UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE Nº 601
Mathématique et Sciences et Technologies de l'Information et de la Communication
Spécialité : Informatique

Par
Clément DALLOUX

FOUILLE DE TEXTE ET EXTRACTION D'INFORMATIONS DANS LES DONNÉES CLINIQUES

Thèse présentée et soutenue à Rennes le 7 décembre 2020
Unité de recherche : IRISA — UMR6074
Thèse Nº :

Rapporteurs avant soutenance :
Dr. Patrick RUCH Professeur, HEG/HES-SO, Genève
Dr. Xavier TANNIER Professeur, Sorbonne Université, Paris

Composition du jury :
Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition ne comprend que les membres présents

Examinateurs :	Dr. Peggy CELLIER	Maître de conférence, INSA, Rennes
	Dr. Claire NÉDELLEC	Directrice de recherche, INRAE, Jouy-en-Josas
	Dr. Patrick RUCH	Professeur, HEG/HES-SO, Genève
	Dr. Xavier TANNIER	Professeur, Sorbonne Université, Paris
Dir. de thèse :	Dr. Vincent CLAVEAU	Chargé de recherche, CNRS, IRISA, Rennes
Co-encadrante :	Dr. Natalia GRABAR	Chargée de recherche, CNRS, STL, Lille

Invité :
Dr. Olivier DAMERON Professeur, Université de Rennes 1, Rennes

Remerciements

Dans ce qui suit, il m'est donné l'opportunité de remercier toutes les personnes, qui, par leur soutien et leur bienveillance m'ont permis d'arriver au bout de cette aventure.

Je tiens tout d'abord à remercier Vincent Claveau et Natalia Grabar, chargés de recherche au CNRS, qui ont encadré mes travaux de recherche durant plus de trois ans et qui par leur disponibilité, confiance ainsi que leurs précieux conseils et encouragements m'ont permis d'apprendre et de progresser constamment et de réaliser ces travaux dans des conditions de travail idéales. Je tiens également à remercier Olivier Dameron, Professeur à l'Université de Rennes 1, d'avoir initialement accepté de diriger cette thèse.

Je tiens aussi à remercier Xavier Tannier, Professeur à Sorbonne Université, et Patrick Ruch, professeur à HEG/HES-SO, de m'avoir fait l'honneur d'être les rapporteurs de cette thèse, ainsi que, Peggy Cellier, maître de conférence à l'INSA, et Claire Nédellec, directrice de recherche à l'INRAE, d'avoir accepté de faire partie de mon jury de thèse.

J'adresse également mes remerciements les plus sincères à tous les membres de l'équipe LinkMedia avec qui j'ai passé de très bons moments quotidiennement pendant plus de trois ans. À Rémi, Cédric, Cheikh et Mikail avec qui j'ai partagé le bureau F318, merci infiniment pour les moments de détente que nous avons partagé, mais aussi et surtout pour l'aide et le soutien que vous m'avez apporté dans les moments difficiles.

À ma famille, et tout particulièrement mes parents qui m'encouragent et me soutiennent depuis tant d'années avec une patience sans faille, merci du fond du cœur. Enfin, à toi Anne-Laure qui m'accompagne depuis près de 11 ans, et qui, par ta gentillesse, ton soutien indéfectible et ton amour inconditionnel m'a permis de tenir bon durant ces longues années d'étude, merci pour tout, je t'aime.

Sommaire

Remerciements **iii**

Introduction Générale **9**

1 État de l'art **13**

1.1 Généralités 15

1.1.1 Traitement automatique de la langue biomédicale 15

1.1.2 Classification automatique de textes 19

1.1.3 Étiquetage de séquences 22

1.2 Classification multi-étiquette de textes cliniques 25

1.2.1 La Classification Internationale des Maladies 25

1.2.2 Jeux de données annotées 26

1.2.3 Systèmes de classification 31

1.3 Détection de la négation et de l'incertitude 35

1.3.1 Négation et incertitude 35

1.3.2 Jeux de données annotées 40

1.3.3 Étiquetage automatique 43

2 Classification multi-étiquette de textes cliniques **59**

2.1 Jeu de données 61

2.2 Nos approches 65

2.2.1 Approche par dictionnaires 65

2.2.2 Approches par apprentissage supervisé 66

2.3 Expériences de classification CIM-10 71

2.3.1 Protocole expérimental 71
2.3.2 Analyse des résultats 74

3 Corpus constitués dans le cadre de la thèse 79
3.1 Corpus biomédicaux français 81
3.1.1 Règles d'annotation 81
3.1.2 ESSAI : corpus français d'essais cliniques 84
3.1.3 CAS : corpus français de cas cliniques 89
3.2 Corpus biomédicaux brésiliens 94
3.2.1 Protocoles d'essais cliniques brésiliens 94
3.2.2 SemClinBr 97

4 Détection de la négation et de l'incertitude 101
4.1 Nos approches 103
4.1.1 Plongements de mots pré-entraînés 103
4.1.2 Approches pour l'étiquetage de séquences 104
4.2 Étiquetage automatique des marqueurs 110
4.2.1 Protocole expérimental 110
4.2.2 Analyse des résultats 111
4.3 Étiquetage automatique de la portée 116
4.3.1 Protocole expérimental 116
4.3.2 Analyse des résultats 117
4.3.3 Analyse des erreurs 124

Conclusion générale 129

Table des matières 135

Table des figures 137

Liste des tableaux 139

Bibliographie **143**

Sources primaires 143

Sources secondaires 144

Résumé/Abstract **163**

Introduction Générale

Avec l'évolution assez récente de la gestion de données dans les hôpitaux et cliniques et la mise en place d'entrepôts de données cliniques (MADEC et al., 2019), les données de santé deviennent de plus en plus agrégées, variées et volumineuses. Elles proviennent en effet de différentes sources de données cliniques, comme par exemple les admissions et les sorties (CURNS et al., 2010 ; CHANTRY et al., 2011 ; MILOJEVIC et al., 2014 ; A. BROWN et al., 2016), les examens biologiques (KALTENBACH et LHERMITTE, 2019), l'imagerie (LAMONTAGNE et al., 2018), les consultations (HOBBS et al., 2016), les examens histopathologiques/anatomopathologiques (TRUSCHNEGG et al., 2016 ; WRENN, CALLAS et ABU-JAISH, 2017), les prescriptions (EL EMAM et al., 2009 ; H. XU et al., 2015), etc. Elles permettent également de suivre l'évolution de la santé de patients de manière plus précise et par conséquent d'améliorer leur prise en charge (SINGH et al., 2009 ; BATES, 2010 ; SITTIG et SINGH, 2012 ; MURPHY et al., 2014 ; KING et al., 2014 ; RUSSO et al., 2016). Ce changement mène à une évolution profonde de la situation clinique face aux données.

Ainsi, si les données de santé étaient précédemment exploitées dans le contexte de soins uniquement, nous assistons actuellement à un changement de positionnement de la communauté scientifique plus global, qui devient alors de plus en plus intéressée par les données de santé. Par exemple, les rares corpus cliniques librement disponibles pour la recherche, comme **MIMIC** et ses dérivés (**I2B2**, **N2C2**, etc.), offrent d'énormes possibilités pour la recherche, comme en témoignent de très nombreux travaux engendrés par ces corpus. Nous présentons ici quelques travaux seulement réalisés autour de **MIMIC-III** (A. E. JOHNSON et al., 2016), qui montrent la variété de besoins et la richesse potentielle de la recherche scientifique : GENTIMIS et al., 2017 prédisent la durée de séjours à l'hôpital à l'aide de réseaux de neurones entraînés sur les données de **MIMIC-III**, PURUSHOTHAM et al., 2018 ; SCHERPF et al., 2019 proposent le même genre de travaux ; VINCENT et al., 2018 proposent, par analyse de **MIMIC-III**, d'étudier la relation entre la pression artérielle moyenne et la mortalité chez les patients présentant un choc distributif, FENG et al., 2018 ; SANDFORT et al., 2019 ; J. XU et al., 2019 proposent aussi, par analyse de **MIMIC-III**, d'étudier les relations entre différents évènements médicaux. Ainsi, les données collectées ou produites au cours du processus de soins cliniques peuvent être exploitées à différents niveaux et dans différents domaines, en particulier en relation avec la recherche clinique et translationnelle (MEYSTRE et al., 2017).

Dans les entrepôts de données, une partie importante des données existe sous forme structurée, c'est-à-dire que l'on peut les stocker et afficher de façon stricte et organisée, comme par exemple le nom d'un patient, sa date de naissance, sa taille,

son poids, les résultats d'un test sanguin ou encore les codes des concepts de la **SNOMED CT** (DONNELLY, 2006), grâce notamment à la mise en place de formulaires, d'éditeurs de modèles d'information clinique (*archetypes*) (MALDONADO et al., 2009), d'outils de codage automatique (STANFILL et al., 2010), etc. Cependant, dans les dossiers patients informatisés, la majeure partie des informations détaillées est toujours stockée sous la forme non structurée : le texte libre. Dans cette situation, l'absence de méthodes et ressources efficaces de traitement automatique du langage naturel (**TALN**) dédiées aux données cliniques, notamment en français, conduit à la mise au point d'outils limités, ce qui limite sensiblement les possibilités d'exploration et d'analyse de données cliniques.

En même temps, un réel besoin existe de la part des cliniciens et biologistes dans l'exploration de données cliniques. Prenons l'exemple du recrutement de patients pour les essais cliniques. Le recrutement de patients pour les essais cliniques est une étape cruciale pour le bon déroulement des essais (KÖPCKE et PROKOSCH, 2014). Le temps de recrutement est limité et le nombre de participants doit être suffisamment important pour permettre une analyse scientifiquement et statistiquement valable. Souvent, soit le nombre de participants initialement prévu n'est pas atteint soit la période de recrutement doit être prolongée pour l'atteindre, ce qui peut retarder le déroulement de l'étude et entraîner un surcoût. Le recrutement est affecté par une multitude de facteurs : la nature de l'essai (prospectif ou rétrospectif), le nombre de patients à filtrer, le nombre de participants potentiels parmi les patients filtrés, le nombre d'hôpitaux participants à l'étude, l'urgence de recruter un patient après avoir découvert son éligibilité, les fonds disponibles, etc. Avec le développement des dossiers patients informatisés et des systèmes d'aide à la décision médicale, de nombreux systèmes d'aide au recrutement pour les essais cliniques ont été proposés mais les problèmes de recrutement persistent (CUGGIA, BESANA et GLASSPOOL, 2011). Afin de faciliter le processus de recrutement, le système a besoin de confronter des critères d'inclusion et d'exclusion aux données des dossiers patients informatisés. Une grande partie de ces données étant textuelles, le développement d'approches de **TALN** est donc indispensable. Ainsi, plusieurs travaux y ont été consacrés (PAKHOMOV, BUNTROCK et CHUTE, 2005 ; L. LI et al., 2008 ; J. ZHANG et al., 2010). De plus, la campagne d'évaluation **N2C2 2018** a dédié l'une de ses tâches à la sélection de patients pour les essais cliniques (STUBBS, FILANNINO et al., 2019). Cette tâche de classification multi-étiquette consistait à déterminer, à l'aide de systèmes de **TALN**, si les patients satisfaisaient ou non les critères d'inclusion dans un essai clinique, tel que calculé à partir de leurs documents cliniques. Le système le plus performant a atteint une F-mesure de 91 %, un score très élevé démontrant l'utilité des approches de **TALN** pour le recrutement de patients (OLEYNIK et al., 2019).

Notre travail de thèse s'inscrit dans le cadre du projet **CominLabs BigClin** [1] (*Big Data Analystics for Unstructured Clinical Data*). Ce projet propose de développer une nouvelle représentation des dossiers patients informatisés reposant sur une annotation sémantique fine grâce aux nouveaux outils de **TALN** dédiés aux textes cliniques en français. Démarré en octobre 2016, ce projet, porté par Marc Cuggia et

1. `https://project.inria.fr/bigclin/`

Vincent Claveau, a impliqué plusieurs équipes de l'**IRISA-Rennes**, l'équipe **HBD-CHU Rennes** et Natalia Grabar de l'équipe **STL-Lille**. Officiellement terminé en septembre 2019, le projet **BigClin** a généré une vingtaine de publications, plusieurs outils disponibles sur la plate-forme **Allgo**[2], ainsi que plusieurs jeux de données annotées présentés dans ce manuscrit. Le rapport pré-final du projet est accessible librement sur le site web du projet.[3] Puisque l'objectif est d'ajouter efficacement ces informations sémantiques aux données structurées existantes pour être analysées dans une infrastructure de données massives, le projet abordait également les problèmes des systèmes distribués (extensibilité, gestion des données incertaines), de la confidentialité, du traitement des flux au moment de l'exécution, etc. Ce projet montre comment la recherche clinique peut tirer parti du **TALN**, de la recherche d'informations (**RI**) et du raisonnement automatisé pour traiter différents cas d'usage. Les objectifs spécifiques étaient de développer des méthodes d'extraction et d'indexation d'informations dédiées aux textes cliniques en français ; d'exploiter les techniques de fouille de données pour traiter conjointement la représentation générée des informations non structurées des dossiers cliniques et des informations cliniques structurées existantes ; de développer des méthodes distribuées pour assurer à la fois l'extensibilité et le traitement en ligne de ces techniques de **TALN**/**RI** et de fouille de données ; d'évaluer la valeur ajoutée de ces méthodes sur des données cliniques et dans des cas d'usages réels, notamment l'épidémiologie et la pharmacovigilance, l'évaluation de la pratique clinique et la recherche sur la qualité des soins de santé, les essais cliniques, etc.

Les travaux présentés dans ce manuscrit s'inscrivent donc dans ce projet plus large. Plus particulièrement, nous proposons de travailler sur deux tâches distinctes : la classification multi-étiquette de textes cliniques et la détection de la négation et de l'incertitude. Afin de résoudre ces deux tâches, nous proposons différentes approches reposant principalement sur des algorithmes d'apprentissage profond entraînés sur des jeux de données annotées. COLLOBERT et al., 2011 ont démontré la supériorité des réseaux de neurones pour plusieurs tâches typiques du **TALN** (étiquetage morpho-syntaxique, reconnaissance d'entités nommées, etc.). Depuis, l'adoption de ces méthodes pour la plupart des tâches est devenue progressivement majoritaire dans le domaine. En effet, lors des conférences **ACL** et **EMNLP** en 2017, près de 70 % des articles acceptés reposaient sur des architectures et algorithmes d'apprentissage profond (YOUNG et al., 2018) et cette tendance semble s'affirmer au fil des années. Les jeux de données français que nous exploitons sont constitués d'une part de textes cliniques du **CHU** de Rennes et d'autre part de textes biomédicaux librement accessibles que nous annotons et rendons accessibles à la communauté scientifique via un formulaire de téléchargement. La création et le partage de jeux de données annotées en français est un enjeu important pour la recherche et le développement d'outils de **TALN** pour le domaine biomédical. En effet, si peu de corpus de textes cliniques anglais annotés sont disponibles pour la recherche, les corpus de textes cliniques français sont bien plus rares encore car la loi interdit leur diffusion. À des fins de recherche, travailler en collaboration avec un établissement

2. `https://allgo.inria.fr/`
3. `https://project.inria.fr/bigclin/`

de santé peut permettre, après autorisation, d'accéder à des données personnelles de santé. Dans le cadre du projet **BigClin**, nous accédons à certaines données du **CHU** de Rennes, sur place, à partir d'un serveur dédié.

La suite de ce manuscrit est composée de 4 chapitres.

Dans le premier chapitre, nous présentons un état de l'art en trois parties. La première partie, *Généralités*, est consacrée à la revue de trois domaines : le traitement automatique de la langue biomédicale, la classification automatique de textes, ainsi que l'étiquetage de séquences. La seconde partie de l'état de l'art est consacrée à la classification multi-étiquette de textes cliniques. Nous y présentons la Classification Internationale des Maladies (**CIM**), les jeux de données annotées et les approches proposées dans la littérature. La dernière partie est consacrée à la détection de la négation et de l'incertitude dans les textes. Nous y présentons ces phénomènes linguistiques, les jeux de données annotées et les approches proposées dans la littérature.

Le second chapitre de ce manuscrit est consacré à la classification multi-étiquette de textes cliniques. Dans ce chapitre, nous commençons par décrire les textes cliniques annotés par le Centre Hospitalier Universitaire de Rennes. Ensuite, nous présentons les approches par apprentissage supervisé et par système expert que nous proposons pour la tâche de classification multi-étiquette. Enfin nous présentons le protocole expérimental ainsi que les résultats obtenus par nos approches sur ces données.

Dans le troisième chapitre, nous présentons les corpus constitués dans le cadre de la thèse. La première partie de ce chapitre est consacrée aux corpus biomédicaux français que nous annotons. Nous y présentons les règles d'annotation ainsi que le contenu de chaque corpus et les résultats du processus d'annotation. La seconde partie de ce chapitre est consacrée aux corpus biomédicaux brésiliens constitués dans le cadre d'échanges avec l'Université Pontificale Catholique du Paraná. Nous y présentons le contenu de chaque corpus ainsi que les résultats des processus d'annotation.

Le quatrième et dernier chapitre du manuscrit est consacré à la détection de la négation et de l'incertitude dans les textes. Nous commençons par y présenter les modèles de plongements de mots ainsi que les approches par apprentissage supervisé retenues pour les tâches d'étiquetage de séquences en question. Ensuite, nous présentons les protocoles expérimentaux et les résultats obtenus par nos approches pour chaque étape du processus de détection et nous les comparons à nos données de référence.

Enfin, dans la conclusion générale, nous revenons sur l'ensemble des contributions de cette thèse.

Chapitre 1

État de l'art

Sommaire

1.1 Généralités 15
1.1.1 Traitement automatique de la langue biomédicale 15
1.1.2 Classification automatique de textes 19
1.1.3 Étiquetage de séquences 22
1.2 Classification multi-étiquette de textes cliniques 25
1.2.1 La Classification Internationale des Maladies 25
1.2.2 Jeux de données annotées 26
1.2.3 Systèmes de classification 31
1.3 Détection de la négation et de l'incertitude 35
1.3.1 Négation et incertitude 35
1.3.2 Jeux de données annotées 40
1.3.3 Étiquetage automatique 43

La classification multi-étiquette de textes cliniques et la détection de la négation et de l'incertitude dans les textes cliniques sont des sujets suscitant l'intérêt de nombreux chercheurs depuis près de vingt ans (LIMA, LAENDER et RIBEIRO-NETO, 1998 ; CHAPMAN et al., 2001). L'intérêt croissant pour ces tâches s'explique d'une part, par la disponibilité relativement récente de jeux de données annotées permettant aux chercheurs de répondre aux défis que ces tâches représentent du point de vue scientifique et, d'autre part, par la nécessité pratique de l'implémentation de solutions dans les systèmes d'informations hospitaliers.

Dans ce manuscrit, nous abordons deux problèmes différents du TAL biomédical : celui de la classification de documents (ou de patients) et celui de l'étiquetage de séquences. Dans les deux cas, ces problèmes se décomposent en trois étapes : les **données** en entrée qui sont des documents pour l'un, et des mots/phrases pour l'autre ; le **classifieur** qui, dans les deux cas, apprend à classer les documents et étiqueter les mots de manière supervisée ; et les **classes** à attribuer qui sont de multiples codes par document pour l'un et deux étapes de classification binaire pour l'autre.

C'est sur la base de ces différentes sous-tâches et de ces applications que nous proposons dans ce chapitre un état-de-l'art. Celui-ci est divisée en trois parties. Dans la section 1.1, nous abordons plusieurs sujets relatifs au domaine d'application et aux tâches que nous abordons dans cette thèse : le traitement automatique de la langue biomédicale, la classification automatique de textes ainsi que l'étiquetage de séquences. Puis, dans la section 1.2, nous présentons la **CIM** ainsi que les données annotées et les méthodes proposées dans la littérature scientifique pour résoudre le problème de la classification multi-étiquette de textes cliniques. Enfin, dans la section 1.3, nous présentons les modalités de l'expression de la négation et de l'incertitude dans les textes ainsi que les données annotées et les méthodes proposées dans la littérature pour résoudre le problème de la détection de ces phénomènes.

1.1 Généralités

1.1.1 Traitement automatique de la langue biomédicale

Le traitement automatique du langage naturel est un domaine de recherche pluridisciplinaire associant, du moins historiquement, l'informatique et la linguistique. Dans ce domaine, les tâches que les systèmes informatiques cherchent à résoudre sont relatives à différentes branches de la linguistique telles que la syntaxe (lemmatisation, étiquetage morpho-syntaxique, analyse syntaxique, etc.), la sémantique (reconnaissance d'entités nommées, traduction automatique, génération automatique de textes, etc.), le discours (résumé automatique, résolution de coréférence, analyse du discours, etc.), le traitement de la parole (reconnaissance automatique de la parole, synthèse vocale, etc.), etc. S'il est possible d'exploiter les textes produits par les systèmes de traduction ou de résumé automatique directement, certaines tâches de **TALN** telles que la lemmatisation, la racinisation ou l'étiquetage morpho-syntaxique sont généralement considérées comme des pré-traitements dont les résultats sont utilisés pour aider à résoudre des tâches plus complexes.

Les méthodes de **TALN** reposent sur deux champs d'études de l'intelligence artificielle : (1) les systèmes experts, raisonnant à partir de faits et règles connus pour répondre à des questions précises, et (2) l'apprentissage automatique, reposant sur des algorithmes et modèles statistiques que les systèmes informatiques utilisent afin d'effectuer une tâche spécifique sans utiliser des instructions explicites mais en apprenant une fonction de prédiction à partir de textes bruts ou annotés.

Ces ensembles de textes sont souvent rassemblés dans un but précis et forment des corpus. Nous distinguons usuellement deux types de corpus : ceux du domaine général et les corpus spécialisés. Les corpus du domaine général incluent des textes de langue générale qui peuvent être de différents genres, tels que le **Open American National Corpus** [1] et le **British National Corpus** [2] ou bien d'un seul genre tels que le jeu de données **20 Newsgroups** [3] (articles de presse) et le **Large Movie Review Dataset** [4] (critiques de films). À l'inverse, les corpus spécialisés contiennent des textes en langue de spécialité d'un ou plusieurs genres, tels que le corpus **GENIA** [5] qui contient 1 999 résumés issus de **MEDLINE** [6], les corpus de données cliniques des campagnes d'évaluation **i2b2** [7] et **n2c2** [8] ou bien le **ACL Anthology Reference Corpus** [9] qui est constitué de 22 878 publications scientifiques de linguistique informatique et de **TALN**.

Le traitement automatique de la langue biomédicale ou **TAL biomédical** peut être considéré comme un sous-domaine du **TALN** où les textes traités comportent

1. http://www.anc.org/
2. http://www.natcorp.ox.ac.uk/
3. http://qwone.com/~jason/20Newsgroups/
4. https://ai.stanford.edu/~amaas/data/sentiment/
5. http://www.geniaproject.org/genia-corpus
6. https://www.nlm.nih.gov/bsd/medline.html
7. https://www.i2b2.org/index.html
8. https://n2c2.dbmi.hms.harvard.edu/
9. https://acl-arc.comp.nus.edu.sg/

les informations et la terminologie biomédicales. Dans un premier temps, nous décrivons brièvement les spécificités des textes biomédicaux, puis nous présentons plusieurs outils pionniers du **TAL biomécal**.

Textes biomédicaux

Qu'il s'agisse d'une langue de spécialité (CABRÉ, 1998), d'une langue spécialisée (LERAT, 1995) ou d'un langage technique (CHARNOCK, 1999), selon les différentes dénominations proposées en linguistique, la langue biomédicale permet à ses initiés de véhiculer de l'information spécialisée entre eux de façon concise et sans ambiguïté. Les termes du vocabulaire biomédical sont donc pour la plupart monosémiques. Les non-initiés ne maîtrisent pas le vocabulaire spécialisé ou le maîtrisent seulement partiellement. De plus, les patrons lexico-syntaxiques spécifiques à ces textes rendent la lecture encore plus difficile. Les personnes non initiées auront donc une compréhension très limitée des informations transmises par les textes biomédicaux. Par conséquent, afin de rendre ces textes compréhensibles pour tout le monde, plusieurs approches ont été proposées pour la simplification automatique des textes biomédicaux ces dix dernières années (KANDULA, CURTIS et ZENG-TREITLER, 2010 ; PENG, TUDOR et al., 2012 ; ABRAHAMSSON et al., 2014 ; KOPTIENT, CARDON et GRABAR, 2019 ; VAN DEN BERCKEN, SIPS et LOFI, 2019). Les corpus de textes biomédicaux sont le plus souvent composés d'articles de revues scientifiques ou de textes cliniques. Comme nous voyons, il existe de différents types de textes biomédicaux (textes cliniques, textes scientifiques, textes pour les spécialistes, textes pour les patients, etc.). Pour décrire ces différents textes, et en particulier, les textes scientifiques, nous nous appuyons sur Kevin Bretonnel COHEN et DEMNER-FUSHMAN, 2014 ainsi que sur les références incluses dans les deux paragraphes suivants.

Les articles de revues scientifiques sont généralement séparés en un résumé et un corps d'article. Ces parties partagent certaines informations. Cependant, la structure et le contenu des résumés et des corps d'articles sont différents. Les corps d'articles contiennent des phrases plus longues que les résumés et utilisent beaucoup plus de texte entre parenthèses. Le texte entre parenthèses présente notamment des difficultés pour l'analyse syntaxique des phrases et pour la détection de la portée de la négation et de l'incertitude (K Bretonnel COHEN, CHRISTIANSEN et HUNTER, 2011). Cependant, il est utile pour plusieurs autres tâches tels que la désambiguïsation d'abréviations, de noms de gènes, de symboles, etc. D'autre part, la voix passive et la négation sont plus fréquemment utilisées dans les corps d'articles, alors que les anaphores pronominales sont plus utilisées dans les résumés. De manière plus ciblée, les mutations sont mentionnées beaucoup plus fréquemment dans les corps d'articles que dans les résumés, tandis que les médicaments et les maladies sont mentionnés un peu plus fréquemment dans les résumés que dans les corps d'articles. En conséquence, les outils de **TALN** performent différemment sur ces textes. Enfin, les corps d'articles du domaine biomédical sont généralement structurés en sections plus ou moins standardisées (il s'agit typiquement d'introduction, de matériel et méthodes, de résultats, de discussion et de conclusion) dont les contenus diffèrent grandement. Par exemple, les sections *matériel et méthodes* semblent être la source de nombreux faux positifs (TANABE et WILBUR, 2002 ; CAMON et al., 2005)

pour les outils de **TALN**. Par conséquent, ces sections ont souvent été ignorées lors de l'évaluation des systèmes d'extraction d'informations, bien que les informations contenues dans ces sections semblent cruciales pour l'interprétation des expériences et de leurs résultats. L'attention principale est donc portée aux résumés d'articles. Cependant, si les résumés décrivent de manière synthétique les articles qu'ils représentent, notamment à l'aide de nombreux mots clés, ce sont les corps d'articles qui contiennent les informations les plus complètes et pertinentes (SHAH et al., 2003). Par conséquent, le traitement des corps d'articles fait l'objet de nombreux articles de **TAL biomédical** (CZARNECKI et al., 2012 ; JAIN et al., 2016 ; TRIEU et al., 2019).

Les documents cliniques tels que le *compte-rendu d'hospitalisation*, le *compte-rendu opératoire* ou le *dossier de soins infirmier* sont également structurés en section. Cependant, la structure employée pour chaque type de documents peut être différente et varier pour un même type de document selon l'hôpital, le professionnel de santé et la date de rédaction. La segmentation de textes cliniques en sections relatives aux informations présentées est un pré-requis pour de nombreuses tâches de **TAL biomédical**. Par exemple, afin de déterminer les problèmes de santé actuels d'un patient, il est nécessaire de faire la différence entre la section *antécédents médicaux* et la section *diagnostic*. Plusieurs travaux de recherche tels que P. S. CHO, TAIRA et KANGARLOO, 2003 ; APOSTOLOVA, CHANNIN et al., 2009 ; TEPPER et al., 2012 ; OROSZ, NOVÁK et PRÓSZÉKY, 2013 sont dédiés à cette tâche. D'autre part, les textes cliniques sont généralement rédigés de manière pragmatique. En conséquence, les phrases nominales, où ni le sujet ni le verbe ne sont indiqués, telles que *Pas de fièvre.*, *Abdomen souple, dépressible.* ou *Tension artérielle à X/X mmHg, fréquence cardiaque à X battements/min, saturation en air ambiant à X %.* sont très fréquemment utilisées. Compte tenue de l'utilité évidente de ces textes pour la recherche, ils sont largement étudiés dans les établissement hospitaliers, lors de campagnes d'évaluation (**i2b2**, **n2c2**, **CLEF eHealth** [10], etc.), et certains sont même en accès libre pour des travaux de recherche (**MIMIC** [11]). Les textes cliniques nécessitent l'application d'approches de **TALN** pour de nombreuses tâches telles que l'anonymisation ou dé-identification qui permet d'exploiter les textes cliniques tout en préservant le secret médical (STUBBS, KOTFILA et Ö. UZUNER, 2015), l'extraction de relations temporelles qui permet de remonter la piste d'effets secondaires (LIN, CHEN et R. A. BROWN, 2013 ; TOURILLE et al., 2017), la sélection de cohorte pour des essais cliniques (STUBBS, FILANNINO et al., 2019), etc.

D'autre part, depuis l'émergence de forums spécialisés en santé, tels que **Doctissimo** et **MedHelp** et de **Twitter**, les textes rédigés en langue biomédicale par des patients sont désormais largement étudiés, notamment à des fins de pharmacovigilance et d'épidémiologie (C. C. YANG et al., 2012 ; SARKER et al., 2015 ; SINNENBERG et al., 2017 ; BIGEARD, GRABAR et THIESSARD, 2018). BIGEARD, 2019 décrit les textes issus des forums **Doctissimo** comme limités dans leur contenu, c'est-à-dire que les informations nécessaires pour détecter le phénomène recherché (non-adhérence médicamenteuse dans le cas de ce travail de recherche) ne sont pas toujours disponibles, et de qualité orthographique variable. En effet, sur les forums et réseaux

10. https://clefehealth.imag.fr/
11. https://mimic.physionet.org/

sociaux, les internautes ont tendance à abréger les termes en écartant les voyelles ou en s'appuyant sur la phonétique des nombres et des lettres. La normalisation de ces abréviations pourrait sembler facile, cependant, la grande variété de formes abrégées pouvant exister pour chaque mot ainsi que l'apparition de nouvelles abréviations à travers le temps rendent la tâche plus difficile. Par exemple, *tomorrow* est rencontré sous de nombreuses formes : *2mo*, *2mrw*, *2moro*, *2morro*, *2morow*, *2morrow*, etc. Ainsi, l'étude d'un échantillon de *tweets* trouvait plus de 4 millions de mots hors vocabulaire créées volontairement et accidentellement par les utilisateurs (F. LIU et al., 2011). Si ces contenus générés par les internautes contiennent des informations intéressantes et présentent un potentiel pour la recherche, leur analyse requiert des méthodes et ressources appropriées.

À l'instar des corpus du domaine général, les corpus de textes biomédicaux sont généralement annotés dans l'optique de résoudre différentes tâches de **TALN**. Ainsi, dans le corpus **GENIA**, les coréférences, évènements biomédicaux, étiquettes morpho-syntaxiques, relations ainsi que les termes du vocabulaire biomédical sont annotés. Dans le corpus **QUAERO Médical du français** [12] (NÉVÉOL, GROUIN et al., 2014), ce sont dix types d'entités cliniques qui sont annotés selon les définitions des groupes sémantiques de l'**UMLS** (BODENREIDER et MCCRAY, 2003) : *Anatomy*, *Chemical* and *Drugs*, *Devices*, *Disorders*, *Geographic Areas*, *Living Beings*, *Objects*, *Phenomena*, *Physiology*, *Procedures*.

Approches de TALN pour le biomédical

Dans la section précédente, nous avons décrit les différents types de textes biomédicaux pour lesquels l'application d'approches de **TALN** permet notamment l'extraction d'informations pertinentes pour les chercheurs du domaine biomédical. Nous avons aussi évoqué plusieurs tâches (anonymisation, simplification, etc.) et activités (épidémiologie, pharmacovigilance, etc.) du domaine biomédical pour lesquelles les approches de **TALN** semblent bénéfiques.

Les premiers outils de **TAL biomédical** sont proposés sous la forme de systèmes à base de règles s'appuyant notamment sur des ressources termino-ontologiques importantes du domaine biomédical. C'est le cas du système de reconnaissance d'entités nommées de FUKUDA et al., 1998 dédié à la détection de mentions de gènes et protéines dans des résumés d'articles récupérés sur **MEDLINE**. C'est aussi le cas de **NegEx** (CHAPMAN et al., 2001), système pionnier dédié à la détection de la négation dans les textes cliniques, et de **MetaMap** [13], programme permettant notamment d'identifier les concepts de l'**UMLS** dans les textes en langue biomédicale.

D'autres programmes, tels que **cTAKES** [14] (SAVOVA et al., 2010) ou **CLAMP** [15] (SOYSAL et al., 2018), reposent, selon la tâche et/ou le choix de l'utilisateur, sur des approches soit à base de règles soit par apprentissage automatique pour annoter les

12. https://quaerofrenchmed.limsi.fr/
13. https://metamap.nlm.nih.gov/
14. https://ctakes.apache.org/
15. https://clamp.uth.edu/

textes cliniques (identification des sections, étiquetage morpho-syntaxique, détection de la négation et de l'incertitude, etc.) et en extraire une multitude d'informations (entités nommées, statut de fumeur des patients, etc.).

Ces dernières années, les systèmes experts, toujours largement utilisés dans le domaine biomédical, cèdent progressivement la place aux systèmes reposant sur l'apprentissage artificiel en raison de la disponibilité de jeux de données annotés volumineux et des avancées du **TALN**. Ainsi, dans le cadre de la campagne d'évaluation de la *Conference on Computational Natural Language Learning* de 2010 (**CoNLL-2010**) (FARKAS et al., 2010), ce sont les approches par machine à vecteurs de support (**SVM**) et champs aléatoires conditionnels (**CRF**) qui sont les plus convaincantes pour la détection de l'incertitude dans les textes biomédicaux. Lors de la campagne **i2b2** de 2012 (W. SUN, RUMSHISKY et O. UZUNER, 2013), **CRF** et **SVM**, souvent en combinaison avec des systèmes à base de règles, obtiennent de bons résultats pour les tâches de détection d'évènements clinique, d'expressions temporelles et de classification des relations temporelles. Lors de la campagne **i2b2** de 2014 (STUBBS, KOTFILA et Ö. UZUNER, 2015), les systèmes d'anonymisation les plus performants combinaient **CRF** et règles expertes. Lors de la campagne **SemEval-2016**, c'est une approche par **SVM** exploitant des descripteurs lexicaux, syntaxiques et morphologiques qui est la plus performante pour l'extraction d'informations temporelles dans des données cliniques (BETHARD et al., 2016). Le système de reconnaissance d'entités nommées biomédicales de HABIBI et al., 2017 combine plongements de mots (**word2vec**) et réseau de neurones récurrents (**LSTM-CRF**) et obtient les meilleurs résultats sur plusieurs jeux de données du domaine. Lors de la campagne d'évaluation **n2c2** de 2018, ce sont les réseaux de neurones récurrents ainsi que les machines à vecteurs de support (**SVM**) qui obtiennent les meilleurs résultats pour les tâches relatives à l'extraction des évènements indésirables liés aux médicaments (HENRY et al., 2020). Cependant, lorsque les données manquent, comme dans le cas de la tâche de sélection de cohorte pour les essais cliniques (seulement 288 patients), les approches à base de règles sont privilégiées (STUBBS, FILANNINO et al., 2019). Enfin, ces dernières années, les systèmes les plus performants proposés pour la classification automatique multi-étiquette de textes cliniques lors des campagnes d'évaluation de **CLEF eHealth** adoptent des approches neuronales *sequence-to-sequence* (NÉVÉOL, ROBERT, GRIPPO et al., 2018) et *transformer* (KELLY et al., 2019).

1.1.2 Classification automatique de textes

La classification automatique de textes a pour objectif d'attribuer une ou plusieurs classes à chaque texte en fonction de son contenu. Il s'agit d'une tâche fondamentale du **TALN** avec de vastes applications, telles que la catégorisation de documents et l'*opinion mining*, deux tâches largement étudiées en raison de la disponibilité de nombreux jeux de données annotés (**AG News corpus** [16], **DBpedia ontology dataset** [17], **IMDb dataset** [18], **Stanford Sentiment Treebank** [19], etc.). Ces deux

16. http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
17. https://wiki.dbpedia.org/Datasets
18. http://ai.stanford.edu/~amaas/data/sentiment/
19. https://nlp.stanford.edu/sentiment/treebank.html

tâches consistent généralement à attribuer une classe à chaque document. Cependant, certaines tâches consistent à attribuer une ou plusieurs classes à chaque document (classification multi-étiquette, *multi-label classification*). C'est le cas du *Toxic Comment Classification Challenge* [20] qui consiste à attribuer à chaque commentaire les différents comportements toxiques exprimés (*toxic*, *severe toxic*, *obscene*, *threat*, *insult* et *identity hate*). Un commentaire peut en effet être toxique et insultant à la fois. Dans ce qui suit, nous présentons les approches proposées pour la classification automatique de textes.

Approches de TALN pour la classification de textes

De nombreuses approches ont été élaborées pour la classification de textes. En effet, de nombreux classifieurs à base d'arbres de décision, de règles expertes, de **SVM**, des k plus proches voisins (**KNN**), de réseaux de neurones et bayésiens, parfois optimisés à l'aide de méta-algorithmes de *boosting* tels qu'**AdaBoost** (FREUND et SCHAPIRE, 1995) ou **BoosTexter** (SCHAPIRE et SINGER, 2000), ont été proposés. D'autre part, avant toute classification, l'une des tâches les plus fondamentales à accomplir est celle de la représentation des documents et de la sélection des descripteurs. Cette tâche est particulièrement importante pour la classification de textes en raison de la grande dimensionnalité des descripteurs et de l'existence de descripteurs non pertinents. Les textes peuvent être représentés de différentes façons. La représentation par sac de mots, très répandue en recherche d'information, consiste à représenter un document par un ensemble de mots ainsi que leurs fréquences dans chaque document, et ce indépendamment de l'ordre réel des mots dans le document. Il est aussi possible de représenter le texte directement sous forme de chaînes de caractères, où chaque document est une séquence de mots. Des méthodes de prétraitement, telles que la suppression des mots vides ainsi que la racinisation, ont longtemps été utilisées afin de sélectionner les descripteurs avant la classification. L'indice de Gini, l'entropie de Shannon, l'information mutuelle, le test du χ^2, l'indexation sémantique latente et bien d'autre méthodes de sélection de descripteurs ont aussi été largement utilisées pour la sélection de descripteurs avant la classification. Aujourd'hui, la plupart des approches proposées pour la classification de textes reposent sur les réseaux de neurones à la fois pour générer la représentation vectorielle des mots (plongements de mots) et pour la classification. Le domaine de la classification de texte est si vaste qu'il est difficile de couvrir l'ensemble des approches proposées dans ce manuscrit. Par conséquent, dans ce qui suit, nous présentons les approches les plus performantes proposées à ce jour sur plusieurs jeux de données.

R. JOHNSON et T. ZHANG, 2016 proposent plusieurs approches pour la catégorisation de documents et l'*opinion mining*. Celles-ci consistent à entraîner des plongements de régions textuelles (*text region embeddings*) en utilisant des couches **LSTM** et **CNN**. Les meilleurs résultats sont obtenus par la combinaison de multiples plongements par **CNN** et **LSTM** : 94,1 % d'*accuracy* sur **IMDB** ainsi qu'un taux d'erreurs

20. `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`

de 2,90 % sur **Yelp** (classification binaire), de 0,62 % sur **DBpedia** et de 6,57 % sur le corpus **AG News**.

ULMFiT (HOWARD et RUDER, 2018), pour *Universal Language Model Fine-tuning*, est une méthode d'apprentissage par transfert (*transfer learning*) qui peut être appliquée à n'importe quelle tâche de **TALN**. Les auteurs proposent aussi plusieurs techniques pour le réglage fin (*fine-tuning*) de modèles de langage (*language models*). **ULMFiT** comprend trois étapes. Dans un premier temps, le modèle de langage est entraîné sur un corpus généraliste afin de capturer des descripteurs généralistes dans différentes couches. Ensuite, ce modèle est affiné sur les données de la tâche cible en utilisant le *fine-tuning* discriminant (**Discr**) et un taux d'apprentissage dynamique (augmentation linéaire au début qui décline linéairement ensuite, *slanted triangular learning rates*, **STLR**) afin d'apprendre les descripteurs spécifiques à la tâche cible. Enfin, le classifieur est affiné sur la tâche cible en utilisant le dégel progressif (*gradual unfreezing*), le **Discr** et le **STLR**, afin de préserver les représentations de bas-niveau et adapter celles de haut niveau. Cette approche obtient de bons résultats sur plusieurs jeux de données : 95,4 % d'*accuracy* sur **IMDB** ainsi qu'un taux d'erreurs de 2,16 % sur **Yelp** (classification binaire), de 0,80 % sur **DBpedia** et de 5,01 % sur le corpus **AG News**.

Basée sur **BERT**, l'approche de C. SUN et al., 2019 consiste à adapter **BERT** à la tâche cible en appliquant plusieurs méthodes de *fine-tuning*. Empiriquement, les auteurs déterminent : la couche de **BERT** la plus pertinente pour la tâche cible, les parties des textes à tronquer, le taux d'apprentissage nécessaire afin d'éviter le problème de l'oubli catastrophique (*catastrophic forgetting*), etc. De plus, plusieurs méthodes de pré-entraînement supplémentaires sont expérimentées afin d'obtenir des modèles pré-entraînés spécifiques aux tâches, spécifiques aux domaines ou de domaines croisés. Des modèles multi-tâches sont aussi obtenus par *fine-tuning*. Si la plupart de ces modèles obtiennent de bon résultats sur plusieurs tâches et différents jeux de données, les meilleurs modèles obtiennent : 95,79 % d'*accuracy* sur **IMDB** et un taux d'erreurs de 1,81 % sur **Yelp** (classification binaire). D'autre part, **BERT** (DEVLIN et al., 2019) obtient un taux d'erreurs 0,64 % sur **DBpedia**.

XLNet (Zhilin YANG et al., 2019) est une méthode de pré-entraînement autorégressif généralisé qui permet d'apprendre des contextes bidirectionnels en maximisant la probabilité attendue sur toutes les permutations de l'ordre de factorisation. Cette méthode dépasse les limites de **BERT** grâce à sa formulation autorégressive. Actuellement, cette approche obtient les meilleurs résultats sur de nombreux jeux de données et tâches : 96,21 % d'*accuracy* sur **IMDB** ainsi qu'un taux d'erreurs de 1,55 % sur **Yelp** (classification binaire), de 0,62 % sur **DBpedia** et de 4,49 % sur le corpus **AG News**.

Les approches que nous venons de présenter sont dédiées à la classification binaire ou multi-classe de textes. En 1.2, nous présentons la tâche de classification à laquelle nous nous intéressons plus spécifiquement dans ce manuscrit. Il s'agit d'une tâche de classification multi-étiquette de textes cliniques. Dans ce qui suit, nous présentons plusieurs approches récemment proposées pour la classification multi-étiquette de textes.

