., Welinder, P., & Weng, L. (2022). Text and Code Embeddings by Contrastive Pre-Training. (Cité en page 33.)

Nguyen, T. H. & Grishman, R. (2015). Relation Extraction : Perspective from Convolutional Neural Networks. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing (pp. 39–48). Denver, Colorado : Association for Computational Linguistics. (Cité en page 57.)

Ni, J., Dinu, G., & Florian, R. (2017). Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In R. Barzilay & M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers) (pp. 1470–1480). Vancouver, Canada : Association for Computational Linguistics. (Cité en page 30.)

Ni, T., Wang, Q., & Ferraro, G. (2021). Explore bilstm-crf-based models for open relation extraction. ArXiv, abs/2104.12333. (Cité en page 63.)

Niklaus, C. M., Cetto, M., Freitas, A., & Handschuh, S. (2018). A Survey on Open Information Extraction. In International Conference on Computational Linguistics. (Cité en pages 52, 60 et 65.)

Nilashi, M., Ibrahim, O., Yadegaridehkordi, E., Samad, S., Akbari, E., & Alizadeh, A. (2018). Travelers decision making using online review in social network sites : A case on tripadvisor. Journal of computational science, 28, 168–179. (Cité en page 22.)

Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence, 194, 151–175. (Cité en page 28.)

Oguz, C. & Vu, N. T. (2021). Few-shot learning for slot tagging with attentive relational network. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume (pp. 1566–1572). Online : Association for Computational Linguistics. (Cité en page 31.)

Pal, H. & Mausam, M. (2016). Demonyms and compound relational nouns in nominal open ie. In AKBC@NAACL-HLT. (Cité en page 63.)

Palacharla, R. K. & Vatsavayi, V. K. (2022). A New Supervised Term Weight Measure Based Approach for Text Classification. Revue d'Intelligence Artificielle, 36(3), 395–407. Number : 3. (Cité en page 67.)

Panchenko, A., Faralli, S., Ruppert, E., Remus, S., Naets, H., Fairon, C., Ponzetto, S. P., & Biemann, C. (2016). TAXI at SemEval-2016 Task 13 : a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling. In SemEval@NAACL-HLT. (Cité en page 52.)

Panchenko, A., Naets, H., Brouwers, L., & Fairon, C. (2013). Recherche et visualisation de mots sémantiquement liés. In TALN-RÉCITAL (pp. 747—-754). (Cité en page 52.)

Panovski, D. & Titus, Z. (2015). Moteur de recherche web avec l'ontologie WOLF. In Traitement et Analyse de l'Information Méthodes et Applications (TAIMA 2015) Hammamet , Tunisia. (Cité en page 22.)

Pantel, P. & Pennacchiotti, M. (2006). Espresso : Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In Association for Computational Linguistics (ACL). (Cité en page 53.)

Park, Y., Byrd, R. J., & Boguraev, B. (2002). Automatic glossary extraction : Beyond terminology identification. In International Conference on Computational Linguistics. (Cité en page 41.)

Patil, N., Patil, A., & Pawar, B. V. (2020). Named Entity Recognition using Conditional Random Fields. Procedia Computer Science, 167, 1181–1188. (Cité en page 28.)

Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation Extraction : A Survey. (Cité en page 52.)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. (Cité en page 26.)

Peng, N., Poon, H., Quirk, C., Toutanova, K., & Yih, W.-t. (2017). Cross-Sentence N -ary Relation Extraction with Graph LSTMs. TACL, 5, 101–115. (Cité en page 57.)

Peng, Q., Zheng, C., Cai, Y., Wang, T., Xie, H., & Li, Q. (2021). Unsupervised cross-domain named entity recognition using entity-aware adversarial training. Neural Networks, 138, 68–77. (Cité en page 29.)

Pennington, J., Socher, R., & Manning, C. D. (2014a). GloVe : Global vectors for word representation. In Proceedings of EMNLP (pp. 1532–1543). (Cité en page 52.)

Pennington, J., Socher, R., & Manning, C. D. (2014b). GloVe : Global vectors for word representation. In Proceedings of EMNLP (pp. 1532—-1543). (Cité en page 56.)

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In North American Chapter of the Association for Computational Linguistics (NAACL). (Cité en page 52.)

Pirkola, A. (2004). Trec 2004 genomics track experiments at uta : The effects of primary keys, bigram phrases and query expansion on retrieval performance. In Text Retrieval Conference. (Cité en page 67.)

Plantié, M., Roche, M., & Dray, G. (2008). Défi DEFT08 : Classification de textes en genre et en thème : Votons utile ! In Défi DEFT08 (pp.8̃). (Cité en page 66.)

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Bjorkelund, A., Uryupina, O., Zhang, Y., & Zhong, Z. (2013a). Towards Robust Linguistic Analysis using OntoNotes. In Seventeenth Conference on Computational Natural Language Learning (pp. 143–152). Sofia, Bulgaria. (Cité en page 36.)

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., O., Zhang, Y., & Zhong, Z. (2013b). Towards robust linguistic analysis using ontonotes. In Conference on Computational Natural Language Learning. (Cité en page 46.)

Priyanshu, A. & Vijay, S. (2022). Adaptkeybert : An attention-based approach towards few-shot & zero-shot domain adaptation of keybert. ArXiv, abs/2211.07499. (Cité en page 44.)

Punyakanok, V., Roth, D., & tau Yih, W. (2008). The importance of syntactic parsing and inference in semantic role labeling. Computational Linguistics, 34, 257–287. (Cité en page 62.)

Pyysalo, S. & Ananiadou, S. (2014). Anatomical entity mention recognition at literature scale. Bioinformatics, 30(6), 868–875. (Cité en page 49.)

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. In System Demonstrations, ACL. (Cité en pages 83 et 128.)

Qian, L., Zhou, G., Kong, F., Zhu, Q., & Qian, P. (2008). Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08, volume 1 (pp. 697–704). Manchester, United Kingdom : Association for Computational Linguistics. (Cité en page 57.)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI blog. (Cité en page 31.)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67. (Cité en page 48.)

Rahimi, A., Li, Y., & Cohn, T. (2019). Massively multilingual transfer for NER. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 151–164). Florence, Italy : Association for Computational Linguistics. (Cité en pages 30 et 49.)

Rajaraman, A. & Ullman, J. D. (2011). Data Mining, (pp. 1–17). Cambridge University Press. (Cité en pages 26, 41 et 42.)

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Mititelu, V. B., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Inurrieta, U., Kovalevskaite, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Escartín, C. P., Zadeh, B. Q., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., & Walsh, A. (2018). Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In LAW-MWE-CxG@COLING. (Cité en pages 41, 43 et 49.)

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Mititelu, V. B., Bhatia, A., Inurrieta, U., Giouli, V., Gungor, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., & Xu, H. (2020). Edition 1.2 of the parseme shared task on semi-supervised identification of verbal multiword expressions. In Workshop on Multiword Expressions. (Cité en pages 43 et 49.)

Rani, P., Gujral, R., & Jain, A. (2022). A hybrid approach for feature selection based on correlation feature selection and genetic algorithm. International Journal of Software Innovation, 10, 1–17. (Cité en page 22.)

Reymonet, A., Thomas, J., & Aussenac-Gilles, N. (2010). Ontologies et Recherche d'Information : une application au diagnostic automobile. 1èmes Journées Francophones d'Ingénierie des Connaissances, (pp.13). (Cité en page 22.)

Riedel, S., Yao, L., & McCallum, A. (2010). Modeling relations and their mentions without labeled text. In J. L. Balcázar, F. Bonchi, A. Gionis, & M. Sebag (Eds.), Machine Learning and Knowledge Discovery in Databases (pp. 148–163). Berlin, Heidelberg : Springer Berlin Heidelberg. (Cité en pages 45 et 64.)

Rose, S. J., Engel, D. W., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents, chapter 1, (pp. 1–20). John Wiley & Sons, Ltd. (Cité en page 42.)

Roth, D. & tau Yih, W. (2004). A linear programming formulation for global inference in natural language tasks. In Conference on Computational Natural Language Learning. (Cité en page 45.)

Safranchik, E., Luo, S., & Bach, S. H. (2020). Weakly supervised sequence tagging from noisy rules. In AAAI Conference on Artificial Intelligence. (Cité en page 33.)

Saied, H. A., Candito, M., & Constant, M. (2018). A transition-based verbal multiword expression analyzer. (Cité en page 41.)

Salaün, O., Piedboeuf, F., Berre, G. L., Hermelo, D. A., & Langlais, P. (2024). Europa : A legal multilingual keyphrase generation dataset. (Cité en pages 49 et 51.)

Salton, G. (1988). Syntactic approaches to automatic book indexing. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 41.)

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523. (Cité en pages 26, 41 et 42.)

Sanchez-Pi, N., Martí, L., & Bicharra Garcia, A. C. (2016). Improving ontology-based text classification : An occupational health and security application. Journal of Applied Logic, 17, 48–58. (Cité en page 22.)

Sandhaus, E. (2008). The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, 6(12), e26752. (Cité en page 45.)

Sang, E. F. T. K. & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition. arXiv :cs/0306050. (Cité en page 36.)

Sang, E. T. K. & Meulder, F. D. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. In Conference on Computational Natural Language Learning. (Cité en pages 27 et 45.)

Sari, Y., Hassan, M. F., & Zamin, N. (2010). Rule-based pattern extractor and named entity recognition : A hybrid approach. In 2010 International Symposium on Information Technology, volume 2 (pp. 563–568). (Cité en page 28.)

Savary, A., Cordeiro, S., & Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly : A position statement. In A. Savary, C. P. Escartín, F. Bond, J. Mitrović, & V. B. Mititelu (Eds.), Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019) (pp. 79–91). Florence, Italy : Association for Computational Linguistics. (Cité en page 41.)

Schneider, N., Hovy, D., Johannsen, A., & Carpuat, M. (2016). SemEval-2016 task 10 : Detecting minimal semantic units and their meanings (DiMSUM). In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 546–559). San Diego, California : Association for Computational Linguistics. (Cité en pages 41 et 49.)

Schneider, N. & Smith, N. A. (2015). A corpus and model integrating multiword expressions and supersenses. In North American Chapter of the Association for Computational Linguistics. (Cité en page 49.)

Seganti, A., Firląg, K., Skowronska, H., Satława, M., & Andruszkiewicz, P. (2021). Multilingual entity and relation extraction dataset and model. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume (pp. 1946–1955). Online : Association for Computational Linguistics. (Cité en page 59.)

Seki, K., Costello, J. C., Singan, V. R., & Mostafa, J. (2004). Trec 2004 genomics track experiments at iub. In Text Retrieval Conference. (Cité en page 67.)

Shang, J., Liu, L., Gu, X., Ren, X., Ren, T., & Han, J. (2018). Learning named entity tagger using domain-specific dictionary. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2054–2064). Brussels, Belgium : Association for Computational Linguistics. (Cité en page 32.)

Sharma, P. & Li, Y. (2019). Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling. Preprints. Publisher : Preprints. (Cité en page 44.)

Shen, Y., Yun, H., Lipton, Z., Kronrod, Y., & Anandkumar, A. (2017). Deep active learning for named entity recognition. In P. Blunsom, A. Bordes, K. Cho, S. Cohen, C. Dyer, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, & S. Yih (Eds.), Proceedings of the 2nd Workshop on Representation Learning for NLP (pp. 252–256). Vancouver, Canada : Association for Computational Linguistics. (Cité en page 33.)

Singh, T. D., Nongmeikapam, K., Ekbal, A., Bandyopadhyay, S., & Ekbal, A. (2009). Named entity recognition for manipuri using support vector machine. In Pacific Asia Conference on Language, Information and Computation. (Cité en page 28.)

Smirnova, A. & Cudré-Mauroux, P. (2019). Relation Extraction Using Distant Supervision : A Survey. ACM Comput. Surv., 51(5), 1–35. Number : 5. (Cité en page 52.)

Snow, R., Jurafsky, D., & Ng, A. (2004). Learning Syntactic Patterns for Automatic Hypernym Discovery. In Advances in Neural Information Processing Systems (NIPS 2004). (Cité en page 53.)

Solawetz, J. & Larson, S. (2021). Lsoie : A large-scale dataset for supervised open information extraction. In Conference of the European Chapter of the Association for Computational Linguistics. (Cité en page 64.)

Speer R., C. J. e. H. C. (2017). Conceptnet 5.5 : An open multilingual graph of general knowledge. In Proceedings of AAAI (pp. 4444—4451). (Cité en page 57.)

Srihari, R. K. & Li, W. (1999). Information extraction supported question answering. In Text Retrieval Conference. (Cité en page 25.)

Stanovsky, G. & Dagan, I. (2016). Creating a large benchmark for open information extraction. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 64.)

Stanovsky, G., Michael, J., Zettlemoyer, L., & Dagan, I. (2018). Supervised open information extraction. In North American Chapter of the Association for Computational Linguistics. (Cité en page 64.)

Sterckx, L., Demeester, T., Deleu, J., & Develder, C. (2015). Topical Word Importance for Fast Keyphrase Extraction. In Proceedings of the 24th International Conference on World Wide Web (pp. 121–122). Florence Italy : ACM. (Cité en page 43.)

Stodden, R., Zadeh, B. Q., & Kallmeyer, L. (2018). TRAPACC and TRAPACCS at PARSEME Shared Task 2018 : Neural Transition Tagging of Verbal Multiword Expressions. In LAW-MWE-CxG@COLING. (Cité en page 43.)

Stoica, G., Platanios, E. A., & Póczos, B. (2021). Re-tacred : Addressing shortcomings of the TACRED dataset. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021 (pp. 13843–13850). : AAAI Press. (Cité en page 59.)

Sun, M., Li, X., Wang, X., Fan, M., Feng, Y., & Li, P. (2018). Logician : A unified end-to-end neural approach for open-domain information extraction. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. (Cité en page 64.)

Suresh kumar, G. & Zayaraz, G. (2015). Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems. Journal of King Saud University - Computer and Information Sciences, 27(1), 13–24. Number : 1. (Cité en page 57.)

Sutton, C. & McCallum, A. (2007). An Introduction to Conditional Random Fields for Relational Learning. In Introduction to Statistical Relational Learning. The MIT Press. (Cité en page 57.)

Takanobu, R., Zhang, T., Liu, J., & Huang, M. (2019). A hierarchical framework for relation extraction with reinforcement learning. In AAAI (pp. 7072–7079). (Cité en page 58.)

Tang, Z.-S., Li, W., & Li, Y. (2021). An improved supervised term weighting scheme for text representation and classification. Expert Syst. Appl., 189, 115985. (Cité en page 67.)

Taniar, D. & Rahayu, J. W. (2006). Web Semantics Ontology. Hershey, PA : IGI Global. (Cité en page 20.)

Taslimipoor, S., Bahaadini, S., & Kochmar, E. (2020). Mtlb-struct @parseme 2020 : Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. ArXiv, abs/2011.02541. (Cité en page 43.)

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., & Navigli, R. (2021). WikiNEuRal : Combined neural and knowledge-based silver data creation for multilingual NER. In Findings of the Association for Computational Linguistics : EMNLP 2021 (pp. 2521–2533). Punta Cana, Dominican Republic : Association for Computational Linguistics. (Cité en pages 49 et 87.)

Tedeschi, S. & Navigli, R. (2022). Multinerd : A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In NAACL-HLT. (Cité en page 46.)

Thomas, A. & Sangeetha, S. (2021). Semi-supervised, knowledge-integrated pattern learning approach for fact extraction from judicial text. Expert Systems, 38. (Cité en page 54.)

Toutanova, K., Haghighi, A., & Manning, C. D. (2008). A global joint model for semantic role labeling. Computational Linguistics, 34, 161–191. (Cité en page 62.)

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03, volume 1 (pp. 173–180). Edmonton, Canada : Association for Computational Linguistics. (Cité en page 28.)

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. Information Retrieval, 2, 303–336. (Cité en pages 40 et 42.)

Turney, P. D. (2005). Measuring semantic similarity by latent relational analysis. In Proceedings of IJCAI (pp. 1136—-1141). (Cité en page 56.)

Upendraa, B. & Sudheer, B. (2016). Knn tfidf based named entity recognition. International Journal of Scientific and Research, 1(12), 35–39. (Cité en page 26.)

Urchade, Z., Tomeh, N., Holat, P., & Charnois, T. (2024). An autoregressive text-to-graph framework for joint entity and relation extraction. (Cité en page 53.)

Ushio, A., Barbieri, F., Sousa, V., Neves, L., & Camacho-Collados, J. (2022). Named entity recognition in Twitter : A dataset and analysis on short-term temporal shifts. In Y. He, H. Ji, S. Li, Y. Liu, & C.-H. Chang (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers) (pp. 309–319). Online only : Association for Computational Linguistics. (Cité en pages 38 et 49.)

Ushio, A. & Camacho-Collados, J. (2021). T-NER : An all-round python library for transformer-based named entity recognition. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations (pp. 53–62). Online : Association for Computational Linguistics. (Cité en page 38.)

Uzuner, O., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5), 552–556. Number : 5. (Cité en page 57.)

Vallet, D., Castells, P., Fernández, M., Mylonas, P., & Avrithis, Y. (2007). Personalized content retrieval in context using ontological knowledge. IEEE Transactions on Circuits and Systems for Video Technology, 17, 336–346. (Cité en page 22.)

Vashishth, S., Jain, P., & Talukdar, P. P. (2018). Cesi : Canonicalizing open knowledge bases using embeddings and side information. Proceedings of the 2018 World Wide Web Conference. (Cité en page 62.)

Velardi, P., Missikoff, M., & Basili, R. (2001). Identification of relevant terms to support the construction of domain ontologies. In Proceedings of the workshop on Human Language Technology and Knowledge Management - HTLKM@ACL (pp. 1–8). : Association for Computational Linguistics. (Cité en pages 41 et 42.)

Vijay, S. & Priyanshu, A. (2022). NERDA-Con : Extending NER models for Continual Learning – Integrating Distinct Tasks and Updating Distribution Shifts. arXiv :2206.14607 [cs]. (Cité en page 39.)

Vlachos, A. (2011). Evaluating unsupervised learning for natural language processing tasks. In ULNLP@EMNLP. (Cité en page 29.)

Vo, D.-T. & Bagheri, E. (2016). Open information extraction. Encycl. Semantic Comput. Robotic Intell., 1, 1630003 :1–1630003 :6. (Cité en pages 60 et 63.)

Voorhees, E. M. (2004). Overview of trec 2004. In Text Retrieval Conference. (Cité en page 66.)

Vrandečić, D. (2009). Ontology Evaluation. In S. Staab & R. Studer (Eds.), Handbook on Ontologies (pp. 293–313). Berlin, Heidelberg : Springer Berlin Heidelberg. (Cité en page 20.)

Walker, C., Strassel, S., Medero, J., & Maeda, K. (2006). Ace 2005 multilingual training corpus ldc2006t06. Linguistic Data Consortium. LDC2006T06 : Web Download. (Cité en pages 45 et 55.)

Wan, X. & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In AAAI Conference on Artificial Intelligence. (Cité en page 43.)

Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., & Kurohashi, S. (2023). Gpt-re : In-context learning for relation extraction using large language models. ArXiv, abs/2305.02105. (Cité en page 55.)

Wang, J., Xu, W., Fu, X., Xu, G., & Wu, Y. (2020). ASTRAL : Adversarial Trained LSTM-CNN for Named Entity Recognition. Knowledge-Based Systems, 197, 105842. (Cité en page 35.)

Wang, K., Ding, Y., & Han, S. C. (2023a). Graph neural networks for text classification : A survey. ArXiv, abs/2304.11534. (Cité en page 66.)

Wang, L., Cao, Z., de Melo, G., & Liu, Z. (2016). Relation Classification via Multi-Level Attention CNNs. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers) (pp. 1298–1307). Berlin, Germany : Association for Computational Linguistics. (Cité en page 57.)

Wang, R., Liu, W., & Mcdonald, C. (2014a). Corpus-independent generic keyphrase extraction using word embedding vectors. In Software engineering research conference, volume 39 (pp. 1–8). (Cité en page 42.)

Wang, X., Zhou, W., Zu, C., Xia, H., Chen, T., Zhang, Y., Zheng, R., Ye, J., Zhang, Q., Gui, T., Kang, J., Yang, J., Li, S., & Du, C. (2023b). Instructuie : Multi-task instruction tuning for unified information extraction. ArXiv, abs/2304.08085. (Cité en page 25.)

Wang, Z., Qu, Y., Chen, L., Shen, J., Zhang, W., Zhang, S., Gao, Y., Gu, G., Chen, K., & Yu, Y. (2018). Label-aware double transfer learning for cross-specialty medical named entity recognition. In M. Walker, H. Ji, & A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers) (pp. 1–15). New Orleans, Louisiana : Association for Computational Linguistics. (Cité en page 30.)