CHALKIDIS et al., 2019 mettent à disposition un nouveau jeu de données multi-étiquette composé de documents législatifs de l'Union européenne en anglais (57 000 documents, 4 271 classes). Plusieurs approches par réseaux de neurones (Zichao YANG et al., 2016 ; WYDMUCH et al., 2018 ; RIOS et KAVULURU, 2018, **BERT**, etc.) sont proposées pour la classification et les résultats sont donnés selon plusieurs critères : pour toutes les classes, pour les classes les plus fréquentes, pour les classes faiblement présentes et pour les classes sans exemple dans le jeu de données d'entraînement (*zero-shot learning*). Selon les trois premiers critères, **BERT** est de loin l'approche la plus efficace, cependant les résultats qu'il montre en *zero-shot learning* sont extrêmement faibles. Dans ce contexte, le **ZERO-BIGRU-LWAN** (Zero Bidirectionnal Gated Reccurent Unit Label-Wise Attention Network), un modèle spécifiquement entraîné pour le *zero-shot learning*, obtient des résultats très largement supérieurs à ceux des autres approches.

CARTA et al., 2019 réalisent deux ensembles d'expérimentations dans le but de comparer plusieurs approches de plongements de mots (**Glove**, **fastText**, **word2vec**, etc.) pour la classification multi-étiquette de commentaires toxiques. Ainsi, un classifieur par régression logistique est entraîné sur les données du *Toxic Comment Classification Challenge* (1) en combinaison de chaque modèle de plongements lexicaux conformément aux jeux de données et (2) en validation croisée à 10 plis. Les résultats montrent que l'utilisation de plongements lexicaux entraînés sur les données du domaine surpassent toujours les modèles pré-entraînés généralistes. Par eilleurs, une baisse de résultats est observée en validation croisée.

1.1.3 Étiquetage de séquences

Dans le domaine du **TALN**, le terme étiquetage de séquences (*sequence labeling*) rassemble les tâches de reconnaissance de motifs et consiste à attribuer une étiquette catégorielle à chaque *token* (caractère, mot, ponctuation, etc.) d'une séquence. Parmi ces tâches, l'étiquetage morpho-syntaxique (*part-of-speech tagging*, **POS tagging**), qui consiste à attribuer à chaque token d'une phrase sa nature (catégorie grammaticale, ponctuation, etc.), ainsi que la reconnaissance d'entités nommées (*named entity recognition*, **NER**), qui consiste à détecter les entités nommées (nom propres et sigles relatifs à des personnes, lieux, entreprises, produits, etc.) dans les textes non-structurés et à les classer selon des classes pré-déterminées, sont sans doute celles qui ont reçu le plus d'attention de la part des chercheurs. Ces deux tâches sont représentatives des deux sous-ensembles de tâches d'étiquetage de séquences, d'étiquetage de *tokens* et d'étiquetage d'entités/portées. En effet, l'étiquetage morpho-syntaxique consiste à attribuer une classe à chaque *token* (*Il* est un pronom personnel, *n'* un adverbe, etc. (exemple 1 plus bas)), tandis que les entités nommées et portées sont souvent composées de plusieurs *tokens* et étiquetées en **BIO** (*begin*, *inside*, *outside*) dans la plupart des corpus (comme par exemple **B-ORG** : *begin organisation* ou **I-ORG** : *inside organisation* (exemple 2 plus bas)) :

1. Il/PRO :PER, n'/ADV, y/PRO :PER, a/VER :pres, pas/ADV, de/PRP, fièvre/NOM, ./SENT
2. Centre/B-ORG, Hospitalier/I-ORG, Universitaire/I-ORG, de/I-ORG,

Rennes/I-ORG, ./O

Dans ce qui suit, nous présentons une selection d'approches de **TALN** proposées pour l'étiquetage morpho-syntaxique ainsi que la **NER**.

Approches de TALN pour l'étiquetage de séquences

Les premiers systèmes de **TALN** pour l'étiquetage morpho-syntaxique sont proposés dès les années 1960. Par exemple, KLEIN et SIMMONS, 1963 proposent leur *Computational Grammar Coder*, une approche mixte, combinant un dictionnaire de moins de 2 000 entrées et un système de règles permettant de reconnaître la structure grammaticale de la phrase et donc d'annoter les mots hors dictionnaires. Testé sur plusieurs pages de textes scientifiques, ce système atteint plus de 90 % de mots correctement étiquetés. Cependant, les auteurs indiquent que l'ambiguïté structurelle dans les langues naturelles est un problème majeur pour le développement de systèmes de **TALN**. Dans les années 1990, Eric Brill (BRILL, 1992 ; BRILL, 1994) propose un système à base de règles qui acquiert automatiquement ses règles et étiquettes avec une *accuracy* (exactitude) comparable à celle des systèmes par apprentissage artificiel qui, dès les années 1980, sont privilégiés pour cette tâche.

Les premières approches par apprentissage artificiel pour le **POS tagging** sont proposées dans les années 1980. Ainsi, JELINEK, 1985 propose un système reposant sur les chaînes de Markov qui atteint une *accuracy* de 97 % sur le jeu de données de test. CUTTING et al., 1992 propose une approche basée sur un modèle de Markov caché (*hidden markov model*, **HMM**). Ce modèle est entraîné sur un lexique ainsi que des textes non-annotés. Il montre de bons résultats avec jusqu'à 96 % des tokens correctement étiquetés. D'autre part, SCHMID, 1994b propose **TreeTagger**, un système d'étiquetage probabiliste qui estime les probabilités de transition à partir d'un arbre de décision afin d'éviter les problèmes rencontrés par les systèmes basés sur les modèles de Markov lorsqu'ils doivent estimer les probabilités de transition à partir de données éparses. Ce système atteint 96,36 % d'*accuracy* sur le **Penn-Treebank**. SCHMID, 1994a propose aussi **Net-Tagger**, un système reposant sur les réseaux de neurones dont les performances sont comparables à celles de CUTTING et al., 1992. Dans un autre travail, LAFFERTY, MCCALLUM, PEREIRA et al., 2001 présentent et exploitent les champs aléatoires conditionnels (*conditional random fields*, **CRF**), un *framework* pour l'élaboration de modèles probabilistes, pour la segmentation et l'étiquetage des données séquentielles. Testées sur le **Penn-Treebank**, les performances des **CRF** sont supérieures à celles des **HMM**. Depuis 2002, les jeux de données (entraînement, validation, test) du **Penn-Treebank** sont standardisées selon le découpage de COLLINS, 2002. Parmi les approches les plus performantes, le système de X. SUN, 2014, qui atteint une *accuracy* de 97.36 %, repose sur une approche par **CRF** et régularisation de structure afin de contrôler le surentraînement structurel. L'approche la plus performante de Z. HUANG, W. XU et YU, 2015 repose sur un réseau de neurones récurrents *long short-term memory* bidirectionnel avec prédiction par **CRF** (BiLSTM-CRF) qui atteint une *accuracy* de 97.55 %. L'approche de MA et HOVY, 2016 repose sur un réseau de neurones **LSTM-CNNs-CRF** bidirectionnel

où le plongement de caractères pour chaque mot est calculé par un réseau de neurones convolutifs (**CNN**). Le vecteur résultant de cette opération est concaténé avec le plongement de mots avant d'être traité par le **BiLSTM**. Ce système atteint une *accuracy* de 97.55 %. **flair** (AKBIK, BLYTHE et VOLLGRAF, 2018), qui atteint une *accuracy* de 97.85 %, repose sur un **BiLSTM-CRF** ainsi que des plongements de chaînes de caractères contextualisés. Enfin, l'approche de BOHNET et al., 2018 repose à la fois sur des plongements de caractères et de mots contextualisés au niveau de la phrase obtenus à partir de **BiLSTM** suivis de perceptrons multicouches (**MLP**), ainsi que sur un **Meta-BiLSTM**. Dans le modèle **Meta-BiLSTM**, pour chaque mot, les plongements de caractères et de mots contextualisés sont concaténés puis traités par un **BiLSTM** afin de créer un plongement contextualisé supplémentaire. Ce plongement est ensuite traité par un **MLP** dont la sortie est traitée par une fonction d'activation linéaire afin de prédire l'étiquette. Ce système atteint une *accuracy* record de 97,96 %

Les premiers articles de recherche concernant la **NER** sont présentés dans les années 1990. Par exemple, RAU, 1991 propose un système à base de règles, de listes d'exceptions et d'une analyse de corpus approfondie pour la détection de noms d'entreprises et de leurs références ultérieures. Afin de détecter ces dernières, le système de règles génère les variations les plus probables de ces noms. Testé sur plus d'un million de mots provenant d'articles de presse financière, le système atteint une *accuracy* de 95 %.

Cependant, les approches par apprentissage artificiel, telles que les **HMM** (BIKEL et al., 1997), arbres de décision (SEKINE, 1998), **SVM** (ASAHARA et MATSUMOTO, 2003) et **CRF** (MCCALLUM et W. LI, 2003), sont rapidement adoptées. Le jeu de données proposé lors de la campagne d'évaluation de **CoNLL-2003** (SANG et DE MEULDER, 2003) est l'un des plus utilisé aujourd'hui. Parmi les approches les plus performantes sur ce jeu de données, **flair** obtient une F-mesure de 93,09 %. Le système de STRAKOVÁ, STRAKA et HAJIC, 2019 combine un BiLSTM-CRF enrichi avec les plongements contextualisés de **BERT** (*Bidirectional Encoder Representations from Transformers*) (DEVLIN et al., 2019), **ELMo** (PETERS et al., 2018) et **flair** afin d'obtenir une F-mesure de 93,38 %. L'approche de JIANG et al., 2019 combine un réseau de neurones récurrents et **I-DARTS** (*Improved Differentiable Architecture Search*), une version améliorée **DARTS** (H. LIU, SIMONYAN et Y. YANG, 2019). **DARTS** est une méthode de *neural architecture search* (**NAS**) visant à automatiser la conception de réseaux de neurones artificiels qui est basée sur le *relâchement* de la représentation de l'architecture neuronale, ce qui permet d'obtenir une **NAS** efficace en utilisant la descente de gradient. **I-DARTS** obtient une F-mesure de 93,47 %. Enfin, BAEVSKI et al., 2019 présente une nouvelle approche pour le pré-entraînement d'un *transformer* bidirectionnel. Ce modèle résout des tâches de type texte à trous, où chaque mot est effacé et doit être prédit à partir du reste du texte. Proposé par VASWANI et al., 2017, le *tranformer* est un modèle d'apprentissage profond basé uniquement sur des mécanismes d'attention. Les plongements de mots contextualisés produits par **BERT** sont, par exemple, issus de *tranformers* bidirectionnels. Cette approche obtient une F-mesure 93,5 %.

1.2 Classification multi-étiquette de textes cliniques

Depuis 2005, le Programme de Médicalisation des Systèmes d'Information (aussi connu comme le **PMSI**), à l'origine présenté comme un outil épidémiologique et de connaissance de l'activité des établissements de santé, est utilisé comme outil d'allocation budgétaire. En effet, le mode de financement des établissements de santé français est désormais basé sur la tarification à l'activité. Ce mode de financement, visant à réduire les inégalités de ressources entre les établissements de santé, consiste à les financer selon la nature et le volume de leurs activités. Par exemple, d'après les tarifs du 1er mars 2019 en Médecine Chirurgie Obstétrique [21], une séance de chimiothérapie pour tumeur est évaluée à 383,11 €.

Afin d'obtenir ce financement, les établissements de santé français doivent rapporter les codes issus de la Classification Internationale des Maladies, 10ème révision (**CIM-10**) correspondant aux actes médicaux, consultations et autres prestations réalisés. L'étiquetage manuel des documents cliniques étant à la fois complexe et chronophage, le développement de systèmes visant à étiqueter automatiquement ces documents intéresse fortement les établissements de santé. Dans ce contexte, dès 2007, le *Computational Medicine Challenge* (PESTIAN et al., 2007) était dédié à l'attribution de codes **CIM-9** à un ensemble de rapports de radiologie en anglais. RUCH et al., 2008 fut le premier travail de ce genre en français.

Dans la suite de cette section, nous décrivons La Classification Internationale des Maladies, les jeux de données annotés, ainsi que les systèmes les plus performants proposés pour la classification automatique de textes cliniques avec la **CIM-10**.

1.2.1 La Classification Internationale des Maladies

La Classification Internationale des Maladies, dont l'appellation complète est Classification Statistique Internationale des Maladies et des Problèmes de Santé Connexes, est une classification médicale hiérarchisée contenant plusieurs milliers de codes relatifs à de nombreux diagnostics et situations cliniques ou sociales. Cette classification est publiée par l'Organisation Mondiale de la Santé (**OMS**) et est utilisée par de nombreux pays Membres à travers le monde.

La **CIM-10** correspond à la 10ème révision de la classification. Elle est composée de 22 chapitres et 14 400 codes (jusqu'à 16 000 avec les sous-classifications facultatives). Actuellement utilisée par le système de santé français, elle permet d'obtenir une représentation structurée et standardisée de textes cliniques et sert notamment d'outil de connaissance de l'activité des établissements de santé. Cette classification est mise à jour quand jugée nécessaire. Par exemple, la mise à jour du 21/10/2019 a permis d'ajouter le code ***U07.0 Affection liée au vapotage*** [22]. Comme indiqué, la classification des codes est hiérarchique. Par exemple, le code ***A00 Choléra*** possède

21. https://www.atih.sante.fr/tarifs-mco-et-had
22. https://www.atih.sante.fr/cim-10-fr-2019-usage-pmsi

plusieurs enfants : ***A000 Choléra à Vibrio cholerae 01, biovar cholerae***, ***A001 Choléra à Vibrio cholerae 01, biovar El Tor***, ainsi que ***A009 Choléra, sans précision***, tel qu'illustré dans la figure 1.1.

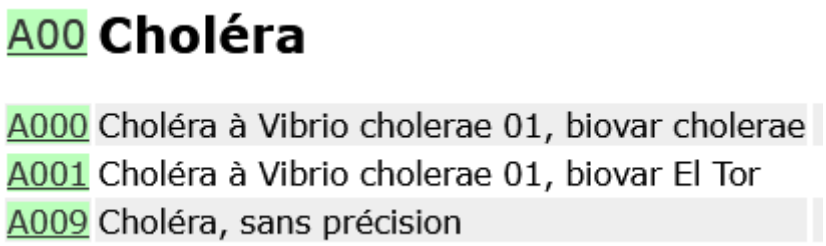


FIGURE 1.1 – Le code **CIM-10** ***A00 Choléra*** et ses enfants, capture d'écran du site : `https://www.aideaucodage.fr/cim-a00`

La **CIM-11**, publiée en mai 2019, rentrera en application à compter du 1er janvier 2022 [23]. Conçue pour être plus simple à utiliser que la **CIM-10**, la **CIM-11** adopte une infrastructure ontologique.

1.2.2 Jeux de données annotées

Par définition, le **PMSI** s'applique aux données hospitalières et cliniques protégées par le secret médical. Par conséquent, malgré l'existence de grands volumes de données annotées, il est très difficile d'y avoir accès et de les partager. Depuis 2013, le **CLEF eHealth Evaluation Lab** propose chaque année une ou plusieurs campagnes d'évaluation sur le thème de l'extraction d'information. Depuis 2016, l'une de ces campagnes d'évaluations consiste à confronter les méthodes de plusieurs équipes de recherche à l'étiquetage automatique de codes **CIM-10**. Les données étant sensibles, elles ne sont accessibles qu'après avoir rempli et signé un formulaire d'accord d'utilisation des données qui contraint le signataire, entre autres, à ne pas redistribuer ou divulguer le contenu du corpus à un tiers non autorisé. Dans cette section, nous décrivons les corpus de données cliniques utilisés lors des campagnes d'évaluation de 2016 à 2018 pour cette tâche. En 2019, la tâche impliquait l'annotation automatique de résumés non-technique d'expérimentations animales en allemand (KELLY et al., 2019). En 2020, la campagne d'évaluation se concentrera sur le codage automatique de textes cliniques en espagnol.

Corpus CépiDC. Utilisé lors de plusieurs compétitions (NÉVÉOL, K Bretonnel COHEN et al., 2016 ; NÉVÉOL, ROBERT, ANDERSON et al., 2017 ; NÉVÉOL, ROBERT, GRIPPO et al., 2018), le corpus **CépiDC** a été fourni par l'Institut national de la santé et de la recherche médicale (**Inserm**). Le corpus est composé de certificats de décès collectés auprès de médecins de ville et d'hôpitaux en France entre 2006 et 2015. Les certificats de décès sont des documents standardisés complétés par les médecins

23. `https://www.who.int/classifications/icd/en/`

pour rendre compte de la mort de patients. Conformément aux normes internationales de l'**OMS**, les certificats de décès sont composés de deux parties. La première partie présente les principaux évènements qui ont conduit au décès du patient. La seconde partie décrit les affections contributives qui ne sont pas directement impliquées dans le décès du patient. Les deux parties sont remplies manuellement par le médecin. Le codage **CIM-10** est, quant à lui, est effectué indépendamment du signalement par le médecin. Les certificats de décès sont stockés dans deux fichiers : un fichier de texte brut dans lequel le médecin écrit le rapport et un fichier contenant les codes correspondants. Le second fichier peut contenir du texte normalisé appuyant les décisions de codage. Ce fichier peut être utilisé pour la création de dictionnaire afin d'assister le codage. Les statistiques présentées dans le tableau 1.1 rendent compte de l'évolution du corpus **CépiDC** à travers les campagnes d'évaluation successives. Notons qu'entre 2016 et 2018, le nombre de certificats a augmenté d'environ 46 %. Par ailleurs, depuis 2017, le corpus **CépiDC** est disponible dans une version alignée. Dans cette version, exemplifiée dans le tableau 1.2, les lignes des certificats, le texte normalisé et les codes leurs correspondant sont alignés.

TABLEAU 1.1 – Statistiques du corpus **CépiDC** par année
(NÉVÉOL, K Bretonnel COHEN et al., 2016 ; NÉVÉOL, ROBERT, ANDERSON et al., 2017 ; NÉVÉOL, ROBERT, GRIPPO et al., 2018)

Année		Entraînement (2006–2012)	Test (2013)	
2016	Certificats	65 844	27 850	
	Lignes	195 204	80 899	
	Tokens	1 176 994	496 649	
	Codes CIM-10 (total)	266 808	110 869	
	Codes CIM-10 (Unique)	3 233	2 363	
	Codes CIM-10 non vu (Unique)	-	224	
		Entraînement (2006–2012)	Development (2013)	Test (2014)
2017	Certificats	65 844	27 850	31 690
	Lignes alignées	195 204	80 899	91 962
	Tokens	1 176 994	496 649	599 127
	Codes CIM-10 (total)	266,808	110,869	131,426
	Codes CIM-10 (Unique)	3,233	2,363	2,527
	Codes CIM-10 non vu (Unique)	-	224	266
		Entraînement (2006–2014)	Test (2015)	
2018	Certificats	125 384	11 931	
	Lignes	368 065	34 918	
	Tokens	1 250 232	84 091	
	Codes CIM-10 (total)	509 103	48 948	
	Codes CIM-10 (Unique)	3 723	1 806	
	Codes CIM-10 non vu (Unique)	-	70	

TABLEAU 1.2 – Exemple de document du corpus de certificats de décès **CépiDC aligné** (NÉVÉOL, ROBERT, GRIPPO et al., 2018)

texte	texte normalisé	codes CIM-10
choc septique	choc septique	A41.9
peritonite stercorale sur perforation colique	peritonite stercorale	K65.9
peritonite stercorale sur perforation colique	perforation colique	K63.1
Syndrome de détresse respiratoire aiguë	syndrome détresse respiratoire aiguë	J80.0
defaillance multivicerale	défaillance multiviscérale	R57.9
HTA	hta	I10.0

Corpus CDC. Utilisé lors d'une campagne d'évaluation (NÉVÉOL, ROBERT, ANDERSON et al., 2017), le corpus **CDC** a été fourni par les Centres pour le contrôle et la prévention des maladies (***Centers for Disease Control and Prevention***, **CDC**). Il se compose de certificats de décès en texte libre collectés aux États-Unis au cours de l'année 2015. Tous les décès rapportés dans ce corpus sont dus à des causes naturelles, c'est-à-dire qu'aucun décès lié à une blessure n'y est inclus. Le tableau 1.3 présente les statistiques liées à ce corpus. Notons que **CDC** contient bien moins de certificats et de codes **CIM-10** uniques que son équivalent français de 2017 mais que la moyenne de tokens par certificat est bien plus élevée : 52 tokens pour **CDC**, 6 pour **CépiDC**.

TABLEAU 1.3 – Statistiques du corpus **CDC** (NÉVÉOL, ROBERT, ANDERSON et al., 2017)

	Entraînement (2015)	Test (2015)
Certificats	13 330	6 665
Lignes	32 714	14 834
Tokens	990 442	42 819
Codes CIM-10 (total)	39 334	18 928
Codes CIM-10 (Unique)	1 256	900
Codes CIM-10 non vu (Unique)	-	157

Corpus KSH-HU. Utilisé lors d'une campagne d'évaluation (NÉVÉOL, ROBERT, GRIPPO et al., 2018), le corpus **KSH-HU** a été fourni par l'Office central de statistiques en Hongrie (*Hungarian Central Statistical Office*, **KSH**). Le corpus se compose d'un échantillon de certificats de décès en texte libre extraits au hasard et recueillis auprès de médecins en Hongrie pour l'année de décès 2016. Il n'y a pas de certificat informatisé dans ce pays : donc contrairement au corpus français, ce corpus ne contient que les décès signalés à l'aide de formulaires papier, puis transcrits informatiquement.

TABLEAU 1.4 – Statistiques du corpus **KSH-HU** (NÉVÉOL, ROBERT, GRIPPO et al., 2018)

	Entraînement (2016)	Test (2016)
Certificats	84 703	21 176
Lignes	324 266	81 291
Tokens	666 839	167 507
Codes CIM-10 (total)	392 020	98 264
Codes CIM-10 (Unique)	3 124	2 011
Codes CIM-10 non vu (Unique)	-	202

Corpus ISTAT-IT. Utilisé lors d'une campagne d'évaluation (NÉVÉOL, ROBERT, GRIPPO et al., 2018), le corpus **ISTAT-IT** a été fourni par l'Institut national italien de statistique (*Istituto nazionale di statistica*, Istat, **ISTAT**). Afin de proposer un corpus

réaliste tout en préservant la confidentialité, le corpus est composé de faux certificats créés à partir de certificats de décès authentiques correspondant à différentes années de codage. En effet, chaque ligne d'un faux certificat provient d'un certificat réel différent, tout en assurant une cohérence thématique et en préservant la chaîne des causes de décès. La cohérence de l'âge et du sexe a également été préservée. Les certificats synthétiques ont ensuite été codés comme s'ils signalaient de véritables décès pour 2016.

TABLEAU 1.5 – Statistiques du corpus **ISTAT-IT**
(NÉVÉOL, ROBERT, GRIPPO et al., 2018)

	Entraînement (2016)	Test (2016)
Certificats	14 502	3 618
Lignes	49 825	12 602
Tokens	666 839	167 507
Codes CIM-10 (total)	60 955	15 789
Codes CIM-10 (Unique)	1 443	903
Codes CIM-10 non vu (Unique)	-	100

1.2.3 Systèmes de classification

Dans cette section, nous décrivons les systèmes les plus performants proposés par les équipes de recherche lors des campagnes d'évaluation de **CLEF eHealth** de 2016 à 2018, ainsi que les résultats obtenus.

Les systèmes de classification sont évalués à l'aide des trois métriques suivantes : la précision, qui quantifie la pertinence de l'étiquetage, le rappel, qui quantifie la sensibilité de l'étiquetage, ainsi que la moyenne harmonique de la précision et du rappel notée F-mesure ou F_1.

$$Précision = \frac{Vrai\ positif}{Vrai\ positif + Faux\ positif}$$

$$Rappel = \frac{Vrai\ positif}{Vrai\ positif + Faux\ négatif}$$

$$F - mesure = 2 \cdot \frac{Précision \cdot Rappel}{Précision + Rappel}$$

CLEF eHealth Evaluation Lab 2016

Lors de l'édition 2016, une équipe se distingue des autres par ses résultats. En effet, l'équipe **Erasmus MC** (VAN MULLIGEN et al., 2016) obtient des résultats qui se positionnent largement au-dessus de ceux des autres équipes (voir tableau 1.6).

Équipe Erasmus MC. Cette équipe propose une approche basée sur les dictionnaires. En effet, deux terminologies de la CIM-10 sont construites sur la base des données d'entraînement fournies par les organisateurs. Le ***Solr text tagger*** est utilisé avec ces terminologies pour indexer les certificats de décès et générer des codes. Un filtrage de score de précision est appliqué pour améliorer la précision finale.

Équipe ERIC-ECSTRA. Arrivée deuxième, cette équipe propose deux approches. La première approche est basée sur l'allocation de Dirichlet latente, méthode probabiliste de la famille des *topic models*. Cette méthode s'appuie sur les structures thématiques latentes afin de prédire les codes **CIM-10** correspondants. La seconde approche est basée sur un classifieur **SVM** avec une représentation des textes par sacs de mots.

TABLEAU 1.6 – 4 meilleurs résultats sur le corpus CépiDC 2016 (NÉVÉOL, K Bretonnel COHEN et al., 2016)

Corpus	Équipe	Précision	Rappel	F-mesure
CépiDC	Erasmus-run2	88,6	**81,3**	**84,8**
	Erasmus-run1	**89,0**	80,3	84,4
	ERIC-ECSTRA-run2	88,2	65,5	75,2
	ERIC-ECSTRA-run1	81,1	61,5	70,0

CLEF eHealth Evaluation Lab 2017

Lors de l'édition 2017, deux équipes se distinguent des autres par leurs résultats (voir tableau 1.7). Les équipes **KFU** (MIFTAHUTDINOV et TUTUBALINA, 2017) et **LIMSI** (ZWEIGENBAUM et LAVERGNE, 2017).

Équipe KFU. Arrivée première sur le corpus **CDC**, cette équipe propose une méthode par apprentissage profond, *sequence-to-sequence*, basée sur les réseaux de neurones récurrents. En séquence d'entrée, le système prend le texte brut et donne en séquence de sortie les codes **CIM-10** prédits. Le corpus et le dictionnaire fournis ont été utilisés pour l'entraînement. Cette équipe ne propose pas de résultats sur les jeux de données **CépiDC**.

Équipe LIMSI. Arrivée première sur les jeux de données français et seconde sur le corpus **CDC**, cette équipe propose deux approches. La première approche utilise une méthode par apprentissage automatique supervisé (**SVM** multi-étiquette, avec unigrammes, bigrammes et année de codage pour le français). La seconde approche utilise une méthode hybride, par dictionnaire et **SVM** multi-étiquette.

Équipe SIBM. Arrivée seconde sur les jeux de données français, cette équipe propose une approche basée sur une extraction de termes effectuée au niveau de la

phrase. L'extraction se fait en trois étapes. Lors de la tokenisation, le texte est découpé en phrases, puis en mots. Les mots vides sont filtrés et la vérification orthographique est effectuée à l'aide de la bibliothèque **Enchant** [24]. Lors de la sélection des codes candidats, une méthode basée sur l'algorithme de codage phonétique **Double Metaphone** (PHILIPS, 2000) est utilisée pour la recherche approximative de termes. Les codes candidats sont alors classés par la combinaison de la plus longue sous-chaîne de caractères en commun et d'algorithmes de correspondance approximative (*fuzzy match*). Le candidat avec le score le plus élevé est conservé en tant que code **CIM-10** correspondant à la phrase. Avec le meilleur résultat officiel sur le corpus **CépiDC**, cette méthode obtient des résultats largement inférieurs à ceux de l'équipe **LIMSI**.

TABLEAU 1.7 – Résultats de la campagne d'évaluation de **CLEF eHealth 2017** (NÉVÉOL, ROBERT, ANDERSON et al., 2017)

Corpus	Équipe	Précision	Rappel	F-mesure
CDC	KFU-run1	89,3	81,1	**85,0**
	KFU-run2	89,1	**81,2**	**85,0**
	LIMSI-run2	**89,9**	80,1	84,7
CépiDC	LIMSI-run2	87,2	**78,4**	**82,5**
	LIMSI-run1	**88,3**	76,0	81,7
	SIBM-run1	85,7	68,9	76,4
CépiDC aligné	LIMSI-run2	85,4	**88,1**	**86,7**
	LIMSI-run1	**86,5**	86,5	86,5
	SIBM-run1	83,5	77,5	80,4

CLEF eHealth Evaluation Lab 2018

Lors de l'édition 2018, une équipe se distingue des autres par ses résultats. En effet, l'équipe **IxaMed** (ATUTXA et al., 2018) obtient les meilleurs résultats sur tous les corpus (voir tableau 1.8).

Équipe IxaMed. Comme **KFU** en 2017, cette équipe utilise une approche neuronale *sequence-to-sequence* qui, indépendamment de la langue, encode une séquence de tokens de taille variable (document, phrase) en une séquence de représentations vectorielles, et décode ces représentations en une séquence de tokens en sortie (codes **CIM-10**).

Équipe IAM-ISPED. COSSIN et al., 2018 utilisent une approche par dictionnaire afin d'assigner des codes **CIM-10** à chaque ligne. Pour chaque *phrase*, leur système utilise trois techniques de recherche : par correspondance parfaite, par correspondance d'abréviations et par distance de Levenshtein. La recherche par correspondance d'abréviations utilise un dictionnaire d'abréviations tandis que la distance de Levenshtein est utilisée pour détecter les fautes de frappe. Deux systèmes sont soumis : un système utilise un dictionnaire de 148 447 termes et 6 392 codes, l'autre un

24. `https://abiword.github.io/enchant/`

dictionnaire de 42 439 termes et 3 539 codes. Cette méthode obtient de bons résultats sur **CépiDC aligné**.

Équipe LSI-UNED. ALMAGRO et al., 2018 ont exploré deux approches alternatives très compétitives pour l'attribution de codes **CIM-10** aux certificats de décès. La première approche par apprentissage supervisé utilise **SVM** et réseau de neurones afin de profiter des données d'entraînement en générant des modèles en *One-Vs-Rest* (**OVR**) pour les codes **CIM-10** les plus fréquents. Cependant, étant donné que peu d'exemples sont disponibles pour représenter certains codes **CIM-10**, une méthode non-supervisée basée sur des techniques de recherche d'information est proposée afin d'obtenir un meilleur rappel.

TABLEAU 1.8 – Résultats de la campagne d'évaluation de **CLEF eHealth 2018** (NÉVÉOL, ROBERT, GRIPPO et al., 2018)

Corpus	Équipe	Précision	Rappel	F-mesure
CépiDC	IxaMed-run1	87,2	**59,7**	**70,9**
	IxaMed-run2	87,7	58,8	70,4
	LSI-UNED-run1	84,2	55,6	67,0
	LSI-UNED-run2	**87,9**	54,0	66,9
CépiDC aligné	IxaMed-run2	84,1	**83,5**	**83,8**
	IxaMed-run1	**84,6**	82,2	83,4
	IAM-run2	79,4	77,9	78,6
	IAM-run1	78,2	77,2	77,7
KSH-HU	IxaMed-run2	**97,0**	**95,5**	**96,3**
	IxaMed-run1	96,8	95,4	96,1
	LSI UNED-run2	94,6	91,1	92,8
	LSI UNED-run1	93,2	92,2	92,7
ISTAT-IT	IxaMed-run1	**96,0**	**94,5**	**95,2**
	IxaMed-run2	94,5	92,2	93,4
	LSI UNED-run1	91,7	87,5	89,5
	LSI UNED-run2	93,1	86,1	89,5

1.3 Détection de la négation et de l'incertitude

Le Dictionnaire de l'Académie française (9ème édition)[25] définit la négation comme l'énoncé qui rejette comme faux, qui déclare faux, une proposition ou un jugement, ou bien, du point de vue de la linguistique, comme la construction utilisée pour nier ou pour exprimer le refus. L'incertitude y est, quant à elle, définie comme le caractère de ce qui est incertain, douteux ou imprévisible. Ces définitions génériques n'indiquent aucune piste pour permettre de repérer ces phénomènes dans les textes, mais laissent supposer une grande variété de réalisations linguistiques.

Ainsi, en 1.3.1, nous décrivons succinctement la négation et l'incertitude du point de vue de la linguistique. Ensuite, en 1.3.2, nous présentons les corpus annotés avec la négation et/ou l'incertitude. Enfin, en 1.3.3, nous présentons les méthodes de détection automatique de la négation et de l'incertitude, par système expert et par apprentissage supervisé, proposées dans la littérature scientifique.

1.3.1 Négation et incertitude

La négation

Du point de vue de la linguistique, la négation est une opération qui établie comme fausse une proposition exprimée. Cette opération a été abordée dans de très nombreux ouvrages. En effet, SEIFERT et WELTE, 1987 recensait près de 3 200 livres et articles traitant de la négation dans de nombreuses langues. Dans MULLER, 1991, l'auteur examine toutes les formes de la négation en français, leur syntaxe, leur sémantique, et établit des comparaisons avec les autres langues romanes. HORN, 2001 propose une synthèse des travaux passés et actuels sur la structure, le sens et l'utilisation de la négation et des expressions négatives du point de vue de la philosophie, de la psychologie et de la linguistique.

Dans les textes, les opérations de négation sont construites à partir de marqueurs que l'on retrouve dans plusieurs catégories grammaticales. Les marqueurs de négation sont tous les mots, groupes de mots et unités morphologiques qui permettent d'exprimer la négation. Dans le tableau 1.9, nous présentons une liste non exhaustive de marqueurs de négation dans les langues avec lesquelles nous travaillons : l'anglais, le français ainsi que le portugais. Nous les classons selon leurs natures grammaticales, affixes compris. Nous constatons que la négation peut être exprimée par une grande variété de marqueurs. En outre, nous constatons que certains marqueurs sont communs à plusieurs langues. C'est notamment le cas pour les préfixes *im-* et *in-* dans les trois langues, pour l'adverbe *jamais* en français et portugais, ainsi que pour les marqueurs *non-* et *absent* en français et anglais. D'autre marqueurs sont spécifiques à chaque langue. C'est le cas du suffixe *-less* en anglais et de la particule *ne* en français, auxquels nous ne trouvons pas d'équivalents dans les deux autres langues. Par ailleurs, certains marqueurs multi-mots, tels que *ni[...]ni* et ses traductions *nem[...]nem* et *neither[...]nor*, fonctionnent de la même manière dans les trois

25. `https://www.dictionnaire-academie.fr/`

langues. En français, certains marqueurs présentent des ambiguïtés liées à l'homonymie et au contexte d'emploi des termes. Dans les exemples présentés ci-dessous, les marqueurs et homonymes sont soulignés.

Dans l'exemple (1), le nom commun *pas* ne doit pas être confondu avec son homonyme adverbe de négation. Dans l'exemple (2), la phrase contient soit l'adverbe *plus* comparatif de *beaucoup de/davantage de* ou l'adverbe de négation *plus*. La particule *ne* n'étant pas utilisée, la désambiguïsation lexicale du mot ne peut être faite que par le contexte ou les prononciations, *plus* (comparatif de *beaucoup de*) étant prononcé \plys\ et *plus* (négation) \ply\. Puisque nous travaillons uniquement avec du texte libre, la désambiguïsation doit donc être faite au niveau du contexte, c'est-à-dire des phrases précédentes ou suivantes. Dans l'exemple (3), le contexte donné dans la phrase précédente confirme l'utilisation de la négation dans la seconde. Dans l'exemple (4), à l'inverse, son homonyme est employé. En outre, certains marqueurs de négation, tels que *à part*, *à l'exception de*, *excepté* ou *en dehors de*, non seulement ne marquent pas lorsqu'ils sont utilisés après une négation, mais annulent aussi la négation qui les précèdent. Dans l'exemple (5), la séquence *de traitement curateur de la maladie* est d'abord niée par le marqueur ***ne... pas***, cependant cette information est elle-même infirmée par la suite de l'énoncé.

1. Elle n'est soulagée que par la marche et doit donc écouter la télévision en faisant les cent pas dans son salon.
2. Il veut plus d'antalgique.
3. Les douleurs ont disparu. Il veut plus d'antalgique.
4. Il a toujours mal. Il veut plus d'antalgique.
5. Il n'y a pas de traitement curateur de la maladie en dehors de l'allogreffe de moelle.

TABLEAU 1.9 – Marqueurs de négation en anglais, français et portugais

Anglais	Adjectifs	absent, impossible, negative, unable
	Adverbes	never, no, not, no longer
	Affixes	dis-, im-, in-, non-, un-, -less, -not, -n't
	Conjonctions	neither[...]nor, rather than
	Locutions	with the exception of
	Noms	absence of, lack of
	Prépositions	except, instead of, without
	Verbes	exclude, fail, lack, miss, rule out
Français	Adjectifs	absent, aucun, négatif, nul
	Adverbes	jamais, ne pas, ne[...]pas, non, pas
	Affixes	a-, dis-, im-, in-, non-,
	Conjonctions	ni, ni[...]ni, sans que
	Locutions	à la place de, à l'exception de
	Noms	absence de, carence,
	Prépositions	excepté, sans, sauf
	Pronoms	personne, rien
	Verbes	annuler, éliminer, exclure, manquer, nier
Portugais	Adjectifs	negativo, nenhum
	Adverbes	jamais, não, nunca, tampouco
	Affixes	a-, des-, im-, in-
	Conjonctions	nem, nem[...]nem
	Noms	ausência de, não
	Prépositions	exceto, sem
	Pronoms	nenhum, ninguém
	Verbes	negar

Bien que la détection des marqueurs de négation présente des difficultés en raison de l'ambiguïté et du contexte, il est bien plus difficile de déterminer leurs portées avec précision. La portée est l'effet de l'opération de négation sur la phrase. Résoudre le problème de la portée de la négation revient à répondre à la question : qu'est-ce qui est réellement nié ? La portée d'un marqueur de négation s'étend sur tout ou partie de la phrase (négation de phrase et négation de constituants). Dans les exemples suivants, les marqueurs sont soulignés et les portées sont en gras.

Dans l'exemple (1), la portée du marqueur s'étend à l'ensemble de la phrase. Dans le second exemple aussi, cependant, cela n'exclut pas que le patient en question a ou non le diabète. Ainsi, dans le cadre d'une application de **TAL biomédical**, il conviendra d'annoter la phrase différemment. Dans le troisième exemple, seule une partie spécifique de la phrase se trouve dans la portée de la négation. Dans l'exemple (4), la portée du marqueur est modifiée par l'adverbe de fréquence *toujours* : la chimiothérapie n'est pas exclue. Enfin, le dernier exemple présente deux instances de négation : la première en rouge, la seconde en bleu. Ici, la portée de la seconde instance est discontinue. En effet, si la proposition qui précède la première

instance n'est pas infirmé par celle-ci, la seconde instance l'infirme en raison de la coréférence entre *il* et cette proposition.

1. **Il** n'**y a** pas **de fièvre**.
2. Pas **d'antécédents familiaux de diabète**.
3. Le greffon rénal se recolore immédiatement mais **la diurèse** ne **reprend** pas **sur table**.
4. La chimiothérapie n'**est** pas **toujours possible**.
5. **Le retrait du matériel d'ostéosynthèse incriminé** n'**est** pas **systématique**, ce qui explique qu'**il** n'**ait** pas **été proposé** à notre patient asymptomatique.

Selon certains linguistes, il serait nécessaire de distinguer la portée de la négation de son foyer, c'est-à-dire de la partie de la portée la plus explicitement niée. Ainsi, dans le premier exemple, le foyer correspond au nom *fièvre*. Dans le cadre de la création d'un corpus annoté, il conviendra de déterminer les éléments à annoter (marqueurs, portées et/ou foyers) ainsi que les règles d'annotation à adopter en fonction des besoins relatifs au domaine d'application.

L'incertitude

Dans cette sous-section, nous clarifions notre compréhension du terme *incertitude*. De nombreux articles scientifiques (SMITHSON, 1989 ; KRAUSE et CLARK, 1993 ; BOUCHON-MEUNIER et H.-T. NGUYEN, 1996 ; SMETS et MOTRO, 1997 ; BRONNER, 1997 ; JOUSSELME, MAUPIN et BOSSÉ, 2003 ; FARKAS et al., 2010) en ont proposé leur classification du point de vue de la linguistique, de la sociologie, de la psychologie, etc.

Nous nous intéressons ici à l'incertitude linguistique telle qu'elle est décrite dans VINCZE, 2015. De façon générale, l'incertitude dénote un manque d'information, c'est-à-dire que le lecteur ne peut pas être sûr de la factualité d'une ou de plusieurs informations. L'incertitude de ces informations diffère donc de leur affirmation ou de leur infirmation. Les théories linguistiques associent généralement le concept logique de modalité épistémique à l'incertitude. En effet, c'est par modalité épistémique que « le locuteur exprime son degré de certitude sur ce qu'il asserte » (QUERLER, 1996). Ainsi, une proposition est incertaine si nos connaissances ne permettent pas de déterminer sa fiabilité en raison d'un manque d'information. En se basant sur les corpus annotés, VINCZE, 2015 indique que le terme *incertitude* peut être utilisé pour couvrir les phénomènes aux niveaux sémantique et discursif à la fois. Ainsi, les propositions sémantiquement incertaines peuvent être définies en termes de sémantique conditionnelle de vérité. Nous ne pouvons pas leur attribuer une valeur de vérité, c'est-à-dire que nous ne pouvons pas affirmer avec certitude qu'elles sont vraies ou fausses, compte tenu de l'état mental actuel du locuteur. Alors que certaines instances d'incertitude peuvent être vraies, fausses ou incertaines (incertitude hypothétique), d'autres instances sont définitivement incertaines (incertitude épistémique). Par exemple, dans la phrase *Il pleut peut-être.*, la proposition n'est ni vraie ni fausse et décrit un monde possible mais qui ne coïncide pas

nécessairement avec le monde réel du locuteur. Il est certain que la proposition est incertaine (incertitude épistémique). Au niveau du discours, l'incertitude s'exprime soit par le manque de sources ou de fiabilité des sources, la spéculation, l'imprécision et la subjectivité.

À l'instar de la négation, l'incertitude est exprimée par l'utilisation de marqueurs dont la portée s'étend sur tout ou partie des *tokens* de la phrase. Dans le tableau 1.10, nous présentons une liste non exhaustive des marqueurs d'incertitude dans les deux langues avec lesquelles nous travaillons : l'anglais et le français. L'incertitude est principalement exprimée par l'emploi de verbes, conjonctions, adjectifs et adverbes.

TABLEAU 1.10 – Marqueurs d'incertitude en anglais et français

Anglais	Adjectifs/Adverbes	problable, likely, possible, unsure, etc.
	Conjonctions	assuming that, or, and/or, either[...]or, if, etc.
	Verbes	may, might, can, would, could, should, assume, suggest, question, presume, suspect, indicate, suppose, seem, appear, favor
Français	Adjectifs/Adverbes	probable, possible, incertain, suspect, peut-être
	Conjonctions	en supposant/admettant que, ou et/ou, soit[...]soit, soit[...]ou, si, etc.
	Verbes	devoir, pouvoir, supposer, suspecter, sembler reste à + verbe à l'infinitif, verbes au conditionnel présent

Contrairement aux marqueurs multi-mots de négation, dont la composition ne varie pas (*à la place de, à l'exception de*, etc.), la composition de plusieurs marqueurs d'incertitude est variable. C'est le cas de l'expression *reste à* suivie d'un verbe à l'infinitif, illustrée dans l'exemple (1), qui est compatible avec un grand nombre de verbes tels que *définir*, *déterminer*, *vérifier*, *établir*, *démontrer*, *prouver*, etc. C'est aussi le cas des verbes au conditionnel présent qui constituent à eux seuls un grand nombre de marqueurs potentiels. En effet, alors qu'en anglais les verbes marquant l'incertitude ne changent pas ou peu de forme, en français, les formes du conditionnel présent de la plupart, sinon de tous les verbes, peuvent marquer l'incertitude. Dans l'exemple (2), au présent de l'indicatif, le verbe *consister* ne marquerait pas l'incertitude. En anglais, *would consist* serait utilisé. Le marqueur *would* étant très fréquemment utilisé, il serait facilement détecté. Cependant, dans les textes en français avec lesquels nous travaillons, le conditionnel présent n'est pas utilisé souvent, à l'exception des verbes *devoir* et *pouvoir*. Équivalent français de *would + verbe à l'infinitif*, *pourrait/pourraient + verbe à l'infinitif* est l'un des marqueurs les plus fréquemment utilisé (comme illustré dans l'exemple 3).