Wang, Z., Song, H., Ren, Z., Ren, P., Chen, Z., Liu, X., Li, H., & de Rijke, M. (2021). Cross-domain contract element extraction with a bi-directional feedback clause-element relation network. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. (Cité en page 55.)

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014b). Knowledge graph and text jointly embedding. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 63.)

Washio, K. & Kato, T. (2018). Filling missing paths : Modeling co-occurrences of word pairs and dependency paths for recognizing lexical semantic relations. In NAACL (pp. 1123–1133). (Cité en page 57.)

Waszczuk, J. (2018). Traversal at parseme shared task 2018 : Identification of verbal multiword expressions using a discriminative tree-structured model. In LAW-MWE-CxG@COLING. (Cité en page 41.)

Wei, C.-H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., Wiegers, T. C., & Lu, Z. (2016). Assessing the state of the art in biomedical relation extraction : overview of the biocreative v chemical-disease relation (cdr) task. Database, 2016. (Cité en page 49.)

White, A. S., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger, R., Rawlins, K., & Durme, B. V. (2016). Universal decompositional semantics on universal dependencies. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 62.)

Wimalasuriya, D. C. & Dejing Dou (2010). Ontology-based information extraction : An introduction and a survey of current approaches. Journal of Information Science, 36(3), 306–323. (Cité en page 22.)

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). Kea : Practical automatic keyphrase extraction. In Proceedings of the Fourth ACM Conference on Digital Libraries, DL '99 (pp. 254–255). New York, NY, USA : Association for Computing Machinery. (Cité en pages 40 et 42.)

Wong, W., Liu, W., & Bennamoun (2007). Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency. In Australasian Data Mining Conference (pp.8̃). (Cité en page 40.)

Wu, F. & Weld, D. S. (2010). Open information extraction using wikipedia. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 61.)

Wu, T.-H., Wu, Z., Kao, B., & Yin, P. (2018). Towards practical open knowledge base canonicalization. Proceedings of the 27th ACM International Conference on Information and Knowledge Management. (Cité en page 62.)

Xia, Y., Zhao, Y., Wu, W., & Li, S. (2023). Debiasing generative named entity recognition by calibrating sequence likelihood. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 27.)

Xie, C., Liang, J., Liu, J., Huang, C., Huang, W., & Xiao, Y. (2021). Revisiting the negative data of distantly supervised relation extraction. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 57.)

Xie, J., Yang, Z., Neubig, G., Smith, N. A., & Carbonell, J. (2018). Neural cross-lingual named entity recognition with minimal resources. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 369–379). Brussels, Belgium : Association for Computational Linguistics. (Cité en page 30.)

Xie, R., Liu, Z., & Sun, M. (2016). Representation learning of knowledge graphs with hierarchical types. In International Joint Conference on Artificial Intelligence. (Cité en page 63.)

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., & Wang, H. (2023). Empirical study of zero-shot ner with chatgpt. (Cité en page 39.)

Xinwei, W. & Hui, Z. (2020). Open information extraction for waste incineration nimby based on bert network in china. Journal of Physics : Conference Series, 1576. (Cité en page 63.)

Xiong, W., Song, Y., Wang, P., & Li, S. (2023). Rationale-enhanced language models are better continual relation learners. (Cité en page 54.)

Xu, K., Feng, Y., Huang, S., & Zhao, D. (2015). Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 536–540). Lisbon, Portugal : Association for Computational Linguistics. (Cité en page 57.)

Xu, K., Yang, Z., Kang, P., Wang, Q., & Liu, W. (2019). Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. Computers in Biology and Medicine, 108, 122–132. (Cité en page 36.)

Xu, W., Deng, Y., Lei, W., Zhao, W., seng Chua, T., & Lam, W. (2022a). Conreader : Exploring implicit relations in contracts for contract clause extraction. ArXiv, abs/2210.08697. (Cité en page 55.)

Xu, W., Li, X., Zhang, W., Zhou, M., Bing, L., Lam, W., & Si, L. (2022b). From clozing to comprehending : Retrofitting pre-trained language model to pre-trained machine reader. ArXiv, abs/2212.04755. (Cité en page 37.)

Xue, L., Qing, S., & Pengzhou, Z. (2018). Relation Extraction Based on Deep Learning. 17th International Conference on Computer and Information Science (ICIS), (pp. 687–691). (Cité en page 52.)

Yahya, M., Whang, S. E., Gupta, R., & Halevy, A. Y. (2014). Renoun : Fact extraction for nominal attributes. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 63.)

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., & Qiu, X. (2021). A unified generative framework for various ner subtasks. (Cité en page 36.)

Yang, L., Yuan, L., Cui, L., Gao, W., & Zhang, Y. (2022). Factmix : Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. ArXiv, abs/2208.11464. (Cité en page 37.)

Yang, Y., Chen, W., Li, Z., He, Z., & Zhang, M. (2018). Distantly supervised NER with partial annotation learning and reinforcement learning. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics (pp. 2159–2169). Santa Fe, New Mexico, USA : Association for Computational Linguistics. (Cité en page 32.)

Yang, Y. & Katiyar, A. (2020). Frustratingly simple few-shot named entity recognition with structured nearest neighbor learning. ArXiv, abs/2010.02405. (Cité en page 27.)

Yao, L., Haghighi, A., Riedel, S., & McCallum, A. (2011). Structured relation discovery using generative models. In Conference on Empirical Methods in Natural Language Processing. (Cité en page 64.)

Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., & Soderland, S. (2007). Textrunner : Open information extraction on the web. In North American Chapter of the Association for Computational Linguistics. (Cité en page 60.)

Yu, H., Mao, X.-L., Chi, Z., Wei, W., & Huang, H. (2020). A robust and domain-adaptive approach for low-resource named entity recognition. In 2020 IEEE International Conference on Knowledge Graph (ICKG) (pp. 297–304). (Cité en page 34.)

Yue, H., Lai, S., Yang, C., Zhang, L., Yao, J., & Su, J. (2024). Towards better graph-based cross-document relation extraction via non-bridge entity enhancement and prediction debiasing. (Cité en page 53.)

Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2023). Gliner : Generalist model for named entity recognition using bidirectional transformer. (Cité en pages 37 et 86.)

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation Classification via Convolutional Deep Neural Network. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers, (pp.1̃0). (Cité en page 57.)

Zeng, X., Li, Y., Zhai, Y., & Zhang, Y. (2020). Counterfactual generator : A weakly-supervised method for named entity recognition. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7270–7280). Online : Association for Computational Linguistics. (Cité en page 31.)

Zhan, J. & Zhao, H. (2019). Span model for open information extraction on accurate corpus. In AAAI Conference on Artificial Intelligence. (Cité en page 64.)

Zhang, D. & Wang, D. (2015). Relation classification via recurrent neural network. CoRR, abs/1508.01006. (Cité en page 58.)

Zhang, J., Shen, D., Zhou, G., Su, J., & Tan, C.-L. (2004). Enhancing hmm-based biomedical named entity recognition by studying special phenomena. Journal of Biomedical Informatics, 37(6), 411–422. Named Entity Recognition in Biomedicine. (Cité en page 28.)

Zhang, K., Yao, Y., Xie, R., Han, X., Liu, Z., Lin, F., Lin, L., & Sun, M. (2021). Open hierarchical relation extraction. In North American Chapter of the Association for Computational Linguistics. (Cité en page 63.)

Zhang, R., Li, Y., Ma, Y., Zhou, M., & Zou, L. (2023). Llmaaa : Making large language models as active annotators. ArXiv, abs/2310.19596. (Cité en page 59.)

Zhang, R., Zhao, P., Guo, W., Wang, R., & Lu, W. (2022). Medical named entity recognition based on dilated convolutional neural network. Cognitive Robotics, 2, 13–20. (Cité en page 35.)

Zhang, S. & Elhadad, N. (2013). Unsupervised biomedical named entity recognition : Experiments with clinical and biological texts. Journal of Biomedical Informatics, 46(6), 1088–1098. (Cité en page 29.)

Zhang, Y., Milios, E. E., & Zincir-Heywood, N. (2007). A comparative study on key phrase extraction methods in automatic web site summarization. J. Digit. Inf. Manag., 5, 323–332. (Cité en pages 42 et 43.)

Zhang, Y., Zhong, V., Chen, D., Angeli, G., & Manning, C. D. (2017a). Position-aware attention and supervised data improve slot filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017) (pp. 35–45). (Cité en pages 54, 55 et 59.)

Zhang, Y., Zhong, V., Chen, D., Angeli, G., & Manning, C. D. (2017b). Position-aware attention and supervised data improve slot filling. In M. Palmer, R. Hwa, & S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 35–45). Copenhagen, Denmark : Association for Computational Linguistics. (Cité en page 59.)

Zhao, S. & Grishman, R. (2005). Extracting Relations with Integrated Information Using Kernel Methods. In Proceedings of the 43rd Annual Meeting on ACL'05 (pp. 419–426). USA. (Cité en page 53.)

Zhao, X., Yang, D., Yang, M., Wang, L., Zhang, R., Cheng, H., Lam, W., SHEN, Y., & Xu, R. (2024). A comprehensive survey on relation extraction : Recent advances and new frontiers. ACM Computing Surveys. Just Accepted. (Cité en pages 52, 54, 55, 56 et 58.)

Zheng, S., Hao, Y., Lu, D., Bao, H., Xu, J., Hao, H., & Xu, B. (2017). Joint entity and relation extraction based on a hybrid neural network. Neurocomputing, 257, 59–66. (Cité en page 57.)

Zhong, L., Wu, J., Li, Q., Peng, H., & Wu, X. (2023). A comprehensive survey on automatic knowledge graph construction. ACM Computing Surveys, 56, 1 – 62. (Cité en page 25.)

Zhou, G. & Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In Annual Meeting of the Association for Computational Linguistics. (Cité en page 28.)

Zhou, W., Smalheiser, N. R., & Yu, C. T. (2006). A tutorial on information retrieval : basic terms and concepts. Journal of Biomedical Discovery and Collaboration, 1, 2 – 2. (Cité en page 67.)

Zhou, W., Zhang, S., Gu, Y., Chen, M., & Poon, H. (2023a). Universalner : Targeted distillation from large language models for open named entity recognition. ArXiv. (Cité en page 48.)

Zhou, W., Zhang, S., Gu, Y., Chen, M., & Poon, H. (2024). Universalner : Targeted distillation from large language models for open named entity recognition. (Cité en page 40.)

Zhou, W., Zhang, S., Poon, H., & Chen, M. (2023b). Context-faithful prompting for large language models. arXiv preprint arXiv :2303.11315. (Cité en page 54.)

Zhu, J., Nie, Z., Liu, X., Zhang, B., & Wen, J.-R. (2009). Statsnowball : a statistical approach to extracting entity relationships. In The Web Conference. (Cité en page 62.)

Zouaq, A., Gagnon, M., & Jean-Louis, L. (2017). An assessment of open relation extraction systems for the semantic web. Inf. Syst., 71, 228–239. (Cité en page 60.)

Śniegula, A., Poniszewska-Marańda, A., & Popović, M. (2019). Study of Named Entity Recognition methods in biomedical field. Procedia Computer Science, 160, 260–265. (Cité en page 29.)

Şükrü Demir İnan Özer, Orman, G. K., & Labatut, V. (2024). Link prediction in bipartite networks. (Cité en page 24.)

CHAPITRE A

Annexes

Sommaire du chapitre

Données 203
Résultats classification de domaine 205

Données

Domaine	Extrait d'article
Commande publique	La violation des principes fondamentaux de la commande publique conduit à l'annulation de la procédure. Sans respect de l'égalité de traitement, de la liberté d'accès et de transparence des procédures, l'équilibre de la compétition qui permet des achats publics au meilleur rapport qualité-prix est rompu. Une telle violation est aussi une atteinte à la société car la confiance dans les élus et la bonne gestion des deniers publics est brisée. Le législateur a instauré un ensemble de règles pour préserver cette confiance et ces équilibres fragiles, dont la violation est punie de peines délictuelles voire criminelles, avec une efficacité remarquable : seuls 0,01 % des élus locaux ont été condamnés pour de tels faits entre 2008 et 2014 (c'est-à-dire 75 élus locaux sur les plus de 600 000 que compte la France). Ce taux de condamnation tombe à 0,002 % pour les fonctionnaires territoriaux sur la même période, selon les chiffres du baromètre 2015 de l'Observatoire SMACL des risques de la vie territoriale.

Domaine	Extrait d'article
Compta et finances locales	Historique. Le budget de l'État a le caractère d'une loi (loi de finances), et s'il peut faire l'objet d'un contrôle de constitutionnalité, il ne peut subir de contrôle de légalité. Le budget de la commune est un acte administratif (délibération) et est soumis à ce titre au contrôle de légalité. Jusqu'en 1982, les décisions prises par les collectivités locales étaient soumises à l'approbation de l'autorité de tutelle. La loi du 2 mars 1982, qui a substitué le contrôle de légalité à l'ancien contrôle d'opportunité, ne doit pas conduire à penser que l'État se désintéresse des finances des…
Élections	Accords de Nouméa et avenir de la Nouvelle Calédonie. Collectivité de l'océan Pacifique située à 19 000 kilomètres de Paris, la Nouvelle-Calédonie se compose de nombreuses îles, regroupées en 33 communes et trois provinces, la Province Nord (dont le siège est à Koné), la Province Sud (Nouméa) et la Province des îles Loyauté (Lifou). La population comptait, en 2012, environ 265 000 habitants, dont les deux tiers en Province Sud. Les accords de Nouméa ont été traduits par une révision constitutionnelle du 20 juillet 1998. Le statut de la Nouvelle-Calédonie figure désormais au titre XIII de la…
État civil & Cimetières	En principe, la mariage doit être célébré dans la « maison commune » de la commune où l'un des deux époux, ou l'un de leurs parents, aura son domicile ou sa résidence établie par au moins un mois d'habitation continue. Des exceptions à cela peuvent exister. Le jour du mariage est déterminé par les parties, dans certaines limites, et l'heure est définie par l'officier d'état civil. L'assistance doit être composée au minimum des futurs époux, de leurs témoins et de l'officier d'état civil. Un interprète peut être présent lors de la cérémonie si besoin.Un certain nombre d'indications pratiques, permettant une bonne préparation et un bon déroulement de la cérémonie de mariage, sont développées. Celles-ci concernent les points à voir avec les époux et la préparation par l'agent de l'état civil chargé du dossier.
Justice	L'assistance à l'audience. Les fonctionnaires du corps des greffiers assistent les magistrats à l'audience et dans tous les cas prévus par la loi. En vertu de cette disposition de l'article R. 812-11123-1 du Code de l'organisation judiciaire, un greffier est affecté à chaque chambre du tribunal correctionnel, qu'elle siège en formation collégiale ou à juge unique.Depuis le décret 2014-64 du 29 janvier 2014, cette disposition est ainsi rédigée : « Le greffe des cours d'appel et des tribunaux de grande instance comprend l'ensemble des services administratifs du siège et du Parquet… »

Domaine	Extrait d'article
Urbanisme	Les principales caractéristiques de la procédure. Si, malgré les innovations introduites par la loi SRU et les évolutions législatives ultérieures, la procédure de modification demeure essentiellement une procédure menée unilatéralement par la commune ou l'EPCI compétent, la procédure de révision engagée par ces collectivités publiques est une procédure qui met en jeu un certain nombre d'acteurs publics ou privés dans le cadre de relations diverses : association, consultation, audition à l'initiative soit de ces collectivités, soit de ces acteurs eux-mêmes.
RH Territoriales	Dans une récente décision, le Conseil d'État précise que la circonstance qu'un agent soit placé en congé de maladie ne fait pas obstacle au déroulé d'une procédure disciplinaire à son égard y compris à l'entrée en vigueur d'une décision de sanction (CE, 3 juillet 2023, n° 459472). M. A., professeur certifié hors classe de lettres dans un établissement privé, a demandé au tribunal administratif de Lyon d'annuler, d'une part, l'arrêté du 1er juin 2018 par lequel la rectrice de son académie a décidé la suspension de sa rémunération à compter du 17 février 2018, en conséquence de la décision du 13 février 2018 portant exclusion de ses fonctions pour une durée de deux ans et, d'autre part, le titre de perception du 24 mai 2018 portant recouvrement d'un trop perçu de 4 792,48 euros, ainsi que la décision du 30 mai 2018 lui transmettant ce titre et la décision du 18 juillet
Santé	Se reporter à la fiche Diplôme d'état d'accompagnant éducatif et social. L'arrêté du 30 août 2021 est relatif au diplôme d'Etat d'accompagnant éducatif et social qui atteste des compétences professionnelles pour exercer les fonctions et les activités telles que définies à l'annexe I « référentiel professionnel » du présent arrêté. Le diplôme d'état d'accompagnant éducatif (AES) remplace les diplômes d'auxiliaire de vie sociale (AVS) et d'aide médico-psychologique (AMP).

TABLE A.1: Tableau des 8 domaines avec un exemple réel de texte. Exemples tirés des données disponibles sur legibase.fr

Résultats classification de domaine

ZeroShot Learning

Nous avons souhaité nous orienter vers une classification *Zero-Shot Learning*. L'intérêt du *Zero-Shot Learning* est de pouvoir reconnaître les domaines des textes sans entraînement ni exemple. Nous nous sommes orientés vers ce type de classification car nos données sont

en constante évolution. Cette année, par exemple, des données ont été ajoutées ainsi qu'un domaine. Il est donc difficile de réapprendre entièrement un modèle qui n'a jamais vu les données du nouveau domaine. Nous souhaitons donc privilégier un apprentissage non supervisé qui ne nécessite pas d'avoir « vu » les données au préalable. Nous avons utilisé deux modèles tirés de *Hugging face*. Les tableaux A.2 et A.3 présentent respectivement les résultats du premier modèle étant un modèle CamemBERT-XNLI et du deuxième étant un modèle DistilCamemBERT-NLI. CamemBERT-XNLI est un modèle CamemBERT construit sur le jeu de données XNLI qui consiste à déterminer si une prémisse entraîne, contredit ou n'entraîne ni ne contredit une hypothèse. DistilCamemBERT-NLI, quant à lui, est un modèle DistilCamemBERT(Delestre & Amar, 2022) affiné pour la tâche d'inférence en langage naturel (NLI) pour la langue française, également connue sous le nom de reconnaissance d'implication textuelle (RTE). Lui aussi est construit sur le jeu de données XNLI.

Nous constatons que les résultats ne sont pas concluants. Nous constatons que certains domaines sont trop précis et donc ne sont pas connus des modèles comme « état civil et cimetières » par exemple. Lorsque nous divisons les domaines composés de plusieurs mots entre chacun des mots, nous n'obtenons pas de bons résultats. Nous avons donc mis de côté cette technique.

Domaine	**Précision**	**Rappel**	**F1-Score**
Commande publique	0,7	0,64	0,67
Comptabilité et finances locales	0,29	0,18	0,22
Justice	0,29	0,7	0,41
Ressources humaines territoriales	0,56	0,45	0,5
Santé	0,5	0,09	0,15
Urbanisme	0,95	0,9	0,93
Élections	0,7	0,58	0,63
État civil & cimetières	0,06	0,06	0,06

TABLE A.2 – Résultats de Classification Zero-Shot Learning en utilisant le modèle « camembert-base-xnli » de Hugging Face

Domaine	**Précision**	**Rappel**	**F1-Score**
Commande publique	0,27	0,36	0,31
Comptabilité et finances locales	0,25	0,18	0,21
Justice	0,25	0,25	0,25
Ressources humaines territoriales	0,17	0,18	0,17
Santé	0,2	0,2	0,2
Urbanisme	0	0	0
Élections	0,33	0,02	0,04
État civil & cimetières	0,36	0,06	0,1

TABLE A.3 – Résultats de Classification Zero-Shot Learning en utilisant le modèle « distilcamembert-base-nli » de Hugging Face

Méthodes classiques : Représentation par sac de mots (*Bag Of words*)

Les résultats sont présentés sous formes de tableaux mais aussi le plus souvent sous forme de matrice des résultats pour déterminer la classe majoritaire attribuée par les prédic-

tions. Cette approche permet de comprendre la répartition des prédictions du modèle entre les différentes classes. En examinant chaque ligne de la matrice, qui représente les prédictions pour une classe réelle spécifique, on identifie la colonne ayant la valeur la plus élevée, indiquant ainsi la classe prédite majoritairement pour cette classe réelle. Cette analyse révèle les tendances et biais potentiels du classifieur, comme une éventuelle tendance à confondre certaines classes entre elles, ou à privilégier une classe particulière. En somme, l'observation des classes majoritairement attribuées aide à diagnostiquer et améliorer la précision et la robustesse du modèle de classification.

TF-IDF

Nous allons comparer les résultats de trois classifieurs différents sur la représentation TF-IDF fournis dans des tableaux différents : le tableau A.4 pour le classifieur multinomialNB, le tableau A.5 pour le classifieur LinearSVC et le tableau A.6 pour le classifieur SGDClassifier. Des matrices de corrélations sont associées à chacun des résultats : la figure A.1 pour le classifieur MultinomialNB, A.2 pour le classifieur LinearSVC et A.3 pour SGDClassifier. Nous pouvons analyser les valeurs de précision, de rappel et de F1-Score pour chaque domaine, ainsi que les moyennes macro et pondérées (Macro AVG et Weighted AVG). Voici un résumé des performances de chaque classifieur :

- MultinomialNB
  - Précision globale : 0,88
  - Rappel global : 0,88
  - F1-Score global : 0,88
- LinearSVC
  - Précision globale : 0,93
  - Rappel global : 0,93
  - F1-Score global : 0,93
- SGDClassifier
  - Précision globale : 0,92
  - Rappel global : 0,92
  - F1-Score global : 0,92

Domaine	**Précision**	**Rappel**	**F1-Score**
Commande publique	0,9	0,96	0,93
Comptabilité et finances locales	0,95	0,75	0,83
Justice	0,88	0,86	0,87
Ressources humaines territoriales	0,79	0,92	0,85
Santé	0,91	0,83	0,87
Urbanisme	0,86	0,95	0,9
Élections	0,87	1	0,93
État civil & cimetières	0,93	0,81	0,86
Macro AVG	0,89	0,88	0,88
Weighted AVG	0,88	0,88	0,88

TABLE A.4 – Résultats de Classification de domaine TF-IDF – CLASSIFICATION MULTINOMIALNB – (ACC = 0,88)

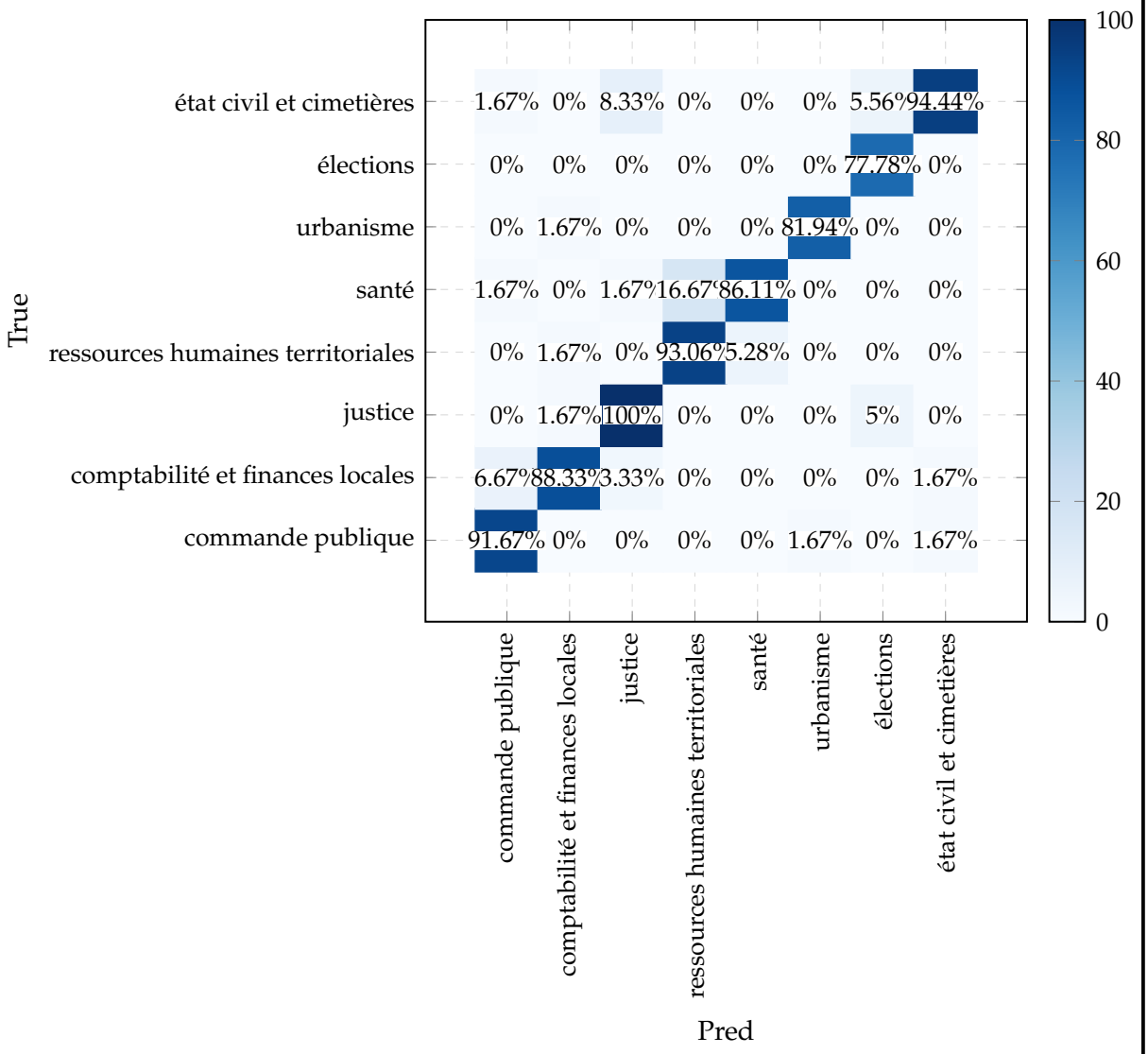

FIGURE A.1 – Matrice de confusion - TF-IDF – CLASSIFICATION MULTINOMIALNB – (ACC = 0,88)

En examinant les détails par domaine :

- Commande publique
  - MultinomialNB : F1-Score = 0,93
  - LinearSVC : F1-Score = 0,96
  - SGDClassifier : F1-Score = 0,97
- Comptabilité et finances locales
  - MultinomialNB : F1-Score = 0,83
  - LinearSVC : F1-Score = 0,91
  - SGDClassifier : F1-Score = 0,91
- Justice
  - MultinomialNB : F1-Score = 0,87
  - LinearSVC : F1-Score = 0,94

  - SGDClassifier : F1-Score = 0,92
- Ressources humaines territoriales
  - MultinomialNB : F1-Score = 0,85
  - LinearSVC : F1-Score = 0,91
  - SGDClassifier : F1-Score = 0,89
- Santé
  - MultinomialNB : F1-Score = 0,87
  - LinearSVC : F1-Score = 0,90
  - SGDClassifier : F1-Score = 0,88
- Urbanisme
  - MultinomialNB : F1-Score = 0,90
  - LinearSVC : F1-Score = 0,94
  - SGDClassifier : F1-Score = 0,93
- Élections
  - MultinomialNB : F1-Score = 0,93
  - LinearSVC : F1-Score = 0,98
  - SGDClassifier : F1-Score = 0,98
- État civil & cimetières
  - MultinomialNB : F1-Score = 0,86
  - LinearSVC : F1-Score = 0,93
  - SGDClassifier : F1-Score = 0,92

Domaine	**Précision**	**Rappel**	**F1-Score**
Commande publique	0,93	0,98	0,96
Comptabilité et finances locales	0,93	0,9	0,91
Justice	0,93	0,94	0,94
Ressources humaines territoriales	0,92	0,9	0,91
Santé	0,91	0,89	0,9
Urbanisme	0,97	0,92	0,94
Élections	0,98	0,98	0,98
État civil & cimetières	0,91	0,93	0,93
Macro AVG	0,93	0,93	0,93
Weighted AVG	0,93	0,93	0,93

TABLE A.5 – Résultats de Classification de domaine TF-IDF – CLASSIFICATION LINEARSVC – RESULTATS (ACC = 0,93)

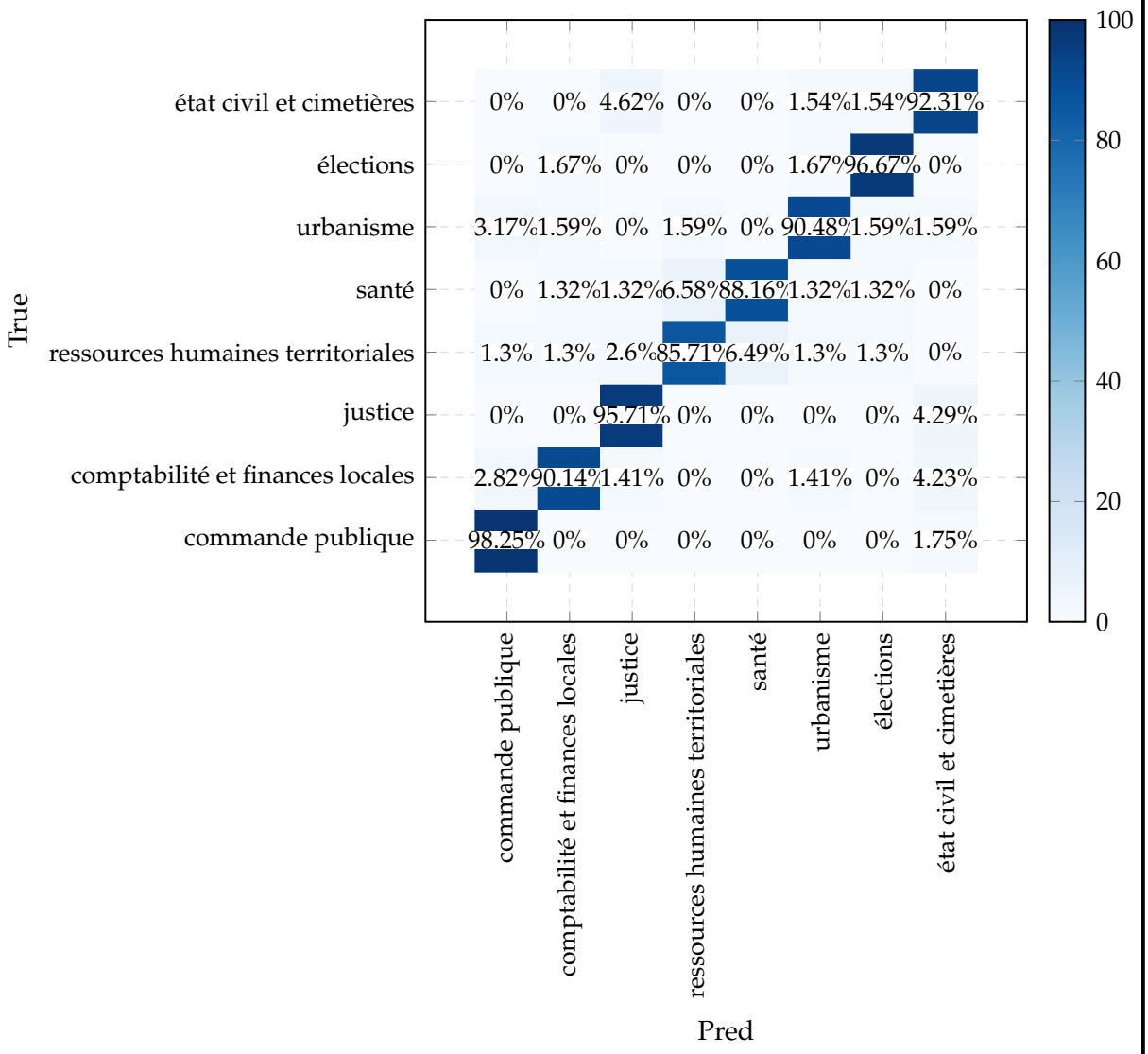

FIGURE A.2 – Matrice de confusion - TF-IDF – CLASSIFICATION LinearSVC – (ACC = 0,88)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,95	0,98	0,97
Comptabilité et finances locales	0,91	0,9	0,91
Justice	0,94	0,9	0,92
Ressources humaines territoriales	0,9	0,88	0,89
Santé	0,88	0,88	0,88
Urbanisme	0,93	0,92	0,93
Élections	0,98	0,98	0,98
État civil & cimetières	0,88	0,95	0,92
Macro AVG	0,92	0,92	0,92
Weighted AVG	0,92	0,92	0,92

TABLE A.6 – Résultats de Classification de domaine TF-IDF – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,92)

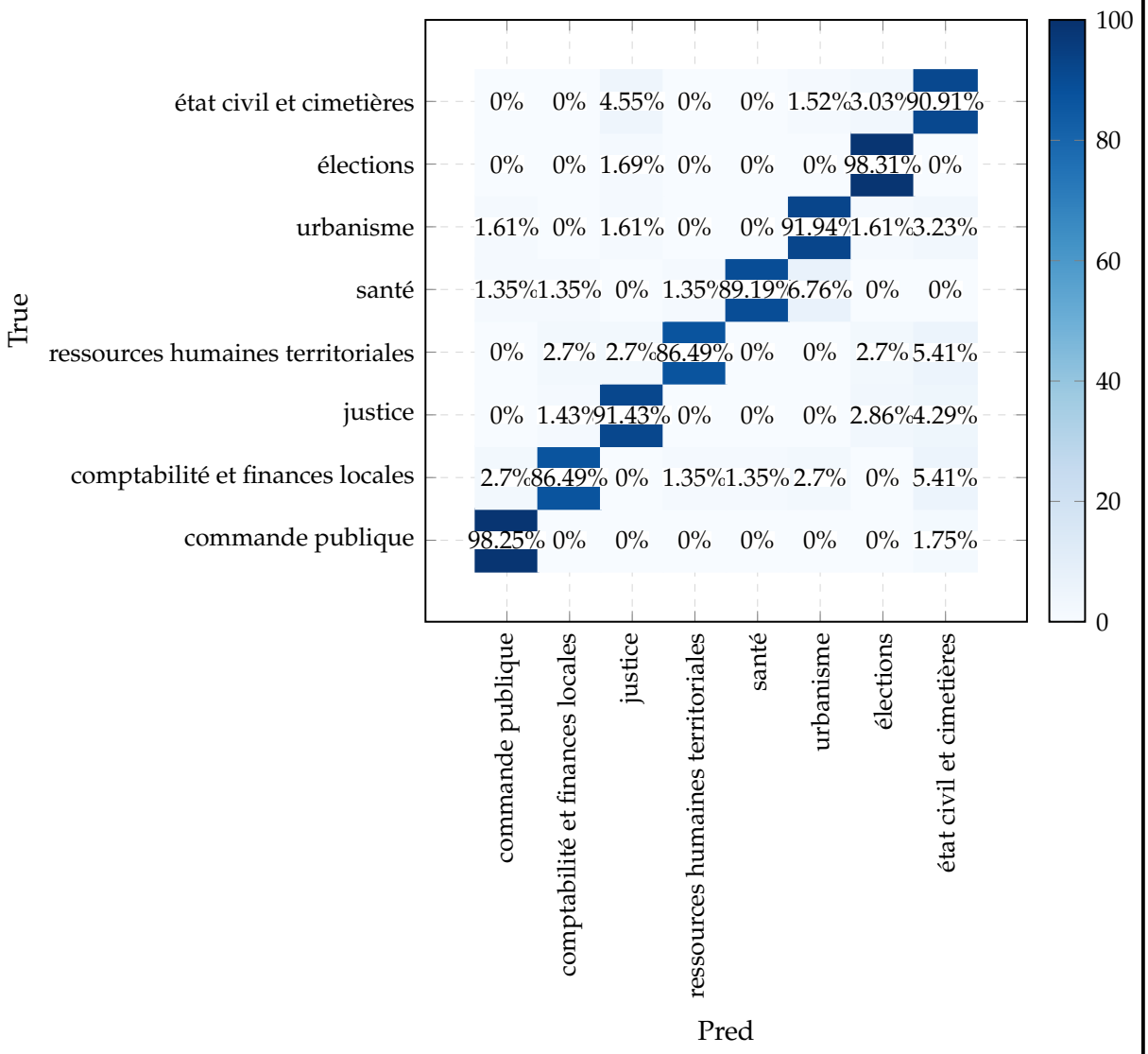

FIGURE A.3 – Matrice de confusion - TF-IDF – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,92)

Le classifieur LinearSVC présente les meilleures performances globales avec un F1-Score moyen (Macro AVG et Weighted AVG) de 0,93, surpassant les autres classifieurs dans presque tous les domaines. Le SGDClassifier suit de près avec des performances très similaires, particulièrement élevé en précision et rappel pour certains domaines spécifiques comme "Élections". Le MultinomialNB montre des performances moindres en comparaison, surtout dans les domaines "Comptabilité et finances locales" et "Ressources humaines territoriales".

Si l'on devait choisir un classifieur basé sur ces résultats, LinearSVC serait le meilleur choix global pour cette tâche de classification de documents par domaine.

Les deux tableaux A.7 et A.8 montrent les résultats après avoir conservé les apostrophes dans les textes, ce qui peut affecter les performances des classifieurs. Nous nous sommes concentrés sur LinearSVC et SGDClassifier uniquement, ceux ayant obtenus les meilleurs résultats pour le TF-IDF. Des matrices de corrélations sont aussi associées à chacun des

résultats : la figure A.4 pour le classifieur LinearSVC et A.5 pour SGDClassifier.

— LinearSVC (avec apostrophes)
  — Précision globale : 0,95. La précision de LinearSVC avec les apostrophes est légèrement supérieure (0,95) par rapport à sans les apostrophes (0,93).
  — Rappel global : 0,94. Le rappel est très similaire, légèrement inférieur avec les apostrophes (0,94 contre 0,93).
  — F1-Score global : 0,94. Globalement stable, avec une légère amélioration en gardant les apostrophes (0,94).
— SGDClassifier (avec apostrophes)
  — Précision globale : 0,94. La précision avec les apostrophes est meilleure (0,94) qu'avant (0,92).
  — Rappel global : 0,94. Identique (0,94).
  — F1-Score global : 0,94. Amélioré à 0,94 comparé à 0,92 sans apostrophes.