1. L'efficacité des inhibiteurs de H-DAC pour ce type de lymphome est déjà montré, mais le bénéfice de l'association Romidepsine+CHOP par rapport au CHOP reste à démontrer.

2. Une reconstruction anatomique complète et directe consisterait en la recanalisation, en ligne droite, de l'ensemble du segment veineux occlus, depuis la jonction poplitée fémorale jusqu'à la veine iliaque commune.
3. Par conséquent, en bloquant l'action de la protéine « MEK », le médicament à l'étude pourrait parvenir à stopper la multiplication des cellules tumorales.

Si les marqueurs de l'incertitude semblent être plus complexes que ceux de la négation, la portée de l'incertitude semble être plus facile à déterminer. Le domaine médical laissant peu de place à l'imprécision et la subjectivité, dans les textes avec lesquels nous travaillons, l'incertitude est le plus souvent causée par la spéculation (exemples 1, 2, et 3) et le manque d'information (exemple 4). La délimitation de la portée est donc celle de la proposition marquée.

1. Possibilité **de subir au moins une leucophérèse**.
2. Une reconstruction anatomique complète et directe consisterait **en la recanalisation, en ligne droite, de l'ensemble du segment veineux occlus, depuis la jonction poplitée fémorale jusqu'à la veine iliaque commune**.
3. Par conséquent, en bloquant l'action de la protéine « MEK », le médicament à l'étude pourrait **parvenir à stopper la multiplication des cellules tumorales**.
4. L'efficacité des inhibiteurs de H-DAC pour ce type de lymphome est déjà montré, mais **le bénéfice de l'association Romidepsine+CHOP par rapport au CHOP** reste à démontrer.

1.3.2 Jeux de données annotées

Depuis plus de dix ans, avec la démocratisation des techniques d'apprentissage supervisé, plusieurs jeux de données de tailles variables, principalement en anglais, ont été annotés de différentes manières avec des informations relatives à la négation et l'incertitude. Ces jeux de données ont permis le développement de nombreux systèmes de détection automatique que nous décrivons dans la sous-section 1.3.3. Dans ce qui suit, nous décrivons les corpus créés.

Genia Event corpus

Le **Genia Event corpus** (J.-D. KIM, OHTA et TSUJII, 2008) comprend 1 000 résumés **MEDLINE**. Il s'agit d'un sous-ensemble du corpus **GENIA** original, qui a été sélectionné à l'aide des trois termes MeSH *human*, *blood cells* et *transcription factors*. Pour chaque phrase, trois types d'informations sont annotés : les termes biomédicaux, qui sont identifiés et assignés à des catégories à l'aide d'une ontologie ; les relations entre les termes formant des évènements, qui sont identifiées et assignées à l'aide d'une ontologie ; les méta-connaissances (le type d'évènement, le niveau de certitude associé à l'occurrence de l'évènement, la polarité et la source).

Speculative Text Corpus

Le **Speculative Text Corpus** (SETTLES, CRAVEN et FRIEDLAND, 2008) est constitué de 850 phrases extraites des résumés **PubMed**. Chaque phrase est classée comme certaine ou incertaine. Cependant, aucun marqueur n'est annoté.

BioScope

Le corpus **BioScope** (VINCZE et al., 2008), est constitué de 3 sources textuelles différentes, à savoir : (1) comptes-rendus d'examens radiologiques, (2) articles de **FlyBase** et **BMC Bioinformatics**, ainsi que (3) résumés d'articles scientifiques issus du corpus **GENIA**. Selon les consignes d'annotations préétablies, cet ensemble de données est annoté au niveau des marqueurs de l'incertitude et de la négation, et au niveau de la phrase pour marquer leur portée linguistique. Les portées discontinues ne sont pas prises en compte. Le tableau 1.11 quantifie la composition du corpus. Ce corpus est disponible au format XML, où chaque phrase et chaque paire portée/marqueur sont indexées par un identifiant unique.

TABLEAU 1.11 – Statistiques du corpus BioScope

	Examens	Articles	Résumés
Documents	1 954	9	1 273
Phrases	6 383	2 670	11 871
Phrases négatives	13,55 %	12,70 %	13,45 %
Marqueurs de négation	877	389	1 848
Phrases incertaines	13,39 %	19,44 %	17,70 %
Marqueurs d'incertitude	1 189	714	2 769

FactBank

Le corpus **FactBank** (SAURÍ et PUSTEJOVSKY, 2009) est constitué de 208 documents et contient un total de 9 488 évènements annotés manuellement pour quatre valeurs de factualité (*certain*, *probable*, *possible* et *underspecified*). Les documents sont issus de **TimeBank** et du **A-TimeML Corpus**.

CoNLL-2010 Shared Task

Dans le cadre de la campagne d'évaluation de la *Conference on Computational Natural Language Learning* de 2010 (**CoNLL-2010**), trois jeux de données ont été mis à la disposition des participants (FARKAS et al., 2010). La première tâche consistait à identifier les phrases qui contiennent des informations incertaines ou douteuses à l'aide de textes où seuls les marqueurs sont annotés. Le premier jeu de données est donc composé de paragraphes issus de **Wikipedia** où apparaissent des *weasel words*, c'est-à-dire, des mots et phrases qui donnent l'impression qu'une déclaration spécifique ou significative a été faite, alors qu'il s'agit en réalité d'une revendication vague et ambiguë. Le deuxième jeu de données est issu de **BioScope**, mais sans les compte-rendus cliniques. En effet, la seconde tâche consistait à identifier la portée de

l'incertitude dans les données biomédicales. Dans ce jeu de données, les marqueurs d'incertitude et leurs portées sont annotés. Pour cette tâche, l'évaluation était très stricte puisque seuls les marqueurs et portées exactement identifiées étaient considérés comme vrais positifs par les organisateurs.

i2b2/VA-2010

La compétition **i2b2/VA-2010** (Ö. UZUNER et al., 2011) présentait trois tâches d'extraction d'informations à partir de dossiers cliniques américains. Une des tâches concernait la détection d'assertions et de leurs portées. Ainsi, à chaque concept médical devait être associée l'une de ces six classes : *present*, *absent*, *possible*, *conditional*, *hypothetical* ou *not associated with the patient*.

*SEM 2012 Shared Task

La première *Joint Conference on Lexical and Computational Semantics* (***SEM-2012**) proposait une *shared task* composée de deux sous-tâches : la détection de la portée et du foyer de la négation (MORANTE et BLANCO, 2012). Le foyer correspond à la partie de la portée la plus explicitement niée. Un jeu de données a été annoté pour chaque tâche. Dans l'exemple suivant, la portée est entre crochets, l'objet entre * et le foyer est souligné.

[John had] never [*said* as much before].

Tâche 1 : Détection de la portée. Le jeu de données **CD-SCO** est composé d'un roman et de trois nouvelles de *Sherlock Holmes* écrits par Sir Arthur Conan Doyle :

— *The Hound of the Baskervilles* pour l'ensemble d'entraînement,
— *The Adventures of Wisteria Lodge* pour l'ensemble d'évaluation,
— *The Adventure of the Red Circle* et *The Adventure of the Cardboard Box* pour l'ensemble de test.

Toutes les occurrences de la négation sont annotées, ce qui donne 1 056 phrases sur 3 899. Pour chaque occurrence, le marqueur et la portée sont annotés, ainsi que l'évènement nié, s'il y en a un. Les marqueurs et les portées peuvent être discontinus dans ce corpus. Le guide d'annotation a également été publié (MORANTE, SCHRAUWEN et DAELEMANS, 2011). Parallèlement, une version chinoise de ce corpus a été annotée (Q. LIU, FANCELLU et WEBBER, 2018).

Tâche 2 : détection du foyer. Dans le jeu de données **PB-FOC**, le foyer de la négation est annoté pour les 3 993 phrases de la section **WSJ** de la **Penn TreeBank** marquées avec **MNEG** dans **PropBank**. Contrairement à **CD-SCO**, toutes les phrases de **PB-FOC** contiennent une négation.

À l'instar de nombreux jeux de données, toutes les annotations sont fournies au format **CoNLL-2005 Shared Task**. Chaque ligne correspondant à un *token* et chaque

annotation sont fournies dans une colonne. Une ligne vide indique la fin d'une phrase. En plus des annotations relatives à la négation, sont également fournies : les lemmes, l'étiquetage morpho-syntaxique (*PoS-tags*) et l'analyse syntaxique de surface (*chunks*).

SFU Review Corpus

Le **SFU Review Corpus** (KONSTANTINOVA et al., 2012) est composé de 400 documents de critiques de films, de livres et de produits de consommation. Les marqueurs de négation et d'incertitude ainsi que leur portée y sont annotés. Les règles d'annotation sont basées sur celles du corpus **BioScope**.

MiPACQ

Le corpus **MiPACQ** (*Multi-source Integrated Platform for Answering Clinical Questions*) (ALBRIGHT et al., 2013) est constitué de données cliniques en anglais annotées par plusieurs couches d'étiquettes syntaxiques et sémantiques. Deux attributs sont disponibles pour chaque entité UMLS détectée. L'attribut *Negation*, qui peut prendre deux valeurs, *true* ou *false* ainsi que l'attribut *Status*, qui peut prendre quatre valeurs, *none*, *possible*, *HistoryOf* ou *FamilyHistoryOf*.

CNESP

Le corpus **CNESP** (*Chinese Negation and Speculation corpus*) (ZOU, Qiaoming ZHU et Guodong ZHOU, 2015) est composé de trois sous-ensemble de données : 19 articles scientifiques, 311 articles du domaine de la finance ainsi que 821 avis sur différents produits pour un total de 16 841 phrases. Parmi ces phrases, plus de 20 % contiennent une instance de négation ou d'incertitude. Les règles d'annotation sont basées sur celles du corpus **BioScope**.

1.3.3 Étiquetage automatique

Afin de détecter automatiquement la négation et l'incertitude dans les textes, la communauté scientifique a principalement proposé des méthodes reposant sur deux champs d'études de l'intelligence artificielle : (1) les systèmes experts, raisonnant à partir de faits et règles connus pour répondre à des questions précises, et (2) la classification par apprentissage automatique supervisé, reposant sur des algorithmes et modèles statistiques que les systèmes informatiques utilisent afin d'effectuer une tâche spécifique sans utiliser des instructions explicites mais en apprenant une fonction de prédiction à partir de données annotées. Dans cette section, nous revenons sur les systèmes de détection automatique proposés par la communauté scientifique. Cependant, de très nombreux systèmes ont été proposés ces vingt dernières années. Par conséquent, cette section n'a pas vocation à être exhaustive, mais à présenter les systèmes pionniers, ainsi que les systèmes les plus performants sur différents corpus annotés.

Exploitation de systèmes experts

Un système expert est un programme qui raisonne à partir d'informations symboliques et utilise des règles expertes afin d'essayer d'atteindre les performances du niveau d'un expert. Pour des tâches spécifiques, il serait possible de construire un système qui raisonne aussi bien que les spécialistes de ces domaines. Partant de cette hypothèse, les premiers systèmes dédiés à la détection de la négation et de l'incertitude sont proposés sous cette forme. Dans cette section, nous présentons donc les systèmes de détection de la négation et de l'incertitude qui n'utilisent aucune méthode par apprentissage automatique.

NegEx. (CHAPMAN et al., 2001) utilisent les expressions régulières pour détecter les phrases négatives, filtrer les phrases faussement négatives et identifier les termes médicaux à l'intérieur de la portée des négations. Les termes médicaux indexés par le système prennent le statut ***negated*** ou ***possible***. Ce système a ensuite été adapté à d'autres langues, telles que le suédois (VELUPILLAI, DALIANIS et KVIST, 2011) et le français (DELÉGER et GROUIN, 2012).

Negfinder. (MUTALIK, DESHPANDE et NADKARNI, 2001) combinent un analyseur lexical, qui utilise des expressions régulières afin de générer un automate fini, et un analyseur syntaxique, qui repose sur un sous-ensemble restreint de la grammaire non contextuelle *Look-Ahead Left-to-right Rightmost derivation* (**LALR**). Ainsi, **NegFinder** permet d'identifier les concepts impactés par la négation dans les textes médicaux lorsqu'ils sont proches du lemme marquant la négation.

ELKIN et al., 2005 proposent une extension des travaux précédents. Le système utilise le **Mayo Vocabulary Server** afin de récupérer un ensemble de concepts médicaux présents dans les fragments de phrases. Ensuite, des règles expertes assignent à chaque concept un statut : positif, négatif ou incertain.

Y. HUANG et LOWE, 2007 proposent d'automatiser la détection de la négation en combinant les expressions régulières et l'analyse grammaticale. Les négations sont classées sur la base des catégories syntaxiques et sont indiquées dans les arbres syntaxiques. Cette approche hybride permet d'identifier les concepts marqués par la négation dans les rapports de radiologie même s'ils sont situés à distance des marqueurs de négation.

ConText. (HARKEMA et al., 2009) proposent un système expert dérivé de **NegEX** et couvrant plusieurs objectifs additionnels. Ainsi, **ConText** détecte la négation, la temporalité ainsi que le sujet concerné par ces informations dans les textes cliniques. Ce système a ensuite été adapté au français (ABDAOUI et al., 2017).

ØVRELID, VELLDAL et OEPEN, 2010 proposent un petit ensemble de règles qui définissent la portée de chaque marqueur. Pour développer ces règles, les auteurs

exploitent les informations contenues dans le guide d'annotation du corpus **BioScope**, ainsi qu'une analyse manuelle des données d'entraînement pour trouver les interactions des constructions pour divers types de marqueurs.

KILICOGLU et BERGLER, 2010 proposent une méthodologie pour la détection de l'incertitude essentiellement basée sur des règles, en exploitant des informations lexicales et syntaxiques. L'information lexicale est représentée par un simple dictionnaire, alors que l'information syntaxique nécessaire est identifiée grâce aux arbres d'analyse en constituants et aux arbres d'analyse en dépendances retournés par le **Stanford Parser**[26].

ScopeFinder (APOSTOLOVA, TOMURO et DEMNER-FUSHMAN, 2011) est un système expert dédié à la détection de la portée de la négation et de l'incertitude dont les règles sont construites à partir des patrons lexico-syntaxiques extraits automatiquement du corpus **BioScope**. La méthode de construction du système fait qu'il est adaptable aux données de différents domaines.

NegBio. (PENG, X. WANG et al., 2018) repose sur des règles définies à partir de graphes de dépendances universelles (*universal dependency graph* ou **UDG**). Le code de ce système est disponible en ligne[27].

Les systèmes experts présentés dans cette section arrivent à obtenir de bons résultats sur les données biomédicales pour lesquelles ils sont développés. Cependant, pour être appliqués efficacement à d'autres données ou d'autres langues, ces systèmes doivent être spécifiquement adaptés et testés. Les systèmes par apprentissage automatique, que nous présentons dans la section qui suit, souffrent moins de ce problème.

Exploitation de l'apprentissage supervisé

La classification par apprentissage automatique supervisé repose sur des algorithmes et modèles statistiques qui, à partir de paires entrée/sortie, apprennent une fonction de prédiction permettant d'attribuer une sortie à toute nouvelle entrée soumise au modèle. Par exemple, dans le cas de la détection automatique des marqueurs, chaque paire correspond au minimum à ***mot/marqueur*** ou ***mot/~~marqueur~~***. Cependant, la plupart des systèmes prennent en entrée plusieurs descripteurs pour chaque mot. Par exemple, les lemmes et l'étiquetage morpho-syntaxique (*part-of-speech tagging* ou **PoS**) sont très couramment utilisés. Dans les paragraphes suivants, la plupart des systèmes que nous décrivons ont été proposés après la publication du corpus BioScope, lors de campagnes d'évaluations ou bien à la suite de la récente résurgence des réseaux de neurones. D'autre part, les systèmes de détection présentés ici ne fonctionnent pas nécessairement que par apprentissage supervisé. En effet, certains systèmes fonctionnent par systèmes experts pour la détection des marqueurs et par apprentissage automatique pour la portée et inversement. Plus

26. https://nlp.stanford.edu/software/lex-parser.shtml
27. https://github.com/ncbi-nlp/NegBio

rarement, les méthodes par apprentissage peuvent être utilisées comme système de secours, quand le système principal échoue. La plupart de méthodes par apprentissage supervisé proposées pour la détection de la négation et/ou de l'incertitude reposent sur les champs aléatoires conditionnels (*conditional random fields* ou **CRF**), les séparateurs à vaste marge (*support vector machine* ou **SVM**), ainsi que les réseaux de neurones artificiels.

LIGHT, QIU et SRINIVASAN, 2004 utilisent un classifieur par **SVM** afin de sélectionner les phrases spéculatives dans des résumés d'articles de **MEDLINE**. Les résumés sont d'abord pré-traités avec **SMART** (*System for the Mechanical Analysis and Retrieval of Text*) (SALTON, 1971) afin d'obtenir la représentation vectorielle des mots. L'implémentation utilise **SVM$^{\text{light}}$** [28] avec les paramètres par défaut et une validation croisée à 10 plis.

(MORANTE, LIEKENS et DAELEMANS, 2008; MORANTE et DAELEMANS, 2009a; MORANTE et DAELEMANS, 2009b; MORANTE, VAN ASCH et DAELEMANS, 2010). Roser Morante ainsi que ses co-auteurs ont proposé plusieurs systèmes par apprentissage pour la détection de la négation et de l'incertitude. Pour la détection des marqueurs, les systèmes utilisent soit **IGTree** (***memory-base learning***) via **TiMBL** [29] soit un classifieur SVM. Pour la détection de la portée, plusieurs méthodes par apprentissage ont été proposées (**IGTree**, **SVM**, **CRF**). Les descripteurs régulièrement utilisés sont les mots, les lemmes, l'étiquetage morpho-syntaxique ainsi que l'analyse syntaxique de surface.

TANG et al., 2010 proposent un système de détection des marqueurs d'incertitude en cascade qui classe chaque *token* selon une représentation en **BIO** (*Begin, Inside, Outsite*) en trois étapes. La première consiste à pré-traiter les données avec le **GENIA Tagger**. La seconde opère deux classifications indépendantes et adaptées à l'étiquetage de séquences (**CRF++ 0.53** [30] et **SVM$^{\text{hmm}}$** [31]) sur les données pré-traitées. La dernière étape consiste à entraîner un **CRF** en utilisant les prédictions de l'étape précédente. Le système de détection de la portée utilise le **GDep Tagger** [32] afin d'obtenir de nombreux descripteurs. Finalement, un classifieur par **CRF** est entraîné dans le but de détecter le début et la fin de chaque portée.

X. LI et al., 2010 détectent les marqueurs d'incertitude à l'aide un classifieur **CRF** (**CRF++ 0.51**) entraîné avec les paramètres par défaut et selon une représentation en **BIO**. Un algorithme glouton est utilisé afin de sélectionner un meilleur ensemble de descripteurs (mot, lemme, **PoS**, *chunking*) pour le classifieur selon les résultats obtenus sur les données de développement. En post-traitement, des règles expertes sont utilisées afin de classer correctement certains marqueurs oubliés ou incorrectement classés. Afin de détecter la portée, un classifieur **CRF** est entraîné sur un ensemble

28. http://svmlight.joachims.org/
29. https://languagemachines.github.io/timbl/
30. https://taku910.github.io/crfpp/
31. https://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html
32. http://www.sagae.org/gdep/

de descripteurs lexicaux et syntaxiques dans le but de détecter le début et la fin de chaque portée.

H. ZHOU et al., 2010 identifient les marqueurs candidats à partir d'une liste de mots clés puis appliquent un **CRF** pour corriger les erreurs d'identification. Les marqueurs récupérés sont utilisés comme descripteurs pour la détection par **CRF**, en plus des mots, lemmes, radicaux, **PoS** et *chunks*. La détection de la portée est effectuée à l'aide d'un classifieur **CRF** qui prédit pour chaque *token* de la phrase s'il est le premier ou dernier *token* de la portée.

J. LI et al., 2010 et Qiaoming ZHU et al., 2010 proposent une approche pour l'apprentissage de la portée de la négation utilisant un système d'analyse sémantique superficielle simplifiée qui combine des règles expertes, l'élagage et un classifieur **SVM**. Étant donné un arbre syntaxique et un prédicat lui appartenant, l'analyse sémantique superficielle reconnaît et assigne à tous les constituants dans la phrase le rôle sémantique leur correspondant. Le marqueur de négation peut être considéré comme le prédicat, alors que les constituants de la portée peuvent être considérés comme les arguments du marqueur de la négation. Deux règles expertes sont adoptées pour assigner le rôle d'argument de la portée sur plusieurs constituants dans un arbre syntaxique donné. Excepté le marqueur de négation et ses ancêtres, tous les constituants dans l'arbre syntaxique, dont le parent couvre le marqueur, sont collectés comme arguments candidats. Finalement, un **SVM** est appliqué pour déterminer si les arguments candidats sont valides.

REI et BRISCOE, 2010 détectent les marqueurs d'incertitude à l'aide d'un classifieur **CRF** exploitant les descripteurs de l'analyseur syntaxique **RASP**. Les prédictions du **CRF** sont ensuite filtrées en utilisant une liste de marqueurs connus, afin d'augmenter la précision tout en conservant le rappel. Leur système de détection de la portée comporte trois étapes : premièrement, un système expert, dont les règles sont basées sur les relations grammaticales de la phrase et l'étiquetage morpho-syntaxique du marqueur, donne une première prédiction. Ensuite, un classifieur CRF affine ces prédictions. Enfin, un petit ensemble de règles de post-traitement est utilisé afin de produire les portées définitives.

(VELLDAL, ØVRELID et al., 2012; READ et al., 2012; LAPPONI et al., 2012; PACKARD et al., 2014). VELLDAL, ØVRELID et al., 2012 traitent l'ensemble de marqueurs en tant que classe fermée. Ce faisant, ils expliquent que l'on peut réduire considérablement le nombre d'exemples et conséquemment le nombre de descripteurs présentés au classifieur sans perdre en rappel. Le système utilise un **SVM** appliqué en utilisant de simples descripteurs n-grammes sur les mots et lemmes, à droite et à gauche des marqueurs candidats. Pour détecter la portée, le système combine deux approches. L'une utilise des règles expertes opérant sur des arbres syntaxiques de dépendance. La seconde utilise un *ranking SVM* qui opère sur des noeuds d'arbres syntaxiques des constituants syntagmatiques.

READ et al., 2012 et **LAPPONI et al., 2012** sont deux systèmes proposés par des chercheurs de l'université d'Oslo dans le cadre de ***SEM-2012 Shared Task**. Adaptation de (VELLDAL, ØVRELID et al., 2012), le système de détection des marqueurs prend désormais en charge de la détection des marqueurs de négation affixaux annotés dans les données d'entraînement et de développement. Dans ce but, le classifieur **SVM** utilise des descripteurs spécifiques. Le premier descripteur utilise les n-grammes du début et de la fin de la base à laquelle un affixe s'accroche. Par exemple, dans le cas du marqueur ***impossible***, les n-grammes [*possi, poss, pos, ...*] et [*sible, ible, ble, ...*] sont utilisés et joints à l'affixe *im-* et l'étiquette **PoS JJ**. Le second descripteur cherche à imiter l'effet d'une recherche dans un lexique pour les chaînes de caractères restantes auxquelles un affixe s'attache aussi, afin de vérifier leurs statuts en tant que formes de bases indépendantes et leurs étiquetages morpho-syntaxiques. Ce descripteur a pour but d'éviter de classer comme marqueur de négation les mots dont la première partie de la chaîne de caractère restante constitue très rarement le début d'un mot (***underlying, underneath, important***). D'autre part, un petit ensemble de règles expertes est développé afin de détecter les marqueurs rares en post-traitement (***neither...nor***, ***on the contrary***).

READ et al., 2012. Dans la continuité de (VELLDAL, ØVRELID et al., 2012), ce système de détection de la portée de la négation combine un système expert et un *ranking SVM*. Les règles expertes sont formulées en fonction de patrons cohérents relevés en alignant constituants syntaxiques et portées. Les règles s'appuient sur la fréquence des chemins allant du marqueur jusqu'au constituant aligné avec la portée sur la base des annotations dans les données d'entraînement (**CD-SCO**). D'autre part, Le classifieur **SVM** apprend une fonction de rang sur les constituants syntaxiques candidats afin de choisir les portées candidates correctes. Les constituants syntaxiques candidats sont générés en plusieurs étapes : sélection d'instances de négation pour lesquelles l'arbre syntaxique contient un constituant qui s'aligne avec la portée ; sélection d'un candidat initial en prenant le constituant le plus petit qui s'étend sur tous les mots dans la portée ; génération d'un candidat subséquent en traversant le chemin vers la racine de l'arbre. Les candidats dont la projection correspond à la portée sont étiquetés comme corrects, les autres comme incorrects.

LAPPONI et al., 2012 détectent la portée de la négation à l'aide d'un classifieur **CRF** et d'un étiquetage en **BIO**, où les marqueurs de négation sont étiquetés comme démarreurs de séquence et les *tokens* de la portée comme *chunks*. En plus de descripteurs fournis pour chaque *token* dans le corpus **CD-SCO**, sont également pris en considération : la distance entre chaque token et le marqueur le plus proche à droite et à gauche, les bigrammes et trigrammes vers l'avant et l'arrière des tokens et des étiquettes morpho-syntaxiques, ainsi que les unigrammes et bigrammes de *PoS-tags* lexicalisés (couple forme/*PoS-tags*). Les représentations syntaxiques sont converties en représentations de dépendances en utilisant le *Stanford Parser*. Les descripteurs récupérés à partir des graphes de dépendances ont pour but de modéliser les relations syntaxiques entre chaque unité et le marqueur le plus proche. Finalement, un algorithme basé sur des règles expertes est appliqué en sortie du **CRF** dans le but d'assigner ou non chaque unité sous la portée d'un marqueur et déterminer les chevauchements des portées.

PACKARD et al., 2014. Ce système de détection de la portée de la négation repose sur deux approches. La première est basée sur un système expert par ***minimal Recursion semantics*** (**MRS**). La **MRS** est un cadre applicatif pour la sémantique computationnelle. Il peut être implémenté selon des formalismes tels que la grammaire syntagmatique guidée par les têtes et la grammaire lexicale-fonctionnelle. La seconde approche reprend le système de (READ et al., 2012) et intervient pour la détection des marqueurs et pour la détection de la portée dans les cas où le *LinGo English Ressource Grammar* (**ERG**) ne peut pas créer de représentation sémantique fiable pour une phrase.

CHOWDHURY, 2012. Après avoir écarté les affixes marquant la négation, ce système collecte automatiquement, dans les données d'entraînement, un vocabulaire de toutes les unités lexicales qui ne sont pas des marqueurs de négation et dont la taille dépasse trois caractères. Ensuite, le système les utilise pour extraire des descripteurs qui pourraient être utiles pour identifier des affixes marqueurs de négation potentiels. Une liste de termes exprimant très probablement la négation est créée à partir des données d'entraînement en se basant sur leurs fréquences. Les marqueurs de négation affixaux sont identifiés si l'unité est prédite en tant que marqueur par le classifieur et possède l'un des affixes collectés dans les données d'entraînement. Un post-traitement supplémentaire est effectué pour annoter certains marqueurs présents dans les données d'entraînement mais oubliés par le classifieur lors de la prédiction sur les données de développement. Finalement, un classifieur **CRF** est entraîné sur les descripteurs collectés et utilisés pour prédire les marqueurs de négation sur les données de test. Le système de détection de la portée de la négation utilise aussi les **CRF**. Le classifieur est entraîné à l'aide des descripteurs fournis dans **CD-SCO** ainsi que par des descripteurs contextuels : lemme du 1er mot de la phrase correspondante, position du *token* par rapport au marqueur, présence d'une conjonction de coordination ou de caractères spéciaux entre le *token* et le marqueur, etc.

ZOU, G. ZHOU et Q. ZHU, 2013 proposent une approche par classifieur **SVM** qui exploite des descripteurs issus de noyaux d'arbres syntaxiques de dépendances et de constituants syntagmatiques (*constituent and dependency trees*) pour la détection de portée de l'incertitude et de la négation. Les résultats obtenus sur le corpus BioScope montrent que ces descripteurs syntaxiques structurés ont l'avantage de capturer les relations potentielles entre les marqueurs et leurs portées.

QIAN et al., 2016 proposent un système basé sur un réseau de neurones convolutifs (*Convolutional Neural Networks*, **CNN**). Les descripteurs pertinents sont récupérés à partir des chemins syntaxiques entre les marqueurs et les unités candidates dans les arbres syntaxiques de dépendances et les arbres syntaxiques de constituants syntagmatiques. La couche convolutionnelle concatène ces descripteurs avec leurs positions relatives en un seul vecteur qui alimente une couche entièrement connectée avec une fonction d'activation softmax pour la prédiction.

FANCELLU, LOPEZ et WEBBER, 2016 proposent deux approches par réseau de neurones pour résoudre le problème de détection de la portée de la négation. La première utilise un réseau de neurones à propagation avant (*feedforward neural network*) et la seconde un réseau de neurones récurrents (*recurrent neural network*, **RNN**) *Long Short-Term Memory* (**LSTM**) bidirectionnel. Le système de base prend en entrée une instance $I(n;c)$, où chaque *token* est représenté par un vecteur n (*word-embedding*) et par un vecteur c qui détermine si le *token* fait partie d'un marqueur (*cue-embedding*). Dans le but d'affiner la prédiction, le système peut aussi utiliser une représentation vectorielle de l'étiquetage morpho-syntaxique, ainsi qu'un modèle *word2vec* entraîné sur **Wikipedia** et le corpus **CD-SCO**. Pour chaque système, la prédiction est assurée par une couche *softmax*.

H. LI et LU, 2018 utilisent les conditional random fields (**CRF**), *semi-Markov **CRF***, ainsi que les *latent-variable **CRF*** pour capturer la portée de la négation. Leur observation clé est que certaines informations utiles telles que les caractéristiques liées aux marqueurs de négation, les dépendances de longue distance ainsi que certaines informations structurelles latentes peuvent être exploitées pour cette tâche.

SERGEEVA et al., 2019 et A. KHANDELWAL et SAWANT, 2020 proposent leurs adaptations de **BERT** (DEVLIN et al., 2019) pour la détection de la négation et de l'incertitude. SERGEEVA et al., 2019 proposent plusieurs **BiLSTM** et *fine-tuning* de **BERT** pour la détection de la portée de la négation et de l'incertitude. Toutes leurs approches utilisent des *cue-embeddings* afin d'indiquer le statut de marqueur de chaque token pendant l'entraînement. A. KHANDELWAL et SAWANT, 2020 détectent les marqueurs de négation ainsi que leur portée à l'aide d'une version ré-entraînée de $\textbf{BERT}_{BASE}$. L'étiquetage des marqueurs est multi-classe (Affixe, marqueur simple, partie d'un marqueur multi-mot, pas un marqueur) tandis que l'étiquetage de la portée est binaire (dans la portée, en dehors de la portée). Lors de l'étiquetage de la portée, les tokens qui sont des marqueurs sont remplacés ou augmentés afin d'indiquer au système qu'ils marquent la négation.

Évaluation des systèmes de classification

Les systèmes d'étiquetage sont généralement évalués à l'aide des trois métriques que nous avons déjà présenté en 1.2.3 : la précision, le rappel et la F-mesure.

BioScope. Étant donné que BioScope n'est pas découpé en ensembles d'entraînement et de test, les modèles sont uniquement entraînés sur le corpus de résumés d'articles, puis testés sur les articles scientifiques et examens cliniques. Une validation croisée à 10 plis permet d'obtenir des résultats sur le corpus de résumés.

Les tableaux 1.12 et 1.13 répertorient les résultats obtenus pour la détection des marqueurs sur le corpus BioScope. Peu de systèmes ont été développés pour la détection des marqueurs du corpus **BioScope**. Cependant, il existe de nombreux résultats sur les jeux de données issus de BioScope lors de la campagne d'évaluation de CoNLL-2010. Cette préférence est pragmatique, les jeux d'entraînement et de

test étant prédéfinis. Concernant la négation, nous constatons que les systèmes de VELLDAL, ØVRELID et al., 2012 et MORANTE et DAELEMANS, 2009a restent efficace hors domaine (Articles, Examens) tandis que le système de Qiaoming ZHU et al., 2010, qui obtient de bons résultats en validation croisée sur les résumés, n'y parvient pas. Concernant l'incertitude, Qiaoming ZHU et al., 2010 s'en sort bien mieux que MORANTE et DAELEMANS, 2009b sur l'ensemble des jeux de données. Cependant, les résultats chutent fortement hors domaine.

TABLEAU 1.12 – Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur le corpus **BioScope**.

Corpus	Systèmes	Précision	Rappel	F-mesure
Résumés	VELLDAL, ØVRELID et al., 2012	93,46	98,73	**96,00**
	Qiaoming ZHU et al., 2010	**94,35**	94,99	94,67
	MORANTE et DAELEMANS, 2009a	84,72	**98,75**	91,20
Examens	VELLDAL, ØVRELID et al., 2012	96,44	95,90	96,17
	Qiaoming ZHU et al., 2010	88,54	86,81	87,67
	MORANTE et DAELEMANS, 2009a	**97,33**	**98,09**	**97,71**
Articles	VELLDAL, ØVRELID et al., 2012	85,22	**98,25**	**91,27**
	Qiaoming ZHU et al., 2010	**87,47**	90,48	88,95
	MORANTE et DAELEMANS, 2009a	87,18	95,72	91,25

TABLEAU 1.13 – Résultats des systèmes de détection des marqueurs d'incertitude au niveau des tokens sur le corpus **BioScope**.

Corpus	Systèmes	Précision	Rappel	F-mesure
Résumés	Qiaoming ZHU et al., 2010	**93,14**	**83,74**	**88,19**
	MORANTE et DAELEMANS, 2009b	90,81	79,84	84,77
Examens	Qiaoming ZHU et al., 2010	**91,77**	**33,33**	**48,90**
	MORANTE et DAELEMANS, 2009b	88,10	27,51	41,92
Articles	Qiaoming ZHU et al., 2010	**82,31**	**73,02**	**77,39**
	MORANTE et DAELEMANS, 2009b	75,35	68,18	71,59

Les tableaux 1.14, 1.15 et 1.16 répertorient les résultats obtenus pour la détection de la portée sur le corpus BioScope. Les résultats présentés dans le tableau 1.16 sont rapportés en pourcentage de portées correctes (**PPC**). Le **PPC** est défini comme le nombre de portées correctement annotées divisé par le nombre de total de portées. Concernant la négation, les résultats au niveau des *tokens* sont globalement élevés, notamment sur le corpus d'examens cliniques bien que les systèmes soient entraînés uniquement sur les résumés. En termes de **PPC**, l'adaptation de **BERT** de SERGEEVA et al., 2019 obtient de loin les meilleurs résultats sur le corpus de résumés mais ne communique pas de résultats comparable pour les autres corpus. H. LI et LU, 2018 obtient le **PPC** le plus élevé sur le corpus d'examens cliniques mais ne parvient pas à égaler Qiaoming ZHU et al., 2010 sur le corpus d'articles scientifiques. Concernant l'incertitude, au niveau des *tokens*, le **CNN** de QIAN et al., 2016 obtient

les meilleurs résultats, loin devant MORANTE et DAELEMANS, 2009b. En termes de **PPC**, SERGEEVA et al., 2019 obtient à nouveau les résultats les plus élevés sur le corpus de résumés. Le **CNN** de QIAN et al., 2016 est le plus performant sur le corpus d'examens cliniques mais ne parvient pas à égaler le **SVM** de ZOU, G. ZHOU et Q. ZHU, 2013 sur le corpus d'articles. Ce dernier obtient d'ailleurs des résultats compétitifs sur les autres corpus.

TABLEAU 1.14 – Résultats des systèmes de détection de la portée de la négation au niveau des tokens sur le corpus **BioScope**.

Corpus	Systèmes	Précision	Rappel	F-mesure
Résumés	H. LI et LU, 2018	NC	NC	**92,1**
	QIAN et al., 2016	89,49	90,54	89,91
	MORANTE et DAELEMANS, 2009a	90,68	90,68	90,67
Examens	H. LI et LU, 2018	NC	NC	**97,5**
	QIAN et al., 2016	91,97	97,03	94,43
	MORANTE et DAELEMANS, 2009a	91,65	92,50	92,07
Articles	H. LI et LU, 2018	NC	NC	83,1
	QIAN et al., 2016	82,08	84,90	83,46
	MORANTE et DAELEMANS, 2009a	84,47	84,95	**84,71**

TABLEAU 1.15 – Résultats des systèmes de détection de la portée de l'incertitude au niveau des tokens sur le corpus **BioScope**.

Corpus	Systèmes	Précision	Rappel	F-mesure
Résumés	QIAN et al., 2016	**95,95**	**95,19**	**95,56**
	MORANTE et DAELEMANS, 2009b	89,71	89,09	89,40
Examens	QIAN et al., 2016	**86,85**	**93,84**	**90,21**
	MORANTE et DAELEMANS, 2009b	79,16	78,13	78,64
Articles	QIAN et al., 2016	**86,78**	**86,59**	**86,69**
	MORANTE et DAELEMANS, 2009b	77,78	77,10	77,44

TABLEAU 1.16 – Résultats des systèmes de détection de la portée au niveau des portées correctement identifiées **BioScope**.

	Systèmes	Résumés	Examens	Articles
Négation	SERGEEVA et al., 2019	**87.03**	NC	NC
	H. LI et LU, 2018	84,1	**94,4**	60,1
	QIAN et al., 2016	77,14	89,66	55,32
	ZOU, G. ZHOU et Q. ZHU, 2013	76.90	85.31	61.19
	Qiaoming ZHU et al., 2010	81,84	89,79	**64,02**
	MORANTE et DAELEMANS, 2009a	73.36	87.27	50.26
Incertitude	SERGEEVA et al., 2019	**89.28**	NC	NC
	QIAN et al., 2016	85,75	**73,92**	59,82
	ZOU, G. ZHOU et Q. ZHU, 2013	84.21	72.92	**67.24**
	Qiaoming ZHU et al., 2010	83,74	68,78	63,49
	MORANTE et DAELEMANS, 2009b	77.13	60.59	47.94

CoNLL-2010 Shared Task. Dans ce qui suit, nous présentons les résultats des meilleurs systèmes sur les jeux de données de la campagne d'évaluation de **CoNLL-2010** qui sont issus de **BioScope**.

L'évaluation officielle de la tâche 1 n'est pas effectuée au niveau des marqueurs mais de la phrase. Cependant, une évaluation complémentaire au niveau des marqueurs (plus précise) est fournie. Dans le tableau 1.17, nous présentons cette dernière. Malgré une amélioration par rapport à son système de 2010, VELLDAL, ØVRELID et al., 2012 ne parvient pas à égaler les résultats du système de TANG et al., 2010, qui obtient un bien meilleur rappel que les autres systèmes.

TABLEAU 1.17 – Résultats des systèmes de détection des marqueurs d'incertitude de **CoNLL-2010**

Système	Précision	Rappel	F-mesure
VELLDAL, ØVRELID et al., 2012	84,8	77,2	80,8
TANG et al., 2010	81,7	**81,0**	**81,3**
J. LI et al., 2010	83,1	78,8	80,9
X. LI et al., 2010	**87,4**	73,4	79,8
REI et BRISCOE, 2010	81,4	77,4	79,3
VELLDAL, ØVRELID et OEPEN, 2010	81,2	76,3	78,7

Concernant la seconde tâche, dont les résultats sont disponibles dans le tableau 1.18, l'évaluation était bien plus stricte puisque seuls les marqueurs et portées exactement identifiées étaient considérés comme vrais positifs par les organisateurs. Il n'existe pas de résultats au niveau des tokens individuels pour la portée. Par conséquent, les résultats sont bien plus faibles. MORANTE, VAN ASCH et DAELEMANS, 2010 obtient les meilleurs résultats lors de la campagne d'évaluation, tandis que VELLDAL, ØVRELID et al., 2012 parvient à améliorer son système et obtient les meilleurs résultats à ce jour.

TABLEAU 1.18 – Résultats des systèmes de détection de la portée de l'incertitude de **CoNLL-2010**, portées et marqueurs exactement identifiée

Équipe	Précision	Rappel	F-mesure
VELLDAL, ØVRELID et al., 2012	62,0	**57,0**	**59,4**
MORANTE, VAN ASCH et DAELEMANS, 2010	59,6	55,2	57,3
REI et BRISCOE, 2010	56,7	54,6	55,6
VELLDAL, ØVRELID et OEPEN, 2010	56,7	54,0	55,3
KILICOGLU et BERGLER, 2010	**62,5**	49,5	55,2
X. LI et al., 2010	57,4	47,9	52,2

CD-SCO. Dans ce qui suit, nous présentons les résultats des meilleurs systèmes sur le corpus **CD-SCO** de la campagne d'évaluation ***SEM-2012**.

Bien que les tâches de cette campagne d'évaluation ne portent pas sur la détection des marqueurs, les systèmes soumis reposent tous sur un sous-système de détection des marqueurs dont la description et les résultats sont donnés. Nous donnons les résultats des trois meilleurs systèmes de la campagne dans le tableau 1.19. Nous constatons que le système de CHOWDHURY, 2012 devance de peu le système de READ et al., 2012 ; LAPPONI et al., 2012, qui obtient d'ailleurs deux résultats différents (deux entraînements du même système). Cependant, l'adaptation de **BERT** de A. KHANDELWAL et SAWANT, 2020 dépasse de peu ces systèmes hybrides.

TABLEAU 1.19 – Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur le corpus **CD-SCO**

Système	Précision	Rappel	F-mesure
A. KHANDELWAL et SAWANT, 2020	NC	NC	**92,94**
CHOWDHURY, 2012	**93,41**	91,29	92,34
READ et al., 2012	91,42	92,80	92,10
LAPPONI et al., 2012	89,17	**93,56**	91,31

Les systèmes d'étiquetage de la portée sont évalués de deux façons : sur les tokens individuels de la portée, ainsi que sur les portées exactement identifiées. Nous donnons les résultats des meilleurs systèmes disponibles à ce jour dans les tableaux 1.20 et 1.20. Pour cette tâche, les systèmes proposés lors de la campagne d'évaluation sont largement dépassés par PACKARD et al., 2014, FANCELLU, LOPEZ et WEBBER, 2016 et plus récemment par H. LI et LU, 2018, autant au niveau des tokens qu'au niveau des portées exactes. A. KHANDELWAL et SAWANT, 2020 ne donnent la F-mesure de leur adaptation de **BERT** qu'au niveau des tokens. Elle est largement plus élevée que celle des autres approches, cependant, la tokenisation spécifique de **BERT** génère plus de tokens, ce qui impacte sans doute les résultats.

TABLEAU 1.20 – Résultats des systèmes de détection de la portée de la négation au niveau des tokens sur le corpus **CD-SCO**

Système	Précision	Rappel	F-mesure
A. KHANDELWAL et SAWANT, 2020	NC	NC	**92,36**
H. LI et LU, 2018 *Semi o*	94,00	85,30	89,40
H. LI et LU, 2018 *Latent io*	**94,80**	83,20	88,60
FANCELLU, LOPEZ et WEBBER, 2016	92,62	85,13	88,72
PACKARD et al., 2014	86,10	**90,40**	88,20
READ et al., 2012	81,99	88,81	85,26
LAPPONI et al., 2012	86,03	81,55	83,73

TABLEAU 1.21 – portée exacte starsem

Système	Précision	Rappel	F-mesure
H. LI et LU, 2018 *Semi o*	100	69,10	81,70
H. LI et LU, 2018 *Latent io*	100	69,50	**82,00**
FANCELLU, LOPEZ et WEBBER, 2016	99,40	63,87	77,70
PACKARD et al., 2014	98,80	65,50	78,70
READ et al., 2012	87,43	61,45	72,17
LAPPONI et al., 2012	85,71	62,65	72,39

Conclusion

Dans ce chapitre, nous avons commencé par présenter des problématiques liées au traitement automatique de la langue biomédicale. Nous avons décrit différents types de textes du domaine biomédical ainsi que les approches de **TALN** proposées dans la littérature scientifique pour les exploiter. Ces dernières années, aux articles scientifiques et documents cliniques se sont rajoutés les textes rédigés par les patients eux-mêmes sur les forums et réseaux sociaux, impliquant de nouvelles tâches et approches pour leur exploitation. Par ailleurs, les systèmes à base de règles et de terminologies, longtemps préférés aux approches par apprentissage artificiel pour de nombreuses tâches, semblent aujourd'hui être en recul par rapport à ces dernières.