Domaine	**Précision**	**Rappel**	**F1-Score**
Commande publique	0.94	0.95	0.94
Comptabilité et finances locales	0.97	0.86	0.91
Justice	0.95	0.90	0.92
Ressources humaines territoriales	1.00	1.00	1.00
Santé	0.99	0.95	0.97
Urbanisme	0.92	0.92	0.92
Élections	0.98	0.96	0.97
État civil et cimetières	0.90	0.98	0.94
macro avg	0.94	0.95	0.94
weighted avg	0.95	0.94	0.94

TABLE A.7 – Résultats de Classification de domaine PRETRAITEMENT EN GARDANT LES APOSTROPHES - TF-IDF – CLASSIFICATION LINEARSVC (ACC = 0.94)

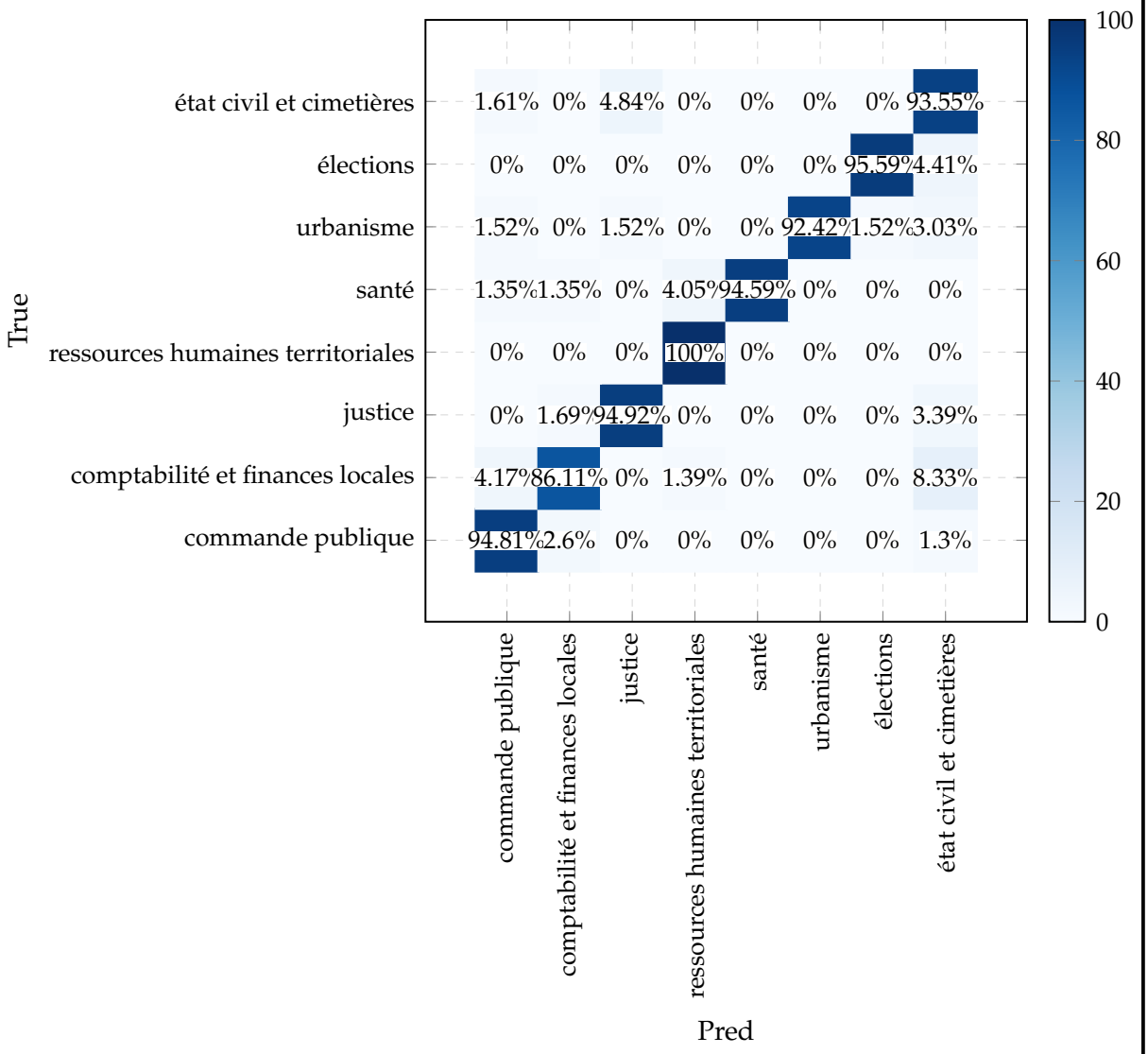

FIGURE A.4 – Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - TF-IDF – CLASSIFICATION LINEARSVC (ACC = 0.94)

Domaine	Précision	Rappel	F1-Score
commande publique	0.94	0.95	0.94
comptabilité et finances locales	0.95	0.83	0.89
justice	0.94	0.94	0.94
ressources humaines territoriales	1.00	0.95	0.97
santé	0.97	0.93	0.95
urbanisme	0.92	0.92	0.92
élections	0.98	1.00	0.99
état civil et cimetières	0.91	0.98	0.95
macro avg	0.94	0.94	0.94
weighted avg	0.94	0.94	0.94

TABLE A.8 – Résultats de Classification de domaine PRETRAITEMENT EN GARDANT LES APOSTROPHES - TF-IDF – CLASSIFICATION SGD (ACC = 0.94)

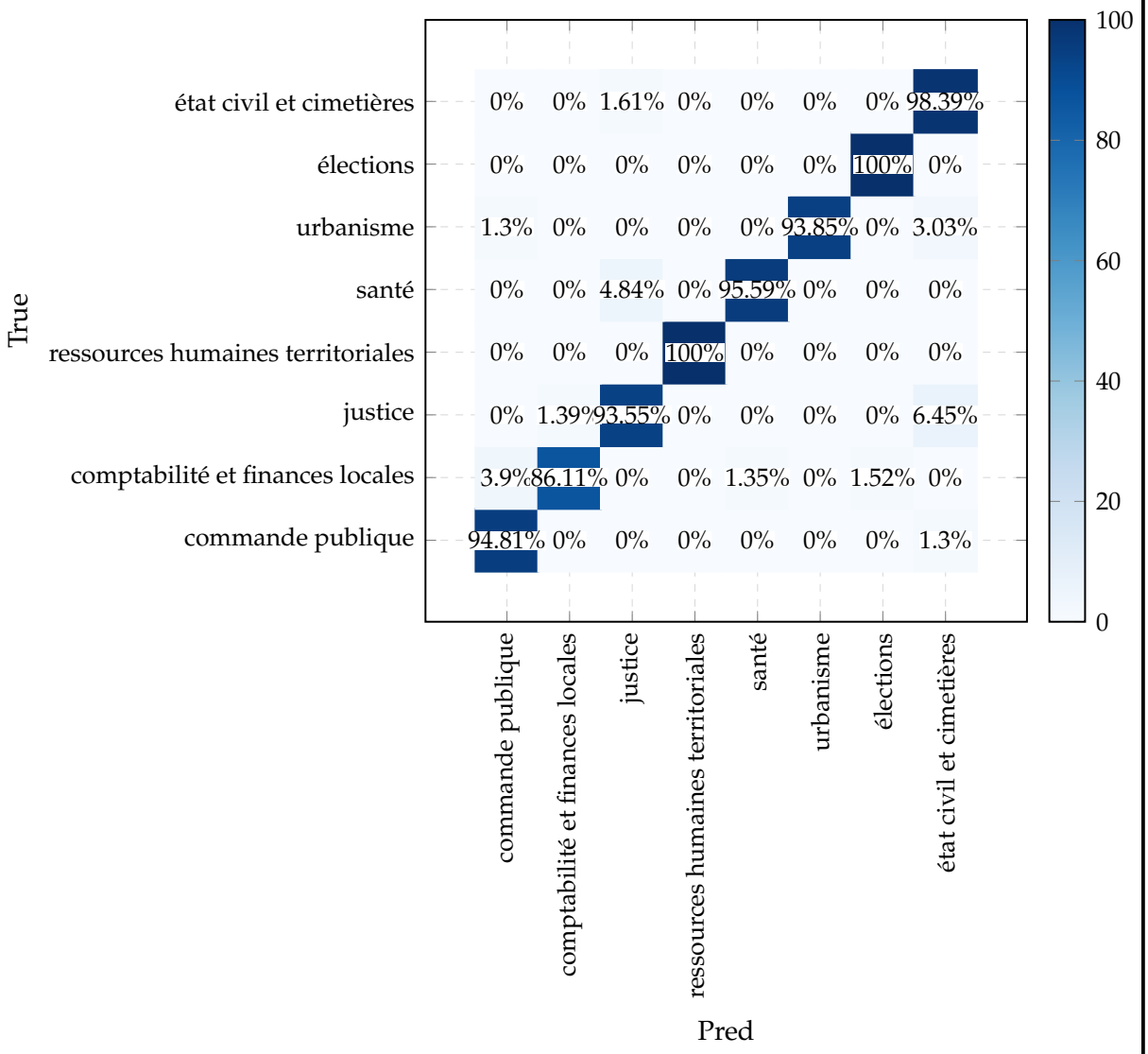

FIGURE A.5 – Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - TF-IDF – CLASSIFICATION SGD (ACC = 0.94)

La conservation des apostrophes dans le texte améliore légèrement les performances de LinearSVC et SGDClassifier en termes de précision et de F1-Score. Pour LinearSVC, l'amélioration est surtout visible dans certains domaines comme "Ressources humaines territoriales" et "Santé". Pour SGDClassifier, les gains sont plus homogènes à travers les différents domaines.

Si l'on devait choisir un modèle de classification après avoir gardé les apostrophes, LinearSVC reste le meilleur choix global avec une légère amélioration par rapport à la version sans apostrophes. Cependant, SGDClassifier montre aussi une amélioration notable et peut être préféré dans certains contextes spécifiques.

ClassicBoW

Les résultats des classifications de documents représentés par la méthode Classic Bag-of-Words (BoW) sont présentés dans les Tables A.9, A.10 et A.11. Chaque table montre les performances obtenues par les classificateurs MultinomialNB, LinearSVC, et SGDClassifier respectivement, tous ayant une précision globale de 0,94 à l'exception de SGDClassifier avec 0,92. L'analyse des résultats par domaine révèle que MultinomialNB excelle particulièrement en "Commande publique" (0,98 de F1-Score) et en "Urbanisme" (0,97 de F1-Score), tandis que LinearSVC surpasse les autres classificateurs en "Comptabilité et finances locales" avec un F1-Score de 0,94. SGDClassifier montre sa force en "Ressources humaines territoriales" avec un F1-Score de 0,92. De manière générale, MultinomialNB offre des performances élevées dans plusieurs domaines clés et est recommandé comme le meilleur choix global pour la méthode Classic BoW, bien que LinearSVC constitue une alternative robuste pour certains domaines spécifiques. Les matrices associées aux résultats des tableaux sont présentes sur la figure A.6 pour le classifieur MultinomialNB, A.7 pour le classifieur LinearSVC et A.8 pour SGDClassifier.

Domaine	Précision	Rappel	F1-Score
Commande publique	0,98	0,97	0,98
Comptabilité et finances locales	0,9	0,89	0,9
Justice	0,94	0,88	0,91
Ressources humaines territoriales	0,85	0,96	0,9
Santé	0,97	0,87	0,92
Urbanisme	0,95	0,99	0,97
Élections	0,95	1	0,97
État civil & cimetières	0,95	0,95	0,95
Macro AVG	0,94	0,94	0,94
Weighted AVG	0,94	0,94	0,94

TABLE A.9 – Résultats de Classification de domaine CLASSICBOW– CLASSIFICATION MULTINOMIALNB (ACC = 0,94)

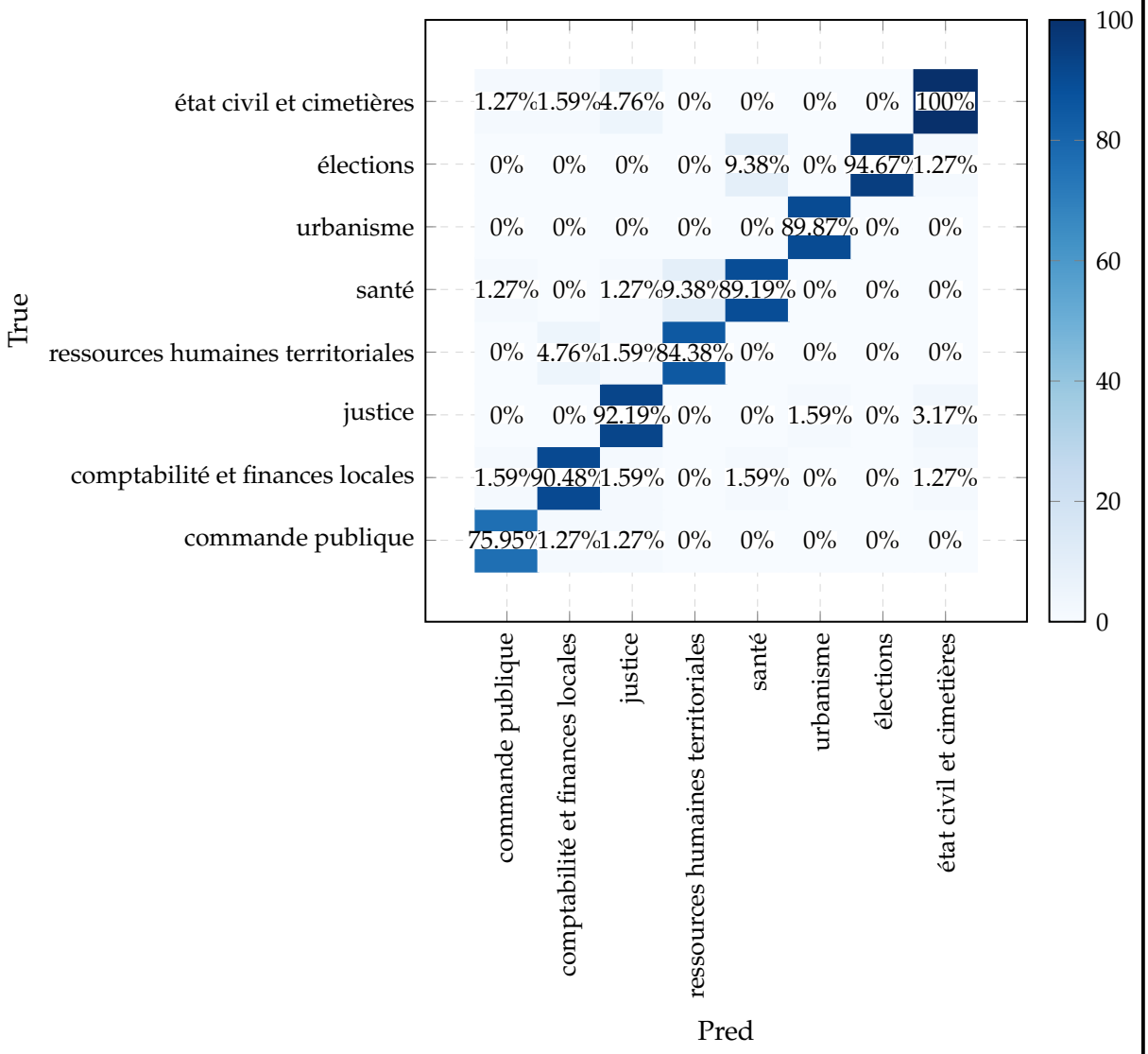

FIGURE A.6 – Matrice de confusion - ClassicBOW – CLASSIFICATION MultinomialNB (ACC = 0.94)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,97	0,95	0,96
Comptabilité et finances locales	0,95	0,94	0,94
Justice	0,88	0,91	0,9
Ressources humaines territoriales	0,89	0,93	0,91
Santé	0,95	0,92	0,93
Urbanisme	0,99	0,92	0,95
Élections	0,97	1	0,98
État civil & cimetières	0,9	0,92	0,91
Macro AVG	0,94	0,94	0,94
Weighted AVG	0,94	0,93	0,93

TABLE A.10 – Résultats de Classification de domaine CLASSICBOW– CLASSIFICATION LINEARSVC – (ACC = 0,94)

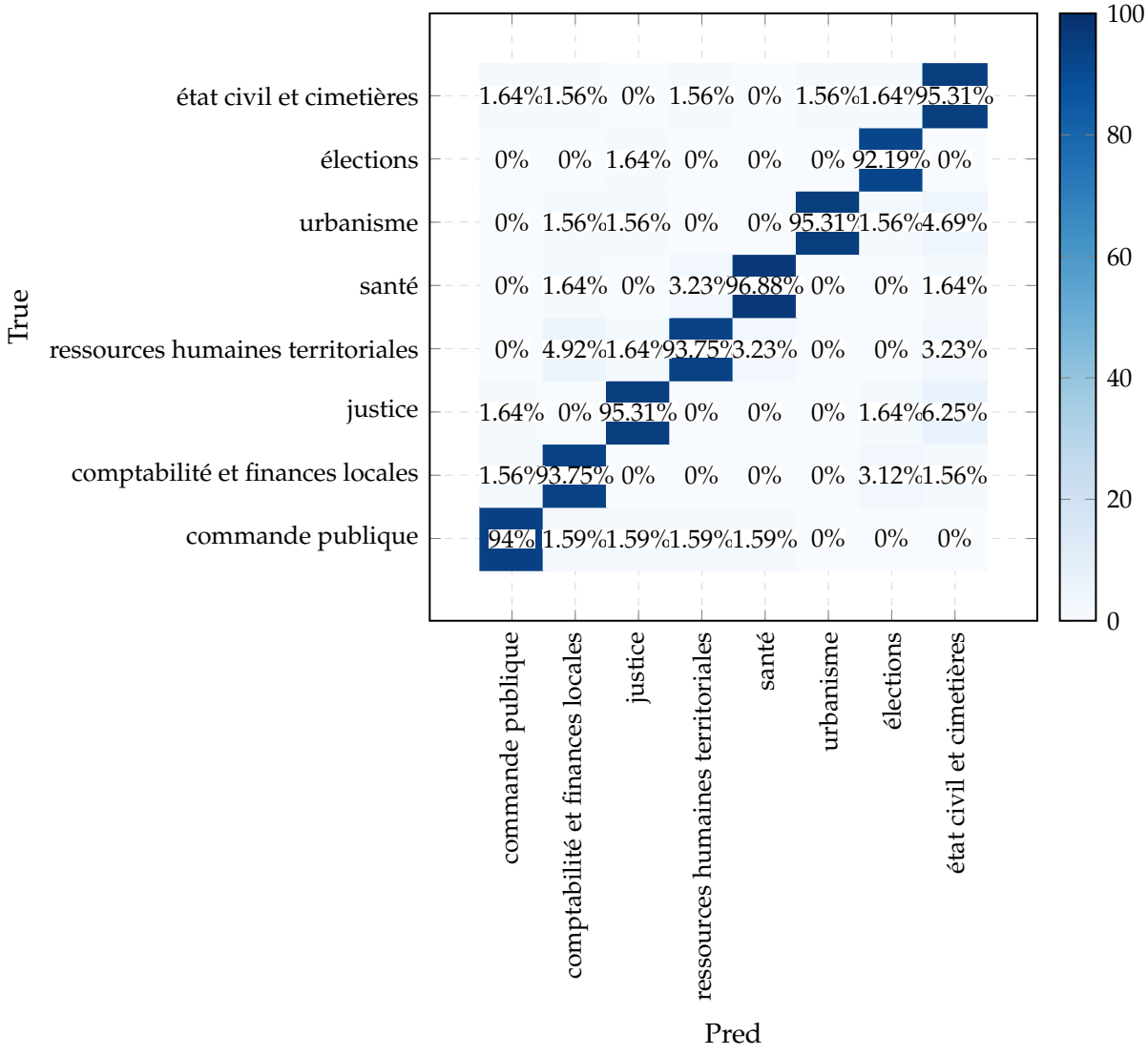

FIGURE A.7 – Matrice de confusion - ClassicBOW – CLASSIFICATION LINEARSVC (ACC = 0.94)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,97	0,95	0,96
Comptabilité et finances locales	0,87	0,86	0,87
Justice	0,92	0,84	0,88
Ressources humaines territoriales	0,9	0,94	0,92
Santé	0,92	0,92	0,92
Urbanisme	0,91	0,94	0,93
Élections	0,97	1	0,98
État civil & cimetières	0,94	0,94	0,94
Macro AVG	0,92	0,92	0,92
Weighted AVG	0,92	0,92	0,92

TABLE A.11 – Résultats de Classification de domaine CLASSICBOW– CLASSIFICATION SGDCLASSIFIER – (ACC = 0,94)

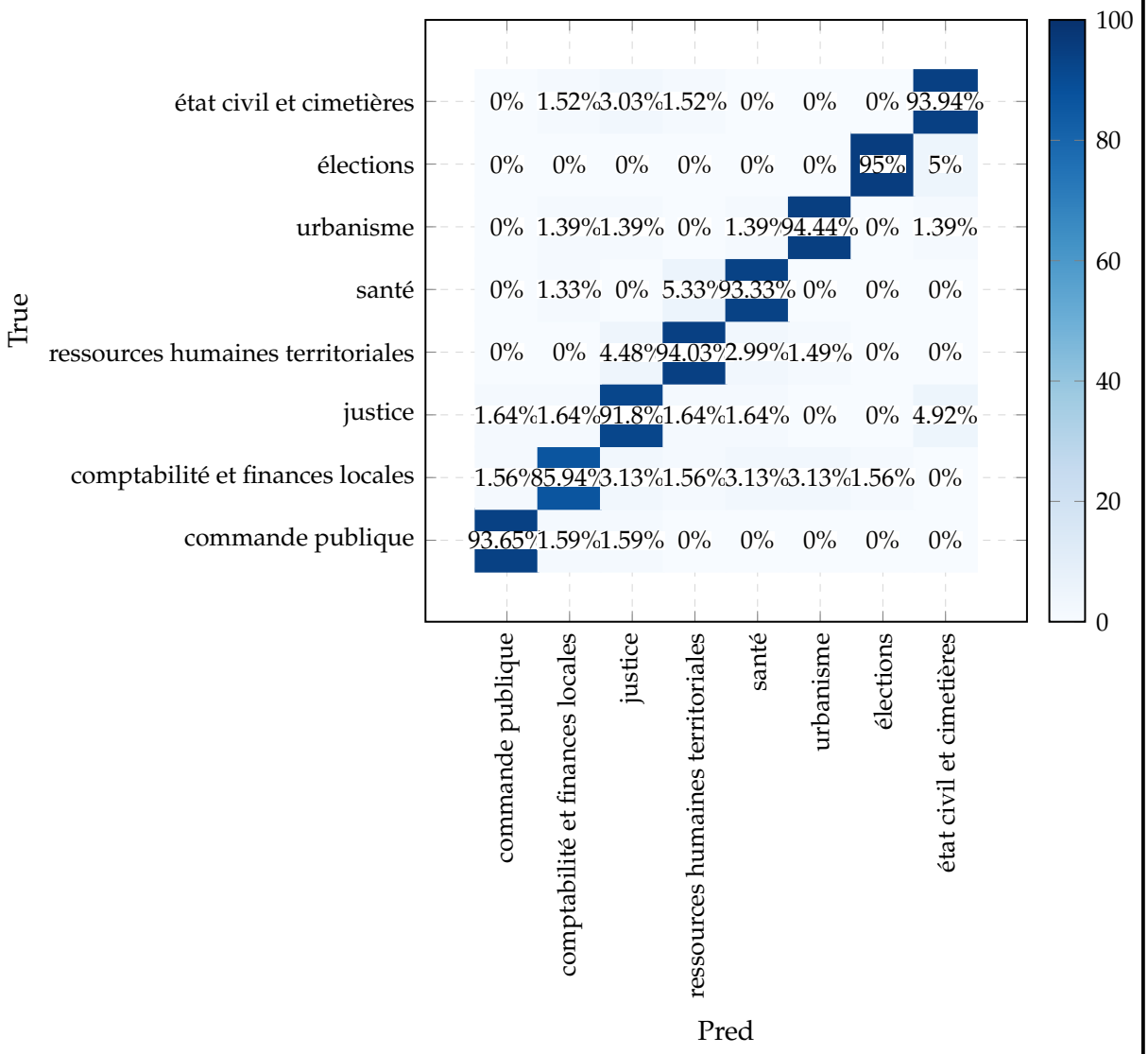

FIGURE A.8 – Matrice de confusion - ClassicBOW – CLASSIFICATION SGDClassifier (ACC = 0.94)

Domaine	Précision	Rappel	F1-Score
Commande publique	1.00	0.99	0.99
Comptabilité et finances locales	0.94	0.97	0.95
Justice	0.91	0.86	0.88
Ressources humaines territoriale	0.94	0.99	0.96
Santé	0.96	0.92	0.94
Urbanisme	0.97	0.95	0.96
Élections	0.99	1.00	0.99
État civil & cimetières	0.91	0.91	0.91
Macro AVG	0.95	0.95	0.95
Weighted AVG	0.95	0.95	0.95

TABLE A.12 – Résultats de Classification de domaine CLASSICBOW - Représentation en gardant les apostrophes – CLASSIFICATION LinearSVC – (ACC = 0,95)

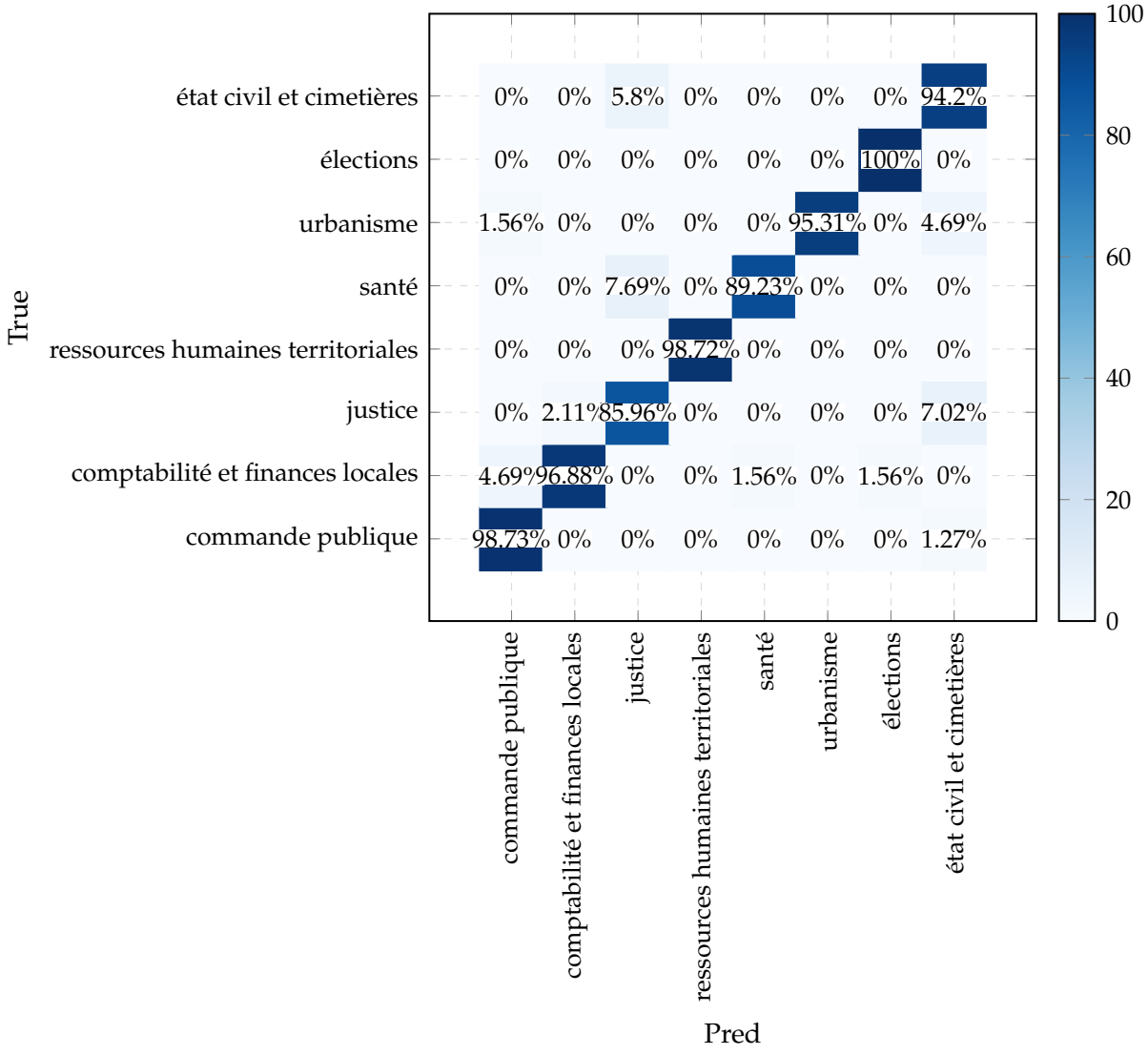

FIGURE A.9 – Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - ClassicBOW – CLASSIFICATION LINEARSVC (ACC = 0.95)

Domaine	Précision	Rappel	F1-Score
Commande publique	0.99	0.91	0.95
Comptabilité et finances locales	0.96	0.84	0.90
Justice	0.83	0.88	0.85
Ressources humaines territoriale	1.00	0.90	0.95
Santé	0.81	0.80	0.81
Urbanisme	0.98	0.95	0.96
Élections	0.95	1.00	0.97
État civil & cimetières	0.88	0.80	0.84
Macro AVG	0.90	0.89	0.90
Weighted AVG	0.91	0.90	0.90

TABLE A.13 – Résultats de Classification de domaine CLASSICBOW - Représentation en gardant les apostrophes – CLASSIFICATION MultinomialNB – (ACC = 0,90)

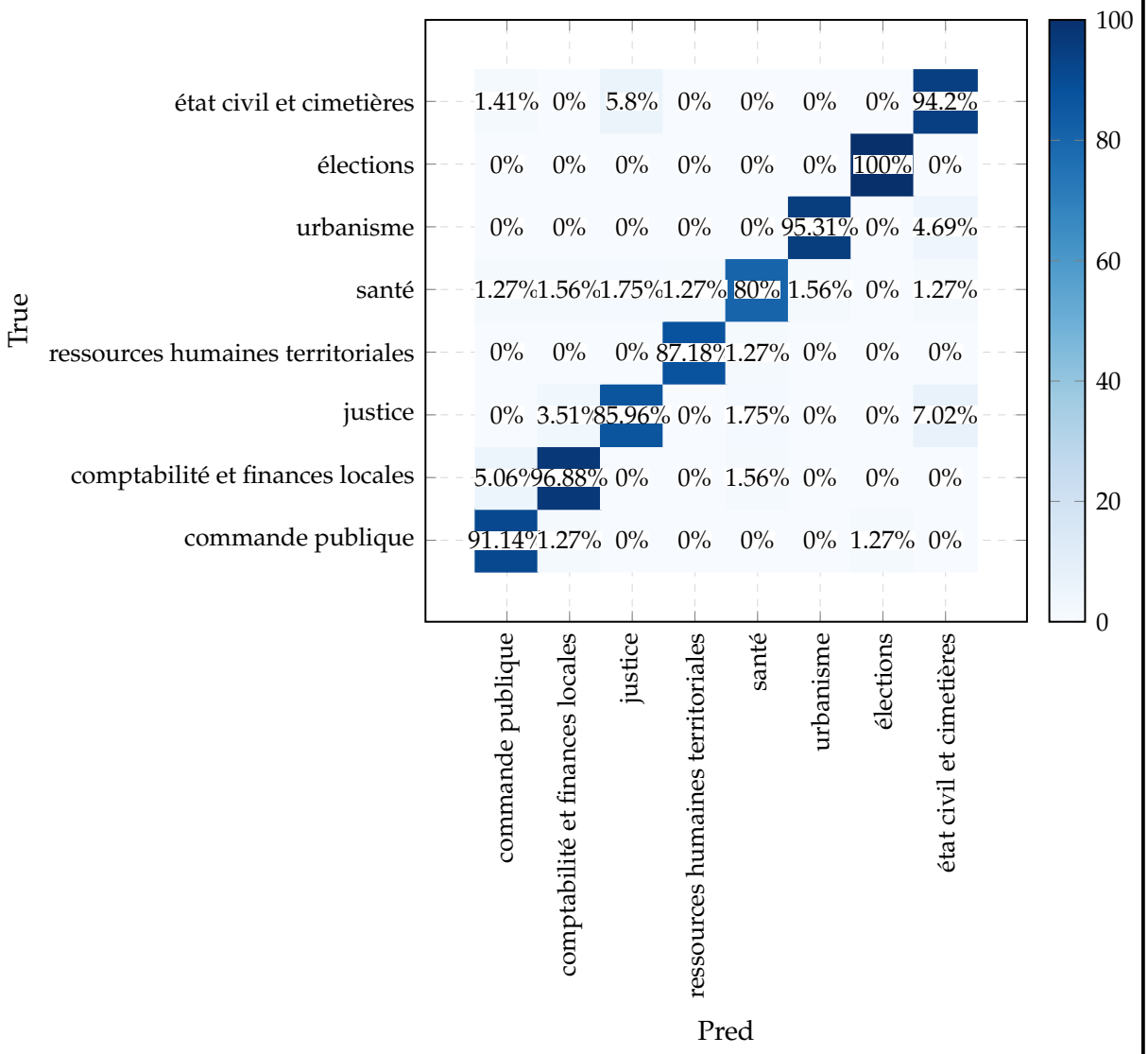

FIGURE A.10 – Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - ClassicBOW – CLASSIFICATION MultinomialNB (ACC = 0.90)

Domaine	Précision	Rappel	F1-Score
Commande publique	1.00	0.99	0.99
Comptabilité et finances locales	0.91	0.98	0.95
Justice	0.89	0.88	0.88
Ressources humaines territoriale	0.96	0.96	0.96
Santé	0.96	0.91	0.93
Urbanisme	0.97	0.95	0.96
Élections	0.99	1.00	0.99
État civil & cimetières	0.91	0.91	0.91
Macro AVG	0.95	0.95	0.95
Weighted AVG	0.95	0.95	0.95

TABLE A.14 – Résultats de Classification de domaine CLASSICBOW - Représentation en gardant les apostrophes – CLASSIFICATION SGDClassifier – (ACC = 0,95)

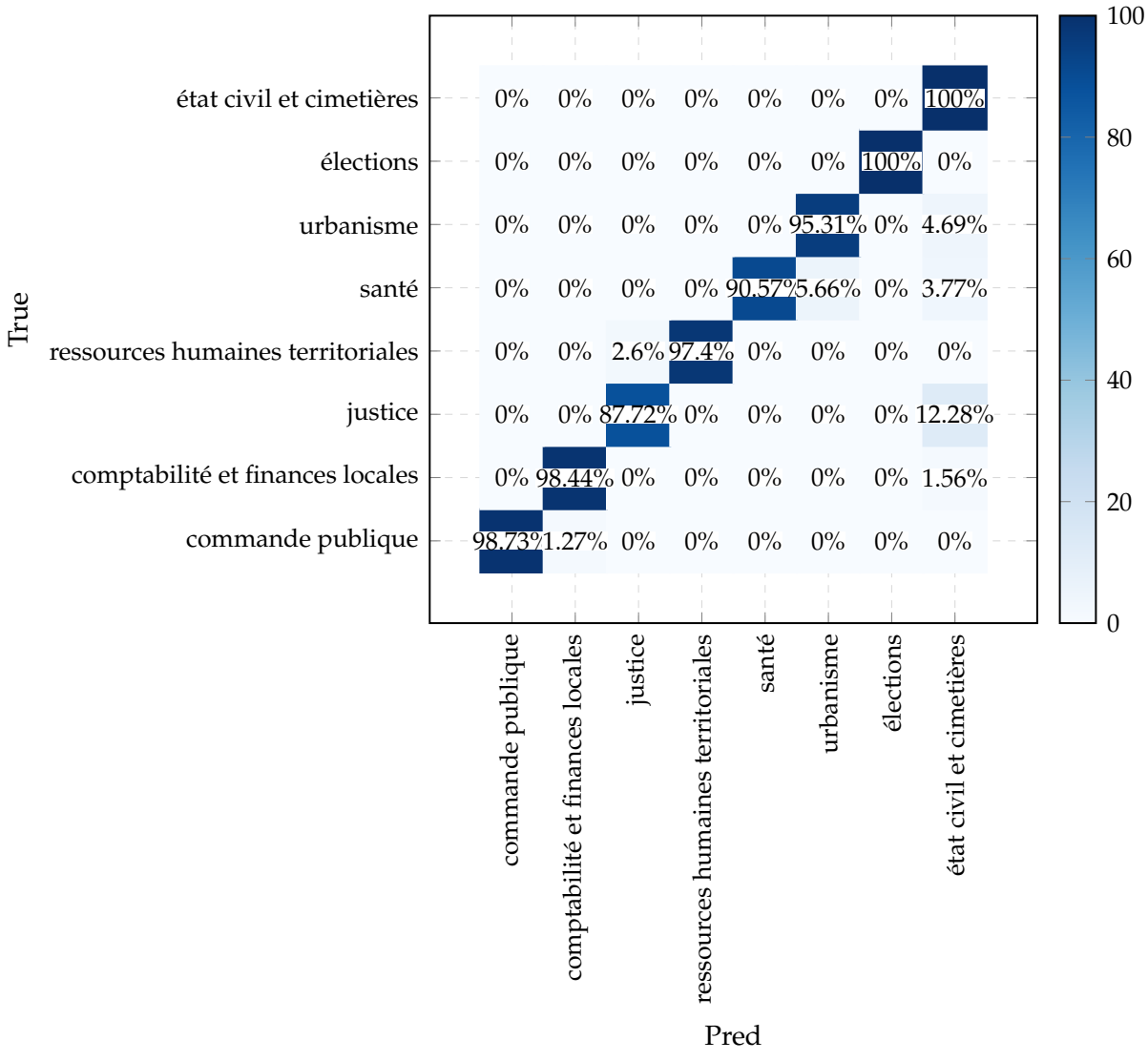

FIGURE A.11 – Matrice de confusion - PRETRAITEMENT EN GARDANT LES APOSTROPHES - ClassicBOW – CLASSIFICATION SGDClassifier (ACC = 0.95)

Les tableaux A.12, A.13 et A.14 donnent les résultats en gardant les apostrophes pour la représentation de documents. Des matrices de corrélations sont aussi associées à chacun des résultats : la figure A.9 pour le classifieur LinearSVC, A.10 pour le classifieur MultinomialNB et A.11 pour SGDClassifier. Voici une analyse et comparaison entre les différents classifieurs :

1. MultinomialNB :
   — **Sans apostrophes :** Les performances sont assez bonnes avec une précision, un rappel et un F1-score tous à 0,94.
   — **Avec apostrophes :** Les performances baissent légèrement, avec des scores de précision et de rappel à 0,90 et 0,89 respectivement.
2. LinearSVC :
   — **Sans apostrophes :** Les performances sont correctes avec un F1-score de 0,93.
   — **Avec apostrophes :** Les performances s'améliorent nettement, avec un F1-score atteignant 0,95.
3. SGDClassifier :
   — **Sans apostrophes :** Les performances sont égales à celles des autres modèles sans apostrophes, avec un F1-score de 0,92.
   — **Avec apostrophes :** Les performances augmentent sensiblement avec un F1-score de 0,95.

L'analyse montre que : (1) MultinomialNB est moins performant lorsqu'on garde les apostrophes. Cela peut être dû au fait que ce modèle se base sur la probabilité des mots individuels, et les apostrophes peuvent introduire de nouvelles formes de mots qui sont moins fréquentes et donc moins bien traitées ; et (2) LinearSVC et SGDClassifier montrent des améliorations significatives lorsqu'on garde les apostrophes. Garder les apostrophes dans la représentation textuelle semble avoir un intérêt pour les modèles LinearSVC et SGDClassifier, en améliorant leurs performances globales.

En examinant les résultats, le modèle LinearSVC avec les apostrophes et SGDClassifier avec les apostrophes sont les plus performants, chacun ayant des scores de Macro AVG et Weighted AVG de 0.95 pour la précision, le rappel et le F1-Score. Ces deux modèles obtiennent les mêmes scores élevés, donc on peut conclure que LinearSVC avec les apostrophes et SGDClassifier avec les apostrophes sont les plus performants pour la tâche de classification des domaines basée sur les représentations textuelles. Les deux modèles LinearSVC et SGDClassifier avec les apostrophes offrent des performances optimales, mais si un seul doit être choisi, on peut préférer le LinearSVC en raison de sa robustesse et de son efficacité généralement observée dans les tâches de classification textuelle.

Plongements lexicaux (*Word Embeddings*), *Sentence Embeddings* et de documents (*Doc Embeddings*)

Sentence BERT

Les résultats obtenus à partir des différentes tables montrent des variations notables dans les performances des modèles de classification en fonction de la méthode et du traitement des données utilisés.

Le **tableau A.15** présente les résultats obtenus avec le modèle *Sentence CamemBERT Base* et la classification *LinearSVC* sur des données prétraitées. On observe une précision globale de 0,86, avec des scores F1 relativement élevés pour les domaines tels que les *Commandes publiques* (0,91) et les *Élections* (0,96). Cependant, certains domaines comme les *Ressources humaines territoriales* (0,72) montrent des scores inférieurs.

Le **tableau A.16** utilise le modèle *Flaubert-Base-Uncased-XNLI-STS* avec *LinearSVC*, atteignant une précision globale légèrement supérieure de 0,87. Les domaines comme les *Comptabilité et finances locales* (0,87) et la *Justice* (0,87) présentent une amélioration par rapport à la table précédente, bien que la *Santé* ait un score F1 de 0,80, légèrement inférieur à d'autres domaines.

Le **tableau A.17** montre une augmentation notable des performances avec le même modèle *Sentence CamemBERT Base* mais sans prétraitement des données. La précision globale atteint 0,92, et des domaines comme les *État civil & cimetières* (0,97) et la *Justice* (0,90) affichent des scores F1 élevés, indiquant que le modèle bénéficie de données non prétraitées pour une meilleure classification.

Enfin, le **tableau A.18** présente les résultats de la classification *SGDClassifier* sur des textes sans prétraitement. Avec une précision globale de 0,91, les performances sont comparables à celles de la *LinearSVC* sans prétraitement. Les scores F1 pour des domaines comme les *État civil & cimetières* (0,95) et les *Élections* (0,94) restent élevés, consolidant l'idée que l'absence de prétraitement peut améliorer les performances globales du modèle.

En résumé, les résultats indiquent que l'utilisation de *Sentence CamemBERT Base* sans prétraitement des données, couplé avec *LinearSVC*, offre les meilleures performances globales, atteignant une précision et un rappel plus élevés que les autres configurations testées. L'importance de choisir le bon modèle et le bon prétraitement est cruciale pour optimiser la performance de la classification de domaine.