Nous avons ensuite présenté plusieurs approches de **TALN** proposées pour la classification automatique de textes afin de résoudre plusieurs tâches telles que l'*opinion mining*, la catégorisation de textes ainsi que la classification multi-étiquette de commentaires toxiques et de documents législatifs. Pour ces tâches, les approches basées sur les réseaux de neurones (**CNN** et **RNN**) et sur les *transformers* (**BERT**, **XLNet**) sont aujourd'hui les plus performantes. Pour l'*opinion mining*, la catégorisation de textes et la classification de commentaires toxiques, la marge de progression est faible. Au contraire, pour les cas extrêmes de classification multi-étiquette (très nombreuses classes, *few-shot learning*, *zero-shot learning*), la marge de progression reste importante.

Dans la sous-section suivante, nous avons présenté plusieurs approches de **TALN** pour l'étiquetage de séquences au travers des deux tâches les plus étudiées. Pour l'étiquetage morpho-syntaxique comme pour la **NER**, les approches basées sur les réseaux de neurones récurrents, les *tranformers* ainsi que les plongements contextualisés sont aujourd'hui les plus performantes. Pour l'étiquetage morpho-syntaxique, la marge de progression est faible et ces méthodes n'améliorent que peu la performance des outils de **TALN**. En effet, sur les données du **Penn-Treebank**, BOHNET et al., 2018 atteint 97,96 % d'*accuracy*, alors que COLLINS, 2002 atteignait 97,11 % d'*accuracy*. À l'inverse, sur les données de la campagne d'évaluation de **CoNLL-2003**, depuis FLORIAN et al., 2003, environ 5 points de F-mesure en plus ont été obtenus par les systèmes de **NER** les plus performants, dont environ 2,5 points de plus depuis LAMPLE et al., 2016.

Dans la section suivante, nous avons abordé la classification multi-étiquette de textes cliniques en nous intéressant aux campagnes d'évaluation **CLEF eHealth** de 2016 à 2018. Nous avons d'abord présenté la Classification Internationale des Maladies. Ensuite, nous avons présenté les jeux de données annotées utilisés lors des campagnes d'évaluation. Ils sont constitués de certificats de décès. Enfin, nous avons présenté les approches proposées pour la classification de ces textes lors des campagnes d'évaluation. Ces dernières années, ce sont des approches neuronales *sequence-to-sequence* qui ont été les plus performantes sur ces jeux de données. Cependant, bien que la tâche abordée lors de ces campagnes d'évaluation et les systèmes proposés soient intéressants, les certificats de décès ne peuvent pas vraiment

être considérés comme du texte libre mais plutôt comme des documents standardisés où le texte est contrôlé. Cela est particulièrement vrai pour les données alignées du corpus **CépiDC**, qui alignent pour chaque ligne le code CIM-10 correspondant, ce qui facilite grandement la classification pour les méthodes phrastiques proposées pendant le défi. Dans le cadre de cette thèse, nous utilisons des documents cliniques contenant principalement du texte libre. Par conséquent, les systèmes proposés pendant ces campagnes d'évaluation ne semblent pas appropriés.

Enfin, la dernière section de cet état de l'art a été consacrée à l'étude de la négation et de l'incertitude. Nous avons d'abord décrit le fonctionnement de ces opérations linguistiques. Comme nous l'avons expliqué, ces opérations linguistiques sont complexes pour plusieurs raisons allant de la variété des marqueurs à la portée discontinue. Ensuite, nous avons présenté les jeux de données annotées et librement accessibles. De tels jeux de données sont peu nombreux et principalement en anglais. Enfin, nous avons présenté les approches proposées pour la détection des marqueurs de négation et d'incertitude et de leur portée ainsi que les résultats qu'elles obtiennent. Les systèmes proposés lors des campagnes d'évaluation se ressemblent (utilisation de **SVM**, **CRF**, etc.), et les systèmes les plus performants sont ceux qui exploitent le mieux les descripteurs lexicaux, syntaxiques ou contextuels. Depuis la résurgence récente des réseaux de neurones, plusieurs approches ont été proposées et obtiennent des résultats intéressants. Dernièrement, plusieurs approches basées sur **BERT** ont été proposées et obtiennent de très bon résultats.

Chapitre 2

Classification multi-étiquette de textes cliniques

Sommaire

2.1 Jeu de données 61
2.2 Nos approches 65
2.2.1 Approche par dictionnaires 65
2.2.2 Approches par apprentissage supervisé 66
2.3 Expériences de classification CIM-10 71
2.3.1 Protocole expérimental 71
2.3.2 Analyse des résultats 74

En 1.2, nous présentions brièvement le **PMSI**, outil épidémiologique et de connaissance de l'activité des établissements de santé, qui est utilisé comme outil d'allocation budgétaire depuis 2005. les établissements de santé français sont actuellement rémunérés en fonction des codes **CIM-10** qu'ils rapportent. Ainsi, le développement d'outils d'aide au codage automatique de textes cliniques français avec la **CIM-10** est devenu une nécessité pour les établissement de santé.

Cette tâche de classification est une thématique de recherche à laquelle les chercheurs en informatique médicale s'intéressent depuis plus de 20 ans (LIMA, LAENDER et RIBEIRO-NETO, 1998). En 1.1.2, nous avons présenté les méthodes de classification automatique de textes proposées par les chercheurs afin de résoudre différentes tâches. En 1.2.1, nous avons présenté la Classification Internationale des Maladies, une classification médicale hiérarchisée publiée par l'Organisation Mondiale de la Santé (**OMS**) et utilisée par de nombreux pays membres. En 1.2.2 et 1.2.3, nous avons présenté les jeux de données annotées ainsi que les méthodes proposées dans le cadre des campagnes d'évaluation du **CLEF eHealth Evaluation Lab** de 2016 à 2018 et leurs résultats. Cependant, les textes cliniques français proposés dans le cadre de ces campagnes, bien d'intéressants, ne correspondent pas à du texte libre mais plutôt à du texte contrôlé dans une structure standardisée. Par conséquent, les approches phrastiques proposées lors de ces campagnes nous semblent trop spécifiques pour être utilisées dans le cadre de nos travaux, et notamment peu à même de de gérer la grande variété d'expressions que l'on peut trouver dans des dossiers cliniques pour évoquer un évènement médicale encodable par la CIM-10.

Dans ce chapitre, qui s'appuie principalement sur DALLOUX, CLAVEAU, CUGGIA et al., 2020, nous proposons plusieurs approches pour la classification multi-étiquette de textes cliniques par apprentissage artificiel que nous entraînons et testons sur des données réelles provenant du Centre Hospitalier Universitaire (**CHU**) de Rennes. Ainsi, dans la section 2.1, nous décrivons le contenu du corpus de textes cliniques du **CHU** à l'aide d'une note clinique factice et de statistiques. Ensuite, dans la section 2.2, nous commençons par présenter l'approche par dictionnaires qui constitue notre point de référence, puis nous présentons les systèmes de classification multi-étiquette par apprentissage supervisé que nous développons. Enfin, dans la section 2.3, nous présentons le protocole expérimental et l'analyse des résultats obtenus pour la tâche classification multi-étiquette de textes cliniques.

2.1 Jeu de données

Le corpus dont nous disposons contient 28 000 textes cliniques créés au **CHU** de Rennes en 2016. Chaque document contient une ou plusieurs notes pour un seul patient issues d'un seul séjour à l'hôpital. Les textes cliniques que nous manipulons sont protégés par le secret médical. Par conséquent, nous ne pouvons pas illustrer ce manuscrit avec des exemples réels. Cependant, bien que la note clinique présentée dans la figure 2.1 soit factice, le format ainsi que le contenu de cette note sont conformes à ceux des textes cliniques de notre jeu de données.

Dans l'en-tête du document, nous retrouvons les statuts (Professeur, Docteur, interne), prénom et nom des praticiens ayant participé aux actes médicaux réalisés sur le patient. Nous y retrouvons aussi la spécialité concernée, ici la néphrologie, ainsi que le lieu et la date de rédaction de la note. Le contenu du compte-rendu d'hospitalisation commence par identifier le patient (prénom, nom, date de naissance), le contexte de son hospitalisation et la date. Ensuite, les antécédents médicaux du patient (*insuffisance rénale*, *hypertension artérielle*, *orchiectomie*, etc.) et traitements habituels (*ARA II*, *LÉVOTHYROX*) sont indiqués. Puis, le traitement ou l'opération proposés au patient sont renseignés, ici une greffe de rein. Le déroulement de l'opération est ensuite décrit. La date de l'opération, le temps opératoire, le docteur responsable et le résultat de l'opération sont donnés. Dans le cas présent, les suites de l'opération, du fait des complications, sont décrites (*extubation*, *diurèse*, *écho-doppler*, etc.). Enfin, des informations concernant le CHU de Rennes (adresse, numéro de téléphone/de fax, services concernés, etc.) sont données. D'autres informations, telles que les résultats d'examens de biologie médicale présentés dans la figure 2.2, sont souvent contenues dans ce type de textes.

Nous constatons que ces textes cliniques sont très différents des certificats de décès présentés dans le tableau 1.2. En effet, là où les certificats de décès sont composés de phrases succinctes parfois très proches de l'intitulé des codes **CIM-10**, les notes cliniques présentent bien plus d'informations relatives aux patients, aux diagnostics et à leurs prises en charge. Elles sont verbeuses et contiennent beaucoup de bruit. En conséquence, et bien que les méthodes de tokenisation puissent différer, nos documents sont beaucoup plus longs que les certificats de décès. D'après les statistiques présentées dans les tableaux 1.1 et 2.1, les certificats contiennent en moyenne 10 tokens et 4 codes par document, tandis que nos textes cliniques sont composés en moyenne de 1 345 tokens par document et deux fois plus de codes. Par ailleurs, nous constatons, d'une part, que le corpus **CépiDC** compte moins de 4 000 codes **CIM-10** uniques alors que notre jeu de données en compte 6 113, et, d'autre part, nous constatons que la majorité des codes sont très peu représentés dans le corpus du CHU. En effet, 3 885 des 5 735 codes présents dans les données d'entraînement apparaissent moins de 10 fois et seuls 382 codes comptent plus de 100 occurrences. La figure 2.3 illustre ce problème de représentation. Enfin, là où l'ensemble de test du corpus **CépiDC** compte 70 codes non vu dans l'ensemble de test, notre jeu de données en compte 212.

Dans la section 2.2, nous présentons les approches qui seront utilisées pour la classification multi-étiquette des textes que nous venons de décrire.

Professeur [PRÉNOM NOM].
Docteur [PRÉNOM NOM]. Praticien Hospitalier.
Docteur [PRÉNOM NOM].
NEPHROLOGIE. PONTCHAILLOU.
[NOM PRÉNOM]. Interne.
Rennes, le [DATE].
COMPTE-RENDU D'HOSPITALISATION.
Monsieur [PRÉNOM NOM] né le [DATE] a été à hospitalisé dans le service de Réanimation Chirurgicale le [DATE] en postopératoire d'une greffe rénale. Monsieur [NOM] présente une insuffisance rénale chronique d'origine hypertensive diagnostiquée en [DATE], il est dialysé 6 fois par semaine. Le patient porte une fistule artério-veineuse radio céphalique droite.
Antécédents médico-chirurgicaux et facteurs de risque cardio-vasculaires : hypertension artérielle, parathyroïdectomie simple, orchiectomie droite pour tumeur.
Traitements habituels : ARA II, LÉVOTHYROX.
Une greffe est proposée au patient. Le bilan est satisfaisant. Le geste a lieu le [DATE] (Docteur [NOM]). La dernière dialyse a eu lieu dans la matinée du [DATE]. Le temps opératoire est sans particularité (ischémie froide 5h30, ischémie chaude 25 minutes). Le greffon rénal se recolore immédiatement mais la diurèse ne reprend pas sur table. Évolution en service de Réanimation Chirurgicale. L'extubation est réalisée immédiatement. Le problème suivant est la diurèse. Le patient nécessitera un remplissage important, la diurèse ne se produira qu'à la suite de l'administration d'un diurétique, moyennement efficace étant donné que la fonction rénale n'est pas corrigée. Un écho-doppler est réalisé le soir de la greffe. Il est normal mais ne décrit que le flux artériel. Étant donné ces difficultés, un second doppler est demandé le jour suivant, afin de voir les axes veineux. Après cet examen, Monsieur [NOM] pourra être pris en charge en Néphrologie.
Copie pour information : Docteur [NOM PRÉNOM], Pontchaillou CHU RENNES.
CENTRE HOSPITALIER UNIVERSITAIRE DE RENNES.
Hôpital PONTCHAILLOU - [ADRESSE] - Standard [#TELEPHONE].
REANIMATION CHIRURGICALE - Centre Urgences Réanimations - Niveau 1.
Secrétariat : [#TELEPHONE] - [#FAX].

FIGURE 2.1 – Note clinique factice

Numération formule sanguine : hémoglobine 14 g/dl, plaquettes 220 Giga/l . leucocytes 2.8 Giga/l. Coagulation : TP 103 %, ratio TCA 1.1, ddimères 0.54 µg/ml, fibrinogène 3.26 g/l. Acideurique : 424 µmol/l. Calcium : 2.32 mmol/l. Ionogramme sanguin : natrémie 141 mmol/l, kaliémie 3.9 mmol/l, urée 4.3 mmol/l. créatinine 75 µmol/l avec une clairance à 98.1 ml/min. Troponine : négative ; HBA1C : 5.2 % ; TSH : 1.40 µUI/ml ;. VS : 24 mm CRP 2.8 mg/l. Bilan lipidique : cholestérol total 5.39 mmol/l, LDL cholestérol 1.30 g/l, HDL. cholestérol : 1.25 mmol/l, triglycérides 2.45 mmol/l. Bilan hépatique : bilirubine totale 15.5 µmol/l, ASAT 38 UI/l, ALAT 34 UI/l . phosphatases alcalines 59 UI/l, Gamma-GT 119 UI/l, CPK 51 UI/l. Bandelette urinaire : négative, pas de protéinurie.

FIGURE 2.2 – Résultats d'examens de biologie médicale

TABLEAU 2.1 – Statistiques descriptives du jeu de données du **CHU**

	Entraînement	Validation	Test	Total
Documents	22 400	2 800	2 800	28 000
Tokens	30 307 091	3 718 715	3 654 599	37 680 405
Codes CIM-10 (total)	182 112	23 010	22 561	227 683
Codes CIM-10 (unique)	5 735	2 824	2 887	6 113
Codes CIM-10 non vu (unique)	-	191	212	-
Codes avec - de 10 exemples	3 885	2 320	2 391	4 018
Codes avec + de 100 exemples	382	32	31	455

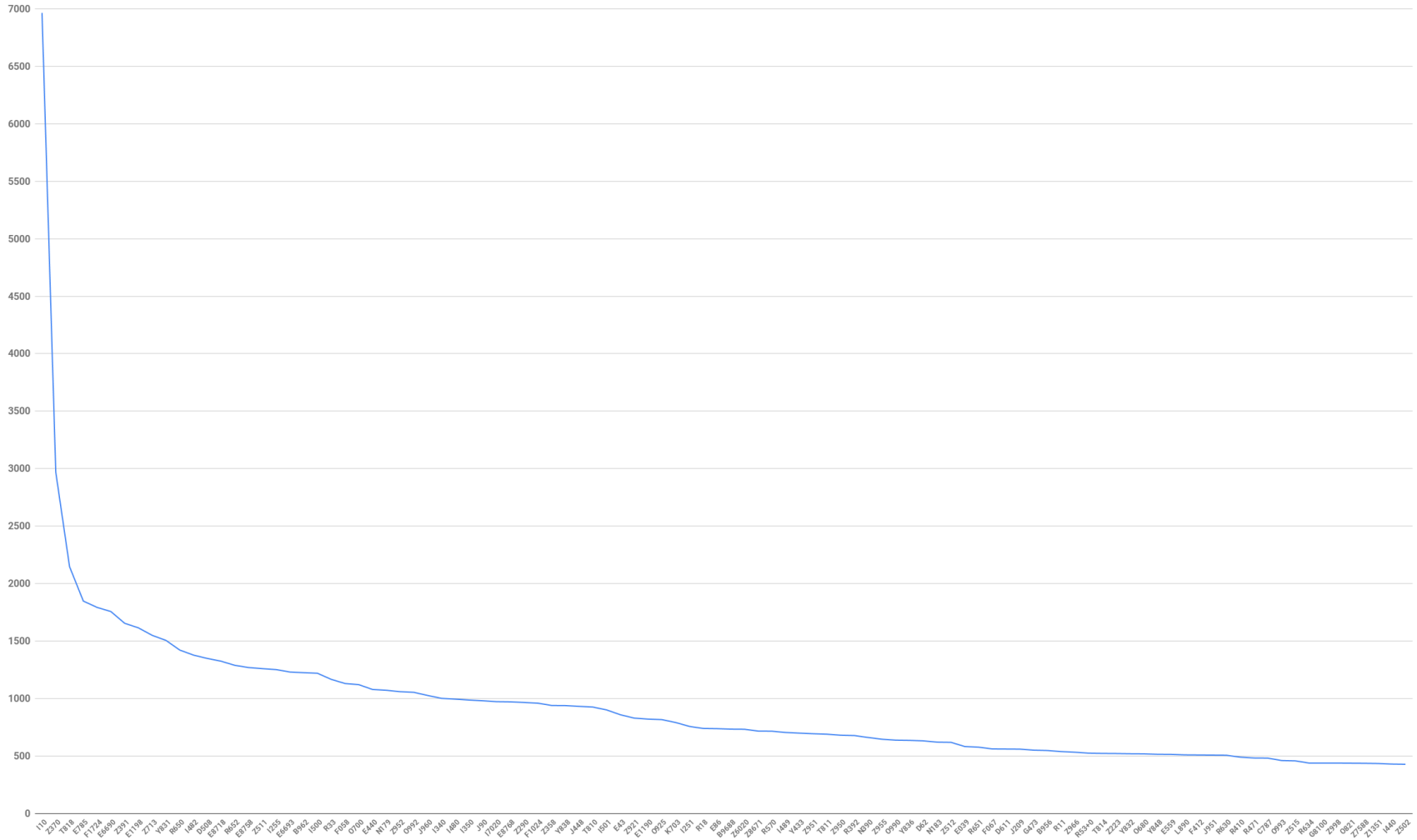


FIGURE 2.3 – Courbe du nombre d'occurrences par code pour les 100 codes **CIM-10** les plus fréquents dans le jeu de donnée du **CHU**

2.2 Nos approches

Dans cette section, nous présentons différentes approches que nous développons pour résoudre la tâche de classification. En 2.2.1, nous présentons l'approche par dictionnaires qui constitue notre point de référence (*baseline*). Si les approches par apprentissage artificiel sont aujourd'hui prédominantes, plusieurs approches de ce type ont été proposées au cours des campagnes d'évaluation du **CLEF eHealth Evaluation Lab**, soit en méthode principale, soit en complément d'une approche par apprentissage (VAN MULLIGEN et al., 2016 ; ZWEIGENBAUM et LAVERGNE, 2017 ; COSSIN et al., 2018). Puis, en 2.2.2, nous présentons différents modèles pour l'apprentissage de plongements de mots et le modèle pré-entraîné que nous utilisons, puis, nous présentons les systèmes de classification par apprentissage supervisé que nous développons.

2.2.1 Approche par dictionnaires

Inspirée de COSSIN et al., 2018, bien que moins sophistiquée, notre approche utilise plusieurs dictionnaires **CIM-10** ainsi que la distance de Levenshtein pour attribuer des codes à chaque phrase de nos textes cliniques. Cette approche servant simplement de point de référence, la segmentation de nos documents en phrases n'est pas non plus très sophistiquée. Ainsi, nous séparons chaque phrase sur la base des signes de ponctuation de fin de phrase et des sauts de ligne.

Les dictionnaires **CIM-10** utilisés sont, d'une part, les codes et leur définition, tels que publiés sur les sites de l'Agence technique de l'information sur l'hospitalisation [1] ou **Wikipedia** [2], et, d'autre part, le *dictionnaire2015.csv* mis à disposition lors des campagnes d'évaluation du **CLEF eHealth Evaluation Lab**. Ce dictionnaire contient 147 342 entrées relatives à 6 291 codes. 3 500 de ces 6 291 codes sont présents dans notre dataset, alors que 2 613 codes présents dans notre jeu de données ne sont pas représentés dans le *dictionnaire2015.csv*.

Dans le cadre de ces travaux, nous avons essayé de retrouver les entrées de ces dictionnaires dans nos textes par correspondance exacte (*exact matching*). Utilisée par COSSIN et al., 2018 sur les certificats de décès lors du **CLEF eHealth Evaluation Lab 2018**, cette approche obtient de bon résultats. Cependant, dans le cas présent, elle ne retourne que quelques codes pour l'intégralité du jeu de données de test. Le contenu de nos documents explique cette absence de résultats. En effet, comme nous l'avons vu en 1.2.2, les certificats de décès sont proprement segmentés en phrases et le contenu de ces documents est très proche des intitulés des codes **CIM-10**. Au contraire, le contenu de nos documents, présenté en 2.1, est très différent. En conséquence, nous décidons d'utiliser une méthode de recherche plus approximative, par distance de Levenshtein. Utilisée par l'approche de COSSIN et al., 2018 afin de détecter les fautes de frappe, la distance de Levenshtein (LEVENSHTEIN, 1966) entre deux chaînes de caractères est le nombre minimum de modifications de caractères (insertions, suppressions ou substitutions) nécessaires pour passer d'une chaîne à

1. https://www.atih.sante.fr/cim-10-fr-2020-usage-pmsi
2. https://fr.wikipedia.org/wiki/Liste_de_codes_CIM-10

l'autre. Ainsi, elle permet de calculer la similarité entre les entrées des dictionnaires et les phrases de nos textes cliniques. Le ratio retourné par cette mesure de similarité nous permet de valider ou non chaque code pour chaque phrase selon un seuil défini empiriquement.

2.2.2 Approches par apprentissage supervisé

Approches pour la représentation vectorielle des mots

Diverses méthodes ont été utilisées pour représenter les mots ou les textes en tant que vecteurs en entrée des classifieurs. Mentionnons par exemple les modèles de sacs de mots, tels que le nombre d'occurrences de chaque token dans chaque document ou **TF-IDF** (*term frequency–inverse document frequency*), auxquels l'allocation de Dirichlet latente (BLEI, NG et JORDAN, 2003) ou l'analyse sémantique latente (DEERWESTER et al., 1990) peuvent être appliquées. Même si ces approches continuent d'être utilisées, de nouveaux modèles proposant de meilleures représentations des relations sémantiques entre les mots ont été proposés. Dans le cadre de nos travaux, nous utilisons plusieurs représentations vectorielles de mots basées sur les plongements de mots (*word embeddings*), dans différentes langues. Ces techniques de modélisation du langage et d'apprentissage de descripteurs, où chaque token est représenté par un vecteur de nombres réels, permettent notamment de réduire la dimensionnalité de l'espace vectoriel, ce qui est crucial en apprentissage automatique pour lutter contre le fléau de la dimension.

Word2vec (MIKOLOV et al., 2013) est un groupe de modèles visant à produire des plongements de mots à partir du texte brut. L'architecture de ces modèles est composée d'un réseau de neurones artificiels formé de deux couches qui est entraîné dans le but de capturer le contexte linguistique des mots. **Word2vec** prend en entrée un grand volume de textes et produit un espace vectoriel, généralement de plusieurs centaines de dimensions, où chaque mot unique du corpus se voit attribuer un vecteur lui correspondant dans l'espace. Les vecteurs de mots sont positionnés dans l'espace vectoriel de sorte que les mots qui partagent des contextes communs dans le corpus soient situés les uns à côté des autres. Les vecteurs de mots peuvent être calculés à l'aide de deux modèles d'apprentissage : le modèle **CBOW** (sacs de mots continus), qui prédit le mot cible à partir des mots du contexte, et le modèle **Skip-Gram**, qui prédit les mots du contexte à partir du mot cible. D'après les auteurs, **CBOW** est plus rapide que **Skip-Gram**, cependant, ce dernier fait un meilleur travail pour les mots peu fréquents.

FastText (BOJANOWSKI et al., 2017) est une bibliothèque logicielle pour l'apprentissage de plongements de mots et la classification de textes. Elle propose de résourde le principal problème de **word2vec** : les mots qui n'apparaissent pas dans le vocabulaire ne peuvent pas être représentés par le modèle. En effet, **word2vec** représente chaque mot rencontré par un vecteur unique sans tenir compte de la structure morphologique des mots. **FastText** répond à cette limitation en représentant chaque mot comme un sac de tous les n-grammes de caractères possibles qu'il contient. Le

mot est complété à l'aide d'un ensemble de symboles uniques (* dans l'exemple qui suit) qui aide à distinguer les préfixes et suffixes. La séquence complète est également ajoutée au sac de n-grammes. Par exemple, si $n = 3$, le mot *prélèvements* donnera le sac de tri-grammes suivant : [*pr, pré, rél, élè, lèv, ève, vem, eme, men, ent, nts, ts* , *prélèvements*]. Le vecteur prend désormais en compte chaque n-gramme de caractères et le vecteur du mot est la somme de tous les vecteurs n-grammes de caractères du mot. Avec un corpus suffisamment grand, tous les n-grammes de caractères possibles peuvent être couverts et, étant donné que les représentations entre les mots sont souvent partagées, les mots rares peuvent également obtenir des représentations fiables. À l'instar de **word2vec**, les vecteurs de mots sont calculés par **CBOW** ou par **Skip-Gram**.

Le modèle de plongements de mots pré-entraînés, que nous utilisons pour cette tâche, est un modèle **fastText** que nous entraînons sur les 28 000 compte-rendus d'hospitalisations de notre jeu de données à l'aide de **Gensim** (REHUREK et SOJKA, 2010). Le modèle est entraîné avec l'algorithme *Skip-Gram* et les paramètres suivants : 300 dimensions, une fenêtre contextuelle de 5 mots avant et après chaque mot, un décompte minimum de cinq occurrences pour chaque mot et un échantillonnage négatif (*negative sampling*).

Approche par réseau de neurones convolutifs

Largement utilisés en vision artificielle et traitement d'images, les réseaux de neurones convolutifs (**CNN**) (LECUN et al., 1989) ont également été utilisés avec succès dans plusieurs tâches de **TALN** (COLLOBERT et al., 2011 ; Y. KIM, 2014 ; CONNEAU, SCHWENK et al., 2017). Les **CNN** reposent principalement sur deux opérations : la convolution et le *pooling*.

La convolution est une opération mathématique permettant de combiner deux signaux pour en former un troisième. Ici, les deux signaux en question sont les entrées du réseau, c'est-à-dire une matrice de vecteurs de mots par document, et les noyaux/filtres. Un filtre est une matrice de taille prédéfinie (taille de filtre/région x dimension des plongements de mots) dont les valeurs sont initialisées aléatoirement et deviennent des paramètres qui seront appris par le réseau. Chaque filtre agit comme une fenêtre coulissante se déplaçant sur la matrice du texte traité. À chaque déplacement, le produit matriciel de Hadamard entre les valeurs des deux matrices est calculé, puis additionné. Les résultats de la convolution d'un filtre sur la matrice d'un document sont donc les moyennes pondérées des vecteurs de mots obtenues après tous les déplacements. Une fonction d'activation non-linéaire, telles que *tahn* (tangente hyperbolique) ou *ReLu* (*rectified linear unit*), est ensuite appliquée sur ces résultats pour générer les *feature maps*.

Dans un **CNN**, la fonction de *pooling* sert à remplacer les valeurs de chaque *feature map* par une ou plusieurs valeurs relatives aux valeurs remplacées. Le *max-pooling* remplace les valeurs concernées par la valeur la plus élevée, par exemple le bloc [9,4,5,7] est remplacé par [9]. L'*average-pooling* fait la moyenne de ces valeurs ([6,25]).

L'architecture de notre **CNN** est présentée dans la figure 2.4. Nous entraînons ce modèle avec plusieurs ensembles d'hyperparamètres pendant 30 périodes d'entraînement chacun. Nous utilisons : (1) quatre tailles de filtre (2, 3, 4, 5), 100, 1 000 ou 2 000 filtres de chaque taille et (2) le max pooling global qui remplace toutes les valeurs de chaque *feature map* par leur maximum. Après le *pooling*, une couche entièrement connectée avec 100, 1000 ou 2000 unités cachées est utilisée afin d'apprendre les combinaisons non-linéaires des paramètres appris lors de l'opération de convolution. La prédiction est effectuée par une couche entièrement connectée avec une fonction d'activation sigmoïde. La fonction d'activation sigmoïde calcule la probabilité de chaque classe pour chaque document indépendamment les unes des autres. En effet, pour chaque classe, la couche sigmoïde sort un nombre entre 0 et 1, et la classe est attribuée si ce nombre est supérieur ou égal à 0,5. Elle permet donc en théorie d'attribuer plusieurs codes à chaque texte, ce qui est évidemment important pour notre application au codage **PMSI**. En outre, plusieurs régularisations par *dropout* sont appliquées au cours de l'entraînement. Le *dropout* ou abandon en français, est une technique de régularisation qui vise à réduire le surapprentissage dans les réseaux de neurones en évitant les co-adaptations complexes sur les données d'entraînement (HINTON et al., 2012). Le terme *dropout* fait référence à l'abandon temporaire et aléatoire d'unités du réseau de neurones et de leurs connexions durant l'entraînement. Nous utilisons un *dropout* de 0,5 (50 % des unités et de leurs connexions sont désactivées temporairement), valeur qui donne le plus haut niveau de régularisation (BALDI et SADOWSKI, 2013).

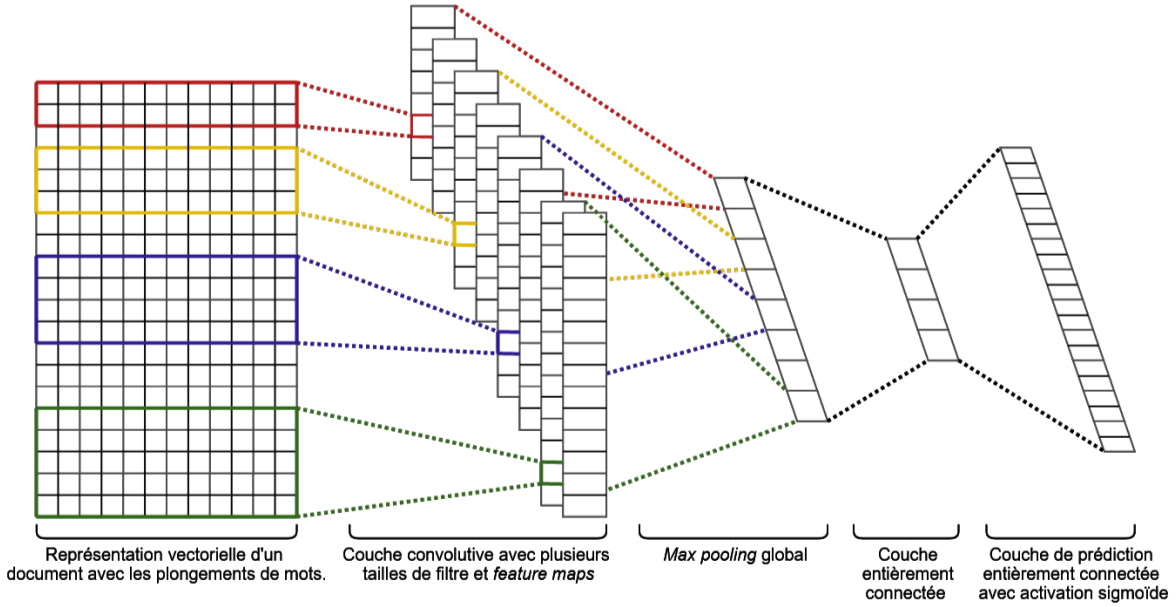


FIGURE 2.4 – Architecture de notre **CNN**

Pour les tâches de classification multi-étiquette extrême telle que celle-ci, les performances des systèmes proposés dépendent directement de leur capacité à détecter les descripteurs pertinents pour chaque classe. Dans le cas présent, nos approches

doivent identifier, dans de longs textes cliniques (1 345 tokens en moyenne), les tokens pertinents pour chaque classe (8 classes par document en moyenne). Par conséquent, une approche par **CNN** nous semblent tout indiquée. En effet, le **CNN** est conçu pour identifier les descripteurs locaux pertinents dans une structure large et les combiner pour produire une représentation vectorielle de taille fixe de la structure, capturant ainsi les aspects locaux qui sont les plus informatifs pour la tâche à accomplir (GOLDBERG, 2017).

Approche par réseau de neurones récurrents bidirectionnel

En théorie, un réseau de neurones récurrents (*recurrent neural network*, **RNN**) est un réseau qui est capable d'adapter sa décision en tenant compte des données vues précédemment, en plus des données actuellement vues. Cette opération est mise en œuvre par le report de l'état de la couche cachée de l'étape précédente à l'étape présente de l'apprentissage et ainsi de suite.

Cependant, la cellule **RNN** « classique » souffre du problème de la disparition du gradient et ne parvient pas à mémoriser à long terme les informations vues précédemment. Plusieurs architectures de cellule ont été développées pour répondre à ce problème. Dans le cadre de cette tâche, nous exploitons la cellule *long short-term memory* (**LSTM**). La cellule **LSTM** (HOCHREITER et SCHMIDHUBER, 1997) est très efficace pour l'apprentissage de dépendances à long terme grâce à son architecture à base de « portes ». La « porte d'oubli » décide quelles informations vont être retirées de l'état de la cellule de l'étape précédente. La « porte d'entrée » décide quelles informations de l'état de la cellule serons mises à jour. La « porte de sortie » décide quelles parties de l'état de la cellule vont être gardées en sortie de la couche récurrente.

Les réseaux de neurones récurrents bidirectionnels (**BiRNN**) (Mike SCHUSTER et PALIWAL, 1997) connectent deux couches cachées de directions opposées (la séquence est lue à l'endroit et à l'envers) à la même sortie. Par conséquent, la couche de sortie peut obtenir simultanément des informations sur les états passés et futurs. Ce contexte additionnel permet généralement aux **RNN** d'apprendre mieux et plus rapidement à résoudre une tâche.

Dans le cas présent, notre **BiRNN**, présenté dans la figure 2.5, utilise des cellules **LSTM** (**BiLSTM**) et les couches récurrentes (avant/arrière) sont constituées de 600 unités cachées chacune. À l'instar du **CNN**, la prédiction est effectuée par une couche entièrement connectée avec une fonction d'activation sigmoïde et les mêmes *dropout* sont appliqués.

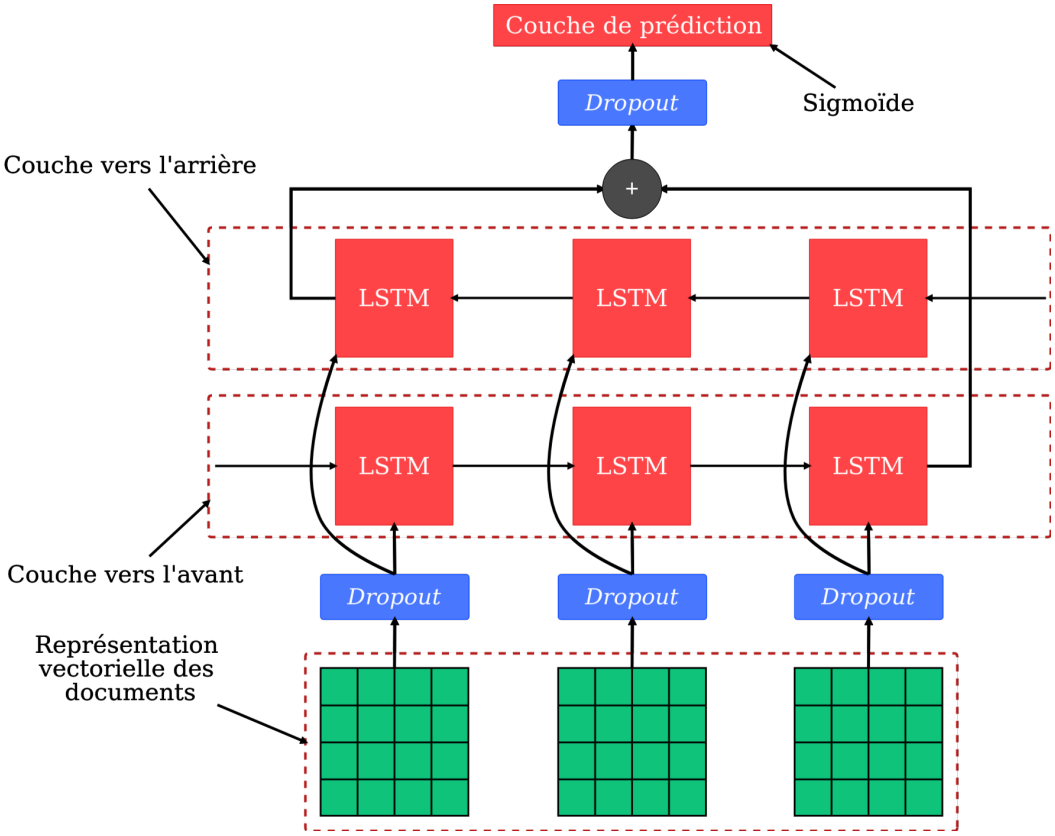


FIGURE 2.5 – Architecture de notre **BiLSTM**

Dans la section 2.3, nous présentons notre protocole expérimental ainsi que les résultats obtenus par les approches que nous venons de présenter sur les données décrites en 2.1.

2.3 Expériences de classification CIM-10

2.3.1 Protocole expérimental

Le but de nos travaux est de concevoir une approche efficace pour la classification multi-étiquette de textes cliniques. Cependant, d'une part, la composition des textes que nous exploitons rend difficile la classification par dictionnaire, et, d'autre part, le manque d'exemples d'entraînement pour la majorité des classes à attribuer rend l'apprentissage supervisé difficile. Dans ce contexte, nous proposons d'aborder ce problème de deux façons. La première revient à utiliser les données d'origine, c'est-à-dire à classifier les textes cliniques avec la totalité des classes ; la seconde simplifie la tâche en réduisant le nombre de classes soumises au système de détection automatique. Pour chaque tâche, le seuil de l'approche par dictionnaires pour l'attribution des codes est défini empiriquement. Cela représente 80 % pour la première et 60 % pour la seconde. Les mesures d'évaluation standards sont utilisées : la précision, le rappel et la F-mesure.

Réduction de niveau hiérarchique

Comme nous l'avons vu dans la sous-section 1.2.1, la classification internationale des maladies est une classification médicale hiérarchisée composée de 22 chapitres. Il serait donc possible de ramener chaque code au niveau du chapitre auquel il appartient et de n'avoir que 22 classes à attribuer. Cependant, une classification à ce niveau hiérarchique n'aurait que peu d'intérêt d'un point de vue applicatif (pour les hôpitaux) puisque l'information apportée serait très générique. La réduction de niveaux hiérarchiques que nous proposons consiste à réduire à deux chiffres les codes à trois et quatre chiffres. Par exemple, les codes **A01.0** (Fièvre typhoïde), **A01.1** (Paratyphoïde A), **A01.2** (Paratyphoïde B), **A01.3** (Paratyphoïde C) et **A01.4** (Paratyphoïde, sans précision) sont tous ramenés au niveau hiérarchique supérieur, **A01** (Fièvres typhoïde et paratyphoïde). Par conséquent, seuls les codes A00 à Z99 sont exploités, ce qui réduit le nombre de classes du corpus à 1 549 au total, dont 1 501 sont présentes dans les données d'entraînement. Dans point de vue pratique, la réduction des codes **CIM-10** à ce niveau hiérarchique présente deux avantages. D'une part, cela permet d'augmenter le nombre d'exemples par classe tout en réduisant le nombre de classes soumises au système de classification automatique. En outre, d'un point de vue applicatif, à l'inverse des chapitres, les codes de ce niveau restent assez spécifiques pour constituer des classes pertinentes dans le cadre d'un outil d'aide au codage. Cependant, nous observons qu'ils ne sont pas suffisamment précis pour être utilisé dans le cadre d'un outil de classification automatique. En effet, dans notre jeu de donnée, les codes à 3 caractères ne constituent qu'environ 10 % du nombre total de codes et 164 des 6113 classes. Il s'agit principalement de codes n'ayant pas de descendants dans la hiérarchie tels que **A35** (Autres formes de tétanos), **A46** (Érysipèle) ou **I10** (Hypertension essentielle (primitive)), code le plus utilisé dans notre jeu de données. Nous présentons quelques statistiques descriptives supplémentaires dans le tableau 2.2. Par rapport aux statistiques présentées dans le tableau 2.1, après la réduction de niveau hiérarchique, nous constatons une baisse d'environ 6 % du nombre total de codes. Cela montre que, dans notre jeu

de données, qui représente un sous-ensemble représentatif des données du CHU de Rennes, peu de codes à 4 caractères et plus appartenant à la même hiérarchie à trois caractères se retrouvent dans les mêmes documents. En conséquence, environ 25 % des classes ont plus de 100 exemples contre environ 7 % sans réduction et environ 38 % des classes ont moins de 10 exemples contre environ 66 % sans réduction. Autre élément démontrant l'amélioration de la représentativité des classes avec la réduction, la courbe présentée dans la figure 2.6 chute bien moins brutalement que celle de la figure 2.3.

TABLEAU 2.2 – Statistiques descriptives du jeu de données du **CHU** pour la seconde tâche

	Entraînement	Validation	Test	Total
Codes CIM-10 (total)	171 346	21 680	21 255	214 281
Codes CIM-10 (unique)	1 501	1 057	1 087	1 549
Codes CIM-10 non vu (unique)	-	31	20	-

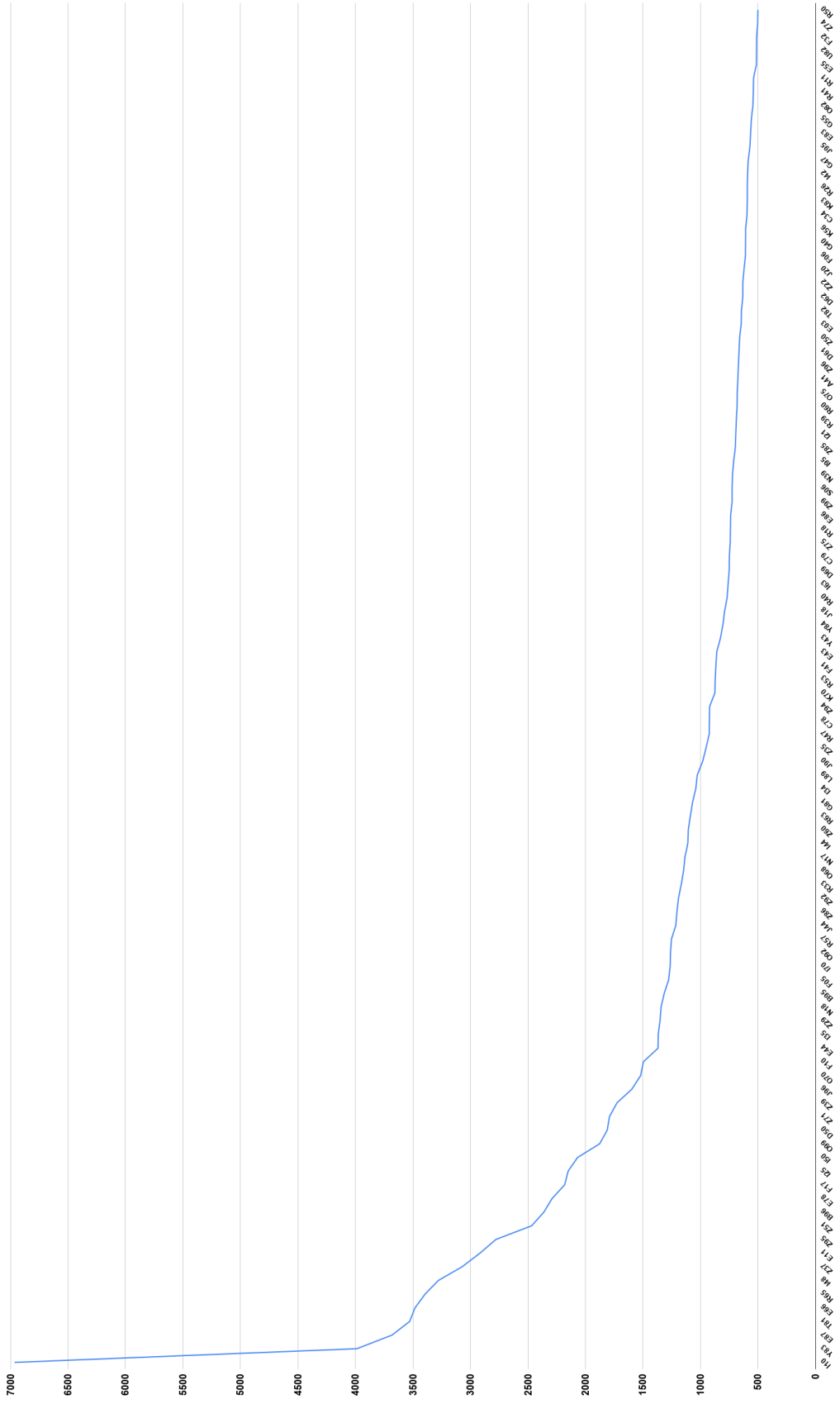


FIGURE 2.6 – Courbe du nombre d'occurrences par code pour les 100 codes **CIM-10** les plus fréquents après la réduction de niveau hiérarchique

2.3.2 Analyse des résultats

Les résultats obtenus par nos approches pour la tâche de classification multi-étiquette de textes cliniques sont présentés dans le tableau 2.3. Notons que seuls les résultats des approches neuronales sont publiés dans DALLOUX, CLAVEAU, CUGGIA et al., 2020.

Nous constatons que les résultats de notre approche par dictionnaires (**DICT**) sont très faibles. La segmentation en phrase étant peu précise et l'approche peu sophistiquée, ces résultats ne nous étonnent pas. Comme nous l'avons vu en 1.2.3, les approches de ce type (VAN MULLIGEN et al., 2016 ; COSSIN et al., 2018) obtiennent de bon résultats sur les certificats de décès. Cependant, les résultats que nous obtenons montrent que ce type d'approches est difficilement adaptable à nos documents pour la tâche en question du fait de leur faible capacité à gérer des textes libres, et potentiellement bruités (abréviations non conventionnelles, fautes de frappe, etc.).

Nous constatons aussi que notre **CNN** est de loin l'approche la plus performante pour les deux tâches avec une F-mesure de 17 points supérieures à celle du **BiLSTM** avec le paramétrage le plus performant. Nos expérimentations montrent également que l'augmentation des valeurs des hyperparamètres est bénéfique. En effet, augmenter le nombre de filtres par taille de filtre permet de gagner 3,33 points de F-mesure (**CNN-100-100** Vs. **CNN-1000-100**). Augmenter le nombre d'unités de la couche entièrement connectée permet aussi d'améliorer les performances du **CNN** (+1,41 point de F-mesure, **CNN-100-100** Vs. **CNN-100-1000**). En combinant ces deux hyperparamètres (**CNN-1000-1000**), le gain total est de **7,53** points de F-mesure. Cependant, en doublant ces paramètres, le gain n'est plus que de 0,78 point de F-mesure pour un temps de calcul bien plus long. Il est donc peu probable qu'augmenter davantage ces hyperparamètres soit bénéfique.

Comme nous nous y attendions, les résultats obtenus pour la seconde tâche sont bien plus élevés. +16,27 points de F-mesure pour le **BiLSTM** et +12,44 points de F-mesure pour le **CNN**. La réduction du nombre de classes nous semble être une option valide pour réduire la complexité de la tâche tout en gardant un niveau de détail suffisamment élevé pour aiguiller les annotateurs hospitaliers. Cependant, d'une part, parmi les 1 500 classes restantes, de nombreuses classes sont toujours sous-représentées dans le jeu de données. Collecter plus de documents annotés pour l'entraînement pourrait permettre de résoudre ce problème mais les codes sous-représentés dans notre corpus sont sans doute peu utilisés par le **CHU** et seraient aussi sous-représentés dans les nouvelles données. Une autre possibilité serait d'utiliser des modèles spécifiquement entraînés pour le *few-shot learning* et le *zero-shot learning*, tels que ceux présentés dans CHALKIDIS et al., 2019, afin de détecter les classes pas ou peu représentées dans nos documents. D'autre part, dans les rares cas, où plusieurs codes appartenant au même code hiérarchiquement supérieur sont présents dans un document, nous perdons possiblement des informations importantes car nous ne pouvons affecter chaque classe qu'une seule fois à chaque document.

Enfin, les résultats que nous obtenons sont bien inférieurs à ceux des systèmes

proposés lors de la campagne d'évaluation de **CLEF eHealth 2018** pour plusieurs raisons. D'une part, comme nous l'avons démontré, les certificats de décès sont des textes très courts et structurés tandis que nos textes cliniques sont plutôt longs, verbeux et bruités, ce qui complexifie la tâche. D'autre part, certains codes pourraient ne pas être expliqués par les données textuelles mais plutôt par des données structurées des dossiers patients informatisés. Il faudrait donc pouvoir incorporer ces données en entrée de nos systèmes de classification. Il est aussi possible que l'annotation servant de référence (pour entraîner et évaluer) manque de régularité. En effet, par gain de temps, les codes non-rémunérateurs, et donc peu prioritaires dans l'usage comptable actuel du PMSI, ne sont pas systématiquement renseignés par certains annotateurs alors qu'ils sont renseignés par d'autres. Ces codes peuvent donc être attribués par nos systèmes à des documents pour lesquels ils n'ont pas été annotés.

TABLEAU 2.3 – Résultats obtenus par nos systèmes pour la tâche de classification multi-étiquette de textes cliniques.

	Système	Filtres	Unités cachées	Précision	Rappel	F-mesure
Tâche 1	DICT	–	–	7,22	3,56	4,76
	BiLSTM	-	600	**65,13**	13,87	22,87
	CNN	100	100	65,10	20,91	31,65
	CNN	1 000	100	53,53	25,97	34,98
	CNN	100	1 000	49,16	24,91	33,06
	CNN	1 000	1 000	50,52	32,00	39,18
	CNN	2 000	2 000	50,29	**33,01**	**39,86**
Tâche 2	DICT			5,22	7,25	6,07
	BiLSTM	-	600	**63,09**	28,37	39,14
	CNN	1 000	1 000	63,06	44,27	52,02
	CNN	2 000	2 000	60,18	**46,25**	**52,30**

Conclusion

Dans ce chapitre, plusieurs approches reposant sur des algorithmes d'apprentissage profond sont proposées pour la classification multi-étiquette de textes cliniques. À l'exception de l'approche par dictionnaires, les approches et résultats présentés dans ce chapitre ont été publiés dans DALLOUX, CLAVEAU, CUGGIA et al., 2020.

Concernant les données, nous avons commencé par décrire le corpus de textes cliniques mis à notre disposition par le **CHU** de Rennes à l'aide d'exemples et de statistiques descriptives. Nous montrons que ces textes diffèrent en tout point (structure, contenu, longueur, bruit, etc.) des certificats de décès utilisés dans le cadre des campagnes d'évaluation présentées en 1.2.2. Cependant, les certificats de décès sont les seuls textes cliniques auxquels il est possible d'accéder assez librement en dehors des hôpitaux, du moins pour des documents cliniques de langue française. Le corpus,que nous exploitons dans notre travail, issu du CHU de Rennes, n'est pas partageable avec la communauté scientifique en raison du secret médical.

Concernant la tâche en elle-même, nous montrons à quel point les classes sont inégalement distribuées dans le corpus : une large majorité de classes est représentées par moins de 10 exemples. Ces déséquilibres de classe sont un obstacle pour les approches par apprentissage. Nous tentons de diminuer ce problème en proposant une réduction du niveau hiérarchique afin de réduire le nombre total de classes et ainsi proposer plus d'exemples par classe, tout en conservant des classes faisant sens pour la pratique hospitalière.

Dans la section suivante, nous présentons les approches retenues pour la tâche à accomplir. Nous avons commencé par présenter notre approche par dictionnaires. Elle constitue notre point de référence et obtient, sans surprise, des résultats très faibles, bien inférieurs aux résultats de la littérature sur d'autres types de textes médicaux. Nous proposons ensuite deux approches basées sur des architectures de réseau de neurones différentes (**BiLSTM**, **CNN**). Pour les deux tâches proposées, le **CNN** obtient des résultats significativement supérieurs au **BiLSTM**. Par ailleurs, nos expérimentations montrent qu'augmenter le nombre de paramètres de notre **CNN** en augmentant le nombre de filtres par taille de région ainsi que le nombre d'unités dans la couche cachée permet d'améliorer grandement ces performances. Cependant, les limites de l'augmentation de ces paramètres sont vite atteintes, et le plafond, en terme de résultats, laisse une grande marge à l'amélioration pour cette tâche de classification rendue complexe par la nature des données et des labels utilisés. Afin de réduire cette marge, plusieurs approches sont envisageables. En effet, d'une part, CHALKIDIS et al., 2019 montrent que, dans le cadre d'une tâche de classification multi-étiquette comparable à la notre, les modèles de langage à base de *transformers* sont efficaces pour prédire les classes faiblement présentes dans les données d'entraînement et que les approches de *zero-shot learning* améliorent grandement la détection des classes absentes des données d'entraînement. D'autre part,

l'utilisation de modèles tel que **Big Bird** (ZAHEER et al., 2020), approche par *transformers* spécifiquement conçue pour traiter les textes longs, nous semble particulièrement appropriée pour cette tâche.

Chapitre 3

Corpus constitués dans le cadre de la thèse

Sommaire

3.1 Corpus biomédicaux français 81
3.1.1 Règles d'annotation 81
3.1.2 ESSAI : corpus français d'essais cliniques 84
3.1.3 CAS : corpus français de cas cliniques 89
3.2 Corpus biomédicaux brésiliens 94
3.2.1 Protocoles d'essais cliniques brésiliens 94
3.2.2 SemClinBr 97

Le projet **CominLabs BigClin**, dans lequel s'inscrit cette thèse, a pour but de développer de nouveaux outils de **TALN** dédiés aux textes cliniques en français. Cependant, les approches de **TALN** actuelles sont principalement basées sur des algorithmes d'apprentissage profond qui nécessitent un grand volume de textes, soit bruts si l'apprentissage est non-supervisé, soit annotés si il est supervisé. Le développement d'un outil de **TALN** basé sur une ou plusieurs approches par apprentissage supervisé n'est donc possible que si des corpus spécifiquement annotés pour les tâches visées sont disponibles. Cependant, dans le domaine médical, les corpus annotés disponibles pour la recherche dans des langues autres que l'anglais sont très peu nombreux. Dans ce qui suit, nous nous intéressons aux corpus de textes biomédicaux en français et en portugais brésilien.

Parmi les corpus de textes biomédicaux français accessibles librement, le **Corpus QUAERO Médical du français** (NÉVÉOL, GROUIN et al., 2014) contient des titres **MEDLINE** et des documents de l'Agence européenne des médicaments annotés avec dix types d'entités cliniques définies selon les groupes sémantiques de l'UMLS (*Anatomy*, *Chemical and Drugs*, *Devices*, *Disorders*, *Geographic Areas*, *Living Beings*, *Objects*, *Phenomena*, *Physiology*, *Procedures*). Les corpus médicaux du **CRTT** (Centre de Recherche en Terminologie et Traduction) sont annotés en parties du discours et lemmatisés [1]. Le **CLEAR - Simple Corpus for Medical French** (GRABAR et CARDON, 2018) contient des versions complexes et simplifiées de documents biomédicaux en français.

Il existe très peu de corpus de données biomédicales en portugais annotés et librement accessibles. Le **QTLP Portuguese Corpus for the Medical Domain** [2] contient des documents collectés en ligne qui ont été automatiquement détectés comme étant en portugais et du domaine médical. Chaque document a été automatiquement classé dans l'une de ces quatre catégories : *Reference*, *News/Journalism*, *Discussion* et *Other*. Les glossaires **MeSpEn** [3] contiennent quarante-six glossaires médicaux bilingues pour différentes paires de langues générées à partir de glossaires médicaux en ligne gratuits et de dictionnaires rédigés par des traducteurs professionnels. Le corpus **EMEA** [4] est un corpus parallèle des langues de l'Union Européenne constitué de documents de l'Agence européenne des médicaments.

Dans ce chapitre, nous décrivons les corpus constitués pendant cette thèse dans le cadre de ces projets. Dans la section 3.1, nous décrivons les corpus de textes médicaux français constitués et annotés par nos soins. Les marqueurs de négation et d'incertitude, ainsi que leur portée, y sont annotés. Dans la section 3.2, nous présentons les corpus de textes médicaux brésiliens constitués et annotés avec nos confrères de l'Université Pontificale Catholique du Paraná (PUCPR) dans le cadre de nos échanges. Les marqueurs de négation, ainsi que leur portée, y sont annotés.

1. https://perso.univ-lyon2.fr/%7Emaniezf/Corpus/Corpus_medical_FR_CRTT.htm
2. http://qt21.metashare.ilsp.gr/
3. https://github.com/PlanTL-SANIDAD/MeSpEn_Glossaries
4. http://opus.nlpl.eu/EMEA.php

3.1 Corpus biomédicaux français

Dans le cadre de cette thèse, nous avons constitué deux corpus de textes médicaux français. Les marqueurs de négation et d'incertitude, ainsi que leur portée, y ont été annotés par nos soins selon les règles que nous présentons en 3.1.1. Ensuite, en 3.1.2, nous décrivons le contenu du corpus **ESSAI** et nous présentons le processus d'annotation et ces résultats. Enfin, en 3.1.3, nous faisons de même pour le corpus **CAS**.

3.1.1 Règles d'annotation

Le but final de l'annotation des marqueurs de négation et de l'incertitude et de leur portée est de déterminer quels évènements de la phrase sont affectés par ces opérations linguistiques. Nous utilisons ici le terme évènement de manière très générale. Il peut s'agir d'un processus, d'une action ou d'un état. Comme nous l'avons expliqué précédemment, la détection des évènements niés et incertains est importante car ce ne sont pas des faits. En effet, si nous concevons un système d'extraction d'informations, ce système doit être capable de discerner les faits des hypothèses et des infirmations.

Le style d'annotation que nous employons est inspiré des corpus **BioScope** et **CD-SCO** présentés dans la sous-section 1.3.2. À l'instar du corpus **CD-SCO** et contrairement au corpus **BioScope**, les marqueurs ne sont pas considérés dans la portée et la portée peut être discontinue. Cependant, contrairement au corpus **CD-SCO** et à l'instar du corpus **BioScope**, les portées annotées n'incluent pas tous les tokens liés aux évènements niés et incertains. Par ailleurs, les versions actuelles de nos corpus annotés ne couvrent pas la négation affixale.

Dans nos corpus, la négation est définie comme l'implication de la non-existence de quelque chose. Cependant, la présence d'un marqueur de négation n'implique pas nécessairement que la phrase soit négative. En outre, l'incertitude linguistique y est définie de la façon suivante : toute phrase qui déclare possible l'existence d'une chose, c'est-à-dire que son existence ou pas n'est pas confirmée, est une phrase spéculative. Dans ce qui suit, nous détaillons les règles établies pour l'annotation des marqueurs et de leur portée.

Annotation des marqueurs

Nous avons vu précédemment que la négation et l'incertitude sont exprimées au moyen de mots ou de combinaisons de mots, appelés marqueurs, qui peuvent être des verbes, des noms, des adverbes, des adjectifs, etc. Dans les exemples ci-dessous, les marqueurs sont soulignés. Dans l'exemple (1), *n'[...]pas* est un marqueur de négation adverbial. Dans l'exemple (2), *aucun[...]n'* est un marqueur de négation adjectival. Dans l'exemple (3), *probablement* est un marqueur d'incertitude adverbial. Dans l'exemple (4), *hypothèse* est un marqueur d'incertitude nominal tandis que *évoquée* un marqueur d'incertitude verbal. Les marqueurs peuvent être composés de

plusieurs mots se suivant, tels que *sans aucun* (5) ou *reste à démontrer* (6), ou être discontinus comme dans les exemples (1) et (2). Certains marqueurs de négation font partie de marqueurs d'incertitude. C'est le cas de la locution adverbiale *sans doute* dans l'exemple (7) :

1. Il n'y a pas de fièvre.
2. Aucun foyer digestif primitif n'était identifié.
3. Conclusion : occlusion probablement sur bride bien tolérée.
4. L'hypothèse d'une toxicité médicamenteuse était évoquée.
5. Un homme de 58 ans, informaticien sans antécédent personnel et sans aucun suivi médical, est amené par les pompiers aux urgences après une agression dans la rue.
6. Le bénéfice de l'association lénalidomide + R-CHOP par rapport au R-CHOP reste à démontrer.
7. Lors de son examen le médecin trouve devant lui un jeune homme grand et fort, au physique puissant, mais à l'expression clairement avachie, ce qui fut sans doute causé par les médicaments.

Les règles d'annotation des marqueurs sont assez simples. Une liste non-exhaustive de marqueurs de négation et d'incertitude a été créée sur la base des marqueurs annotés dans les corpus **BioScope** et **CD-SCO**. Pour l'incertitude, certains marqueurs spécifiques au français, tels que l'emploi du conditionnel et *reste à + verbe à l'infinitif*, ont été ajoutés à cette liste. Il était demandé aux annotateurs d'annoter les marqueurs de cette liste. Cependant, les annotateurs avaient la possibilité d'annoter également les mots ou combinaisons de mots en dehors de la liste ; si ces mots marquaient, selon eux, la négation ou l'incertitude. Selon nous, il était nécessaire de procéder ainsi car il est difficile de recenser l'intégralité des formes, flexions et combinaisons de mots que les marqueurs d'incertitude peuvent prendre. La négation étant une opération linguistique clairement définie, les marqueurs de négation sont plus faciles à identifier. Cependant, les textes médicaux peuvent également comporter des marqueurs spécifiques, qui doivent alors être identifiés.

Par ailleurs, certains marqueurs de négation font partie de concepts biomédicaux tels que *lymphome non hodgkinien*, *lymphocytes ni T ni B* ou *cancer bronchique non à petites cellules*. Dans les premières versions de nos corpus, ceux-ci étaient annotées. Cependant, à la suite d'échanges avec plusieurs chercheurs en informatique médicale, nous avons décidé de retirer ces annotations étant donné que ces concepts biomédicaux sont identifiés par leur propre **CUI** (*Concept Unique Identifier*) dans l'**UMLS** (*Unified Medical Language System*), respectivement C0024305, C0024265 et C0007131, qui correspondent à leurs représentations sémantiques complètes. En conséquence, il était demandé de ne pas annoter les marqueurs faisant partie de concepts biomédicaux présents dans la version actuelle de l'**UMLS** en français. Inversement, quand aucun **CUI** ne couvre la séquence complète, le marqueur de négation doit être annoté. Par exemple, le marqueur *non* dans la séquence *mésothéliome pleural malin non résécable* est annoté puisque le **CUI** le plus proche de cette séquence, C0812413, ne couvre que *mésothéliome pleural malin*. Finalement, l'**UMLS** étant, notamment en

français, une ressource très incomplète et en évolution permanente, ces cas ne représentaient qu'un faible pourcentage des instances de négation.

Annotation de la portée

La portée d'un marqueur de négation ou d'incertitude s'étend sur la ou les parties de la phrase qu'il annule ou met en doute. La portée des marqueurs de négation et d'incertitude a été étudiée dans de nombreuses publications dont CULIOLI, 1988 ; MULLER, 1991 ; HORN, 2001 ; VINCZE et al., 2008 ; FARKAS et al., 2010 ; VINCZE, 2015 et peut généralement être déterminée sur la base de modèles syntaxiques dépendant du marqueur. L'annotation de la portée dépend donc du marqueur, du contexte et de la structure syntaxique de la phrase. L'annotation de nos corpus se base sur les observations suivantes :

— La portée des verbes, adjectifs et adverbes s'étend généralement vers la droite du marqueur,

— La portée d'un verbe s'étend soit jusqu'à la fin de la proposition qui suit ou de la phrase, ou, dans le cas où ce verbe clos la phrase, sur une ou plusieurs propositions qui le précèdent, voire sur toute la phrase ([...] *sont* ***exclus/suspectés***),

— La portée des adjectifs attributs couvre généralement toute la phrase (*il est* ***possible*** *que* [...]), tandis que la portée des adjectifs épithètes s'étend aux syntagmes nominaux auxquels ils s'attachent (***aucune*** *intervention supplémentaire*),

— La portée des adverbes phrastiques (***probablement***, ***apparemment***), s'étend sur toute la phrase, tandis que la portée des autres adverbes se termine à la fin de la proposition ou de la phrase,

— La portée des conjonctions s'étend sur les propositions qu'elles interconnectent.

Conformément à ces observations, nous n'annotons pas systématiquement tous les tokens des phrases négatives et spéculatives que nous rencontrons. Plus spécifiquement, dans les phrases négatives et/ou spéculatives, tout token ne faisant pas partie d'une séquence explicitement niée ou incertaine ne doit pas être inclus dans la portée. Dans les exemples ci-dessous, les marqueurs sont soulignés et les portées sont marquées en gras.

Dans l'exemple (1), *mésothéliome pleural malin* ne doit pas être inclus dans la portée puisque la phrase ne peut pas être reformulée comme suit : *Pas de mésothéliome pleural malin*. À l'inverse, dans l'exemple (2), *une autopsie* fait partie de la portée puisque nous pouvons reformuler la phrase comme suit : *Pas d'autopsie*. Par ailleurs, dans l'exemple (3), la portée de la négation est modifiée par le token *toujours*, ce qui est pris en compte dans l'annotation. En effet, sans l'occurrence de *toujours*, *Le traitement de référence* ferait partie de la portée. Dans l'exemple (4), la portée de l'adjectif *probable* s'étend à la droite du marqueur jusqu'à la fin de la phrase. Dans l'exemple (5), la portée du marqueur d'incertitude *évoqué* s'étend sur l'intégralité des tokens le précédent.

1. Mésothéliome pleural malin non **résécable**.
2. **Une autopsie** n'**était** pas **réalisée**.

3. Le traitement de référence est chirurgical mais celui-ci n'**est** pas **toujours possible**.
4. Comme pour beaucoup de maladies, il est probable **que cette différence de susceptibilité individuelle repose sur une différence dans les gènes**.
5. **Le diagnostic d'ulcère solitaire du rectum était** évoqué.

Nous venons de présenter les règles retenues pour l'annotation des marqueurs de négation et d'incertitude ainsi que leur portée. Dans les sous-sections 3.1.2 et 3.1.3, nous présentons les corpus de données biomédicales constitués dans le cadre de la thèse. Plus précisément, pour chaque corpus, nous décrivons les données (sources, contenu, statistiques), ainsi que le processus d'annotation.

3.1.2 ESSAI : corpus français d'essais cliniques

Un essai clinique est une étude scientifique pratiquée sur l'être humain afin de développer les connaissances du domaine médical. Les essais cliniques consistent généralement à vérifier les données pharmacocinétiques (absorption, distribution, métabolisme et excrétion), pharmacodynamiques (mécanisme d'action) et thérapeutiques (efficacité et tolérance) d'un médicament expérimental ou d'un traitement connu mais utilisé différemment. Par ailleurs, les essais cliniques peuvent aussi concerner de nouvelles méthodes de prélèvement, de diagnostic, ou bien d'opération chirurgicale. Dans ces documents, les négations fournissent des informations utiles concernant le déroulement de l'étude, la spécification de la cohorte cible et le recrutement des patients. Les incertitudes exprimées concernent principalement les objectifs de l'étude et les évolutions éventuelles de la condition des patients.

Premier corpus constitué dans le cadre de cette thèse, le corpus **ESSAI**, contient des protocoles d'essais cliniques en français collectés à partir du registre de l'Institut national du cancer [5]. Chaque protocole se compose de deux documents : le résumé et la description détaillée.

Comme le montre la figure 3.1, le résumé de l'essai présente l'objectif principal de l'étude. Dans notre exemple, il s'agit d'étudier la tolérance et l'efficacité d'un traitement. Ensuite, sont décrits les examens, techniques ou traitements expérimentaux qui seront réalisés. Dans notre exemple, il s'agit de la technique de CHIP. Enfin, les traitements et leurs administrations sont évoqués. Souvent, d'autres informations sont données, telles que le suivi des patients ou le mode de délivrance des résultats aux patients.

5. https://www.e-cancer.fr

Résumé
L'objectif de cet essai est d'étudier la tolérance et l'efficacité d'une chimio-hyperthermie intra-péritonéale (CHIP) chez des patientes opérés d'un cancer de l'ovaire et traitées par chimiothérapie.
La technique de CHIP est une combinaison d'une intervention chirurgicale et d'une chimiothérapie locale. Elle consiste à baigner la cavité abdominale avec une chimiothérapie à forte concentration pour augmenter l'effet sur les cellules cancéreuses.
Dans cet essai, les patientes seront traitées par de l'oxaliplatine pendant 30 min au cours d'une seconde intervention chirurgicale à visée exploratoire, réalisée après 6 cures de chimiothérapie.

FIGURE 3.1 – Protocole d'essai clinique en français : Résumé

Comme le montre la figure 3.2, la description détaillée est un document bien plus long que le résumé. Ce document commence par le titre de l'essai, ici ***CHIPOVAC***, suivi d'une brève description de l'étude. La seconde partie, le résumé scientifique ou schéma thérapeutique, contient plus ou moins les mêmes informations que le résumé, parfois reformulées et complétées par de nouvelles informations telles que *essai de phase 2, non randomisé et multicentrique*. La partie suivante présente l'objectif principal de l'essai, i.e. *Étudier la tolérance et les résultats carcinologiques de l'oxaliplatine*. Souvent, un ou plusieurs objectifs secondaires de l'étude sont présentés. Ensuite, les critères d'inclusion, tels que *Age* $\geq$ *18 ans et* $\leq$ *65 ans* ou *Consentement éclairé signé*, ainsi que les critères de non-inclusion, tels que *Insuffisance hépatique ou rénale* ou *Allergie connue aux sels de platine*, sont présentés. Enfin, le critère d'évaluation principal de l'étude est donné. Dans notre exemple, il s'agit de *Morbidité chirurgicale.*.

CHIPOVAC : Essai de phase 2 évaluant l'oxaliplatine, administré lors d'une chimiothérapie hyperthermique intra-péritonéale avec chirurgie, en traitement de consolidation chez des patientes ayant un cancer de l'ovaire.
[essai clos aux inclusions]
Résumé scientifique / schéma thérapeutique
Il s'agit d'un essai de phase 2, non randomisé et multicentrique.
Les patientes incluses dans cet essai auront eu une chirurgie et 6 cures de chimiothérapie.
Une exérèse complète de la carcinose péritonéale est pratiquée suivie d'une chimiothérapie hyperthermique intra-péritonéale avec de l'oxaliplatine, pendant 30 min à température efficace.
Objectif principal
Étudier la tolérance et les résultats carcinologiques de l'oxaliplatine.
Critères d'inclusion
Age $\geq$ 18 ans et $\leq$ 65 ans.
Cancer de l'ovaire de stade IIIc (FIGO) traité par chirurgie et 6 cures de chimiothérapie.
Données hématologiques : polynucléaires neutrophiles.
Chirurgie antérieure effectuée en 1, 2 ou 3 temps comportant au minimum une hystérectomie, une annexectomie bilatérale, une omentectomie totale, et un curage pelvien et lombo-aortique.
Chimiothérapie antérieure comprenant 6 cures par un sel de platine.
Exploration chirurgicale réalisée après 6 cures de chimiothérapie. Peut être associée à des gestes chirurgicaux pour obtenir une carcinose résiduelle macroscopique millimétrique ou nulle. Aucune résection digestive ne doit être réalisée au moment de ce second-look.
Consentement éclairé signé.
Critères de non-inclusion
Insuffisance hépatique ou rénale.
Score ASA 3.
Allergie connue aux sels de platine.
Critère d'évaluation principal
Morbidité chirurgicale.

FIGURE 3.2 – Description détaillée du protocole d'essai clinique

Dans le tableau 3.1, nous présentons les statistiques relatives au corpus **ESSAI**. Deux versions du corpus sont actuellement disponibles : la première est annotée avec les informations relatives à la négation et la seconde avec les informations relatives à l'incertitude.

La version du corpus annotée avec la négation contient 7 247 phrases et 163 425 tokens, soit 22,55 tokens par phrase en moyenne, un vocabulaire de 8 283 tokens, 981 phrases marquées par la négation, 1 064 instances de négation et 56 de marqueurs de négation différents. Les marqueurs de négation les plus fréquents sont : *non*/*non-* avec 310 occurrences, *ne*/*n'*[...]*pas* avec 221 occurrences, *sans* avec 180 occurrences et

absence de/d' avec 131 occurrences. Les marqueurs de négation les moins fréquents sont : *en dehors d'*, *exclus*, *hormis*, *à l' exclusion des*, etc. avec une seule occurrence.

La version du corpus annotée avec l'incertitude est moins volumineuse que la précédente et ne contient que 6 601 phrases et 151 070 tokens, soit 22,89 tokens par phrase en moyenne, un vocabulaire de 7 906 tokens, 631 phrases marquées par l'incertitude, 754 instances d'incertitude et 76 marqueurs d'incertitude différents. Les marqueurs d'incertitude les plus fréquents sont : *si* avec 190 occurrences, les formes fléchies du verbe *pouvoir* avec 120 occurrences, *en cas de/d'* avec 110 occurrences, ainsi que *et/ou* avec 108 occurrences. Les marqueurs d'incertitude les moins fréquents sont : *a priori*, *pour être sûr*, *supposer*, *suspicion*, etc. avec une seule occurrence.

TABLEAU 3.1 – Statistiques relatives au corpus ESSAI

	Négation	Incertitude
Phrases	7 247	6 601
Tokens	163 425	151 070
Vocabulaire	8 283	7 906
Phrases marquées	981	631
Marqueurs (total)	1 064	754
Marqueurs (unique)	62	76

Processus d'annotation

Premier corpus constitué dans le cadre de cette thèse, le corpus **ESSAI**, a impliqué deux annotateurs. Ces annotateurs ont participé à l'annotation manuelle des marqueurs de négation et de leur portée selon les règles d'annotation établies en 3.1.1. Cependant, à ce jour, les marqueurs d'incertitude et leur portée n'ont été annotés manuellement que par un seul annotateur. Par conséquent, aucun accord inter-annotateurs ni processus d'arbitrage n'a été réalisé pour l'incertitude.

Les accords inter-annotateurs présentés dans ce manuscrit sont calculés avec le Kappa de Cohen (J. COHEN, 1960) qui permet de mesurer l'accord entre deux annotateurs en prenant en compte la possibilité que l'accord se produise par hasard, ce qui rendrait la mesure plus robuste que le simple calcul de l'accord en pourcentage. LANDIS et KOCH, 1977 proposent d'interpréter les valeurs retournées par cette mesure de la façon suivante : < 0 : désaccord, 0 – 0,2 : accord très faible, 0,21 – 0,4 : accord faible, 0,41 – 0,60 : accord modéré, 0,61 – 0,80 : accord fort, 0,81 – 1,00 : accord presque parfait. FLEISS, LEVIN et PAIK, 1981 considèrent excellents les Kappa supérieurs à 0,75, moyen à bon les Kappa de 0,40 à 0,75 et faible les Kappa inférieurs à 0,40. Ces ordres de grandeurs ne font pas nécessairement consensus dans la communauté scientifique. Cependant, nous estimons qu'un Kappa de Cohen supérieur à 0,75 doit être le minimum attendu d'un processus réussi lors de l'annotation de séquences. Nous présentons les accords inter-annotateurs obtenus pour le corpus **ESSAI** dans le tableau 3.2.

	Kappa de Cohen
Marqueurs de négation	0,9001
Portées des négations	0,8089

TABLEAU 3.2 – **ESSAI** : Accords inter-annotateurs obtenus pour les marqueurs de négation et leur portée

L'accord inter-annotateurs obtenu pour les marqueurs de négation est très élevé puisque le Kappa de Cohen est de 0,9001. La plupart des désaccords entre annotateurs sont liés au préfixe *anti-*. En effet, 19 occurrences de ce préfixe ont été annotées comme marquant la négation par le second annotateur. Parmi celles-ci, nous retrouvons les termes *anti-angiogéniques*, *anti-tumorale* ou encore *anti-cancéreuses*. Durant le processus d'arbitrage, il a été décidé de ne pas les annoter, étant donné que le sens de *anti-* dans ces concepts médicaux ne correspond pas à la négation du mot préfixé mais plutôt au soin ou bien à la neutralisation de ce que désigne le mot préfixé. D'autre part, la seule occurrence du marqueur *hormis* dans la séquence *hormis tumeur cérébrale et hémopathies* a été oubliée par le premier annotateur. Elle est ajoutée à la version finale du corpus.

L'accord inter-annotateurs obtenu pour la portée des marqueurs de négation est plutôt élevé. En effet, le Kappa de Cohen est de 0,8089. Cependant, la mesure baisse de près de 0.1 par rapport aux marqueurs de négation. Nous observons plusieurs raisons à cette baisse :

— les portées associées aux occurrences de *anti-* annotées par erreur comportent plus d'un token (anti-**tumorale systémique**),

— les portées des marqueurs oubliés comportent plus de tokens que les marqueurs (hormis **tumeur cérébrale et hémopathies**),

— des désaccords apparaissent dans l'annotation des frontières de la portée de certains marqueurs.

Nous donnons quelques exemples par rapport au dernier cas. Dans l'exemple (1), le second annotateur fait l'erreur d'inclure *chez des patients immunodéprimés* dans la portée. C'est une erreur car le fait est que les patients sont immunodéprimés. Dans l'exemple (2), c'est le premier annotateur qui fait l'erreur d'annoter la séquence *le rôle de la plupart des anomalies* dans la portée du marqueur *n'*[...]*pas*. Dans l'exemple (3), le premier annotateur a annoté une portée discontinue, incluant le sujet *des patients éligibles* en amont de la phrase. C'est une erreur car le sujet n'est pas nié. Cette erreur d'annotation date de la première version des annotations, lorsque les règles d'annotation étaient en cours d'élaboration.

(AR : annotation retenue, A1 : premier annotateur, A2 : second annotateur)

1. AR : **cela** n'**a** pas **été démontré** chez des patients immunodéprimés.
   A1 : **cela** n'**a** pas **été démontré** chez des patients immunodéprimés.
   A2 : cela n'**a** pas **été démontré chez des patients immunodéprimés**.
2. AR : le rôle de la plupart des anomalies n'**est** pas **bien défini actuellement**.

A1 : **le rôle de la plupart des anomalies** n'**est** pas **bien défini actuellement**.

A2 : le rôle de la plupart des anomalies n'**est** pas **bien défini actuellement**.

3. AR : [...] des patients éligibles [...] n'**ayant** pas **satisfait aux critères du protocole sur la progression de la maladie**.

   A1 : [...] **des patients éligibles** [...] n'**ayant** pas **satisfait aux critères du protocole sur la progression de la maladie**.

   A2 : [...] des patients éligibles [...] n'**ayant** pas **satisfait aux critères du protocole sur la progression de la maladie**.

Un extrait du corpus dans sa forme finale est visible dans le tableau 3.3. À l'instar du corpus **CD-SCO**, le corpus est au format de la campagne d'évaluation de **CoNLL-2005**, où chaque information est présentée dans une colonne séparée. De gauche à droite, les informations données pour chaque token sont : le numéro identifiant la phrase (**ID**), la position du token dans la phrase (**PT**), son lemme, l'étiquetage morpho-syntaxique (PoS-tag), ainsi que les annotations relatives à la négation, marqueur (**M–neg**) et portée (**P–neg**). L'étiquetage morpho-syntaxique et les lemmes sont obtenus avec **TreeTagger** (SCHMID, 1994b).

TABLEAU 3.3 – Extrait tiré du corpus ESSAI

ID	PT	Token	Lemme	PoS-tag	M–neg	P–neg
1871	0	Aucun	aucun	PRO :IND	B_cue_neg	_
1871	1	traitement	traitement	NOM	_	B_scope_neg
1871	2	expérimental	expérimental	ADJ	_	I_scope_neg
1871	3	ne	ne	ADV	I_cue_neg	_
1871	4	sera	être	VER :futu	_	I_scope_neg
1871	5	administré	administrer	VER :pper	_	I_scope_neg
1871	6	.	.	SENT	_	_

3.1.3 CAS : corpus français de cas cliniques

Le corpus **CAS** (GRABAR, CLAVEAU et DALLOUX, 2018), contient des cas cliniques en français, tels que ceux publiés dans la littérature scientifique, le matériel juridique ou de formation. Il a été collecté en utilisant du matériel disponible gratuitement dans des sources en ligne. Les cas cliniques actuellement collectés sont publiés dans différentes revues et sites Web de pays francophones (par exemple, France, Belgique, Suisse, Canada, pays africains, pays tropicaux). Ces cas cliniques sont liés à diverses spécialités médicales (par exemple, cardiologie, urologie, oncologie, obstétrique, pneumologie, gastro-entérologie).

Le but des cas cliniques est de décrire des situations cliniques pour de vrais patients dépersonnalisés ou pour de faux patients. Les cas cliniques courants font généralement partie des programmes d'enseignement utilisés pour la formation des étudiants en médecine, tandis que les cas rares sont généralement partagés par le biais de publications scientifiques pour illustrer des situations cliniques moins courantes. Quant aux cas cliniques retrouvés dans les sources juridiques, ils rapportent

généralement des situations qui se sont compliquées pour diverses raisons : médecin, équipe soignante, institution, système de santé et leurs interactions. A l'instar des documents cliniques, le contenu des cas cliniques dépend des situations cliniques illustrées et des troubles, mais aussi de la finalité des cas présentés : description des diagnostics, traitements ou procédures, évolution, antécédents familiaux, audience attendue, etc.

La figure 3.3 montre un exemple typique d'un cas clinique du corpus CAS. Le document commence par l'introduction de la patiente (femme de 44 ans), expliquant pourquoi elle a été hospitalisée (1). Les phrases suivantes (2) ajoutent des informations sur les antécédents médicaux de la patiente concernant la maladie de Crohn. La section suivante (3) décrit les résultats d'examen et de laboratoire obtenus pour cette patiente. Enfin, la dernière phrase (4) décrit le traitement choisi, ses effets sur la patiente et l'issue du processus de soins. Dans les cas cliniques, la négation est fréquemment utilisée pour faire la description des signes et symptômes du patient et pour le diagnostic des patients. Elle peut également être utilisée pour la description de l'évolution du patient.

(1) Une femme de 44 ans était hospitalisée en juillet 1999 pour une diarrhée évoluant depuis la veille, faite de 8 selles diurnes et 3 selles nocturnes, glairo-sanglantes et impérieuses. La diarrhée était associée à des douleurs hypogastriques. Les symptômes étaient apparus deux jours après le début d'un traitement d'une pharyngite par de la pristinamycine à la dose de 1 g x 2/jour et de la prednisone à la dose de 40 mg/jour.

(2) Dans ses antécédents, la malade avait eu deux résections iléales pour une maladie de Crohn sténosante de l'intestin grêle diagnostiquée en 1977. La dernière intervention avait été réalisée en octobre 1998 pour un abcès de l'anse iléale préanastomotique. Trente-cinq cm d'iléon et 5 cm de côlon droit avaient été réséqués. Depuis cette date, la maladie de Crohn était quiescente sous un traitement par mésalazine à la dose de 3 g/jour.

(3) L'examen clinique à l'admission était sans particularité, hormis une douleur provoquée à la palpation de la fosse iliaque gauche et de l'hypogastre. Les examens biologiques montraient une hyperleucocytose à 11,4 G/L et un syndrome inflammatoire avec une vitesse de sédimentation à 50 mm à la première heure. L'hémoglobinémie était normale à 14 g/dL. La coproculture sur milieu enrichi en ampicilline mettait en évidence quelques colonies de Klebsiella oxytoca . La coproculture avec ensemencement de milieux sélectifs pour Salmonella , Shigella et Campylobacter spp. était négative, ainsi que la recherche dans les selles par méthode immunoenzymatique de toxine de Clostridium difficile . L'examen tomodensitométrique abdominal montrait un épaississement de la paroi du côlon gauche.

(4) En 24 heures, sous repos digestif et après arrêt des antibiotiques, l'évolution était favorable avec un retour à l'état clinique antérieur à la colite aiguë.

FIGURE 3.3 – Exemple de cas clinique en français

Dans le tableau 3.4, nous présentons les statistiques relatives au corpus **CAS**. Dans le cadre de notre thèse, deux versions du corpus annoté sont actuellement

disponibles : la première est annotée avec les informations relatives à la négation et la seconde avec les informations relatives à l'incertitude. Les deux versions contiennent les mêmes documents. Le corpus annoté est ainsi composé de 200 cas cliniques, 3 811 phrases et 87 487 tokens, soit 22,96 tokens par phrase en moyenne et d'un vocabulaire de 10 500 tokens. 832 phrases sont marquées par la négation, 945 instances de négation sont disponibles et 74 marqueurs de négation sont identifiés. 226 phrases sont marquées par l'incertitude, 945 instances d'incertitude sont disponibles et 80 marqueurs d'incertitude sont identifiés.

Les marqueurs de négation les plus fréquents sont : *ne/n'*[...]*pas* avec 299 occurrences, *sans* avec 264 occurrences et (*ne/n'*) *aucun* (*ne/n'*) avec 85 occurrences et *non* avec 70 occurrences. Les marqueurs de négation les moins fréquents sont : *absents*, *exclure* et *exclus* avec une seule occurrence.

Les marqueurs d'incertitude les plus fréquents sont : les formes fléchies du verbe *évoquer* avec 37 occurrences, *?* avec 31 occurrences, les formes fléchies du verbe *pouvoir* avec 24 occurrences et *aurait* avec 11 occurrences. Les marqueurs d'incertitude les moins fréquents sont : *affirmait*, *consisterait*, *apparemment*, *certainement* et 11 autres marqueurs ne comptant qu'une seule occurrence.

TABLEAU 3.4 – Statistiques relatives au corpus CAS

	Négation	Incertitude
Documents	200	
Phrases	3 811	
Tokens	87 487	
Vocabulaire	10 500	
Phrases marquées	832	226
Marqueurs (total)	945	243
Marqueurs (unique)	74	80

Processus d'annotation

Second corpus constitué et annoté dans le cadre de cette thèse, le corpus **CAS**, a impliqué deux annotateurs. Cependant, à l'instar du corpus **ESSAI**, les marqueurs d'incertitude et leur portée n'ont été annotés que par un seul annotateur. Par conséquent, aucun accord inter-annotateurs ni processus d'arbitrage n'ont été réalisés pour l'incertitude. À l'inverse du corpus **ESSAI**, le corpus **CAS** a d'abord été annoté automatiquement en trois étapes :

1. annotation du corpus avec **TreeTagger** afin d'obtenir les lemmes et l'étiquetage morpho-syntaxique,
2. annotation automatique des marqueurs de la négation et de l'incertitude (séparément) par un **BiLSTM-CRF** entraîné sur les données du corpus **ESSAI**,
3. annotation automatique de la portée des marqueurs détectés à l'étape précédente par un **BiLSTM-CRF** entraîné sur les données du corpus **ESSAI**.