Domaine	**Précision**	**Rappel**	**F1-Score**
Commande publique	0,93	0,89	0,91
Comptabilité et finances locales	0,93	0,78	0,85
Justice	0,78	0,88	0,83
Ressources humaines territoriales	0,7	0,74	0,72
Santé	0,79	0,79	0,79
Urbanisme	0,92	0,93	0,93
Élections	0,97	0,95	0,96
État civil & cimetières	0,88	0,94	0,91
Macro AVG	0,86	0,86	0,86
Weighted AVG	0,87	0,86	0,86

TABLE A.15 – Résultats de Classification de domaine SENTENCE CAMEMBERT BASE – CLASSIFICATION LINEARSVC – (ACC = 0,86)

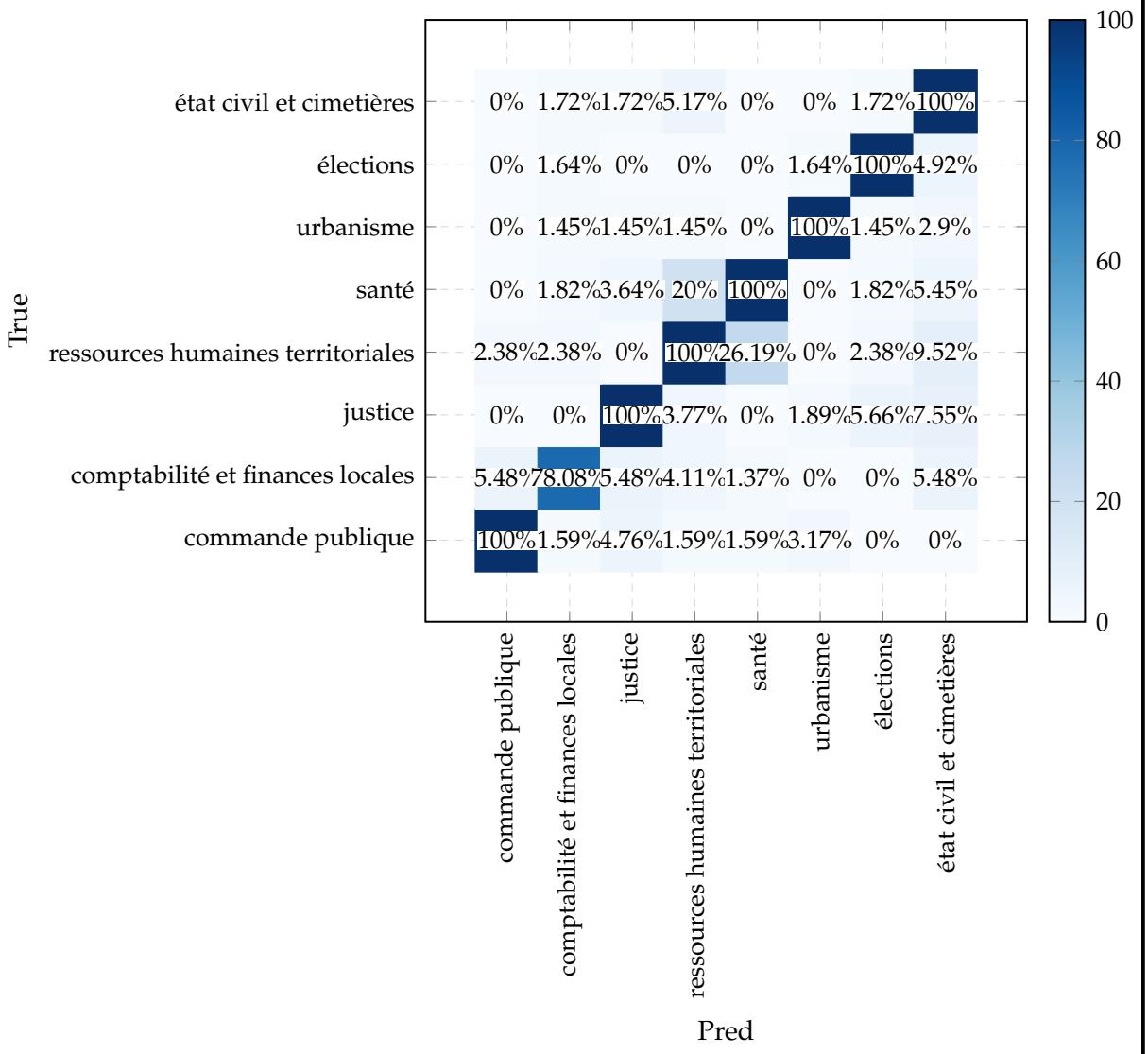

FIGURE A.12 – Matrice de confusion - Sentence CamemBERT – CLASSIFICATION LINEARSVC (ACC = 0.86)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,89	0,88	0,89
Comptabilité et finances locales	0,88	0,87	0,87
Justice	0,85	0,9	0,87
Ressources humaines territoriales	0,81	0,87	0,84
Santé	0,84	0,76	0,8
Urbanisme	0,92	0,86	0,89
Élections	0,89	0,96	0,92
État civil & cimetières	0,88	0,83	0,85
Macro AVG	0,87	0,86	0,87
Weighted AVG	0,87	0,87	0,87

TABLE A.16 – Résultats de Classification de domaine FLAUBERT-BASE-UNCASED-XNLI-STS – CLASSIFICATION LINEARSVC – (ACC = 0,87)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,91	0,97	0,94
Comptabilité et finances locales	0,93	0,86	0,89
Justice	0,94	0,86	0,9
Ressources humaines territoriales	0,88	0,92	0,9
Santé	0,95	0,9	0,92
Urbanisme	0,9	0,95	0,92
Élections	0,93	0,96	0,94
État civil & cimetières	0,95	0,99	0,97
Macro AVG	0,92	0,93	0,92
Weighted AVG	0,92	0,92	0,92

TABLE A.17 – Résultats de Classification de domaine sur les textes SANS PRETRAITEMENT : SENTENCE CAMEMBERT BASE – CLASSIFICATION LINEARSVC – (ACC = 0,92)

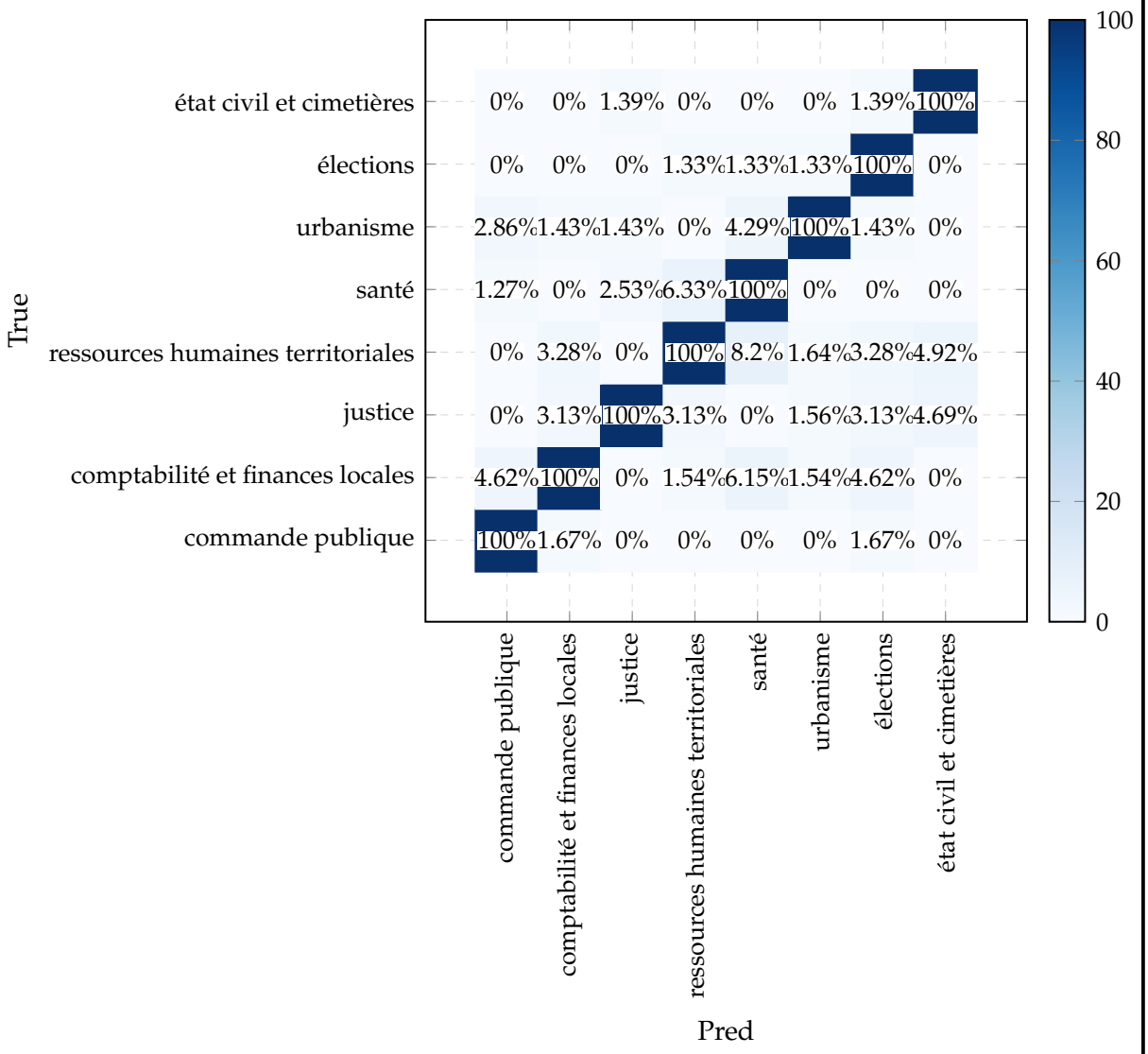

FIGURE A.13 – Matrice de confusion - Sentence CamemBERTsur un dataset sans prétraitement – CLASSIFICATION LINEARSVC (ACC = 0.94)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,92	0,94	0,93
Comptabilité et finances locales	0,9	0,86	0,88
Justice	0,9	0,86	0,88
Ressources humaines territoriales	0,86	0,89	0,87
Santé	0,95	0,86	0,9
Urbanisme	0,9	0,95	0,92
Élections	0,93	0,96	0,94
État civil & cimetières	0,92	0,97	0,95
Macro AVG	0,91	0,91	0,91
Weighted AVG	0,91	0,91	0,91

TABLE A.18 – Résultats de Classification de domaine sur les textes SANS PRETRAITEMENT : SENTENCE CAMEMBERT BASE – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,91)

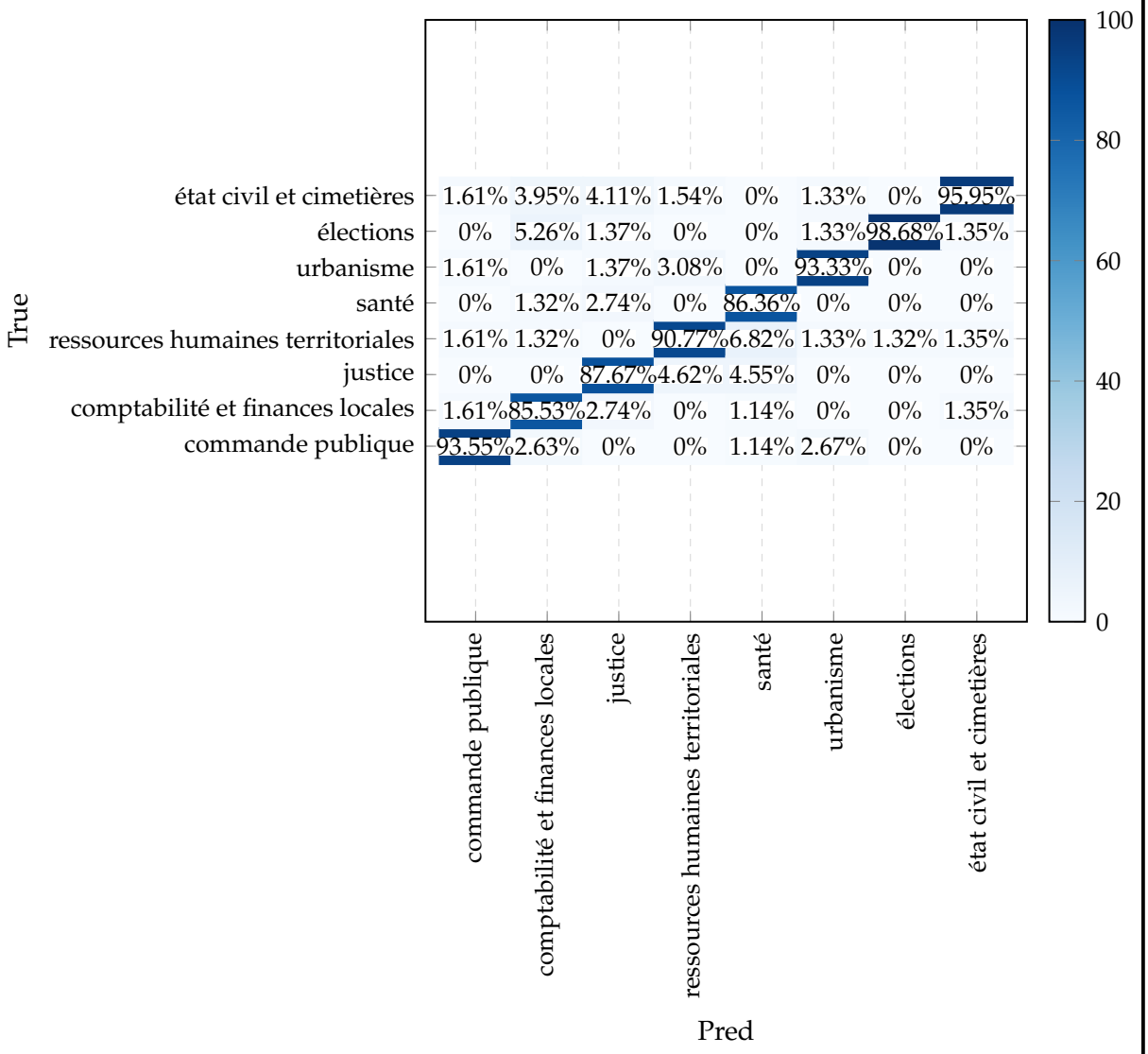

FIGURE A.14 – Matrice de confusion - Sentence CamemBERT sur un Dataset sans pré-traitement – CLASSIFICATION SGDClassifier (ACC = 0.91)

Domaine	Précision	Rappel	F1-Score
Commande publique	0.93	0.96	0.94
Comptabilité et finances locales	0.95	0.94	0.94
Justice	0.91	0.85	0.87
Ressources humaines territoriales	0.87	0.88	0.87
Santé	0.82	0.84	0.83
Urbanisme	0.93	0.91	0.92
Élections	0.94	0.98	0.96
État civil & cimetières	0.94	0.91	0.93
Macro AVG	0.91	0.91	0.91
Weighted AVG	0.91	0.91	0.91

TABLE A.19 – Résultats de Classification de domaine SENTENCE CAMEMBERT BASE – Représentation n°2 – CLASSIFICATION LINEARSVC – (ACC = 0,91)

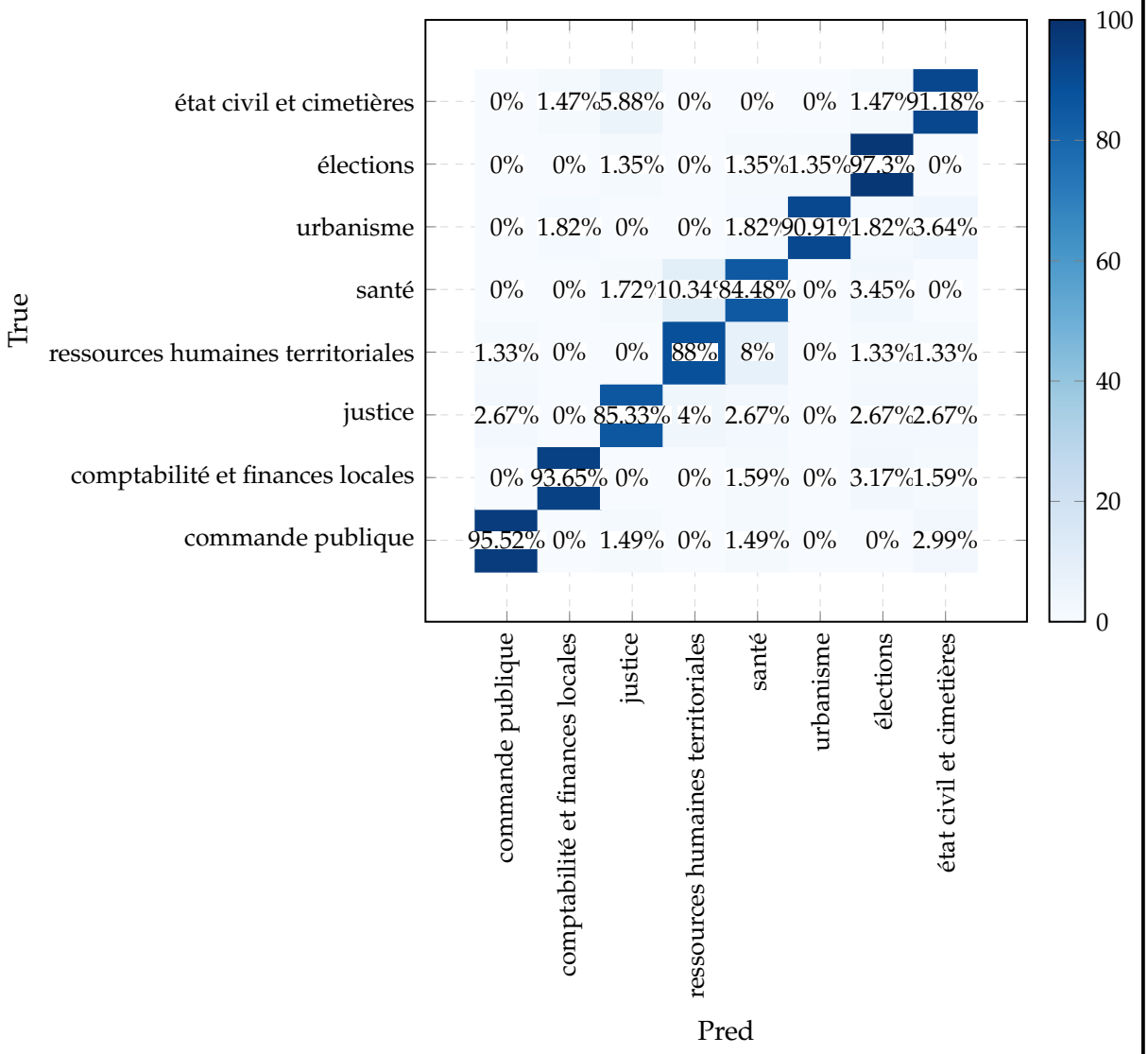

FIGURE A.15 – Matrice de confusion - Sentence CamemBERT – Représentation n°2 – CLASSIFICATION LINEARSVC (ACC = 0.91)

Domaine	Précision	Rappel	F1-Score
Commande publique	0.93	0.96	0.94
Comptabilité et finances locales	0.94	0.97	0.95
Justice	0.81	0.91	0.86
Ressources humaines territoriales	0.83	0.89	0.86
Santé	0.85	0.76	0.80
Urbanisme	0.94	0.85	0.89
Élections	0.96	0.97	0.97
État civil & cimetières	0.97	0.84	0.90
Macro AVG	0.90	0.89	0.90
Weighted AVG	0.90	0.90	0.90

TABLE A.20 – Résultats de Classification de domaine SENTENCE CAMEMBERT BASE – Représentation n°2 – CLASSIFICATION SGDClassifier – (ACC = 0,90)

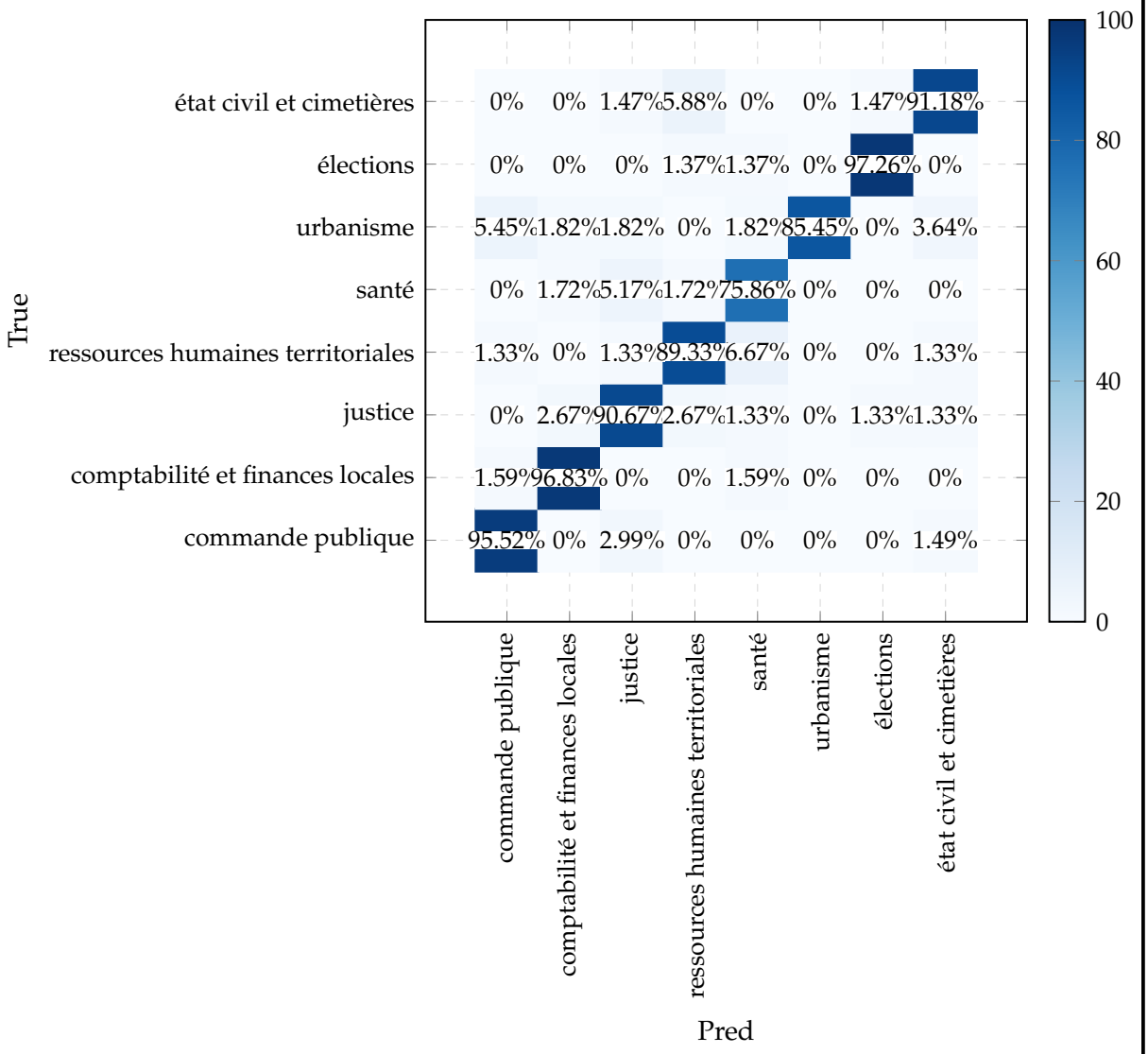

FIGURE A.16 – Matrice de confusion - Sentence CamemBERT – Représentation n°2 – CLASSIFICATION SGDClassifier (ACC = 0.90)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,89	0,88	0,89
Comptabilité et finances locales	0,88	0,87	0,87
Justice	0,85	0,9	0,87
Ressources humaines territoriales	0,81	0,87	0,84
Santé	0,84	0,76	0,8
Urbanisme	0,92	0,86	0,89
Élections	0,89	0,96	0,92
État civil & cimetières	0,88	0,83	0,85
Macro AVG	0,87	0,86	0,87
Weighted AVG	0,87	0,87	0,87