Le système ainsi que les descripteurs utilisés pour l'annotation automatique sont présentés dans les sections 4.1, 4.2 et 4.3. À la suite de cette annotation automatique, les deux annotateurs ont manuellement vérifié chaque phrase afin de corriger les erreurs d'annotation et d'annoter les instances de négation oubliées par le système de détection automatique selon les règles d'annotation établies en 3.1.1. Nous présentons dans le tableau 3.5 les accords inter-annotateurs calculés avec le Kappa de Cohen pour le corpus CAS.

	Kappa de Cohen
Marqueurs de négation	0,9933
Portées des négations	0,8461

TABLEAU 3.5 – CAS : Accords inter-annotateurs obtenus pour les marqueurs de négation et leur portée

L'accord inter-annotateurs obtenu pour les marqueurs de négation est extrêmement élevé, le Kappa de Cohen étant de 0,9933. Dans le cas présent, la majorité des marqueurs a été annotée automatiquement et seuls quelques marqueurs sont oubliés par l'un ou l'autre des annotateurs. Nous présentons ici quelques exemples. Dans l'exemple (1), la seule occurrence du marqueur *infirmé* était oublié par l'un des annotateurs. N'apparaissant pas dans le corpus ESSAI, le système de détection automatique n'était pas entraîné à la détection de ce marqueur. Dans l'exemple (2), l'adverbe *non* de la locution adverbiale *non seulement* ne marque pas la négation mais a tout de même été annoté par le BiLSTM-CRF. Lors de la vérification manuelle l'un des annotateurs n'a pas repéré cette erreur. Cependant, aucun des annotateurs n'a repéré l'instance de négation *disparition complète non seulement du prurit mais également de l'ictère* dont le marqueur est *disparition* et qui peut être reformulée par : *plus de prurit ni d'ictère*. Ce marqueur est repéré lors de l'arbitrage et ajouté au corpus. Dans l'exemple (3), la préposition *sans* de la locution adverbiale *sans doute* ne marque pas la négation mais à tout de même été annotée par le BiLSTM-CRF. Cette erreur apparaît à plusieurs reprises. Lors de la vérification manuelle l'un des annotateurs n'a pas repéré ces erreurs.

1. **Le diagnostic d'abcès** était évoqué, mais infirmé au cours d'une exploration chirurgicale, l'incision ne montrant que des débris nécrotiques sans **pus**.
2. Les suites ont été simples, avec cette fois une disparition complète non seulement du prurit mais également de l'ictère.
3. Lors de son examen le médecin trouve devant lui un jeune homme grand et fort, au physique puissant, mais à l'expression clairement avachie, ce qui fut sans doute causé par les médicaments.

L'accord inter-annotateurs obtenu pour la portée des marqueurs de négation est aussi plus élevée que pour le corpus ESSAI. En effet, le Kappa de Cohen est de 0,8461. Cependant, à l'instar du corpus **ESSAI**, le Kappa de Cohen est bien plus bas que pour l'annotation des marqueurs. Dans le cas présent, les désaccords entre

annotateurs concernant la portée des marqueurs, en dehors de ceux liés aux marqueurs oubliés et erreurs d'annotation non repérées, sont liés à l'annotation de la ponctuation et du sujet de la phrase.

Dans l'exemple (1), le premier annotateur considère que la virgule fait partie de la portée. À l'inverse du second annotateur, il annotera tout le corpus de cette façon. Lors du processus d'arbitrage, il est décidé de ne pas inclure la ponctuation dans la portée. Dans l'exemple (2), le premier annotateur fait l'erreur d'annoter le sujet de la phrase, *le patient*, dans la portée.

(AR : annotation retenue, A1 : premier annotateur, A2 : second annotateur)

1. AR : pas **de tabagisme**, **d'histoire de reflux gastro-œsophagien ou d'ingestion de caustique**.

   A1 : pas **de tabagisme, d'histoire de reflux gastro-œsophagien ou d'ingestion de caustique**.

   A2 : pas **de tabagisme**, **d'histoire de reflux gastro-œsophagien ou d'ingestion de caustique**.

2. AR : le patient n'**a eu** aucune **récidive clinique**

   A1 : **le patient** n'**a eu** aucune **récidive clinique**

   A2 : le patient n'**a eu** aucune **récidive clinique**

Un extrait du corpus dans sa forme finale est visible dans le tableau 3.6. Nous conservons le même format que pour le corpus **ESSAI**.

TABLEAU 3.6 – Extrait tiré du corpus CAS

ID	PT	Token	Lemme	PoS-tag	M–neg	P–neg
1873	0	Il	il	PRO :PER	_	_
1873	1	n'	ne	ADV	B_cue_neg	_
1873	2	y	y	PRO :PER	_	B_scope_neg
1873	3	a	avoir	VER :pres	_	I_scope_neg
1873	4	pas	pas	ADV	I_cue_neg	_
1873	5	de	de	PRP	_	I_scope_neg
1873	6	fièvre	fièvre	NOM	_	I_scope_neg
1873	7	.	.	SENT	_	_

3.2 Corpus biomédicaux brésiliens

Dans le cadre du projet franco-brésilien **FIGTEM** et à l'occasion d'une mission d'un mois au Brésil à l'Université Pontificale Catholique du Paraná (**PUPCR**), nous avons participé à l'élaboration d'un corpus constitué de protocoles d'essais cliniques brésiliens. Les marqueurs de négation ainsi que leur portée y ont été annotés. Nous décrivons le contenu de ce corpus ainsi que le processus d'annotation en 3.2.1. À la suite de cette collaboration, un second corpus constitué de textes cliniques brésiliens a été constitué et annoté par les équipes du partenaire brésilien et mis à notre disposition. Nous décrivons le contenu de ce corpus ainsi que le processus d'annotation en 3.2.2.

3.2.1 Protocoles d'essais cliniques brésiliens

Le corpus constitué dans le cadre du projet **FIGTEM** contient les protocoles d'essais cliniques qui ont été collectés sur le site web brésilien dédié aux essais cliniques [6]. La figure 3.4 présente un protocole d'essai clinique brésilien. Sa structure est similaire au protocole français présenté dans la figure 3.2 en 3.1.2. Chaque protocole indique :

— le titre public de l'essai (*Título Público*),

L'influence de la semelle proprioceptive associée à l'acupuncture chez la femme.

— le titre scientifique (*Título Científico*),

L'influence de la semelle proprioceptive associée à l'acupuncture sur l'équilibre, la posture, l'activité musculaire, la flexibilité et le profil énergétique des méridiens chez la femme.

— le résumé scientifique/schéma thérapeutique (*Texto*),

Le groupe G1 (n=15) a utilisé des semelles de chaussures courantes et recevra les informations pour utiliser la semelle commune 4 heures par jour. Les groupes G2 (n=15) et G3 (n=15) recevront la semelle fabriquée individuellement, après évaluation posturale et clinique et devront l'utiliser 4 heures par jour. [...]

— les critères d'inclusion (*Critério de Inclusão*),

Étudiantes universitaires ; sexe féminin ; âge entre 18-30 ans ; sédentaires [...]

— les critères d'exclusion (*Critério de Exclusão*),

Les étudiantes qui ont une scoliose en S ;[...] ; fractures récentes des membres inférieurs [...]

— un résumé du schéma thérapeutique de l'étude (*Design do Estudo*).

Essai de traitement, contrôlé randomisé, parallèle, ouvert, avec deux bras.

6. `http://ensaiosclinicos.gov.br/`

Título Público :
A influência da palmilha proprioceptiva associada a acupuntura em mulheres.
Título Científico :
A influência da palmilha proprioceptiva associada à acupuntura sobre o equilíbrio, postura, atividade muscular, flexibilidade e perfil energético dos meridianos em mulheres.
Texto :
O grupo G1 (n=15) utilizaram palmilhas comum para sapato e receberá a informação para utilizar a palmilha comum por 4 horas por dia
O grupo G2 (n=15) e G3 (n=15) receberá a palmilha confeccionada individualmente, após avaliação postural e clínica e terão que utilizá-las por 4 horas por dia
O grupo G2 - também receberá aplicação da acupuntura no meridiano tendino muscular uma vez por semana durante 12 atendimentos com duração de meia hora cada atendimento. Primeiramente será realizado assepsia do acupunto específico com álcool 70%. A inserção da agulha será entre 1,5 a 30 mm de profundidade, tamanho 25x30mm, individualmente e descartável após o uso. O tratamento com este meridiano segue da seguinte maneira : aplicação no meridiano da bexiga ; no acupunto B67 (localizado no leito ungueal lateral do 5ª dedo do pé) + ID18 (localizado no processo do arco zigomático) + tonificação acuponto B67. Aplicação do meridiano da vesícula biliar ; no acuponto VB44 (localizado no leito ungueal lateral do 4º dedo do pé) + ID18 + tonificação do VB34 (localizado na cabeça do 4º e 5º metatarso do pé). Aplicação do meridiano estômago ; acuponto E45 (localizado no leito ungueal do 2 dedo do pé) + ID18 + E41 (localizado na linha da articulação do tornozelo entre o tendão do tibial anterior e extensor do hálux).
Os grupos serão avaliados na pré, imediatamente após, 30, 60 e 90 dias após a intervenção.
O número total de indivíduos para os grupos serão 45, sendo 15 para cada grupo.
Critério de Inclusão :
Estudantes universitárias ; sexo feminino ; idade entre 18-30 anos ; sedentárias ;
ter acesso ao Whatsapp ; aplicativo para smartphones ; afim de para facilitar a à comunicação.
Critério de Exclusão :
Estudantes que apresentam escoliose em S ; alterações vestibulares ; auditiva e oculares não corrigida ; fraturas recentes em membro inferior ; prótese no membro inferior ; gestante ; IMC > 25.
Design do Estudo :
Ensaio clinico tratamento, randomizado-controlado, paralelo, aberto, com dois braços.

FIGURE 3.4 – Un protocole d'essai clinique en portugais brésilien

Comme le montre le tableau 3.7, ce corpus est constitué de 3 228 phrases et de plus de 48 000 tokens, soit près de 15 tokens par phrase. Par ailleurs, le corpus est constitué d'un vocabulaire de 6 453 tokens uniques, 643 phrases marquées par la négation (environ 20 % de toutes les phrases) et 819 instances de négation. Un total

de 56 marqueurs de négation sont identifiés dans le corpus. Les marqueurs de négation les plus fréquents sont : *não* (*non*, *pas*) avec 378 occurrences, *sem* (*sans*) avec 102 occurrences, ainsi que le préfixe *in* avec 88 occurrences. Les marqueurs les moins fréquents sont *negarem* (*nier*), *impedem* (*empêcher*), *recusa em* (*refus de*), etc. avec une seule occurrence de chaque marqueur dans le corpus.

TABLEAU 3.7 – Statistiques relatives aux corpus de données médicales brésiliennes

	Essais cliniques
Phrases	3 228
Tokens	48 204
Vocabulaire	6 453
Phrases marquées	643
Marqueurs (total)	819
Marqueurs (uniques)	56

Processus d'annotation

L'annotation des protocoles d'essais cliniques brésiliens a impliqué trois étudiants de la **PUCPR**. Le processus d'annotation a été basé sur des règles d'annotation proches de celles utilisées avec le corpus français. Cependant, à l'instar du corpus **CD-SCO**, les marqueurs affixaux ont été annotés, ce qui augmente considérablement le nombre total d'instances de négation (819). Par ailleurs, le manque de disponibilité des annotateurs a présenté des difficultés. En effet, par manque de temps, chaque annotateur s'est occupé d'une partie différente du corpus. Il était donc impossible de calculer un accord inter-annotateurs et aucun processus d'arbitrage n'a eu lieu. En conséquence, de nombreuses irrégularités entre annotateurs sont à déplorer.

Un extrait du corpus dans sa forme finale est visible dans le tableau 3.8. À l'exception des deux dernières colonnes, qui n'adoptent pas une annotation en **BIO** à l'instar du corpus **CD-SCO**, nous conservons le même format que pour les corpus **ESSAI** et **CAS**. L'étiquetage morpho-syntaxique est obtenu avec le **RDRPOSTagger** (D. Q. NGUYEN et al., 2014). La racinisation est réalisée avec le *Portuguese Snowball Stemmer* du *Natural Language Toolkit* [7] (**NLTK**).

7. https://www.nltk.org/

TABLEAU 3.8 – Extrait tiré du corpus de protocoles d'essais cliniques brésiliens

ID	PT	Token	Racine	PoS-tag	M–neg	P–neg
1712	0	Ausencia	ausenc	VERB	Ausencia	_
1712	1	de	de	ADP	de	_
1712	2	amputação	amput	NOUN	_	amputação
1712	3	de	de	ADP	_	de
1712	4	membro	membr	NOUN	_	membro
1712	5	superior	superior	ADJ	_	superior
1712	6	.	.	PUNCT	_	_

3.2.2 SemClinBr

Le corpus **SemClinBr** est constitué de textes cliniques qui ont été fournis par trois hôpitaux brésiliens et qui sont liés à plusieurs spécialités médicales, telles que la cardiologie, la néphrologie ou l'endocrinologie. 1 000 documents de nature diverse (comptes rendus d'hospitalisation, notes de soins infirmiers, dossiers ambulatoires, évolution clinique, etc.) ont été annotés manuellement avec les marqueurs de négation et leur portée.

L'exemple présenté dans la figure 3.5 suit la structure de notes **SOAP** (*Subjective*, *Objective*, *Assessment*, *Plan*), qui correspond à l'organisation typique des notes cliniques créées par les professionnels de santé. Dans le cas présent, il s'agit de :

— *Subjective : paciente nega queixas, nega dor, dispnéia. Restrita ao leito. Dieta via nasoenteral.*,

le patient nie les plaintes, nie la douleur, la dyspnée. Restreint au lit. Régime naso-entéral.

— *Objective : paciente em BEG, hipocorada 2+/4, hidratada, eupnéica, afebril.* [...],

patient en BEG, pâleur 2 +/4, hydraté, eupnéique, afébrile. [...]

— *Assessment : sepse pulmonar em D8 tazocin (não foi realizado doses de ATB por 2 dias). Disglicêmica.*,

septicémie pulmonaire dans la tazocine D8 (les doses d'ATB n'ont pas été données pendant 2 jours). Dysglycémique.

— *Plan : plano de manter ATBterapia até D10, solicito exs lab, mantenho demais condutas e cuidados de enfermaria.*.

prévoyez de maintenir la thérapie ATB jusqu'au 10ème jour, solliciter un laboratoire extérieur, maintenir les autres pratiques et soins infirmiers.

Sepse pulmonar em D8 tazocin (paciente não recebeu por 2 dias Atb).
Previamente HAS/ DM/ DPOC.
AVE há 2 anos (hemiparesia à E).
Escara em região sacral e perna E.
S : paciente nega queixas, nega dor, dispnéia.
Restrita ao leito.
Dieta via nasoenteral.
O : paciente em BEG, hipocorada 2+/4, hidratada, eupnéica, afebril.
Dados vitais estáveis.
C/P : mucosas úmidas e hipocoradas sem lNM .
CPP : MV+ simétrico bilateralmente , sem RA .
Prec : BCRNF sem sopro .
Abd : globoso, flácido, indolor à palpação.
Ausência de VCM, massas palpáveis ou sinais de irritação peritoneal.
Membros : sem edema, panturrilhas livres, escara em região sacral e perna E.
A : sepse pulmonar em D8 tazocin (não foi realizado doses de ATB por 2 dias).
Disglicêmica.
P : plano de manter ATBterapia até D10, solicito exs lab, mantenho demais condutas e cuidados de enfermaria.

FIGURE 3.5 – Exemple de texte clinique en portugais brésilien

Comme le montre le tableau 3.9, **SemClinBr** est constitué de 9 808 phrases et de 156 166 tokens, soit près de 16 tokens par phrase. D'autre part, le corpus est constitué d'un vocabulaire de 15 127 tokens uniques, 1 758 phrases marquées par la négation (environ 18 % de toutes les phrases) et 2 264 instances de négation. Un total de 55 marqueurs de négation sont identifiés dans le corpus. Les marqueurs de négation les plus fréquents sont : *sem* (*sans*) avec 1024 occurrences, *nega* (*nier*) avec 408 occurrences, *não* (*non*, *pas*) avec 288 occurrences, ainsi que le préfixe *a* avec 270 occurrences. Les marqueurs les moins fréquents sont *livres de* (*sans*) et *cancelada* (*annulé*) avec une seule occurrence, *nunca* (*jamais*) avec 3 occurrences et *nem* (*ni*) avec 4 occurrences.

TABLEAU 3.9 – Statistiques relatives aux corpus de données médicales brésiliennes

	Textes cliniques
Documents	1 000
Phrases	9 808
Tokens	156 166
Vocabulaire	15 127
Phrases marquées	1 758
Marqueurs (total)	2 264
Marqueurs (uniques)	55

Processus d'annotation

Le processus d'annotation de **SemClinBr** a impliqué sept étudiants ainsi qu'une infirmière. L'accord inter-annotateurs obtenu pour l'annotation des marqueurs et de leur portée est plutôt élevé, un Kappa de Cohen de 0,7414. Finalement, l'infirmière ainsi qu'un médecin ont pris part au processus d'arbitrage. Le processus d'annotation est décrit en détail dans OLIVEIRA et al., 2020.

Un extrait de **SemClinBr** dans la forme que nous exploitons est visible dans le tableau 3.10. Nous conservons le même format que pour le corpus d'essais cliniques brésiliens. Les mêmes outils que pour le précédent corpus sont utilisés pour l'étiquetage morpho-syntaxique et la racinisation.

TABLEAU 3.10 – Extrait tiré du corpus de textes cliniques brésiliens

ID	PT	Token	Lemme	PoS-tag	M–neg	P–neg
427	0	Abertura	abertur	NOUN	_	_
427	1	ocular	ocul	ADJ	_	_
427	2	espontanea	espontan	ADJ	_	_
427	3	,	,	PUNCT	_	_
427	4	não	nã	ADV	não	_
427	5	responsivo	respons	ADJ	_	responsivo
427	6	.	.	PUNCT	_	_

Conclusion

L'intérêt pour la détection automatique par apprentissage supervisé de la négation et de l'incertitude s'est accru ces dix dernières années avec la disponibilité de corpus de textes en anglais annotés (**BioScope**, **CD-SCO**, etc.). Cependant, pour beaucoup d'autres langues, le manque de textes annotés, du domaine général comme de domaines spécialisés, entrave le développement de telles approches.

Dans ce chapitre, deux corpus de textes biomédicaux français ont été présentés. Le corpus **ESSAI** est constitué de protocoles d'essais cliniques et le corpus **CAS** est constitué de 200 cas cliniques. Ces corpus ont été annotés par nos soins avec les marqueurs de négation et d'incertitude ainsi que leur portée. Pendant cette thèse, différentes versions de ces corpus annotés avec les marqueurs de négation et leur portée ont été présentées dans plusieurs publications. Dans DALLOUX, CLAVEAU et GRABAR, 2017, nous présentions la première version du corpus **ESSAI**, dans DALLOUX, CLAVEAU, GRABAR et MORO, 2018 et DALLOUX, GRABAR et CLAVEAU, 2019, une seconde version était présentée et dans DALLOUX, CLAVEAU et GRABAR, 2019 et DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020, nous présentions les versions d'**ESSAI** et **CAS** les plus proches de celles actuellement disponibles. À l'inverse, les versions annotées avec les marqueurs d'incertitude ainsi que leur portée, que nous avons décrites dans ce chapitre, n'ont connu qu'une seule version à ce jour (DALLOUX, CLAVEAU et GRABAR, 2019), sans inter-annotation ni processus d'arbitrage. Il faudra y remédier afin que ces annotations puissent être considérées comme finales. Puisque ces corpus ne contiennent pas de données sensibles, nous pouvons les rendre librement accessibles. Un formulaire de téléchargement permet d'accéder à ces données [8]. Depuis octobre 2019, nous avons reçu plusieurs dizaines de demandes de téléchargement, notamment par des chercheurs de l'**AP-HM** [9] et de l'**AP-HP** [10].

Dans la seconde partie de ce chapitre, deux corpus de textes biomédicaux brésiliens ont été présentés. Le premier est constitué de protocoles d'essais cliniques et le second de textes cliniques provenant de trois hôpitaux brésiliens. Ces corpus, annotés avec les marqueurs de négation ainsi que leur portée, ont été élaborés en collaboration avec la **PUCPR** dans le cadre du projet **FIGTEM**. Nous les utilisons dans le cadre de plusieurs publications (DALLOUX, CLAVEAU, GRABAR et MORO, 2018 ; DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020). À notre connaissance, ces corpus ne sont pas, pour le moment, librement accessibles.

8. http://people.irisa.fr/Clement.Dalloux/
9. http://fr.ap-hm.fr/
10. https://www.aphp.fr/

Chapitre 4

Détection de la négation et de l'incertitude

Sommaire

4.1 Nos approches 103
4.1.1 Plongements de mots pré-entraînés 103
4.1.2 Approches pour l'étiquetage de séquences 104
4.2 Étiquetage automatique des marqueurs 110
4.2.1 Protocole expérimental 110
4.2.2 Analyse des résultats 111
4.3 Étiquetage automatique de la portée 116
4.3.1 Protocole expérimental 116
4.3.2 Analyse des résultats 117
4.3.3 Analyse des erreurs 124

La détection d'évènements incertains et niés dans les textes est une thématique de recherche à laquelle les chercheurs en traitement automatique du langage naturel (**TALN**) s'intéressent depuis près de 20 ans (CHAPMAN et al., 2001), notamment dans le domaine biomédical dans lequel s'inscrit le projet **BigClin**. En effet, dans ce domaine, beaucoup d'informations sont apportées sous forme non structurée, c'est-à-dire sous la forme de texte libre, et, d'après de nombreux articles du domaine (CHAPMAN et al., 2001 ; ELKIN et al., 2005 ; DENNY et PETERSON, 2007 ; GINDL, KAISER et MIKSCH, 2008), la détection de ces phénomènes linguistiques joue un rôle prépondérant dans de nombreuses tâches d'extraction d'informations. Dans les textes cliniques, la négation est souvent utilisée pour exclure un diagnostic ou une prise de médicaments. Quant à l'incertitude, elle prend le plus souvent la forme d'hypothèses formulées par les médecins avec prudence. Dans le cas d'essais cliniques, par exemple, la détection de ces opérations linguistiques peut être cruciale dans la sélection, ou non, d'un patient. Il faut donc être capable de déterminer la présence, l'absence ou la possibilité de maladies et co-morbidités, de la prise d'un médicament, d'une grossesse au moment du recrutement, etc.

Sur la base des jeux de données présentés dans le chapitre précédent ainsi que des corpus **BioScope** et **CD-SCO** présentés en 1.3.2, nous proposons dans ce chapitre plusieurs approches d'étiquetage de séquences par apprentissage artificiel pour la détection de la négation et de l'incertitude dans les textes. Ainsi, dans la section 4.1, nous présentons les plongements de mots pré-entraînés ainsi que les systèmes d'étiquetage de séquences que nous développons. Ensuite, dans la section 4.2, nous présentons le protocole expérimental ainsi que l'analyse des résultats obtenus pour la tâche de détection des marqueurs. Enfin, dans la section 4.3, nous présentons le protocole expérimental, l'analyse des résultats obtenus pour la tâche de détection de la portée et nous reprenons l'analyse des erreurs publiée dans DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020.

4.1 Nos approches

Dans cette section, nous présentons les plongements de mots pré-entraînés et les systèmes d'étiquetage de séquences qui constituent les différentes approches par apprentissage artificiel que nous développons afin de résoudre plusieurs tâches : l'étiquetage automatique des marqueurs de négation et d'incertitude ainsi que l'étiquetage automatique de la portée de ces marqueurs, et ce, pour plusieurs langues (français, anglais, portugais brésilien).

4.1.1 Plongements de mots pré-entraînés

En 2.2.2, nous avons présenté différentes méthodes permettant d'obtenir des plongements de mots pré-entraînés (**word2vec**, **fastText**). Les modèles pré-entraînés que nous utilisons en entrée de plusieurs de nos systèmes d'étiquetage sont entraînés avec ces méthodes. Dans ce qui suit, nous décrivons les modèles utilisés pour chaque langue.

Modèles pour le français. Nous entraînons un modèle **word2vec** (**w2v-fr**) ainsi qu'un modèle **fastText** (**fT-fr**) à partir des corpus suivants : les articles de **Wikipedia** en français (2017), les corpus **ESSAI** et **CAS**, le corpus **CRTT**, ainsi que le corpus **QUAERO**. Les paramètres d'entraînement principaux sont : 100 dimensions, la méthode **Skip-Gram**, une fenêtre contextuelle de (-5 ; +5) mots, un nombre minimum de cinq occurrences pour chaque mot ainsi que l'échantillonnage négatif (*negative sampling*).

Modèles pour le portugais brésilien. Nous utilisons deux modèles pré-entraînés téléchargeables sur le site web du **NILC** [1] (*Núcleo Interinstitucional de Linguística Computacional*) (HARTMANN et al., 2017). Un modèle **word2vec** de 100 dimensions entraîné avec la méthode **Skip-Gram** (**w2v-pt**), ainsi qu'un modèle **fastText** de 100 dimensions entraîné avec la méthode **Skip-Gram** (**fT-pt**).

Modèles pour l'anglais. Nous utilisons deux modèles **fastText** lors de l'entraînement et de l'évaluation de nos systèmes d'étiquetage sur les jeux de données du corpus **CD-SCO**. **fT-O** est le modèle **fastText** original de 300 dimensions téléchargeable sur le site web dédié [2]. Ce modèle contient un million de vecteurs de mots entraînés sur **Wikipedia 2017**, l'**UMBC webbase corpus** ainsi que le **statmt.org news dataset**. **fT-D** est un modèle de 100 dimensions que nous entraînons avec l'algorithme **CBOW** sur l'intégralité des romans de Conan Doyle disponibles sur le site du projet Gutenberg [3]. Les paramètres d'entraînement principaux sont : une fenêtre contextuelle de (-5 ; +5) mots, un nombre minimum de cinq occurrences pour chaque mot ainsi que le softmax hiérarchique. Le corpus **CD-SCO** étant composé d'un roman et de trois nouvelles de *Sherlock Holmes*, ce modèle est motivé par l'hypothèse qu'utiliser un modèle entraîné sur des données spécifiques, non seulement au domaine

1. `http://nilc.icmc.usp.br/embeddings`
2. `https://fasttext.cc/`
3. `https://www.gutenberg.org/`

mais aussi à l'auteur, permettra de surpasser les modèles entraînés sur des données génériques. Entraîné sur un volume de données bien plus faible que le modèle original, ce modèle est aussi beaucoup plus léger.

Initialisation aléatoire. Pour toutes les langues avec lesquelles nous travaillons, nous utilisons l'initialisation aléatoire des poids de nos vecteurs de mots en entrée de nos réseaux de neurones. Avec cette méthode, les poids sont initialisés très près de zéro, mais de façon aléatoire. Par rapport à une initialisation des poids à zéro, cela permet de briser la symétrie et chaque neurone n'effectue plus le même calcul. En outre, les plongements des autres descripteurs, que nous utilisons pour la détection des marqueurs (lemmes, parties du discours) et de leur portée (parties du discours et statut de marqueur), sont toujours initialisés aléatoirement. Pendant l'entraînement, les poids des plongements sont mis à jour.

4.1.2 Approches pour l'étiquetage de séquences

En 1.1.3, nous avons présenté les approches de **TALN** les plus efficaces à ce jour pour l'étiquetage de séquences. Dans ce qui suit, nous présentons les différentes approches par apprentissage supervisé que nous développons dans le but d'étiqueter les marqueurs de négation et d'incertitude ainsi que leur portée. Nous commençons par décrire l'approche par **CRF** qui constitue notre point de référence (baseline). Ensuite, nous présentons nos approches par **BiRNN**. Enfin, nous présentons nos approches reposants sur les *transformers*, derniers travaux en date.

Approche par CRF

Les champs aléatoires conditionnels font partie des modèles d'apprentissage automatique de la famille des modèles graphiques non orientés. Modèles discriminant des champs aléatoires de Markov, les **CRF** cherchent à modéliser la probabilité conditionnelle $p(Y|X)$, pour une séquence d'entrées X et une séquence d'étiquettes Y. LAFFERTY, MCCALLUM, PEREIRA et al., 2001 définissent les **CRF** de la manière suivante :

> Soit $G = (V, E)$ un graphe tel que $Y = (Y_v)_{v \in V}$ de sorte que Y est indexé par les sommets de G. Alors (X, Y) est un champ aléatoire conditionnel si, conditionnées à X, les variables aléatoires de Markov Y_v obéissent à la propriété de Markov par rapport au graphe : $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$

Les **CRF** modélisent donc conjointement la séquence complète d'étiquettes associées à une séquence d'entrées. En d'autre termes, ils permettent de déterminer la séquence d'étiquettes la plus probable pour la séquence d'entrées.

Dans le cadre de nos travaux, nous utilisons les *linear-chain CRF* qui permettent de modéliser les dépendances séquentielles entre les étiquettes successives. D'autre

part, les **CRF** permettent d'utiliser une fenêtre contextuelle de descripteurs autour de chaque token. Dans la figure 4.1, nous présentons un *linear-chain CRF* dédié à la détection des marqueurs de négation. Dans le cas présent, la fenêtre contextuelle devrait permettre d'identifier plus précisément les marqueurs multi-mots. Dans cet exemple, la fenêtre contextuelle prend en compte le token précédent et ses descripteurs et le token suivant et ses descripteurs. Ainsi, le token *pas* est représenté par les descripteurs suivants :

— le token, le lemme et la partie du discours correspondant au token *a*,
— le token, le lemme et la partie du discours correspondant au token *pas*,
— le token, le lemme et la partie du discours correspondant au token *de*.

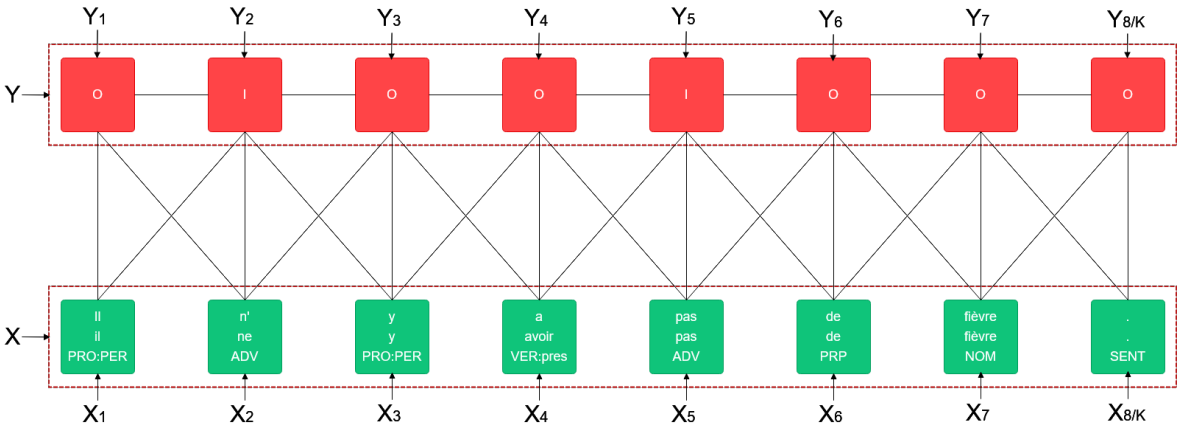


FIGURE 4.1 – Architecture d'un *Linear-chain CRF* pour l'étiquetage des marqueurs de négation avec une fenêtre contextuelle de (-1/+1)

Les systèmes basés sur les *linear-chain CRF* ont longtemps été parmi les plus performants pour les tâches d'étiquetage de séquences dans le domaine du **TALN** jusqu'à la résurgence récente des réseaux de neurones. Dans ce manuscrit, l'approche par **CRF** sert de point de référence (*baseline*) pour la détection des marqueurs de négation et d'incertitude dans les corpus en français et en portugais brésilien. Pour les corpus **BioScope** et **CD-SCO**, les résultats des systèmes présentés en 1.3.3 constituent nos points de référence.

Ce système de détection des marqueurs, basé sur les *linear-chain CRF*, utilise l'algorithme d'optimisation **L-BFGS** (***Limited-memory BFGS***), qui calcule approximativement l'algorithme **Broyden-Fletcher-Goldfarb-Shanno** en utilisant une quantité limitée de mémoire, comme algorithme du gradient. Pour la régularisation, une norme L_1 de 0.1 ainsi qu'une norme L_2 de 0.01 sont appliquées. Les descripteurs utilisés pour chaque token sont ceux présentés dans la figure 4.1 : les tokens, les lemmes et les parties du discours. La fenêtre contextuelle optimale (-4/+4) est définie empiriquement.

Approches par réseaux de neurones récurrents bidirectionnels

En 2.2.2, nous avons présenté notre approche par réseau de neurones récurrents bidirectionnel pour la classification multi-étiquette de textes cliniques. Dans le cas présent, en plus de la cellule **LSTM**, nous exploitons la cellule *gated recurrent unit* (**GRU**). La cellule **GRU** (K. CHO et al., 2014) est une variante de la cellule **LSTM** où plusieurs « portes » et états sont fusionnés. Le modèle créé par un **RNN** à cellules **GRU** est donc plus simple et, en pratique, cette approche réduit la taille du modèle ainsi que le temps de calcul nécessaire tout en conservant des résultats équivalents à ceux d'un **RNN** à cellules **LSTM**. La portée de la négation et de l'incertitude peut s'étendre avant et/ou après le marqueur, de façon continue comme discontinue. Les **BiRNN** nous semblent donc particulièrement appropriés pour résoudre la tâche de détection de la portée.

Dans le cadre de nos travaux, nous utilisons quatre **BiRNN**, dont une vue d'ensemble de l'architecture est présentée dans la figure 4.2. Chaque **BiRNN** exploite soit des cellules **LSTM** ou **GRU** et la prédiction est effectuée soit par une couche entièrement connectée avec une fonction d'activation softmax soit par une couche **CRF**. La fonction softmax est souvent utilisée dans les réseaux de neurones afin de faire correspondre la sortie non-normalisée d'un réseau à une distribution de probabilité sur les classes de sortie prédites. Elle permet donc d'attribuer une classe à chaque token. Les couches récurrentes (avant/arrière) sont constituées de 400 unités cachées chacune. À l'instar du **BiLSTM** présenté en 2.2.2, plusieurs régularisations par *dropout* sont appliquées au cours de l'entraînement. Les descripteurs utilisés pour chaque token sont : le token, le lemme et la partie du discours pour la détection des marqueurs et le token, la partie du discours et le statut de marqueur pour la détection de la portée.

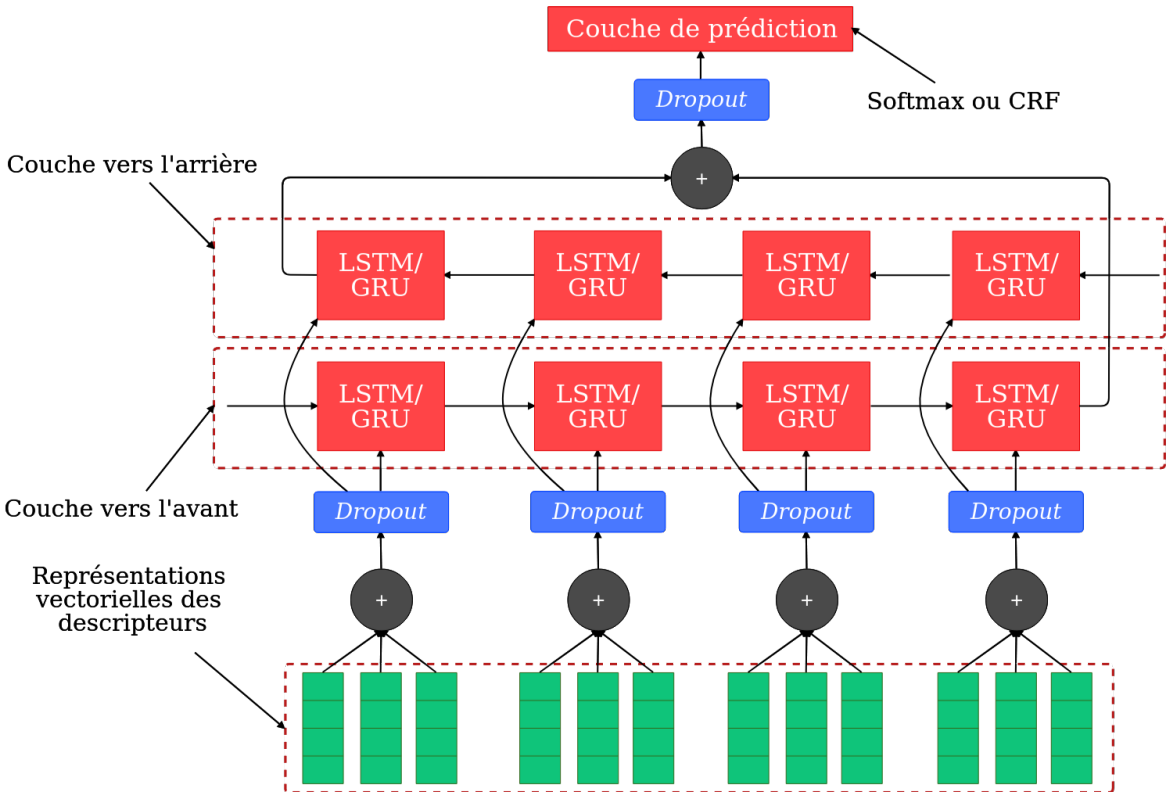


FIGURE 4.2 – Architecture de nos **BiRNN**

Approches reposants sur les *transformers*

Depuis VASWANI et al., 2017, les modèles de langage basés sur le *transformer* sont omniprésents dans le domaine du **TALN**. En effet, comme nous l'avons vu en 1.1.2 et 1.1.3, les *transformers* surpassent les **RNN** et **CNN** sur de nombreuses tâches de **TALN**. Ces performances sont dues à plusieurs éléments. Premièrement, le *transformer* est non-séquentiel, c'est-à-dire que les phrases sont traitées en entier et pas mot par mot. Cela permet aux *transformers* de ne pas souffrir de problèmes de dépendance à long terme (ou du moins à l'échelle de la phrase). Autre élément clé du *transformer* est que la couche d'attention multi-tête (*Multi-Head Attention layer*) est basée sur le mécanisme d'attention introduit par BAHDANAU, K. CHO et BENGIO, 2015. Dans un réseau de neurones, le mécanisme d'attention est utilisé pour quantifier l'interdépendance entre les mots en entrée et en sortie et entre les mots en entrée eux-mêmes afin d'identifier les mots pertinents de la phrase en entrée et de leur donner plus de poids. L'architecture du *transformer* repose donc sur la couche d'attention multi-tête qui consiste essentiellement en de multiples couches d'attention apprenant conjointement différentes représentations à partir de différentes positions. Cela est non seulement plus efficace en termes de représentation, mais aussi en termes de calcul par rapport à la convolution et aux opérations récursives. Enfin, les plongements positionnels remplacent la récurrence. En effet, pour se souvenir de la position de chaque token dans une phrase, les plongements positionnels utilisent des poids fixes ou appris qui encodent les informations relatives à la position spécifique de chaque token.

BERT pour *Bidirectional Encoder Representations from Transformers* (DEVLIN et al., 2019), est un ensemble de modèles de langage monolingues et multilingues pré-entraînés à l'aide de *transformers* bidirectionnels de façon non supervisée qui peuvent ensuite être ré-entraînés soit sur du texte brut afin de les adapter à un domaine particulier ou de façon supervisée afin d'effectuer une tâche en particulier (**NER**, étiquetage morpho-syntaxique, catégorisation de documents, etc.). Les bonnes performances de **BERT** peuvent être expliquées par plusieurs éléments : les nouvelles tâches de pré-entraînement appelées *Masked Language Model* (**MLM**) et *Next Sentense Prediction* (**NSP**), un large volume de données et une puissance de calcul conséquente pour l'entraînement. La **MLM** permet au modèle d'apprendre le contexte de chaque mot à partir des mots apparaissant à la fois avant et après un mot donné. La **MLM** transforme le texte en tokens et utilise la représentation de chaque token en entrée et en sortie pour l'entraînement. Dans les modèles les plus populaires, 15 % des tokens sont masqués, c'est-à-dire cachés pendant l'entraînement, et la fonction objectif à pour but de prédire correctement l'identité de chaque token. À la différence des méthodes bidirectionnelles par concaténation telle qu'**ELMo**, la **MLM** permet d'effectuer un apprentissage directement bidirectionnel à partir du texte. Quant à la **NSP**, elle permet au modèle d'apprendre les relations entre les phrases en prédisant si la phrase suivante d'une paire de phrases est vraiment la phrase suivante ou non. Dans ce but, le modèle est entraîné sur 50 % de paires correctes et 50 % de paires aléatoires. Les fonctions objectifs de la **MLM** et de la **NSP** sont entraînées par **BERT** de façon simultanée.

RoBERTa pour *A Robustly Optimized BERT Pretraining Approach* (Y. LIU, OTT et al., 2019) est une version ré-entraînée de **BERT** qui introduit plusieurs changements dans la méthodologie d'entraînement : la suppression de la **NSP**, l'introduction du masquage dynamique et l'augmentation du volume de données d'entraînement. Le masquage dynamique consiste à changer les tokens masqués à chaque période d'entraînement. Le pré-entraînement de **RoBERTa** prend une journée en utilisant 160 GB de données et 1 024 **GPU Tesla V100**.

Ces modèles sont entraînés sur des données en anglais ou des données multilingues. Récemment, plusieurs modèles monolingues français, tels que **CamemBERT** (L. MARTIN et al., 2020) et **FlauBERT** (LE et al., 2020), ont vu le jour. Dans ce qui suit, nous décrivons brièvement **CamemBERT**, puis nous présentons les deux modèles que nous ré-entraînons pour la détection de la négation et de l'incertitude.

CamemBERT (L. MARTIN et al., 2020) est un *transformer* multicouche bidirectionnel basé sur **RoBERTa**. La principale différence entre **CamemBERT** et **RoBERTa** est l'utilisation du masquage de mots entiers et l'utilisation de la tokenisation **SentencePiece** (KUDO et RICHARDSON, 2018) au lieu de **WordPiece** (M. SCHUSTER et NAKAJIMA, 2012). Deux modèles de **CamemBERT** sont disponibles, $\textbf{CamemBERT}_{BASE}$ et $\textbf{CamemBERT}_{LARGE}$, qui utilisent respectivement les architectures originales de $\textbf{BERT}_{BASE}$ et $\textbf{BERT}_{LARGE}$.

CamemBERT$_{Fine-tuned}$ est notre modèle simplement ré-entraîné de **CamemBERT**$_{BASE}$ pour les tâches d'étiquetage de séquences. Pour l'étiquetage de la portée, la présence de chaque marqueur est signalée par l'ajout du token *[marqueur]* dans la séquence. Par exemple, la séquence *Pas de fièvre.* devient *[marqueur] Pas de fièvre.* Ce modèle est entraîné avec un taux d'apprentissage de 3e-5 ainsi qu'un epsilon de 1e-8.

CASmemBERT est notre version adaptée au domaine de **CamemBERT**$_{BASE}$. Premièrement, nous ré-entraînons ce modèle de langage sur plus de 4 000 cas cliniques pendant 10 périodes d'entraînement à l'aide du script *run_language_modeling.py* [4]. Puis, nous entraînons ce modèle pour les tâches d'étiquetage de séquences de la même manière et avec les mêmes paramètres que pour **CamemBERT**$_{Fine-tuned}$.

Dans la section 4.2, nous présentons notre protocole expérimental ainsi que les résultats obtenus pour l'étiquetage des marqueurs de négation et d'incertitude par les approches que nous venons de présenter sur les données décrites dans le chapitre 3, ainsi qu'en 1.3.2.

4. `https://github.com/huggingface/transformers/blob/master/examples/language-modeling/run_language_modeling.py`

4.2 Étiquetage automatique des marqueurs

La détection des marqueurs est la première étape du processus de détection de la négation et de l'incertitude dans les textes. En 1.3.1, nous avons présenté les marqueurs de négation et d'incertitude, piliers des opérations de négation et de l'expression de l'incertitude. La négation et l'incertitude sont exprimées par de nombreux marqueurs que l'on retrouve dans la plupart des catégories grammaticales dans les langues avec lesquelles nous travaillons (français, portugais brésilien, anglais). L'étiquetage automatique des marqueurs peut être envisagé de différentes manières. En 1.3.3, nous avons présenté les approches proposées pour cette tâche. Ces approches reposent soit sur des ensembles de marqueurs accompagnés de règles expertes, soit sur des algorithmes d'apprentissage supervisé apprenant une fonction de prédiction à partir de données annotées. Certaines de ces approches sont hybrides, c'est-à-dire qu'elles reposent sur une combinaison de ces deux types d'approches, afin de tirer le meilleur de chacune. En 4.1, nous avons décrit les approches par apprentissage supervisé que nous retenons pour la détection des marqueurs et de leur portée. En 4.2.1, nous indiquons lesquelles de ces approches sont retenues pour cette tâche ainsi que les jeux de données sur lesquels ces approches sont entraînées et testées. Nous rappelons aussi les mesures d'évaluation utilisées. Enfin, en 4.2.2, nous présentons et analysons les résultats obtenus pour chaque corpus.

4.2.1 Protocole expérimental

Les approches que nous retenons pour la détection des marqueurs sont le **CRF**, le **BiLSTM-CRF**, **CamemBERT**$_{Fine-tuned}$, ainsi que **CASmemBERT**, approches que nous décrivons en 4.1.2. Les descripteurs utilisés en entrée du **CRF** et du **BiLSTM-CRF** sont les tokens, les lemmes et les parties du discours. Les corpus utilisés pour l'évaluation de nos approches sont les corpus **ESSAI** et **CAS** présentés en 3.1, les corpus brésiliens présentés en 3.2, ainsi que les corpus **BioScope** et **CD-SCO** présentés en 1.3.2.

Pour les corpus français et brésiliens, tous les systèmes de détection sont entraînés et testés sur les mêmes versions randomisées des corpus. Les jeux de données d'entraînement sont toujours constitués de 80 % des données et les jeux de données de test de 20 % des données. Les jeux de données de validation sont toujours constitués de 20 % des jeux de données d'entraînement. Le corpus **CAS** n'ayant pas suffisamment d'instances d'incertitude, nous l'utilisons en combinaison avec le corpus **ESSAI**. **CamemBERT**$_{Fine-tuned}$ et **CASmemBERT** sont entraînés et testés sur le corpus **CAS**.

Étant donné que **BioScope** n'est pas découpé en jeux de données d'entraînement et de test, à l'instar des approches proposées dans la littérature, notre **BiLSTM-CRF** est uniquement entraîné sur le corpus de résumés d'articles, puis testé sur les articles scientifiques et examens cliniques. Une validation croisée à 10 plis permet d'obtenir des résultats sur le corpus de résumés.

Pour le corpus **CD-SCO**, nous entraînons deux modèles : (1) le modèle **BiLSTM-CRF** qui est entraîné, validé et testé avec les jeux de données d'entraînement, de validation et de test de la campagne d'évaluation, et (2) le modèle **BiLSTM-CRF (fT-D)** qui est entraîné, validé et testé de la même manière mais le modèle **fastText fT-D** est utilisé en entrée.

Afin d'obtenir des résultats comparables pour chaque corpus, nous reprenons les mêmes méthodes d'évaluation que dans l'état de l'art. Les mesures d'évaluation utilisées sont la précision, le rappel et la F-mesure au niveau des tokens. Cependant, étant donné que les modèles pré-entraînés basés sur **BERT** utilisent leur propre algorithme de tokenisation, le nombre final de tokens n'est donc pas identique à celui d'origine pour tous les jeux de données. Dans le cas présent, cela pourrait avoir un léger impact positif comme négatif au moment de l'évaluation. Ces résultats sont donc à prendre avec du recul.

4.2.2 Analyse des résultats

Les résultats présentés dans cette sous-section sont issus de plusieurs de nos publications. Concernant les marqueurs de négation, la publication de référence est DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020, alors que pour les marqueurs d'incertitude, il s'agit de DALLOUX, CLAVEAU et GRABAR, 2019. En ce qui concerne les résultats obtenus sur les corpus anglais, **BioScope** et **CD-SCO**, ils ne font pas partie de travaux publiés sur le sujet. Les versions de nos corpus utilisées dans ces publications ne sont pas nécessairement celles présentées dans le chapitre précédent, étant donné que nos corpus ont été périodiquement augmentés et que des annotations ont été harmonisées et corrigées.

ESSAI et CAS

Le tableau 4.1 présente les résultats obtenus sur les deux corpus français. D'une part, nous constatons que la détection des marqueurs de négation en français semble être une tâche plutôt simple. En effet, les F-mesures obtenues pour l'étiquetage des marqueurs de négation sont très élevées (de 93,92 % à 97,58 %). D'autre part, la tâche d'étiquetage des marqueurs d'incertitude semble être plus complexe : les F-mesures obtenues sont bien plus faibles (de 86,88 % à 92,12 %). Nous avions constaté en 3.1 que les marqueurs d'incertitude prennent des formes bien plus variées que les marqueurs de négation et nous disposons de bien moins d'instances d'incertitude que de négation dans nos corpus. Étant donné la multitude de formes que les marqueurs d'incertitude peuvent prendre, ainsi que les nombreux contextes dans lesquels ils apparaissent et expriment ou non l'incertitude, il faudra sans doute annoter bien plus d'exemples pour atteindre des performances équivalentes à celles de la négation. En ce qui concerne l'approche, c'est notre **BiLSTM-CRF** qui est la plus performante pour la détection des marqueurs de négation. Nous entraînons toutes nos approches sur le corpus **CAS** et, dans le cas présent, les approches par *transformers* ne parviennent pas à surpasser le **BiLSTM-CRF**. Cependant, sans représentation des lemmes et des parties du discours, ces approches parviennent à obtenir de très bons résultats. Pour les marqueurs d'incertitude, à l'inverse, ce sont

nos approches par *transformers* qui sont les plus performantes sur la combinaison **ESSAI+CAS**, notamment grâce à des rappels très élevés.

TABLEAU 4.1 – Résultats des systèmes de détection des marqueurs au niveau des tokens sur les corpus en français.

	Corpus	Système	Précision	Rappel	F-mesure
Négation	ESSAI	CRF	96,05	91,89	93,92
		BiLSTM-CRF	**99,09**	**93,70**	**96,32**
	CAS	CRF	**97,05**	97,37	97,21
		BiLSTM-CRF	96,99	98,17	**97,58**
		CamemBERT$_{Fine-tuned}$	94,97	97,82	96,38
		CASmemBERT	95,26	**98,64**	96,92
Incertitude	ESSAI	CRF	**91,43**	82,76	86,88
		BiLSTM-CRF	91,00	**83,84**	**87,27**
	ESSAI+CAS	CRF	**93,93**	88,82	91,30
		BiLSTM-CRF	91,22	87,92	89,54
		CamemBERT$_{Fine-tuned}$	87,90	**95,02**	91,32
		CASmemBERT	89,68	**9**4,70	**92,12**

Corpus brésiliens

Le tableau 4.2 présente les résultats obtenus sur les deux corpus brésiliens. Ces résultats semblent confirmer la supériorité du **BiLSTM-CRF** face au **CRF**. En outre, comme pour le français, les deux approches fonctionnent moins bien sur les essais cliniques que sur les textes cliniques. Cependant, nous constatons que les F-mesures présentées ici sont largement inférieures à celles obtenues sur les corpus français. En effet, les résultats sur les essais cliniques brésiliens sont inférieurs de près de 6 points par rapport aux essais cliniques français. La F-mesure obtenue par notre **BiLSTM-CRF** sur les textes cliniques est plus convaincante (92,63) mais inférieure d'environ 5 points à celle obtenue sur le corpus **CAS**. Nous identifions les principales raisons de l'infériorité de ces résultats.

D'une part, certains marqueurs sont sous-représentés. Dans le corpus d'essais cliniques, 56 marqueurs différents sont identifiés et les marqueurs *não*, *sem* et *in* représentent près de 70 % des instances de négation. Les 53 autres marqueurs se partagent les 251 instances de négation restantes. Dans le corpus de textes cliniques 55 marqueurs différents sont identifiés et les marqueurs *sem*, *nega*, *não* et *a* représentent près de 88 % des instances de négation. Les 51 autres marqueurs se partagent les 274 instances de négation restantes. Parmi les instances de négation restantes, certains marqueurs n'apparaissent qu'une ou deux fois. C'est le cas de *Falta de*, *negarem*, *impedem* ou encore *recusa em* dans le corpus d'essais cliniques. La division de ces corpus en jeux de données d'entraînement, de validation et de test étant aléatoire, si les instances de négation qui contiennent ces marqueurs se retrouvent dans le jeu de données de validation ou de test, elles ne seront pas détectées par nos systèmes. C'est le cas des marqueurs que nous citons et de bien d'autres.

D'autre part, le corpus d'essais cliniques n'a pas bénéficié d'un processus d'arbitrage. Les désaccords entre annotateurs n'ont donc pas été résolus et affectent à la fois la précision et le rappel. Par exemple, quelques occurrences de *anti-* ont été annotées en tant que marqueurs, mais la plupart ne l'ont pas été, ce qui a provoqué des erreurs de rappel. Un autre exemple lié au manque d'inter-annotation et d'arbitrage : plusieurs occurrences de marqueurs, tels que *não*, *nehum* ou *sem*, n'ont pas été annotées, ce qui a provoqué des erreurs de précision.

TABLEAU 4.2 – Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur les corpus en portugais brésilien.

Corpus	Système	Précision	Rappel	F-mesure
Essais cliniques	CRF	90,67	86,08	88,31
	BiLSTM-CRF	**93,51**	**87,80**	**90,57**
SemClinBr	CRF	88,60	90,41	89,49
	BiLSTM-CRF	**94,64**	**90,71**	**92,63**

BioScope

Le tableau 4.3 présente les résultats des systèmes de détection des marqueurs de négation sur le corpus **BioScope**. Dans le cas présent, notre **BiLSTM-CRF** ne parvient pas à surpasser les résultats des systèmes de détection des marqueurs de négation les plus performants. Cependant, nous obtenons une F-mesure de 95,27 % en validation croisée à 10 plis (écart-type des F-mesures de 2,3 points) sur le corpus de résumés d'articles scientifiques sans post-traitement. Cette F-mesure est très proche de celle du système hybride de VELLDAL, ØVRELID et al., 2012 qui est de 96,00 %. Les F-mesures obtenues par notre système sur les deux autres corpus sont inférieures à celles des systèmes concurrents. Sur le corpus d'examens radiologiques, notre F-mesure reste élevée mais est inférieure de 2,5 points à celle de MORANTE et DAELEMANS, 2009a. Sur le corpus d'articles scientifiques, notre F-mesure est inférieure d'environ 10,5 points à celle de VELLDAL, ØVRELID et al., 2012. Ces baisses sont principalement dues au protocole expérimental globalement utilisé sur le corpus BioScope. En effet, les systèmes sont uniquement entraînés sur le corpus de résumés d'articles, puis testés sur les articles scientifiques et examens cliniques. Par conséquent, les systèmes hybrides tels que MORANTE et DAELEMANS, 2009a et VELLDAL, ØVRELID et al., 2012, reposants à la fois sur des listes de marqueurs et des méthodes par apprentissage, obtiennent de meilleurs rappels et donc de meilleures F-mesure.

TABLEAU 4.3 – Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur le corpus **BioScope**.

Corpus	Système	Précision	Rappel	F-mesure
Résumés	BiLSTM-CRF	93,39	97,23	95,27
	VELLDAL, ØVRELID et al., 2012	**93,46**	**98,73**	**96,00**
Examens	BiLSTM-CRF	96,49	93,96	95,21
	MORANTE et DAELEMANS, 2009a	**97,33**	**98,09**	**97,71**
Articles	BiLSTM-CRF	**90,21**	73,09	80,75
	VELLDAL, ØVRELID et al., 2012	85,22	**98,25**	**91,27**

Le tableau 4.4 présente les résultats des systèmes de détection des marqueurs d'incertitude sur le corpus **BioScope**. Dans le cas présent, notre **BiLSTM-CRF** obtient des résultats état-de-l'art pour deux des trois jeux de données. Sur le corpus de résumés d'articles, notre F-mesure est de 91,40 % (écart-type des F-mesures de 3,1 points), soit environ 3,2 points supérieures à celle de Qiaoming ZHU et al., 2010. Ce bon résultat est expliqué par une amélioration conséquente du rappel, environ 5,7 points. Cela est sans doute dû à la supériorité des cellules **LSTM** pour l'apprentissage de dépendances à long terme entre les descripteurs. D'autre part, sur les deux autres corpus, les résultats chutent fortement. Cependant, sur le corpus d'examens radiologiques, avec une F-mesure de 50,24 %, notre système reste le plus performant à ce jour. Sur le corpus d'articles scientifiques, comme pour la négation, notre approche ne parvient pas à égaler les résultats du système le plus performant (environ -2 points).

TABLEAU 4.4 – Résultats des systèmes de détection des marqueurs d'incertitude au niveau des tokens sur le corpus **BioScope**.

Corpus	Système	Précision	Rappel	F-mesure
Résumés	BiLSTM-CRF	93,48	**89,42**	**91,40**
	Qiaoming ZHU et al., 2010	93,14	83,74	88,19
Examens	BiLSTM-CRF	86,60	**35,39**	**50,24**
	Qiaoming ZHU et al., 2010	**91,77**	33,33	48,90
Articles	BiLSTM-CRF	**88,60**	65,66	75,42
	Qiaoming ZHU et al., 2010	82,31	**73,02**	**77,39**

CD-SCO

Le tableau 4.5 présente les résultats des systèmes de détection des marqueurs de négation sur le corpus **CD-SCO**. Conformément à notre hypothèse, le modèle **fastText fT-D** permet d'améliorer les résultats de notre **BiLSTM-CRF**. En effet, nous gagnons près de 2 points de F-mesure, ce qui nous permet de surpasser de peu le système hybride proposé par CHOWDHURY, 2012 lors de la campagne d'évaluation, mais sans aucun pré-traitement sur les données ni post-traitement sur les prédictions de notre système. Cependant, **NegBERT** (A. KHANDELWAL et SAWANT, 2020),

adaptation de **BERT** pour la détection de la négation, obtient des résultats légèrement supérieurs aux nôtres (+0,55 point de F-mesure).

TABLEAU 4.5 – Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur le corpus **CD-SCO**

Système	Précision	Rappel	F-mesure
BiLSTM-CRF (fT-D)	**93,61**	91,21	92,39
BiLSTM-CRF	92,37	88,64	90,47
NegBERT	NC	NC	**92,94**
CHOWDHURY, 2012	93,41	91,29	92,34
READ et al., 2012	91,42	**92,80**	92,10

Dans la section 4.3, de la même manière, nous présentons notre protocole expérimental ainsi que les résultats obtenus pour l'étiquetage de la portée par les approches présentées en 4.1 et sur les mêmes données.

4.3 Étiquetage automatique de la portée

Une fois les marqueurs étiquetés, le plus dur reste à faire. En effet, la détection de la portée, seconde étape du processus de détection de la négation et de l'incertitude dans les textes, est un défi bien plus complexe pour plusieurs raisons évoquées en 1.3.1. À l'instar des marqueurs, l'étiquetage automatique de la portée peut être réalisé à l'aide de systèmes experts, d'algorithmes d'apprentissage supervisé ou d'une combinaison des deux. Les approches proposées pour cette tâche ont aussi été présentées en 1.3.3. En 4.1, nous avons décrit les approches par apprentissage supervisé que nous retenons pour la détection des marqueurs et de leur portée. En 4.3.1, nous indiquons lesquelles de ces approches sont retenues pour cette tâche ainsi que les corpus sur lesquels ces approches sont entraînées et testées. Nous rappelons aussi les mesures d'évaluation utilisées, certaines étant spécifiques à cette tâche. Ensuite, en 4.3.2, nous présentons et analysons les résultats obtenus pour chaque corpus. Enfin, en 4.3.3, nous reprenons l'analyse des erreurs publiée dans DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020.

4.3.1 Protocole expérimental

Les approches que nous retenons pour la détection de la portée sont les quatre **BiRNN**, **CamemBERT**$_{Fine-tuned}$, ainsi que **CASmemBERT**. Afin de nous comparer aux systèmes experts, nous exploitons aussi l'implémentation de **French ConText** du *SIFR BioPortal Annotator*[5]. Deux évaluations des annotations produites par **French Context** sont proposées.

Les corpus utilisés pour l'évaluation de nos approches sont les corpus **ESSAI** et **CAS** présentés en 3.1, les corpus brésiliens présentés en 3.2, ainsi que les corpus **BioScope** et **CD-SCO** présentés en 1.3.2. Pour ces corpus, le protocole expérimental est le même que celui décrit en 4.2.1, à l'exception des éléments décrits ci-dessous :

— le **BiGRU-S** et le **BiGRU-CRF** sont entraînés et testés sur le corpus **ESSAI**,
— **French ConText** est uniquement testé sur le corpus **CAS**,
— **w2v-fr**, **fT-fr**, **w2v-pt** et **fT-pt** sont utilisés en entrée de nos **BiLSTM**,
— **fT-O** est utilisé en entrée de notre **BiLSTM-CRF** sur le corpus **CD-SCO**.

Afin d'obtenir des résultats comparables pour chaque corpus, nous reprenons les mêmes méthodes d'évaluation que dans l'état de l'art. Les mesures d'évaluation utilisées sont la précision, le rappel et la F-mesure au niveau des tokens ainsi qu'au niveau des portées exactement identifiées pour tous les corpus excepté **BioScope** pour lequel le pourcentage de portées correctes (**PPC**) est utilisé dans la littérature. Le **PPC** est défini comme le nombre de portées correctement annotées divisé par le nombre de total de portées.

Rappel : les modèles pré-entraînés basés sur **BERT** utilisent leur propre algorithme de tokenisation, le nombre final de tokens n'est donc pas identique à celui

5. `http://bioportal.lirmm.fr/annotator`

d'origine pour tous les jeux de données. Dans le cas présent, cela a sans doute un impact positif au moment de l'évaluation au niveau des tokens et un impact négatif au niveau des portées exactement identifiées. Ces résultats sont donc à prendre avec du recul.

4.3.2 Analyse des résultats

Les résultats présentés dans cette sous-section sont issus de deux de nos publications. Concernant la portée de la négation, la publication de référence est DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020, alors que pour la portée de l'incertitude, il s'agit de DALLOUX, CLAVEAU et GRABAR, 2019. Cependant, les résultats obtenus sur le corpus **BioScope** n'ont jamais été publiés auparavant. Les versions de nos corpus utilisées dans ces publications ne sont pas nécessairement celles présentées dans le chapitre précédent, étant donné que nos corpus ont été périodiquement augmentés et que des annotations ont été harmonisées et corrigées.

ESSAI et CAS

Le tableau 4.6 présente les résultats de nos systèmes de détection de la portée de la négation pour les corpus **ESSAI** et **CAS**. Notre première constatation est que les systèmes basés sur des cellules LSTM donnent de meilleurs résultats que ceux basés sur les cellules GRU. Nous nous y attendions car l'examen des performances des **RNN** sur plusieurs tâches de différents domaines (JOZEFOWICZ, ZAREMBA et SUTSKEVER, 2015) indique que les cellules **GRU** donnent toujours de meilleurs résultats que les cellules **LSTM**, sauf pour les tâches de **TALN**.

Quant à **French ConText**, cette approche obtient des résultats très faibles. French ConText* compare directement les annotations de référence à celles retournées par le système expert, tandis que pour French ConText**, nous réannotons le jeu de données de test afin de ne prendre en compte que les concepts médicaux annotés dans chaque portée. Par exemple, dans l'instance de négation : ne **trouvait** pas **de lésion obstructive intestinale**, la portée : *trouvait*[...]*de lésion obstructive intestinale* devient : *de lésion obstructive intestinale*. Dans les deux cas, les F-mesures obtenues sont très faibles, 29,92 % et 43,94 %. Ces mauvais résultats sont dus non seulement à la différence entre nos annotations et les concepts retournés par le système, mais aussi aux erreurs d'annotation commises par le système. Dans l'exemple que nous avons donné, **French ConText** n'inclut que le token *lésion* dans la portée.

Nous constatons aussi que nos **BiLSTM** obtiennent des résultats stables au niveau des tokens de la portée. En effet, l'écart entre la F-mesure la plus faible et la plus forte n'est que d'environ un point sur les deux corpus. Cette évaluation ne permet donc pas de déterminer avec certitude la supériorité d'une approche par **BiRNN** par rapport aux autres. Cependant, dans la plupart des cas, nos **BiLSTM-CRF** obtiennent de meilleures F-mesures que nos **BiLSTM-S** lorsque l'évaluation est plus stricte, c'est-à-dire au niveau de la correspondance exacte de la portée. En effet, évalué sur le corpus **ESSAI**, notre **BiLSTM-CRF (w2v-fr)** obtient une F-mesure de 76,51 %, soit plus de 4 points de plus que le **BiLSTM-S** le plus efficace. Sur le corpus

CAS, c'est le **BiLSTM-CRF (fT-fr)** qui obtient une F-mesure de 88,00 %, soit près de deux points de plus que le **BiLSTM-S** le plus efficace. Nous nous attendions à de tels résultats, car les **CRF** sont généralement plus efficaces que la couche softmax pour l'étiquetage de séquences.

Les résultats obtenus par nos approches sur le corpus **CAS** sont bien plus élevés que pour le corpus **ESSAI**, alors que ce dernier contient plus d'instances de négation. Les accords inter-annotateurs obtenus pour le corpus **CAS** sont plus élevés, l'annotation de ce corpus est donc plus consistante. Cela explique au moins en partie ces bons résultats, notamment au niveau de la correspondance exacte de la portée. En effet, afin d'annoter exactement une portée, il est nécessaire que les annotations soient régulières dans le corpus. Cependant, ces résultats sont aussi dus au contenu des documents. En effet, les cas cliniques sont rédigés de manière claire et concise et ne contiennent pas ou peu d'énumérations et de portées discontinues, tandis que les protocoles d'essais cliniques semblent rédigés moins consciencieusement et contiennent beaucoup d'éléments entre parenthèses (plus de 1 500 dans le corpus) et de phrases complexes contenant plusieurs marqueurs et portées à l'instar de la phrase suivante :

— Cet essai s'adresse aux patients présentant un lymphome de Hodgkin en échec de traitement (après autogreffe de moelle ou n'**ayant** pas **pu en bénéficier** du fait de l'absence de **sensibilité de la maladie à la chimiothérapie** et dans les 2 cas n'**ayant** pas **reçu préalablement d'adcetris**).

Récemment, nous avons testé les modèles de langage à base de *transformers* pour l'étiquetage de séquences. Entraînés et testés sur le corpus **CAS**, les deux modèles que nous utilisons surpassent aisément nos **BiRNN** autant au niveau des tokens que de la correspondance exacte de la portée. En effet, avec une F-mesure de 95,21 % au niveau des tokens de la portée (+4,41 points) et une F-mesure de 90,32 % au niveau de la correspondance exacte de la portée (+2,32 points), **CASmemBERT** surpasse de loin le **BiLSTM-CRF (fT-fr)**. Au niveau des tokens, la tokenisation spécifique à ces modèles avantage sans doute ces approches car elle produit plus de tokens. Les résultats seraient sans doute plus faibles à tokenisation identique. Cependant, les résultats en termes de portées exactes n'étant pas impactés positivement ou négativement par cette tokenisation, nous pouvons en conclure que ces approches sont plus performantes que les **BiRNN** pour cette tâche. En 1.1.3, nous constations que les approches par *transformers* surpassent les **RNN** sur plusieurs tâches d'étiquetage de séquences. Il n'est donc pas étonnant que ces modèles surpassent tous les autres. Cela est sans doute dû aux avantages du *transformer* que nous décrivons en 4.1.2. Il s'agit typiquement d'une meilleure représentation des relations entre les mots et de l'absence de problèmes de dépendances à long terme. Notons aussi que **CASmemBERT** surpasse **CamemBERT**$_{Fine-tuned}$. L'adaptation au domaine du modèle de langage a bien plus d'impact sur l'étiquetage de la portée que sur l'étiquetage des marqueurs. Cela est sans doute dû au fait que les marqueurs sont composés de peu de mots utilisés couramment, tandis que les portées sont composées, en partie, de mots spécifiques au domaine biomédical.

TABLEAU 4.6 – Précision (P), Rappel (R) et F-mesure (F_1) de l'étiquetage des tokens de la portée et des portées exactes de la négation. *Évaluation classique. **Évaluation seulement sur les concepts.

		Tokens de la portée			Portées exactes		
Corpus	Système	P	R	F_1	P	R	F_1
ESSAI	BiGRU-S	81,52	86,25	83,82	100	52,68	69,01
	BiGRU-CRF	83,12	84,42	83,76	100	57,07	72,67
	BiLSTM-S	**86,21**	82,85	**84,50**	100	55,61	71,47
	BiLSTM-S (w2v-fr)	83,54	83,68	83,61	100	56,59	72,27
	BiLSTM-S (fT-fr)	80,79	**86,41**	83,51	100	56,59	72,27
	BiLSTM-CRF	84,65	84,09	84,37	100	59,51	74,62
	BiLSTM-CRF (w2v-fr)	83,86	83,10	83,48	100	**61,95**	**76,51**
	BiLSTM-CRF (fT-fr)	82,38	84,84	83,59	100	59,51	74,61
CAS	BiLSTM-S	93,72	87,30	90,40	100	73,21	84,54
	BiLSTM-S (w2v-fr)	93,03	88,69	90,81	100	75,59	86,10
	BiLSTM-S (fT-fr)	91,50	88,69	90,08	100	72,02	83,74
	BiLSTM-CRF	91,87	88,59	90,20	100	68,45	81,27
	BiLSTM-CRF (w2v-fr)	91,47	88,29	89,85	100	76,19	86,49
	BiLSTM-CRF (fT-fr)	**94,82**	87,10	90,80	100	78,57	88,00
	CamemBERT$_{Fine-tuned}$	90,99	96,67	93,74	100	80,12	88,96
	CASmemBERT	93,13	**97,37**	**95,21**	100	**82,35**	**90,32**
	French ConText*	88,78	17,99	29,92			
	French ConText**	88,78	29,19	43,94			

Le tableau 4.6 présente les résultats des systèmes de détection de la portée de l'incertitude pour les corpus **ESSAI** et **CAS**. À l'instar de la négation, l'évaluation au niveau des tokens ne permet pas de déterminer avec certitude la supériorité d'un **BiRNN** par rapport aux autres. C'est aussi le cas au niveau de la correspondance exacte de la portée pour le corpus **ESSAI**. En effet, bien que nous constatons que les plongements de mots pré-entraînés améliorent le rappel et par conséquent la F-mesure de nos systèmes, les différences de performances entre la prédiction par **softmax** ou **CRF** sont trop faibles pour déterminer la supériorité d'un système. Cependant, l'ajout des exemples du corpus **CAS** à ceux du corpus **ESSAI** permet à nos **BiLSTM-CRF** d'obtenir de meilleures F-mesures que nos **BiLSTM-S** au niveau de la correspondance exacte de la portée. Pour ces corpus, l'annotation de l'incertitude n'a pas encore bénéficiée d'inter-annotation ni de processus d'arbitrage. Par conséquent, l'écart creusé par nos **BiLSTM-CRF** peut être dû à une annotation plus consistante des instances d'incertitude du corpus **CAS**. Comme pour la négation, il est aussi possible que les instances d'incertitude du corpus **CAS** soient moins complexes et donc encore plus faciles à étiqueter pour nos **BiLSTM-CRF**. Nous entraînons et testons également nos modèles par *transformers* sur la combinaison **ESSAI+CAS**. Ici aussi, ces modèles surpassent largement nos **BiRNN**. **CASmemBERT** est à nouveau le système le plus performant. Cependant, à l'inverse de la négation, les résultats obtenus par **CamemBERT**$_{Fine-tuned}$ sont équivalents en termes de tokens (-0,36 point) et plus éloignés en termes de portées exactes (-1,88 point).

TABLEAU 4.7 – Précision (P), Rappel (R) et F-mesure (F_1) de l'étiquetage des tokens de la portée et des portées exactes de l'incertitude.

		Tokens de la portée			Portées exactes		
Corpus	System	P	R	F_1	P	R	F_1
ESSAI	BiLSTM-S	89.27	82.14	85.56	100	52.76	69.07
	BiLSTM-S (w2v-fr)	88.61	84.42	**86.47**	100	56.69	72.36
	BiLSTM-S (fT-fr)	85.84	**84.58**	85.21	100	**58.27**	**73.63**
	BiLSTM-CRF	89.77	83.03	86.27	100	51.97	68.39
	BiLSTM-CRF (w2v-fr)	87.85	83.77	85.76	100	55.91	71.72
	BiLSTM-CRF (fT-fr)	**91.04**	79.61	84.94	100	57.48	73.00
ESSAI+CAS	BiLSTM-S	88.90	83.94	86.35	100	57.56	73.06
	BiLSTM-S (w2v-fr)	86.20	85.19	85.69	100	58.14	73.53
	BiLSTM-S (fT-fr)	85.15	87.46	86.29	100	59.30	74.45
	BiLSTM-CRF	89.49	81.16	85.12	100	56.98	72.59
	BiLSTM-CRF (w2v-fr)	88.48	85.04	86.73	100	65.12	78.87
	BiLSTM-CRF (fT-fr)	89.15	83.14	86.04	100	62.79	77.14
	CamemBERT$_{Fine-tuned}$	**91,07**	88,42	89,73	100	72,99	84,39
	CASmemBERT	90,85	**89,34**	**90,09**	100	**75,86**	**86,27**

Corpus brésiliens

Le tableau 4.8 présente les résultats des systèmes de détection de la portée de la négation pour les corpus brésiliens. Contrairement aux résultats obtenus sur les corpus français, pour le corpus d'essais cliniques (**ESSAI-Br**), les résultats sont à la fois faibles et très instables. En effet, la F-mesure au niveau des tokens obtenue par notre **BiLSTM-S (fT-pt)** est de 78,20 %, tandis que le deuxième meilleur score est 74,87 % et le plus faible de 71,02 %. De plus, ni les scores de précision (de 68,47 % à 78,02 %) ni les scores de rappel (de 67,25 % à 80,74 %) au niveau des tokens ne sont particulièrement stables. La même constatation peut être faite pour les F-mesures au niveau des portées exactes. En effet, malgré une F-mesure bien plus élevée au niveau des tokens, le **BiLSTM-S (fT-pt)** n'obtient pas la meilleure F-mesure au niveau des portées exactes. Sur ce corpus, nos systèmes ne parviennent pas à apprendre à annoter les portées aussi efficacement que pour le français. Le manque d'inter-annotation et d'arbitrage durant le processus d'annotation n'a pas permis, non seulement de corriger les erreurs d'annotation et les oublis, mais aussi d'harmoniser les annotations produites par les trois annotateurs.

Au contraire, les résultats obtenus sur le corpus **SemClinBr**, qui a bénéficié d'une inter-annotation et d'un processus d'arbitrage, sont stables et relativement élevés. Au niveau des tokens de la portée, notre **BiLSTM-CRF (fT-fr)** est le système le plus efficace avec une F-mesure de 84,78 %, soit environ 1,3 point de plus que le deuxième meilleur score. Au niveau des portées exactes, la F-mesure la plus élevée (83,25 %) est obtenue par le **BiLSTM-CRF (w2v-pt)** et le **BiLSTM-CRF (fT-fr)**. La combinaison de la prédiction par **CRF** et des plongements de mots pré-entraînés permet d'obtenir les F-mesures les plus élevées.

TABLEAU 4.8 – Précision (P), Rappel (R) et F-mesure (F_1) de l'étiquetage des tokens de la portée et des portées exactes de la négation.

		Tokens de la portée			Portées exactes		
Corpus	Système	P	R	F_1	P	R	F_1
ESSAI-Br	BiLSTM-S	74,40	70,77	72,54	100	30,23	46,43
	BiLSTM-S (w2v-pt)	73,10	75,66	74,35	100	**37,21**	**54,24**
	BiLSTM-S (fT-pt)	75,82	**80,74**	**78,20**	100	36,43	53,41
	BiLSTM-CRF	**78,02**	67,25	72,24	100	24,81	39,75
	BiLSTM-CRF (w2v-pt)	73,97	75,80	74,87	100	31,01	47,34
	BiLSTM-CRF (fT-pt)	68,47	73,76	71,02	100	29,46	45,51
SemClinBr	BiLSTM-S	83,50	**83,15**	83,32	98,76	68,19	80,68
	BiLSTM-S (w2v-pt)	82,46	81,88	82,17	98,73	66,76	79,66
	BiLSTM-S (fT-pt)	83,67	81,56	82,60	**99,59**	69,80	82,08
	BiLSTM-CRF	83,07	81,32	82,19	98,75	67,91	80,48
	BiLSTM-CRF (w2v-pt)	85,97	81,17	83,50	98,82	**71,92**	**83,25**
	BiLSTM-CRF (fT-pt)	**88,72**	81,17	**84,78**	98,82	**71,92**	**83,25**

BioScope

Nous rappelons que, dans le protocole expérimental habituellement utilisé pour le corpus **BioScope**, les systèmes sont uniquement entraînés sur le corpus de résumés d'articles, puis testés sur les articles scientifiques et examens cliniques. Pour le corpus de résumés d'articles, les résultats sont obtenus par validation croisée à 10 plis.

Le tableau 4.9 présente les résultats des meilleurs systèmes de détection de la portée des marqueurs de négation au niveau des tokens sur le corpus **BioScope**. Dans le cas présent, notre **BiLSTM-CRF** obtient les meilleurs résultats publiés à ce jour sur les corpus de résumés (+ 0,45 point) et d'articles (+ 2,10 points), tandis que sur le corpus d'examens radiologiques, l'approche de H. LI et LU, 2018 devance notre système (+ 0,39 point). L'écart-type des F-mesures obtenues pour chaque pli du corpus de résumés est d'environ 2,52 points. Par ailleurs, les résultats obtenus sur le corpus d'examens radiologiques sont très élevés malgré le fait que le système soit entraîné sur le corpus de résumés. Ces résultats indiquent que la plupart des instances de négation de ce corpus sont très peu complexes. À l'inverse, bien que nous obtenons des résultats état-de-l'art sur le corpus d'articles scientifiques, les résultats chutent d'environ 5.7 points de F-mesure (86,81 %) par rapport au résultats de la validation croisée (92,55 %).

TABLEAU 4.9 – Précision, Rappel et F-mesure de l'étiquetage au niveau des tokens de la portée de la négation.

Corpus	Systèmes	Précision	Rappel	F-mesure
Résumés	BiLSTM-CRF	92,40	92,71	**92,55**
	H. LI et LU, 2018	ND	ND	92,10
Examens	BiLSTM-CRF	95,52	98,76	97,11
	H. LI et LU, 2018	ND	ND	**97,50**
Articles	BiLSTM-CRF	**84,78**	**88,93**	**86,81**
	MORANTE et DAELEMANS, 2009a	84,47	84,95	84,71

Le tableau 4.10 présente les résultats des meilleurs systèmes de détection de la portée des marqueurs d'incertitude au niveau des tokens sur le corpus **BioScope**. Comme pour la négation, notre **BiLSTM-CRF** obtient les meilleurs résultats publiés à ce jour sur deux des trois corpus, tandis que, sur le corpus d'examens radiologiques, l'approche de QIAN et al., 2016 devance notre système de près de 2,4 points de F-mesure. En effet, malgré des résultats inférieurs en validation croisée, le système de QIAN et al., 2016 s'adapte mieux aux données cliniques. Les descripteurs issus d'arbres syntaxiques utilisés par leur système permettent sans doute de mieux représenter les portées de ce corpus. Ce n'est pas le cas sur le corpus d'articles scientifiques, pour lequel nous obtenons une F-mesure de plus de 3 points supérieure à celle de QIAN et al., 2016. Sur le corpus de résumés d'articles, l'écart-type des F-mesures obtenues pour chaque pli lors de la validation croisée est d'environ 0,97 points.

TABLEAU 4.10 – Précision, Rappel et F-mesure de l'étiquetage des tokens de la portée et des portées exactes de l'incertitude.

Corpus	Systèmes	Précision	Rappel	F-mesure
Résumés	BiLSTM-CRF	95,45	**97,02**	**96,23**
	QIAN et al., 2016	**95,95**	95,19	95,56
Examens	BiLSTM-CRF	81,73	**94,89**	87,82
	QIAN et al., 2016	**86,85**	93,84	**90,21**
Articles	BiLSTM-CRF	**87,85**	**91,88**	**89,82**
	QIAN et al., 2016	86,78	86,59	86,69

Le tableau 4.11 présente les résultats des meilleurs systèmes de détection de la portée de la négation et de l'incertitude en pourcentage de portées correctes sur le corpus **BioScope**. Sur le corpus de résumés d'articles, les écarts-types des **PPC** obtenus par notre système lors des validations croisées sont de 3,10 % pour la négation et 4,38 % pour l'incertitude. Dans les deux cas, notre système ne parvient pas à égaler les résultats des systèmes les plus performants.

L'approche de SERGEEVA et al., 2019, basée sur **BERT**, devance largement les autres systèmes sur le corpus de résumés, démontrant une fois de plus la supériorité des approches par *transformers*. Les auteurs ne donnent pas de résultats en

termes de tokens à cause de la tokenisation spécifique de **BERT**. Avec l'évaluation en **PPC**, cette tokenisation ne peut pas avantager leur système. Cependant, les résultats de cette approche sur les autres corpus ne sont pas comparables aux autres car le protocole expérimental n'est pas le même.

Sur le corpus d'examens radiologiques, pour la négation, c'est à nouveau le système de H. LI et LU, 2018 qui est le plus performant (+1,2 points par rapport notre approche), tandis que pour l'incertitude, c'est à nouveau le système de QIAN et al., 2016 qui est le plus performant (+5,5 points par rapport notre approche). Sur le corpus d'articles scientifiques, pour la négation, c'est le système de Qiaoming ZHU et al., 2010 qui est le plus performant (+1,8 point par rapport notre approche), tandis que pour l'incertitude, c'est le système de ZOU, G. ZHOU et Q. ZHU, 2013 qui est le plus performant (+5 points par rapport notre approche).

TABLEAU 4.11 – Résultats en pourcentage de portées correctes.

	Systèmes	Résumés	Examens	Articles
Négation	BiLSTM-CRF	79.80	93,18	62,22
	SERGEEVA et al., 2019	**87.03**	NC	NC
	H. LI et LU, 2018	84,10	**94,40**	60,10
	Qiaoming ZHU et al., 2010	81,84	89,79	**64,02**
Incertitude	BiLSTM-CRF	82,35	68,42	62,34
	SERGEEVA et al., 2019	**89.28**	NC	NC
	QIAN et al., 2016	85,75	**73,92**	59,82
	ZOU, G. ZHOU et Q. ZHU, 2013	84.21	72.92	**67.24**

Par manque de temps, aucun modèle de plongements de mots pré-entraînés n'a été utilisé dans le cadre de nos expérimentations sur BioScope. Étant donné les gains de performance obtenus avec nos modèles pré-entraînés sur les corpus français et brésiliens, nous pourrions sans doute gagner en performance en utilisant des plongements de mots adaptés au domaine en entrée de notre **BiLSTM-CRF**.

CD-SCO

Le tableau 4.12 présente les résultats des systèmes de détection de la portée des marqueurs de négation sur le corpus **CD-SCO**. Dans le cas présent, notre **BiLSTM-CRF fT-D** obtient la F-mesure la plus élevée à ce jour au niveau des portées exactes (82,41 %). Notre **BiLSTM-CRF** obtient des résultats légèrement inférieur, tandis que la F-mesure du **BiLSTM-CRF fT-O** au niveau des portées exactes chute. Ces résultats confirment notre hypothèse, selon laquelle le modèle **fastText fT-D** spécifique au domaine et à l'auteur car entraîné sur les romans et nouvelles de Conan Doyle, permet d'augmenter sensiblement les résultats. Récemment, **NegBERT** (A. KHANDELWAL et SAWANT, 2020), une adaptation de **BERT** pour la détection de la négation, surpasse largement notre système au niveau des tokens. Cependant, les résultats en termes des portées exactes ne sont pas donnés. Il aurait été intéressant de voir la progression des résultats à ce niveau étant donné que la tokenisation spécifique de **BERT** est un avantage au niveau de l'évaluation en termes de tokens.

TABLEAU 4.12 – Précision (P), Rappel (R) et F-mesure (F_1) de l'étiquetage des tokens de la portée et des portées exactes de la négation.

	Tokens de la portée			Portées exactes		
Système	P	R	F_1	P	R	F_1
BiLSTM-CRF fT-D	94.38	**87.25**	**90.67**	99.46	**70.34**	**82.41**
BiLSTM-CRF fT-O	94.37	84.40	89.11	99.41	64.26	78.06
BiLSTM-CRF	**95.19**	84.40	89.47	99.46	69.58	81.88
NegBERT	NC	NC	**92,36**	NC	NC	NC
H. LI et LU, 2018	94,00	85,30	89,40	100	69,10	81,70
H. LI et LU, 2018	94,80	83,20	88,60	100	69,50	82,00
FANCELLU, LOPEZ et WEBBER, 2016	92,62	85,13	88,72	99,40	63,87	77,70

4.3.3 Analyse des erreurs

Dans DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020, nous analysions les erreurs fréquemment commises par nos **BiRNN** dans le cadre de la tâche d'étiquetage de la portée de la négation sur les jeux de données de test des corpus français et brésiliens. Nous présentions également plusieurs erreurs résolues par notre **BiLSTM-CRF fT-D** par rapport à l'approche de FANCELLU, LOPEZ et WEBBER, 2016 sur le corpus **CD-SCO**. Dans ce qui suit, nous reprenons cette analyse et nous l'étendons en comparant ces erreurs aux prédictions de **CASmemBERT**.

Afin d'étudier les types d'erreurs fréquentes, les phrases contenant au moins une erreur de prédiction ont été examinée manuellement et les causes des erreurs ont été annotées. Dans ce qui suit, pour chaque langue, nous présentons plusieurs exemples d'erreurs fréquemment commises par nos systèmes. Dans les exemples ci-dessous, les marqueurs de négation sont soulignés, les portées sont en gras et les prédictions sont entre crochets.

Dans le premier exemple, notre **BiLTSM-CRF** échoue à étiqueter l'adjectif *rénale*. Dans de nombreux cas, les portées associées au marqueur *sans* n'incluent qu'un seul token, ce qui peut être à l'origine de cette erreur qui impacte le rappel. Cette erreur n'est pas commise par **CASmemBERT** dont la prédiction est exacte.

1. Le patient sortira du service de réanimation guéri et sans [**insuffisance**] **rénale** après huit jours de prise en charge et cinq séances d'hémodialyse.