TABLE A.21 – Résultats de Classification de domaine FLAUBERT-BASE-UNCASED-XNLI-STS – Représentation n°2 – CLASSIFICATION LINEARSVC – (ACC = 0,87)

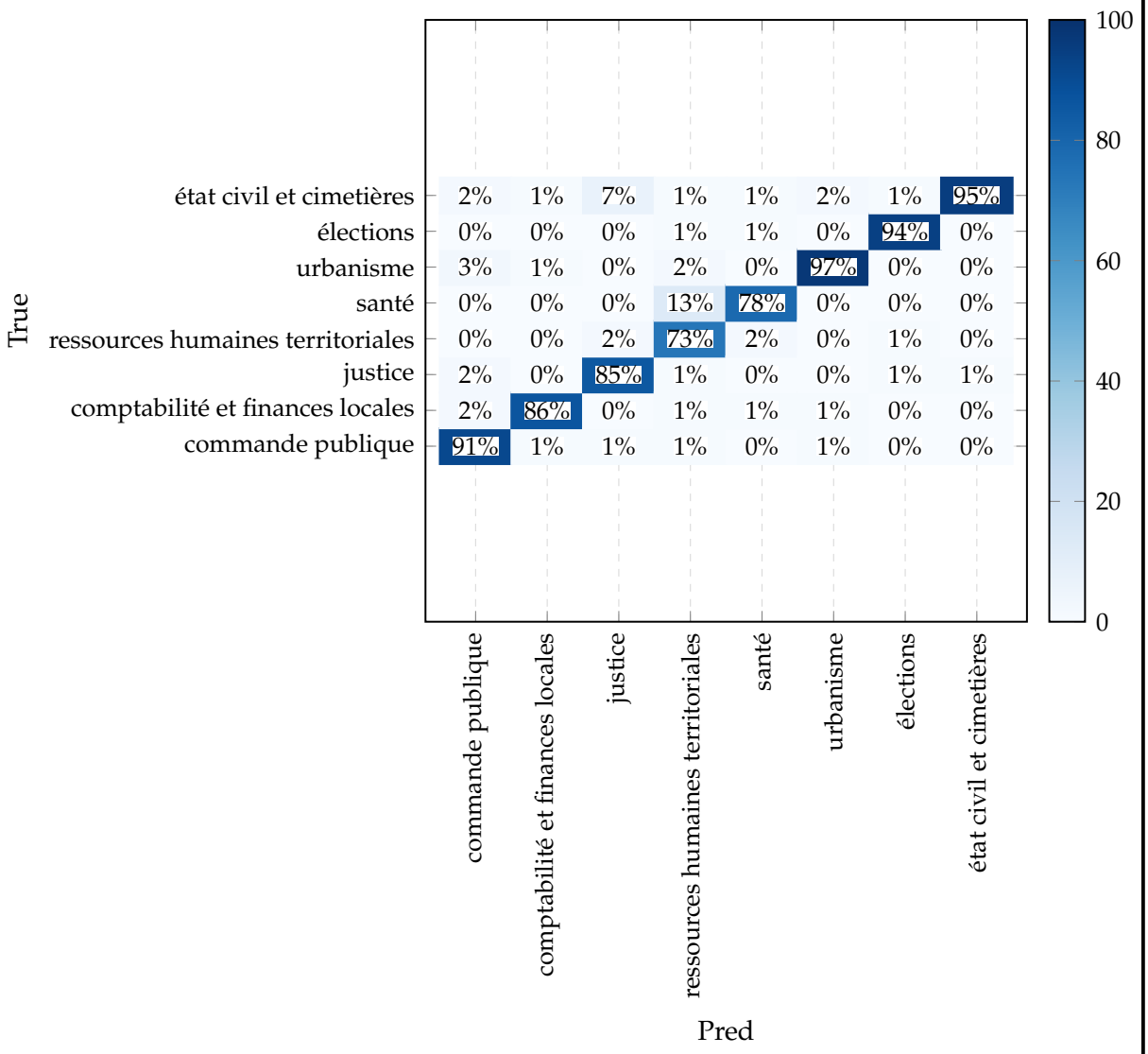

FIGURE A.17 – Matrice de confusion - Sentence FlauBERT – CLASSIFICATION LINEARSVC (ACC = 0.87)

Domaine	Précision	Rappel	F1-Score
Commande publique	0.91	0.88	0.90
Comptabilité et finances locales	0.88	0.88	0.88
Justice	0.97	0.84	0.90
Ressources humaines territoriales	0.74	0.71	0.72
Santé	0.76	0.78	0.77
Urbanisme	0.84	0.99	0.90
Élections	0.95	0.93	0.94
État civil & cimetières	0.92	0.97	0.94
Macro AVG	0.87	0.87	0.87
Weighted AVG	0.87	0.87	0.87

TABLE A.22 – [Résultats de Classification de domaine FLAUBERT-BASE-UNCASED-XNLI-STS – Représentation n°2 – CLASSIFICATION SGDClassifier – (ACC = 0,87)

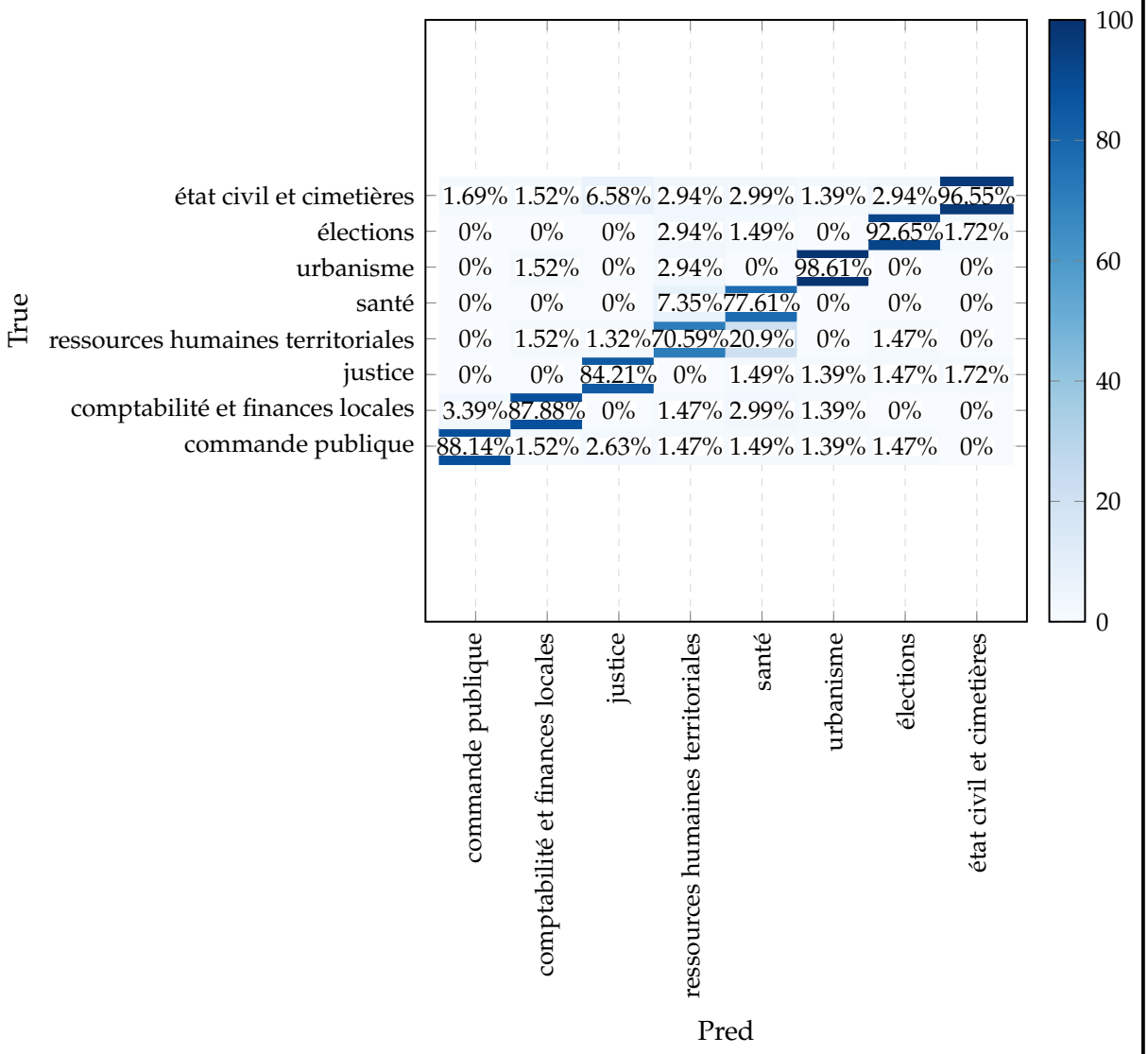

FIGURE A.18 – Matrice de confusion - Sentence FlauBERT – CLASSIFICATION SGDClassifier (ACC = 0.87)

Plongements lexicaux (*Word Embeddings*)

Les plongements lexicaux, ou Word Embeddings, constituent une méthode avancée pour la représentation des mots. Nous avons utilisé Word2Vec avec l'implémentation de gensim, en construisant le vocabulaire soit uniquement à partir du jeu d'entraînement, soit à partir du corpus complet, permettant une représentation des mots en contexte général.

Pour ce qui est des différentes représentations, nous avons exploré trois approches distinctes :

1. **Représentation n°1** : Utilisation de Word2Vec avec un vocabulaire construit sur le corpus complet et détection des trigrams via 'models.phrases.Phrases'. Le vecteur du document est créé par la moyenne des vecteurs des mots, suivie de la concaténation de chacune des moyennes.

2. **Représentation n°2** : Cette approche utilise également Word2Vec avec un vocabulaire construit sur le corpus complet et la détection des trigrams. Cependant, la création du vecteur du document diffère, en utilisant la somme des vecteurs de chacun des mots, puis en calculant la moyenne en divisant par le nombre de mots.

3. **Représentation n°3** : Elle suit le même procédé que les représentations précédentes, mais avec la détection des bigrams au lieu des trigrams. La création du vecteur du document se fait par la somme des vecteurs de chacun des mots, suivie du calcul de la moyenne par le nombre de mots.

   Ces différentes approches permettent de capturer des informations contextuelles et de fournir une représentation vectorielle riche pour les documents, améliorant ainsi la performance des modèles de classification ultérieurs.

Domaine	Précision	Rappel	F1-Score
Commande publique	0	0	0
Comptabilité et finances locales	0,21	0,14	0,17
Justice	0,22	0,28	0,25
Ressources humaines territoriales	0,16	0,15	0,16
Santé	0,61	0,14	0,23
Urbanisme	0,13	0,3	0,18
Élections	0,07	0,03	0,04
État civil & cimetières	0,21	0,52	0,3
Macro AVG	0,2	0,19	0,17
Weighted AVG	0,21	0,19	0,17

TABLE A.23 – Résultats de Classification de domaine WORD2VEC Représentation N°1 – VOCABULAIRE ENTRAINE SUR LE DATASET ENTIER - CLASSIFICATION LINEARSVC - (ACC = 0,19)

Domaine	Précision	Rappel	F1-Score
Commande publique	0.16	0.16	0.16
Comptabilité et finances locales	0.11	0.11	0.11
Justice	0.17	0.08	0.11
Ressources humaines territoriales	0.14	0.23	0.17
Santé	0.27	0.19	0.22
Urbanisme	0.10	0.10	0.10
Élections	0.14	0.01	0.02
État civil & cimetières	0.20	0.45	0.28
Macro AVG	0.16	0.17	0.15
Weighted AVG	0.16	0.16	0.15

TABLE A.24 – Résultats de Classification de domaine WORD2VEC Représentation N°1 – VOCABULAIRE ENTRAINE SUR LE DATASET ENTIER EN GARDANT LES APOSTROPHES - CLASSIFICATION LINEARSVC - (ACC = 0,16)

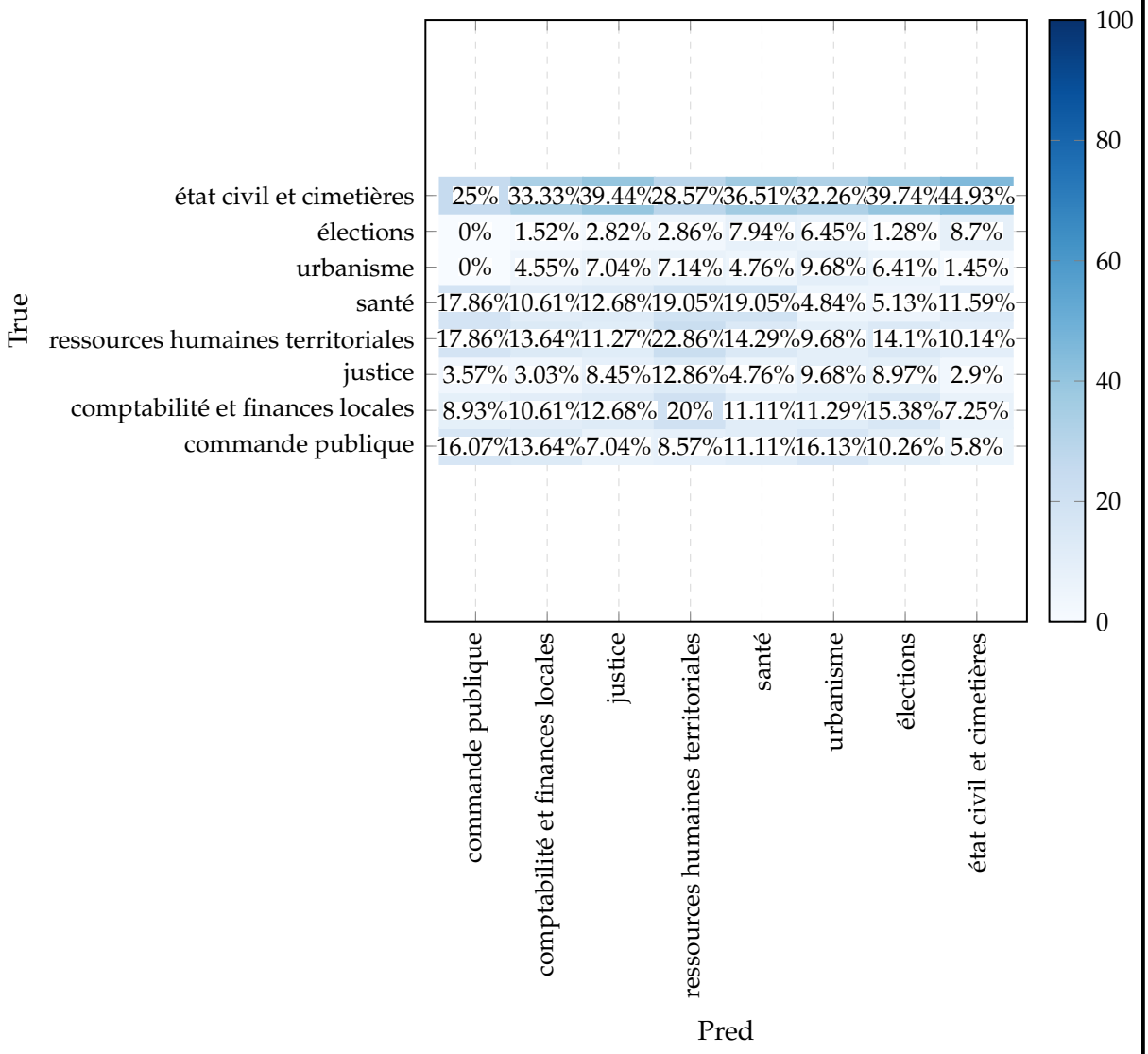

FIGURE A.19 – Matrice de confusion - WORD2VEC Représentation N°1 – VOCABULAIRE ENTRAINE SUR LE DATASET ENTIER EN GARDANT LES APOSTROPHES - CLASSIFICATION LINEARSVC - (ACC = 0,16)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,07	0,02	0,03
Comptabilité et finances locales	0,19	0,39	0,25
Justice	0,16	0,76	0,26
Ressources humaines territoriales	0	0	0
Santé	0,08	0,03	0,05
Urbanisme	0	0	0
Élections	0	0	0
État civil & cimetières	0	0	0
Macro AVG	0,06	0,15	0,07
Weighted AVG	0,07	0,16	0,08

TABLE A.25 – Résultats de Classification de domaine WORD2VEC MOYENNE DES VECTEURS DE MOTS (REP N°2) – TRIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,16)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,37	0,68	0,48
Comptabilité et finances locales	0,25	0,44	0,32
Justice	0,06	0,01	0,02
Ressources humaines territoriales	0,33	0,16	0,22
Santé	0,46	0,32	0,38
Urbanisme	0,16	0,05	0,07
Élections	0,32	0,53	0,40
État civil & cimetières	0,54	0,61	0,57
Macro AVG	0,31	0,35	0,31
Weighted AVG	0,32	0,35	0,31

TABLE A.26 – Résultats de Classification de domaine WORD2VEC MOYENNE DES VECTEURS DE MOTS (REP N°2) – TRIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION SGDCLASSIFIER – En gardant les apostrophes (ACC = 0,35)

Domaine	Précision	Rappel	F1-Score
Commande publique	0	0	0
Comptabilité et finances locales	0	0	0
Justice	0,15	0,86	0,26
Ressources humaines territoriales	0	0	0
Santé	0	0	0
Urbanisme	0	0	0
Élections	0,09	0,14	0,11
État civil & cimetières	0,87	0,53	0,66
Macro AVG	0,14	0,19	0,13
Weighted AVG	0,14	0,20	0,13

TABLE A.27 – Résultats de Classification de domaine WORD2VEC MOYENNE DES VECTEURS DE MOTS (REP N°2) – TRIGRAMS –VOCABULAIRE DATASET ENTIER – CLASSIFICATION LINEARSVC – (ACC = 0,2)

Domaine	Précision	Rappel	F1-Score
Commande publique	0.37	0.68	0.48
Comptabilité et finances locales	0.25	0.44	0.32
Justice	0.06	0.01	0.02
Ressources humaines territoriales	0.33	0.16	0.22
Santé	0.46	0.32	0.38
Urbanisme	0.29	0.40	0.34
Élections	0.32	0.53	0.40
État civil & cimetières	0.54	0.61	0.57
Macro AVG	0.31	0.35	0.31
Weighted AVG	0.32	0.35	0.31

TABLE A.28 – Résultats de Classification de domaine WORD2VEC MOYENNE DES VECTEURS DE MOTS (REP N°2) – TRIGRAMS – VOCABULAIRE DATASET ENTIER EN GARDANT LES APOSTROPHES – CLASSIFICATION LINEARSVC – (ACC = 0,35)

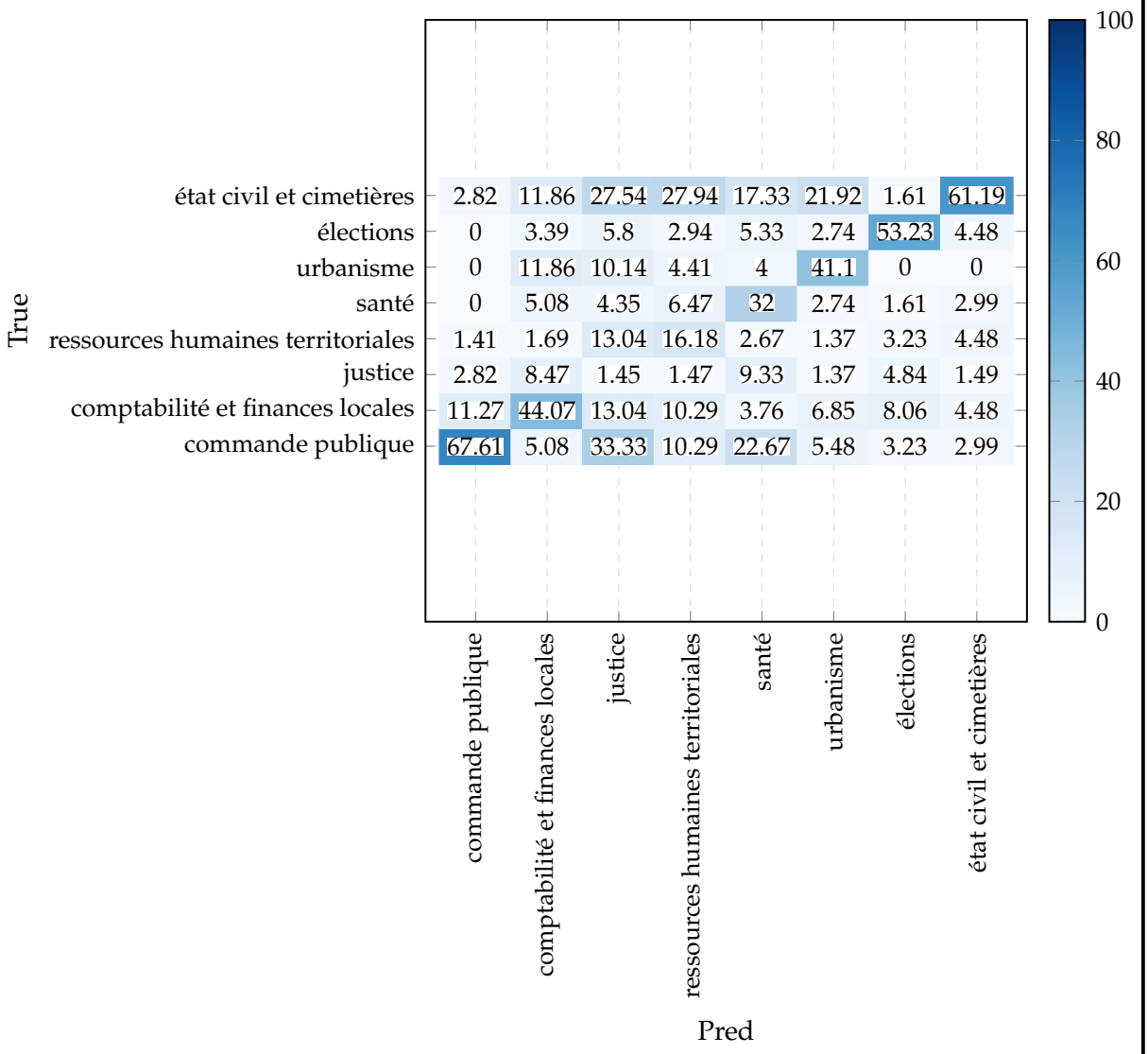

FIGURE A.20 – Matrice de confusion - WORD2VEC Représentation N°2 – VOCABULAIRE ENTRAINE SUR LE DATASET ENTIER EN GARDANT LES APOSTROPHES - CLASSIFICATION LINEARSVC - (ACC = 0,35)

Domaine	Précision	Rappel	F1-Score
Commande publique	0	0	0
Comptabilité et finances locales	0	0	0
Justice	0	0	0
Ressources humaines territoriales	0	0	0
Santé	0	0	0
Urbanisme	0,34	0,45	0,39
Élections	0	0	0
État civil & cimetières	0,16	0,97	0,27
Macro AVG	0,06	0,18	0,08
Weighted AVG	0,06	0,19	0,08

TABLE A.29 – Résultats de Classification de domaine WORD2VEC Représentation N°3 : MOYENNE DES VECTEURS DE MOTS – BIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION SGDCLASSIFIER – (ACC = 0,19)

Domaine	Précision	Rappel	F1-Score
Commande publique	0	0	0
Comptabilité et finances locales	0,5	0,01	0,03
Justice	0	0	0
Ressources humaines territoriales	0,21	0,6	0,31
Santé	0	0	0
Urbanisme	0	0	0
Élections	0,14	0,83	0,25
État civil & cimetières	0	0	0
Macro AVG	0,11	0,18	0,07
Weighted AVG	0,11	0,17	0,07

TABLE A.30 – Résultats de Classification de domaine WORD2VEC Représentation N°3 : MOYENNE DES VECTEURS DE MOTS – BIGRAMS – VOCABULAIRE DATASET ENTIER – CLASSIFICATION LinearSVC – (ACC = 0,17)

Domaine	Précision	Rappel	F1-Score
Commande publique	0,12	0,03	0,05
Comptabilité et finances locales	0	0	0
Justice	0	0	0
Ressources humaines territoriales	0	0	0
Santé	0,13	0,48	0,21
Urbanisme	0,22	0,49	0,3
Élections	0,1	0,03	0,04
État civil & cimetières	0,33	0,43	0,38
Macro AVG	0,11	0,18	0,12
Weighted AVG	0,12	0,19	0,13

TABLE A.31 – Résultats de Classification de domaine WORD2VEC Représentation N°2 : MOYENNE DES VECTEURS DE MOTS – TRIGRAMS – VOCABULAIRE DATASET ENTIER – COSINUS SIMILARITY – (ACC = 0,19)

Les résultats des différentes représentations de vecteurs Word2Vec pour la classification des domaines à l'aide de divers modèles sont présentés dans une série de tableaux. Ces tableaux montrent la précision, le rappel et le F1-score pour chaque domaine ainsi que les moyennes macro et pondérées. Voici une analyse détaillée des résultats pour identifier les meilleures représentations pour chaque modèle.

Le tableau A.24 présente les résultats de la classification utilisant la représentation Word2Vec N°1 avec le modèle LinearSVC. Les scores de précision, rappel et F1-Score sont globalement faibles, avec une précision moyenne pondérée de 0,21 et un rappel de 0,19. Ce modèle semble avoir des difficultés particulières avec des domaines comme *Commande publique* et *Élections*, où les scores sont très bas.

Dans le tableau A.25, la représentation N°2 est utilisée avec le modèle SGDClassifier. Ici, les scores sont encore plus faibles, avec une précision moyenne pondérée de seulement 0,07 et un rappel de 0,16. Cela suggère que cette représentation et ce modèle ne sont pas optimaux pour la classification des domaines.

Le tableau A.26 montre une amélioration significative en gardant les apostrophes dans la représentation N°2 avec le modèle SGDClassifier, atteignant une précision moyenne pondérée de 0,32 et un rappel de 0,35. Les scores pour des domaines comme *Commande publique* et *État civil & cimetières* sont particulièrement améliorés.

Le tableau A.27 utilise la même représentation N°2 avec LinearSVC, mais les scores restent bas, avec une précision moyenne pondérée de 0,14 et un rappel de 0,20, indiquant que ce modèle n'est pas bien adapté à cette représentation.

Le tableau A.29 présente les résultats pour la représentation N°3 avec SGDClassifier, où les scores sont également faibles, avec une précision moyenne pondérée de 0,06 et un rappel de 0,18.

Le tableau A.30 montre que la même représentation N°3 avec LinearSVC n'améliore pas les résultats, avec des scores encore faibles (précision moyenne pondérée de 0,11 et rappel de 0,17).

Enfin, le tableau A.31 montre les résultats pour la représentation N°2 avec la similarité cosinus, où les scores sont légèrement meilleurs mais toujours insuffisants pour une bonne classification, avec une précision moyenne pondérée de 0,12 et un rappel de 0,19.

En conclusion, la meilleure représentation semble être la représentation Word2Vec N°2 en gardant les apostrophes, utilisée avec le modèle SGDClassifier, comme présenté dans le tableau A.26. Cette configuration atteint une précision moyenne pondérée de 0,32 et un rappel de 0,35, ce qui est significativement supérieur aux autres configurations même si cela reste faible.

Plongements de documents

L'algorithme de plongements de documents utilise la technique de Doc2Vec pour transformer un document en un vecteur de caractéristiques. La première étape consiste à entraîner le modèle Doc2Vec sur un corpus de documents d'apprentissage. Ensuite, pour inférer un vecteur de test à partir d'un nouveau document, on utilise la méthode intégrée 'model.infer_vector(tokens)', qui génère un vecteur représentatif pour ce document. Après avoir obtenu ce vecteur, l'algorithme identifie les 10 vecteurs les plus similaires à celui-ci dans le modèle en utilisant la fonction 'model.docvecs.most_similar([new_vector])'. Enfin, il calcule la somme des probabilités associées à ces vecteurs similaires et sélectionne la somme la plus élevée. Cette approche permet de trouver les documents les plus semblables à un document donné en termes de contenu sémantique.

Domaine	Précision	Rappel	F1-Score
Commande publique	0.90	0.97	0.93
Comptabilité et finances locales	0.97	0.82	0.89
Justice	0.90	0.92	0.91
Ressources humaines territoriales	0.77	0.76	0.77
Santé	0.78	0.79	0.78
Urbanisme	0.89	0.94	0.92
Élections	0.97	0.99	0.98
État civil & cimetières	0.91	0.92	0.92
Macro AVG	0.89	0.89	0.89
Weighted AVG	0.89	0.88	0.89

TABLE A.32 – Résultats de Classification de domaine Doc2VEC sans les apostrophes – Most Similar – (ACC = 0,89)

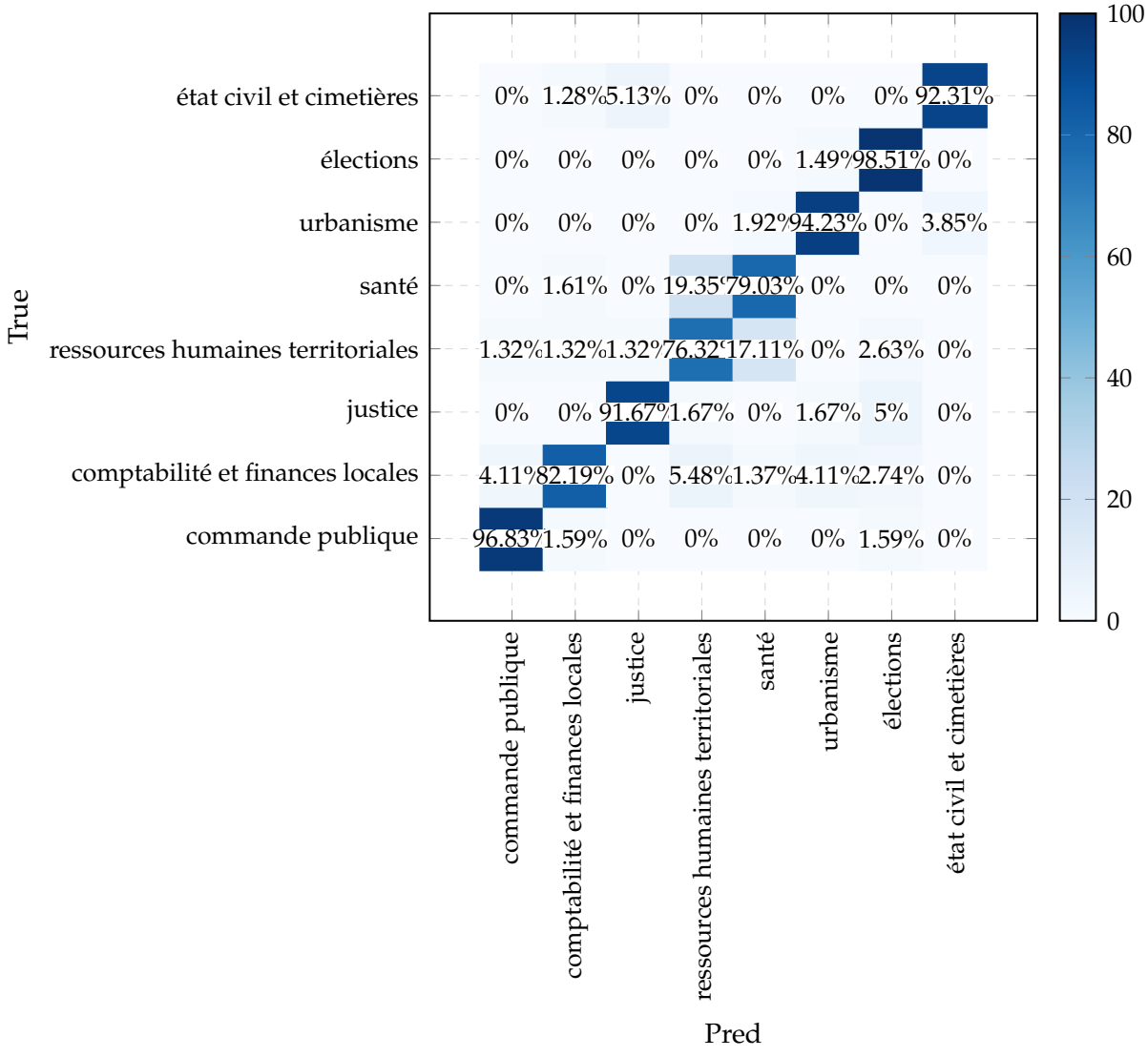

FIGURE A.21 – Matrice de confusion - Doc2Vec – MostSimilar – (ACC = 0,89)

Domaine	Précision	Rappel	F1-Score
Commande publique	0.94	0.96	0.95
Comptabilité et finances locales	0.60	0.81	0.68
Justice	0.89	0.82	0.85
Ressources humaines territoriales	0.71	0.79	0.75
Santé	0.83	0.75	0.78
Urbanisme	0.95	0.89	0.92
Élections	0.96	1.00	0.98
État civil & cimetières	0.91	0.96	0.94
Macro AVG	0.88	0.88	0.88
Weighted AVG	0.88	0.87	0.87

TABLE A.33 – Résultats de Classification de domaine Doc2VEC en gardant les apostrophes – Most Similar – (ACC = 0,87)

Les résultats des classifications de domaine en utilisant Doc2Vec sont présentés dans deux tableaux distincts, chacun offrant une perspective différente sur la performance des représentations de documents. Les tableaux A.32 et A.33 accompagnés de leurs matrices de corrélations respectivement A.21 et A.22 illustrent les performances des modèles en fonction de la gestion des apostrophes dans les représentations Doc2Vec.

Le tableau A.32 affiche les résultats pour la classification avec Doc2Vec sans garder les apostrophes. Ce modèle montre des performances remarquables avec des scores élevés dans presque tous les domaines. En particulier, les domaines tels que *Commande publique*, *Comptabilité et finances locales*, et *Élections* obtiennent des précisions et des rappels très élevés, respectivement de 0,90 et 0,97 pour *Commande publique* et de 0,97 et 0,99 pour *Élections*. La précision moyenne pondérée est de 0,89, avec un rappel moyen de 0,88, et un F1-Score moyen de 0,89, ce qui indique une performance robuste et cohérente pour cette configuration.

En revanche, le tableau A.33 montre les résultats de la même méthode Doc2Vec, mais cette fois-ci en gardant les apostrophes. Bien que les scores soient globalement bons, ils sont légèrement inférieurs à ceux obtenus sans les apostrophes. La précision moyenne pondérée est de 0,88, le rappel moyen est de 0,87, et le F1-Score moyen est de 0,87. Les domaines comme *Comptabilité et finances locales* et *Justice* affichent des résultats moins impressionnants comparés aux scores obtenus sans les apostrophes, avec des précisions respectives de 0,60 et 0,89 contre 0,97 et 0,90 dans le premier tableau.

En conclusion, la représentation Doc2Vec sans apostrophes, comme montré dans le tableau A.32, semble offrir des performances globalement meilleures pour la classification des domaines. Les scores plus élevés en précision, rappel et F1-Score indiquent une meilleure capacité de ce modèle à classer correctement les documents dans les domaines spécifiques. En revanche, bien que garder les apostrophes (tableau A.33) n'entraîne pas une dégradation drastique des performances, il semble que la configuration sans apostrophes soit légèrement plus efficace pour cette tâche de classification.

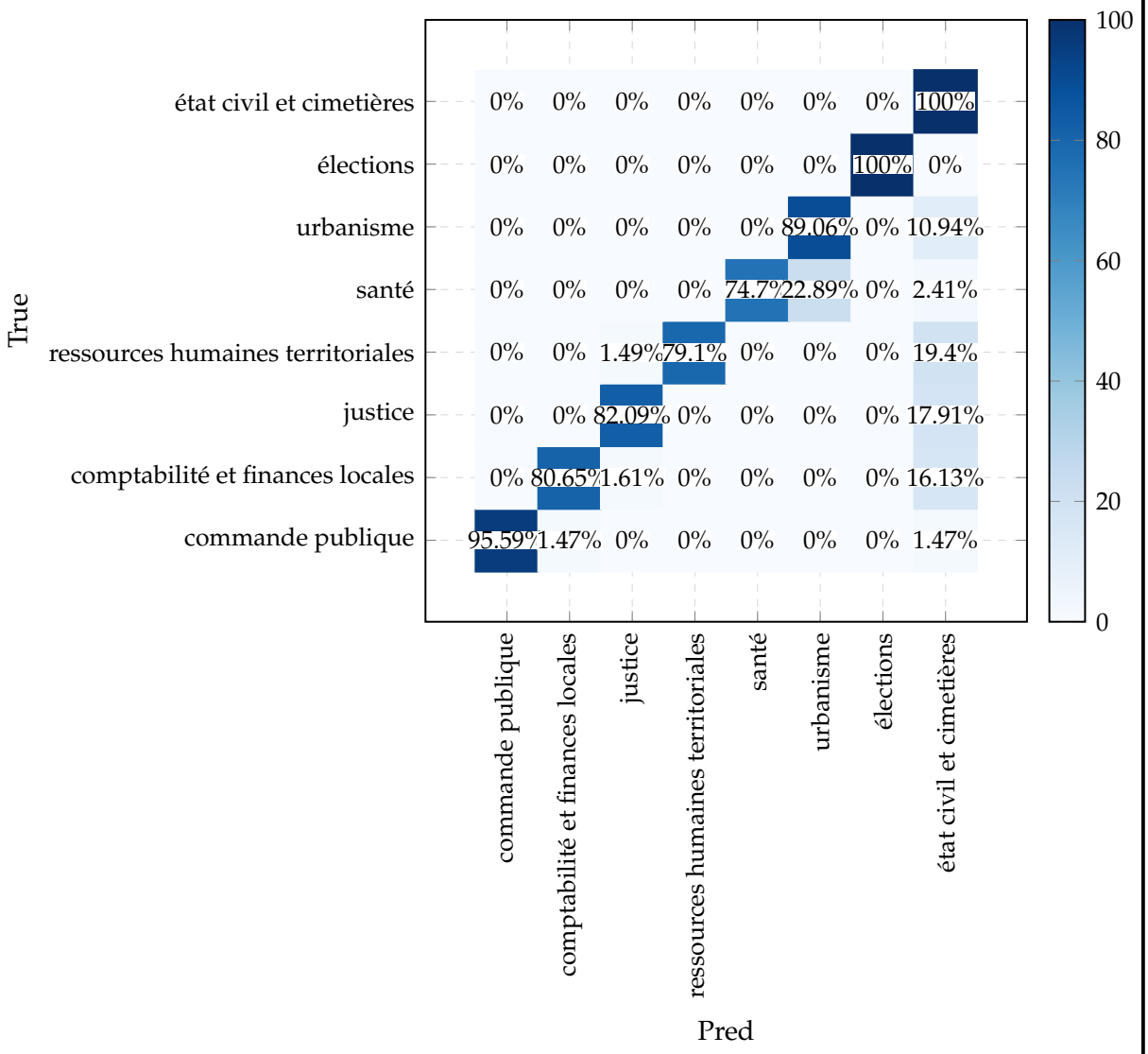

FIGURE A.22 – Matrice de confusion - Doc2Vec - Représentation en gardant les apostrophes – MostSimilar – (ACC = 0,87)