Ce type d'erreur se produit également avec les prépositions manquées (impact sur le rappel) ou incorrectement incluses dans la portée (impact sur la précision). Cela se produit surtout avec la préposition française *de* et la brésilienne *da* comme dans l'exemple suivant dans lequel *da última consulta* a été incorrectement prédit dans la portée du marqueur *nega* :

2. NEGA [**INTERCORRÊNCIAS** DA ÚLTIMA CONSULTA] PRA CÁ

   Nie les affections intercurrentes de la dernière consultation à maintenant

L'erreur présentée dans l'exemple suivant affecte la précision. Ici, le **BiLSTM-CRF** prédit à tort que quasiment tous les tokens de la phrase sont dans la portée. Dans les données de référence, le marqueur *aucun* se trouve souvent au début des phrases et dans les phrases avec plusieurs instances de négation. Principalement entraîné sur ce type d'exemples, le modèle peut tenter de reproduire ces structures et ainsi provoquer des erreurs de prédiction. Cependant, ici aussi, **CASmemBERT** ne commet pas d'erreur et annote parfaitement la portée. Les plongements de mots contextuels du modèle permettant sans doute de différencier les portées associées aux différentes positions du marqueur *aucun* dans les phrases.

3. [Les colorations spéciales (PAS, coloration de Ziehl-Neelsen, coloration de Grocott]) ne [**mettaient en évidence**] aucun [**agent pathogène**].

L'exemple suivant est plus complexe et les erreurs de prédiction affectent à la fois la précision et le rappel. Dans le cas présent, nous avons deux instances de négation avec les mêmes marqueurs : *n'[...]pas*. Elles sont identifiées par deux couleurs différentes. La séquence *Le retrait du matériel d'ostéosynthèse incriminé* fait partie de la seconde instance de négation mais se trouve avant la première. La portée de la seconde instance est donc discontinue. Étant donné que dans la plupart des cas, la portée de *n'[...]pas* s'étend à la droite du marqueur jusqu'au premier signe de ponctuation, notre système ne parvient pas à l'annoter. C'est aussi pour cette raison que la séquence *à notre patient asymptomatique* est annotée par erreur. Il s'agit de la seule instance de négation au subjonctif passé dans le corpus **CAS**. Ici, **CASmemBERT** fait mieux mais n'annote pas parfaitement la portée (4'). Deux erreurs sont commises pour la seconde instance : Les tokens *à* et *to* (tokenisation **SentencePiece**) sont étiquetés par erreur.

4. **Le retrait du matériel d'ostéosynthèse incriminé** n'[**est**] pas [**systématique**], ce qui explique qu'il n'[**ait**] pas [**été proposé** à notre patient asymptomatique].

4'. [**Le retrait du matériel d'ostéosynthèse incriminé**] n'[**est**] pas [**systématique**], ce qui explique qu'il n'[**ait**] pas [**été proposé** à] notre patient asymp[to]matique.

Les erreurs présentées dans les deux exemples suivants sont dues à des erreurs d'annotation détectées dans le cadre de cette analyse. Ainsi, dans le cinquième exemple, le modèle prédit correctement chaque token. Cependant, la virgule suivant le marqueur *pas* a été annotée à tort dans la portée. Cette erreur affecte faiblement le rappel au niveau des tokens mais fortement au niveau des portées exactes. Ici, au contraire, **CASmemBERT** fait des erreurs en considérant le second *ne* de l'instance comme marqueur alors qu'il ne l'est pas. Ainsi, le token *prend* et la séquence de tokens *de l'alcool* sont étiquetés dans la portée par erreur. Dans le sixième exemple, ce sont les tokens *ce* et *est* qui ont été annotés par erreur. Cette erreur s'est produite lors de la pré-annotation automatique du corpus **CAS** et n'a pas été corrigée. En effet, la particule *n'* a été annotée comme marqueur de négation et sa portée a été annotée en conséquence. Lors de l'étape d'annotation manuelle, l'annotation de *n'* comme

marqueur a été retirée, cependant *ce* et *est* n'ont pas été retirés de la portée. L'impact de cette erreur est le même que pour l'exemple précédent. Pour cet exemple, la prédiction de **CASmemBERT** est identique.

5. La patiente ne [**fume**] pas, ne prend que très rarement de l'alcool et n'[**a**] pas [**d'allergie aux médicaments**].

5'. La patiente ne [**fume**] pas, ne [prend] que très rarement [de l'alcool] et n'[**a**] pas [**d'allergie aux médicaments**].

6. **Ce** n'**est** que 48 heures après la dernière dose que les troubles visuels et les hallucinations disparaissent complètement, sans [**laisser de séquelles**].

Les deux exemples suivants sont issus du corpus d'essais cliniques brésiliens. Dans les deux cas, les irrégularités dans l'annotation posent problème. Dans le septième exemple, d'une part, notre approche ne parvient pas à annoter tous les concepts médicaux de l'énumération et, d'autre part, la préposition *de* est prédite dans la portée car les prépositions sont généralement annotées dans ce corpus. Par ailleurs, la préposition *de* de *Ausencia de* fait partie du marqueur mais n'est pas annotée. Annotée correctement, la portée s'étendrait de *diagnóstico* à *cerebral*. Dans le huitième exemple, d'une part, le système échoue à inclure *diagnosticadas* dans la portée et, d'autre part, l'adjectif *outras* est inclus dans la portée car il fait partie du syntagme marqué et devrait être annoté. Annotée correctement, la portée s'étendrait de *apresentem* à *diagnosticadas*.

7. Ausencia de diagnóstico [de **doenças neuromusculares**], [**trauma**], **tumores** ou [**abscessos raquimedulares**], **hemiplegia**/ **paresia**, **lesão de plexo** ou [**encefalopatia**] **cerebral**.

   Absence de diagnostic de maladies neuromusculaires, traumatismes, tumeurs ou abcès rachidiens, hémiplégie / parésie, lésion du plexus ou encéphalopathie cérébrale.

8. que não apresentem [outras **doenças neurológicas**] ou [**ortopédicas**] **diagnosticadas**.

   N'a pas d'autres maladies neurologiques ou orthopédiques diagnostiquées.

Les trois exemples suivants sont issus du corpus **CD-SCO**. Les erreurs illustrées dans ces exemples sont commises par le **BiLSTM** de FANCELLU, LOPEZ et WEBBER, 2016 mais pas par notre système. Les exemples 10 et 11 sont obtenus en utilisant leur système [6]. Dans le neuvième exemple, la portée prédite par leur système inclut le début du syntagme verbal *I knew* dans la portée, sans doute car *that* est omis (*I knew* [*that*] *you could*) et ne délimite pas les propositions. Dans le dixième exemple, leur système fait l'erreur d'inclure le premier syntagme de la phrase, *just the word*, dans la portée. Dans le jeu de données d'entraînement, la portée des instances de négation dont le marqueur est *nothing* s'étend quasiment toujours, en partie, à la gauche du marqueur à l'instar des trois exemples ci-dessous :

— *...and yet **holmes had said** nothing,...*

6. `https://github.com/ffancellu/NegNN`

— *And then **you heard** nothing **until you read the reports of the death in the paper**?*

— *...**although he would walk in his own ground,** nothing **would induce him to go out upon the moor at night**.*

Leur **BiLSTM** tente donc de reproduire cette structure. Dans le dernier exemple, leur système échoue à étiqueter l'auxiliaire *can* ainsi que le pronom personnel *me* dans la portée.

9. You felt so strongly about it that [I knew **you could**] not [**think of Beecher without thinking of that also**].

   Votre indignation était telle que j'étais sûr que vous ne pouviez pas penser à Beecher sans penser également à cela.

10. [Just the word], nothing [**more**].

    Juste le mot, rien de plus.

11. "Well, **can** [**you give**] **me** no [**further indications**]?"

    Eh bien, ne pouvez-vous me donner aucune autre indication ?

Notre **BiLSTM-CRF fT-D** ne commet pas ces erreurs. Les principales différences entre notre système et le leur pouvant expliquer ceci sont : (1) la prédiction par **CRF** qui traite la séquence en entier et obtient généralement de meilleurs résultats que la couche softmax pour l'étiquetage de séquences, (2) l'utilisation du *dropout* qui minimalise le surentraînement et (3) les plongements de mots adaptés au domaine (**fastText fT-D**) qui offrent une meilleure représentation vectorielle des mots en entrée du système.

Conclusion

Dans ce chapitre, plusieurs approches reposant sur des algorithmes d'apprentissage artificiel et profond sont proposées pour l'étiquetage des marqueurs de négation et d'incertitude et/ou de leur portée. Les approches et résultats présentés dans ce chapitre ont, pour la plupart, fait l'objet de publications (DALLOUX, CLAVEAU et GRABAR, 2019 ; DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020), elles-mêmes précédées par d'autres travaux (DALLOUX, 2017 ; DALLOUX, CLAVEAU et GRABAR, 2017 ; DALLOUX, CLAVEAU, GRABAR et MORO, 2018 ; GRABAR, CLAVEAU et DALLOUX, 2018 ; DALLOUX, GRABAR et CLAVEAU, 2019).

Nous avons commencé par décrire les modèles de plongements de mots pré-entraînés que nous utilisons en entrée de nos systèmes d'étiquetage. D'après nos résultats, l'utilisation de ces modèles permet d'améliorer les performances de ces systèmes. Par exemple, sur le corpus **CD-SCO**, le modèle **fastText fT-D** permet de gagner environ 2 points de F-mesure pour l'étiquetage des marqueurs, 1,2 point pour l'étiquetage de la portée au niveau des tokens et environ 0,5 point au niveau des portées exactes. Sur les corpus français, nos **BiRNN** utilisant les modèles pré-entraînés obtiennent des résultats équivalents ou supérieurs à ceux de l'initialisation aléatoire. Nous faisons la même constatation sur les corpus brésiliens.

Nous avons ensuite présenté les approches développées pour l'étiquetage de séquences : l'approche par **CRF** constituant notre point de référence pour la détection des marqueurs sur les corpus français et brésiliens, les approches par **BiRNN** utilisées à la fois pour l'étiquetage des marqueurs et de leur portée et les approches par *transformers* que nous utilisons à la fois pour l'étiquetage des marqueurs et de leur portée. Pour l'étiquetage des marqueurs, les résultats obtenus par notre **BiLSTM-CRF** sont quasiment toujours supérieurs à ceux du **CRF** et les approches basées sur **CamemBERT** obtiennent des résultats soit proches soit supérieurs. Pour l'étiquetage de la portée, lorsqu'ils sont entraînés avec des plongements de mots pré-entraînés, ce sont nos **BiLSTM-CRF** qui sont les **BiRNN** les plus performants. Cependant, **CASmemBERT** surpasse significativement tous ces modèles. Nous rejoignons en cela les conclusions établies sur les corpus anglais, où, lorsque proposées, les approches basées sur **BERT** sont aussi les plus performantes.

Après l'analyse de ces résultats, nous avons repris l'analyse des erreurs publiées dans DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020 et l'avons étendu en ajoutant les prédictions de **CASmemBERT** pour plusieurs exemples. Cette analyse nous permet d'isoler les cas difficiles, voire uniques, posant problème à notre **BiRNN** le plus performant ainsi que les erreurs d'annotation. Nous y isolons aussi les cas résolus par notre approche par rapport au **BiRNN** de FANCELLU, LOPEZ et WEBBER, 2016.

Conclusion générale

Dans cette thèse, nous avons apporté des contributions à deux tâches de **TALN** : la classification multi-étiquette de textes cliniques et la détection de la négation et de l'incertitude. Dans cette conclusion générale du manuscrit, nous rappelons la nature de ces contributions, nous les discutons et proposons des pistes de travail pour les deux tâches abordées.

Classification multi-étiquette de textes cliniques

Dans le chapitre 2, nous avons présenté les travaux réalisés en collaboration avec l'équipe **Données massives en santé** (**DMS**) du **CHU** de Rennes pour la tâche de classification multi-étiquette de textes cliniques (DALLOUX, CLAVEAU, CUGGIA et al., 2020). Entreprise à la fois à des fins de recherche ainsi que dans la perspective de la création d'un outil d'aide au codage des textes cliniques, cette tâche est particulièrement importante pour les établissements de santé. En effet, l'automatisation de cette tâche complexe et chronophage est une thématique de recherche à laquelle les chercheurs en informatique médicale s'intéressent depuis plus de 20 ans. Cependant l'automatisation totale de cette tâche dans un futur proche nous semble irréalisable car il est important de prendre en compte que, pour les établissements de santé, la classification de ces documents ne peut pas être approximative. En d'autres termes, le codage final d'un texte ne peut pas contenir de codes erronés et, autant que possible, aucun code valide ne doit manquer, ce qu'aucun système de classification automatique ne permet (il faut noter que le codage manuel n'est pas non plus exhaustif). Mettre en place un système d'aide au codage est cependant possible. Le codage manuel est très précis mais souffre de deux problèmes que l'outil informatique peut aider à résoudre : il est coûteux, et son rappel est largement améliorable.

Dans ce but, à l'aide du corpus de textes cliniques et d'un serveur équipé de **GPUs** mis à notre disposition par le **CHU** de Rennes, nous avons proposé plusieurs approches s'appuyant sur des algorithmes d'apprentissage neuronal. Comparé aux corpus de textes cliniques français utilisés lors des campagnes d'évaluation du **CLEF eHealth Evaluation Lab**, d'une part, notre corpus contient bien moins de documents et nous avons bien plus de classes à attribuer et, d'autre part, les textes cliniques avec lesquels nous travaillons sont bien plus longs et bruités. Ainsi, même si la cible (codes PMSI) est similaire, il s'agit bien de deux tâches très différentes. De plus, dans les données cliniques, la plupart des classes sont sous-représentées. Pour la classification, nous proposons deux approches neuronales : un **BiLSTM** et un **CNN**. Le **CNN** surpasse largement le **BiLSTM**. Si la cellule **LSTM** est plus efficace

que la cellule **RNN** de base pour mémoriser à les dépendances à long terme, la longueur des textes cliniques avec lesquels nous travaillons (1 345 tokens en moyenne) semble impacter négativement les performances. En effet, lors d'expérimentations préliminaires sur les textes de la campagne d'évaluation de **CLEF eHealth 2017** qui sont très courts, le **BiLSTM** obtenait des résultats équivalents à ceux du **CNN**. Le **CNN** souffre moins de ce problème notamment grâce à la connectivité locale (*sparse connectivity*) (GOODFELLOW, BENGIO et COURVILLE, 2016) qui permet de réduire le nombre de paramètres à estimer tout en rendant leur estimation statistique plus robuste. En **TALN**, cela permet d'identifier les tokens pertinents pour la tâche de classification à accomplir parmi plusieurs centaines/milliers en entrée. Nous entraînons plusieurs **CNN** afin de déterminer les hyperparamètres les plus efficaces et gagnons plus de 8 points de F-mesure par rapport au **CNN** de base. Cependant, les résultats que nous obtenons restent relativement faibles. Afin d'obtenir de meilleurs résultats, nous proposons de réduire le nombre de classes à attribuer par le système tout en augmentant le nombre d'exemples par classe. Cela est rendu possible par la nature des classes. En effet, la **CIM-10** est une classification médicale hiérarchisée qui contient plus de 10 000 codes relatifs aux actes médicaux, consultations et autres prestations réalisables. Nous réduisons donc les codes des niveaux hiérarchiques les plus bas à leur niveau hiérarchique à 3 caractères, ce qui permet de passer de 6 113 à 1 549 classes. Les résultats que nous obtenons alors sont bien plus élevés. Il nous semble donc que cette réduction est une option valable car elle permet de réduire considérablement la complexité de la tâche tout en conservant un niveau de détail suffisant pour aiguiller les annotateurs.

Il est évident qu'il existe une grande marge de progression sur ce sujet. En termes d'approches, les *transformers* nous semblent tout indiqués. Cependant, ces modèles ont une contrainte sur la longueur maximale de la séquence après tokenisation. Par exemple, $\textbf{BERT}_{BASE}$ ne peut traiter que des séquences de 512 tokens maximum. Comme nous venons de le rappeler, la plupart des textes cliniques à notre disposition dépassent cette limite. Plusieurs solutions sont envisageables. La solution la plus simple serait de pré-entraîner un modèle de langage en augmentant cette limite à 1024 voire 2048. Cependant, il est très long d'entraîner un modèle de zéro et augmenter cette limite le prolongerait davantage. Une autre solution serait de développer des techniques de segmentation spécifiques afin de retirer des textes cliniques les sections ne contenant pas d'informations relatives aux codes **CIM-10** annotés. Par exemple, retirer les sections relatives aux praticiens et au **CHU** du début et de la fin du document présenté dans la figure 2.1 permettrait de réduire sa taille tout en conservant les informations relatives aux évènements médicaux.

Par ailleurs, même si les approches basées sur **BERT** semblent relativement efficaces pour prédire les classes faiblement présentes dans les données d'entraînement (CHALKIDIS et al., 2019), il sera tout de même nécessaire de collecter davantage de textes annotés. En effet, d'une part, le corpus dont nous disposons ne couvre que 6 113 des 14 400 codes de la **CIM-10** ou, avec la réduction, 1 549 des 2 038 codes à trois caractères et, d'autre part, 1 621 des 6 113 codes présents ne comptent qu'une seule occurrence. Ainsi, tel qu'indiqué dans le tableau 2.1, de nombreux codes se retrouvent absents des données d'entraînement. Par conséquent, afin d'améliorer la

représentativités des classes rares, il sera non seulement nécessaire d'augmenter le volume global de données annotées, mais aussi de collecter ces données en fonction des codes sous-représentés. Cependant, si augmenter le volume global de données ne présenterait pas d'obstacles majeurs, de tels documents étant produits tous les jours dans les établissements de santé, certains codes étant rares par nature (codes relatifs aux maladies rares, d'utilisation particulière, sous-utilisés par les annotateurs), il sera sans doute difficile de collecter suffisamment d'exemples pour chaque code. Aussi, GOBEILL, RUCH et MEYER, 2020 montrent que la prédiction des codes est sensible à la temporalité. En d'autre termes, pour une année précise, augmenter le volume de données global permet d'améliorer les performances de leur approche, mais ajouter des documents d'années passées n'a aucun effet. Le développement d'approches de *few/zero-shot learning* pourraient aussi permettre de réduire l'impact de ces problèmes de sous/non-représentation.

Par ailleurs, approuvée par l'**Assemblée mondiale de la Santé** en mai 2019, la **CIM-11** deviendra la **CIM** de référence à partir de janvier 2022[7]. Dès son adoption pour la tarification à l'activité ou l'épidémiologie, il sera nécessaire de ré-entraîner les outils d'aide au codage par apprentissage, ce qui implique la disponibilité massive de données annotées. Afin de ne pas être pris de court, il serait préférable de procéder à l'annotation anticipé de textes cliniques actuels ou passés avec cette nouvelle classification. Pour les documents annotés avec la **CIM-10**, la question de la conversion automatique de leur annotation vers la **CIM-11** se pose aussi.

Enfin, nos travaux relatifs à cette tâche sont peu avancés pour plusieurs raisons. D'une part, les données et capacités de calcul ont été mises à notre disposition plus de deux ans après le début du projet. En troisième année de thèse, le temps de travail qu'il était possible de consacrer à cette tâche était très limité. Par ailleurs, il était impossible d'accéder à ces ressources à distances et les capacités de calcul mises à notre disposition étaient partagées avec le reste de l'équipe et largement inférieures à celles de l'**IRISA**[8]. Le manque de temps ainsi que devoir se rendre au **CHU** et partager les capacités de calcul nécessaires à l'entraînement de nos modèles nous a grandement compliqué la tâche.

Détection de la négation et de l'incertitude

Dans les chapitres 3 et 4, nous présentons nos contributions pour la tâche de détection de la négation et de l'incertitude dans les textes (DALLOUX, 2017; DALLOUX, CLAVEAU et GRABAR, 2017; DALLOUX, CLAVEAU, GRABAR et MORO, 2018; GRABAR, CLAVEAU et DALLOUX, 2018; DALLOUX, GRABAR et CLAVEAU, 2019; DALLOUX, CLAVEAU et GRABAR, 2019; DALLOUX, CLAVEAU, GRABAR, OLIVEIRA et al., 2020). Pré-traitement indispensable pour de nombreuses tâches d'extraction d'informations dans le domaine médical, la détection de ces opérations

7. https://www.who.int/classifications/icd/en/
8. http://igrida.gforge.inria.fr/

linguistiques est une thématique de recherche étudiée depuis près de 20 ans. Cependant, jusqu'à nos travaux, il n'existait pas de corpus français annotés librement accessibles pour la recherche.

Dans le chapitre 3, nous avons présenté les différents corpus constitués dans le cadre de cette thèse. Nos deux corpus français, **ESSAI** et **CAS**, sont constitués respectivement de protocoles d'essais cliniques et de cas cliniques. Les marqueurs de négation et d'incertitude ainsi que leur portée y sont annotés. Pour la négation, les accords inter-annotateurs obtenus sont très élevés. Cependant, les marqueurs d'incertitude ainsi que leur portée n'ont été annotés que par un seul annotateur. Par conséquent, aucun accord inter-annotateur ni processus d'arbitrage n'a été réalisé. Afin de valider définitivement ces annotations il faudra au moins impliquer un second annotateur et les harmoniser. Un formulaire de téléchargement permet d'accéder à ces données [9]. Nous présentons aussi deux corpus brésiliens. Le premier est constitué de protocoles d'essais cliniques et a été annoté avec les marqueurs de négation ainsi que leur portée à l'occasion d'une mission d'un mois au Brésil à l'Université Pontificale Catholique du Paraná. Trois annotateurs ont été impliqués. Cependant, aucun accord inter-annotateur ni processus d'arbitrage n'a pu être réalisé. Par conséquent, les annotations sont très irrégulières. Le second corpus est constitué de textes cliniques provenant de plusieurs hôpitaux brésiliens. Les marqueurs de négation ainsi que leur portée y ont été annotés par plusieurs étudiants ainsi qu'une infirmière. L'accord inter-annotateurs obtenu est plutôt élevé. À notre connaissance, ces corpus ne sont pas librement accessibles. Il est donc pour le moment impossible de reproduire les travaux réalisés avec ces corpus.

Dans le chapitre 4, nous avons présenté les travaux réalisés pour la tâche de détection de la négation et de l'incertitude. Cette tâche peut être divisée en deux sous-tâches, l'étiquetage des marqueurs et l'étiquetage de la portée. Afin de résoudre ces tâches d'étiquetage de séquences, nous proposons plusieurs approches par apprentissage profond que nous entraînons sur les corpus dont nous disposons. Ces approches sont basées sur les réseaux de neurones récurrents bidirectionnels ainsi que sur **CamemBERT** (L. MARTIN et al., 2020), une adaptation française de **BERT** (DEVLIN et al., 2019). Pour ces deux tâches et sur la plupart des corpus, ces approches obtiennent des résultats soit proches de l'état de l'art soit supérieurs. Les approches par *transformers* sont particulièrement performantes pour la détection de la portée. Profondément bidirectionnelles grâce à la **MLM** présentée en 4.1.2, ces approches sont aussi plus robustes grâce à leurs *tokenizers* par apprentissage non-supervisé indépendants de la langue. Par exemple, **SentencePiece** (KUDO et RICHARDSON, 2018), le *tokenizer* de **CamemBERT**, implémente plusieurs algorithmes permettant de segmenter les phrases en *subword units* tels que le *byte pair encoding* (SENNRICH, HADDOW et BIRCH, 2016) ou le *unigram language model* (KUDO, 2018). Entraîné directement à partir de phrases brutes, **SentencePiece** traite les phrases comme des séquences de caractères Unicode. Il n'y a pas de logique dépendante de la langue. *WordPiece* (M. SCHUSTER et NAKAJIMA, 2012), le *tokenizer* originellement utilisé par **BERT**, nécessite une pré-segmentation et son implémentation du *byte pair encoding* est légèrement différente de l'originale (GAGE, 1994). Très

9. `https://clementdalloux.fr/`

utilisés en traduction automatique, de nombreux modèles de langage sont multilingues (DEVLIN et al., 2019 ; CONNEAU, K. KHANDELWAL et al., 2020 ; Y. LIU, GU et al., 2020) et nous disposons de corpus annotés en français, anglais, portugais brésilien, chinois, etc. Le développement d'un étiqueteur multilingue sur la base de ces modèles est donc envisageable. Si l'intérêt d'un tel système pour l'étiquetage des marqueurs nous semble limité, l'étiquetage de la portée pourrait en bénéficier en raison des similarités de entre les opérations de négation de certaines langues.

Cependant, il reste une marge de progression importante, notamment pour l'étiquetage des marqueurs d'incertitude et l'étiquetage de la portée en général. En effet, les résultats pour l'étiquetage des marqueurs d'incertitude sont bien plus faibles que pour la négation. Cela est dû à leur diversité ainsi qu'à leur composition. Par exemple, nos jeux de données ne contiennent pas suffisamment d'exemples de certains marqueurs. Il faudrait donc annoter bien plus d'instances d'incertitude que de négation afin d'obtenir des résultats équivalents avec les descripteurs et approches que nous utilisons. Présenter une plus grande diversité d'instances permettrait aussi d'améliorer les résultats pour la détection de la portée étant donné que certaines tournures de phrases peu utilisées posent problème à nos classifieurs. En effet, bien que les résultats en termes de tokens soient très encourageants, en termes de portées exactes, des progrès restent possibles. Ainsi, il est non seulement nécessaire d'augmenter le volume global d'exemples présenté à nos étiqueteurs, mais aussi de sélectionner les exemples ajoutés à nos corpus en fonction de la rareté des marqueurs et/ou des portées associées. Cependant, l'incertitude peut être marquée par un très grand nombre de verbes au conditionnel présent et arriver à couvrir l'intégralité de ces marqueurs dans nos corpus nous semble complexe, voire impossible. Afin de fournir des descripteurs plus précis, nous prévoyons de remplacer les annotations de **TreeTagger** par celles de **Tagex** [10], un étiqueteur spécialement entraîné sur des textes biomédicaux français. En identifiant les verbes au conditionnel présent de façon plus précise, nos systèmes détecteront plus facilement les marqueurs d'incertitude. Une autre possibilité serait de considérer tout verbe au conditionnel présent comme marqueur par défaut. Des descripteurs syntaxiques supplémentaires (analyse syntaxique de surface, arbres syntaxiques, etc.) sont aussi envisagés.

Valorisation, transfert

Ces travaux sont valorisés de plusieurs manières. Outre les publications scientifiques citées précédemment, nous avons mis à disposition de la communauté nos codes, modèles de plongements de mots et données annotées ainsi qu'un outil sous la forme d'application web. D'une part, les scripts permettant d'entraîner et tester nos **BiRNN** dédiés à la détection de la négation et de l'incertitude sont disponibles sur mon **GitHub** [11]. Les modèles de plongements de mots que nous utilisons dans ce contexte sont aussi disponibles au téléchargement [12]. D'autre part, notre système

10. `https://allgo.inria.fr/app/tagex`
11. `https://github.com/Kureman/NegBiRNNs`
12. `https://clementdalloux.fr/?page_id=112`

de détection de la négation en français par **BiLSTM-CRF** est utilisable sur la plateforme Allgo [13]. Par ailleurs, nos corpus annotés ont été téléchargés à de nombreuses reprises par des chercheurs de différentes institutions telles que le **CEA LIST** [14], l'Université de Copenhague et le **FBK-ICT** [15]. Ils ont aussi été utilisés par plusieurs étudiants en Master 2 Sciences du langage de l'Université d'Orléans dans le cadre de projets de **TALN**. Enfin, nos approches les plus récentes, basées sur **CamemBERT**, et les données d'entraînement associées sont désormais utilisées par des chercheurs de l'**AP-HM** [16] et de l'**AP-HP** [17].

13. https://allgo.inria.fr/
14. http://www-list.cea.fr/
15. https://ict.fbk.eu/
16. http://fr.ap-hm.fr/
17. https://www.aphp.fr/

Table des matières

Remerciements **iii**

Introduction Générale **9**

1 État de l'art **13**
1.1 Généralités 15
1.1.1 Traitement automatique de la langue biomédicale 15
1.1.2 Classification automatique de textes 19
1.1.3 Étiquetage de séquences 22
1.2 Classification multi-étiquette de textes cliniques 25
1.2.1 La Classification Internationale des Maladies 25
1.2.2 Jeux de données annotées 26
1.2.3 Systèmes de classification 31
1.3 Détection de la négation et de l'incertitude 35
1.3.1 Négation et incertitude 35
1.3.2 Jeux de données annotées 40
1.3.3 Étiquetage automatique 43

2 Classification multi-étiquette de textes cliniques **59**
2.1 Jeu de données 61
2.2 Nos approches 65
2.2.1 Approche par dictionnaires 65
2.2.2 Approches par apprentissage supervisé 66
2.3 Expériences de classification CIM-10 71
2.3.1 Protocole expérimental 71
2.3.2 Analyse des résultats 74

3 Corpus constitués dans le cadre de la thèse **79**
3.1 Corpus biomédicaux français 81
3.1.1 Règles d'annotation 81
3.1.2 ESSAI : corpus français d'essais cliniques 84
3.1.3 CAS : corpus français de cas cliniques 89
3.2 Corpus biomédicaux brésiliens 94
3.2.1 Protocoles d'essais cliniques brésiliens 94
3.2.2 SemClinBr 97

4 Détection de la négation et de l'incertitude **101**
4.1 Nos approches 103

4.1.1 Plongements de mots pré-entraînés 103
4.1.2 Approches pour l'étiquetage de séquences 104
4.2 Étiquetage automatique des marqueurs 110
4.2.1 Protocole expérimental 110
4.2.2 Analyse des résultats 111
4.3 Étiquetage automatique de la portée 116
4.3.1 Protocole expérimental 116
4.3.2 Analyse des résultats 117
4.3.3 Analyse des erreurs 124

Conclusion générale **129**

Table des matières **135**

Table des figures **137**

Liste des tableaux **139**

Bibliographie **143**
Sources primaires 143
Sources secondaires 144

Résumé/Abstract **163**

Table des figures

1.1 Le code **CIM-10** ***A00 Choléra*** et ses enfants, capture d'écran du site : `https://www.aideaucodage.fr/cim-a00` 26

2.1 Note clinique factice 62
2.2 Résultats d'examens de biologie médicale 63
2.3 Courbe du nombre d'occurrences par code pour les 100 codes **CIM-10** les plus fréquents dans le jeu de donnée du **CHU** 64
2.4 Architecture de notre **CNN** 68
2.5 Architecture de notre **BiLSTM** 70
2.6 Courbe du nombre d'occurrences par code pour les 100 codes **CIM-10** les plus fréquents après la réduction de niveau hiérarchique 73

3.1 Protocole d'essai clinique en français : Résumé 85
3.2 Description détaillée du protocole d'essai clinique 86
3.3 Exemple de cas clinique en français 90
3.4 Un protocole d'essai clinique en portugais brésilien 95
3.5 Exemple de texte clinique en portugais brésilien 98

4.1 Architecture d'un *Linear-chain CRF* pour l'étiquetage des marqueurs de négation avec une fenêtre contextuelle de (-1/+1) 105
4.2 Architecture de nos **BiRNN** 107

Liste des tableaux

1.1 Statistiques du corpus **CépiDC** par année (NÉVÉOL, K Bretonnel COHEN et al., 2016 ; NÉVÉOL, ROBERT, ANDERSON et al., 2017 ; NÉVÉOL, ROBERT, GRIPPO et al., 2018) 28
1.2 Exemple de document du corpus de certificats de décès **CépiDC aligné** (NÉVÉOL, ROBERT, GRIPPO et al., 2018) 29
1.3 Statistiques du corpus **CDC** (NÉVÉOL, ROBERT, ANDERSON et al., 2017) 30
1.4 Statistiques du corpus **KSH-HU** (NÉVÉOL, ROBERT, GRIPPO et al., 2018) 30
1.5 Statistiques du corpus **ISTAT-IT** (NÉVÉOL, ROBERT, GRIPPO et al., 2018) 31
1.6 4 meilleurs résultats sur le corpus CépiDC 2016 (NÉVÉOL, K Bretonnel COHEN et al., 2016) 32
1.7 Résultats de la campagne d'évaluation de **CLEF eHealth 2017** (NÉVÉOL, ROBERT, ANDERSON et al., 2017) 33
1.8 Résultats de la campagne d'évaluation de **CLEF eHealth 2018** (NÉVÉOL, ROBERT, GRIPPO et al., 2018) 34
1.9 Marqueurs de négation en anglais, français et portugais 37
1.10 Marqueurs d'incertitude en anglais et français 39
1.11 Statistiques du corpus BioScope 41
1.12 Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur le corpus **BioScope**. 51
1.13 Résultats des systèmes de détection des marqueurs d'incertitude au niveau des tokens sur le corpus **BioScope**. 51
1.14 Résultats des systèmes de détection de la portée de la négation au niveau des tokens sur le corpus **BioScope**. 52
1.15 Résultats des systèmes de détection de la portée de l'incertitude au niveau des tokens sur le corpus **BioScope**. 52
1.16 Résultats des systèmes de détection de la portée au niveau des portées correctement identifiées **BioScope**. 53
1.17 Résultats des systèmes de détection des marqueurs d'incertitude de **CoNLL-2010** 53
1.18 Résultats des systèmes de détection de la portée de l'incertitude de **CoNLL-2010**, portées et marqueurs exactement identifiée 54
1.19 Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur le corpus **CD-SCO** 54
1.20 Résultats des systèmes de détection de la portée de la négation au niveau des tokens sur le corpus **CD-SCO** 55
1.21 portée exacte starsem 55

2.1 Statistiques descriptives du jeu de données du **CHU** 63
2.2 Statistiques descriptives du jeu de données du **CHU** pour la seconde tâche 72
2.3 Résultats obtenus par nos systèmes pour la tâche de classification multi-étiquette de textes cliniques. 75

3.1 Statistiques relatives au corpus ESSAI 87
3.2 **ESSAI** : Accords inter-annotateurs obtenus pour les marqueurs de négation et leur portée 88
3.3 Extrait tiré du corpus ESSAI 89
3.4 Statistiques relatives au corpus CAS 91
3.5 CAS : Accords inter-annotateurs obtenus pour les marqueurs de négation et leur portée 92
3.6 Extrait tiré du corpus CAS 93
3.7 Statistiques relatives aux corpus de données médicales brésiliennes 96
3.8 Extrait tiré du corpus de protocoles d'essais cliniques brésiliens 97
3.9 Statistiques relatives aux corpus de données médicales brésiliennes 98
3.10 Extrait tiré du corpus de textes cliniques brésiliens 99

4.1 Résultats des systèmes de détection des marqueurs au niveau des tokens sur les corpus en français. 112
4.2 Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur les corpus en portugais brésilien. 113
4.3 Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur le corpus **BioScope**. 114
4.4 Résultats des systèmes de détection des marqueurs d'incertitude au niveau des tokens sur le corpus **BioScope**. 114
4.5 Résultats des systèmes de détection des marqueurs de négation au niveau des tokens sur le corpus **CD-SCO** 115
4.6 Précision (P), Rappel (R) et F-mesure (F_1) de l'étiquetage des tokens de la portée et des portées exactes de la négation. *Évaluation classique. **Évaluation seulement sur les concepts. 119
4.7 Précision (P), Rappel (R) et F-mesure (F_1) de l'étiquetage des tokens de la portée et des portées exactes de l'incertitude. 120
4.8 Précision (P), Rappel (R) et F-mesure (F_1) de l'étiquetage des tokens de la portée et des portées exactes de la négation. 121
4.9 Précision, Rappel et F-mesure de l'étiquetage au niveau des tokens de la portée de la négation. 122
4.10 Précision, Rappel et F-mesure de l'étiquetage des tokens de la portée et des portées exactes de l'incertitude. 122
4.11 Résultats en pourcentage de portées correctes. 123
4.12 Précision (P), Rappel (R) et F-mesure (F_1) de l'étiquetage des tokens de la portée et des portées exactes de la négation. 124

Bibliographie

Sources primaires

DALLOUX, Clément (juin 2017), « Identifying uncertainty and negation's cues and scope : State of the art », in : *RECITAL 2017 - 18ème Rencontre des Étudiants Chercheurs en Informatique en Traitement Automatique des Langues*, Actes de la Rencontre des Jeunes Chercheurs en Traitement Automatique des Langues, RECITAL, joint à la conférence TALN 2017, Orléans, France, p. 1-14, URL : https://hal.archives-ouvertes.fr/hal-01659646.

DALLOUX, Clément, Vincent CLAVEAU, Marc CUGGIA et al. (2020), « Supervised Learning for the ICD-10 Coding of French Clinical Narratives », in : *MIE 2020*, Geneva, Switzerland.

DALLOUX, Clément, Vincent CLAVEAU et Natalia GRABAR (déc. 2017), « Détection de la négation : corpus français et apprentissage supervisé », in : *SIIM 2017 - Symposium sur l'Ingénierie de l'Information Médicale*, Toulouse, France, p. 17-24, URL : https://hal.archives-ouvertes.fr/hal-01659637.

— (sept. 2019), « Speculation and negation detection in french biomedical corpora », in : *RANLP 2019 - Recent Advances in Natural Language Processing*, Varna, Bulgaria, p. 1-10, URL : https://hal.archives-ouvertes.fr/hal-02284444.

DALLOUX, Clément, Vincent CLAVEAU, Natalia GRABAR et Claudia MORO (mai 2018), « Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien (Negation scope : sequence labeling by supervised learning in French and Brazilian-Portuguese) », in : *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, Rennes, France : ATALA, p. 409-418, URL : https://www.aclweb.org/anthology/2018.jeptalnrecital-court.24.

DALLOUX, Clément, Vincent CLAVEAU, Natalia GRABAR, Lucas Emanuel Silva OLIVEIRA et al. (2020), « Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora », in : *Natural Language Engineering*, p. 1-21, DOI : 10.1017/S1351324920000352.

DALLOUX, Clément, Natalia GRABAR et Vincent CLAVEAU (déc. 2019), « Détection de la négation : corpus français et apprentissage supervisé », in : *Revue des Sciences et Technologies de l'Information - Série TSI : Technique et Science Informatiques*, p. 1-21, URL : https://hal.archives-ouvertes.fr/hal-02402913.

GRABAR, Natalia, Vincent CLAVEAU et Clément DALLOUX (2018), « CAS : French Corpus with Clinical Cases », in : *LOUHI 2018 : The Ninth International Workshop on Health Text Mining and Information Analysis*.

Sources secondaires

ABDAOUI, Amine et al. (2017), « French ConText : Détecter la négation, la temporalité et le sujet dans les textes cliniques Français », in : *4e édition du Symposium sur l'Ingénierie de l'Information Médicale*, p. 7-16.

ABRAHAMSSON, Emil et al. (2014), « Medical text simplification using synonym replacement : Adapting assessment of word difficulty to a compounding language », in : *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, p. 57-65.

AKBIK, Alan, Duncan BLYTHE et Roland VOLLGRAF (2018), « Contextual string embeddings for sequence labeling », in : *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1638-1649.

ALBRIGHT, Daniel et al. (2013), « Towards comprehensive syntactic and semantic annotations of the clinical narrative », in : *Journal of the American Medical Informatics Association* 20.5, p. 922-930.

ALMAGRO, Mario et al. (2018), « MAMTRA-MED at CLEF eHealth 2018 : A Combination of Information Retrieval Techniques and Neural Networks for ICD-10 Coding of Death Certificates. », in : *CLEF 2018 Online Working Notes. CEUR-WS.*

APOSTOLOVA, Emilia, David S CHANNIN et al. (2009), « Automatic segmentation of clinical texts », in : *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, p. 5905-5908.

APOSTOLOVA, Emilia, Noriko TOMURO et Dina DEMNER-FUSHMAN (2011), « Automatic extraction of lexico-syntactic patterns for detection of negation and speculation scopes », in : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, Association for Computational Linguistics, p. 283-287.

ASAHARA, Masayuki et Yuji MATSUMOTO (2003), « Japanese named entity extraction with redundant morphological analysis », in : *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, p. 8-15.

ATUTXA, Aitziber et al. (2018), « IxaMed at CLEF eHealth 2018 Task 1 : ICD10 Coding with a Sequence-to-Sequence Approach. », in : *CLEF 2018 Online Working Notes. CEUR-WS.*

BAEVSKI, Alexei et al. (2019), « Cloze-driven pretraining of self-attention networks », in : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 5360-5369.

BAHDANAU, Dzmitry, Kyunghyun CHO et Yoshua BENGIO (2015), « Neural machine translation by jointly learning to align and translate », in : *ICLR 2015*.

BALDI, Pierre et Peter J SADOWSKI (2013), « Understanding dropout », in : *Advances in neural information processing systems*, p. 2814-2822.

BATES, David W (2010), « Getting in Step : Electronic Health Records and their Role in Care Coordination », in : *Journal of General Internal Medicine* 3.25, p. 174-176.

BETHARD, Steven et al. (2016), « Semeval-2016 task 12 : Clinical tempeval », in : *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 1052-1062.

BIGEARD, Elise (oct. 2019), « Detection and analysis of drug non-adherence in social media », Theses, Université Charles de Gaulle - Lille III, URL : https://tel.archives-ouvertes.fr/tel-02478927.

BIGEARD, Elise, Natalia GRABAR et Frantz THIESSARD (2018), « Detection and analysis of drug misuses. A study based on social media messages », in : *Frontiers in pharmacology* 9, p. 791.

BIKEL, Daniel M. et al. (mars 1997), « Nymble : a High-Performance Learning Namefinder », in : *Fifth Conference on Applied Natural Language Processing*, Washington, DC, USA : Association for Computational Linguistics, p. 194-201, DOI : 10.3115/974557.974586, URL : https://www.aclweb.org/anthology/A97-1029.

BLEI, David M, Andrew Y NG et Michael I JORDAN (2003), « Latent dirichlet allocation », in : *Journal of machine Learning research* 3.Jan, p. 993-1022.

BODENREIDER, Olivier et Alexa T MCCRAY (2003), « Exploring semantic groups through visual approaches », in : *Journal of biomedical informatics* 36.6, p. 414-432.

BOHNET, Bernd et al. (juill. 2018), « Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings », in : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Melbourne, Australia : Association for Computational Linguistics, p. 2642-2652, DOI : 10.18653/v1/P18-1246, URL : https://www.aclweb.org/anthology/P18-1246.

BOJANOWSKI, Piotr et al. (2017), « Enriching word vectors with subword information », in : *Transactions of the Association for Computational Linguistics* 5, p. 135-146.

BOUCHON-MEUNIER, Bernadette et Hung-T NGUYEN (1996), *Les incertitudes dans les systèmes intelligents*, Presses Universitaires de France (« Que sais-je ? »)

BRILL, Eric (1992), « A simple rule-based part of speech tagger », in : *Proceedings of the third conference on Applied natural language processing*, Association for Computational Linguistics, p. 152-155.

BRONNER, Gerald (1997), *L'incertitude.* Presses Universitaires de France (« Que sais-je ? »)

BROWN, Anna et al. (2016), « Comparison of dementia recorded in routinely collected hospital admission data in England with dementia recorded in primary care », in : *Emerging themes in epidemiology* 13.1, p. 11.

CABRÉ, Maria Térésa (1998), *Terminologie : théorie, méthode et applications*, Les presses de l'Université d'Ottawa, Armand Colin.

CAMON, Evelyn B et al. (2005), « An evaluation of GO annotation retrieval for BioCreAtIvE and GOA », in : *BMC bioinformatics* 6.S1, S17.

CARTA, Salvatore et al. (2019), « A Supervised Multi-class Multi-label Word Embeddings Approach for Toxic Comment Classification. », in : *11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR-2019)*, Vienna, Austria, p. 105-112.

CHALKIDIS, Ilias et al. (juill. 2019), « Large-Scale Multi-Label Text Classification on EU Legislation », in : *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, Florence, Italy : Association for Computational Linguistics, p. 6314-6322, DOI : 10.18653/v1/P19-1636, URL : https://www.aclweb.org/anthology/P19-1636.

CHANTRY, Anne A et al. (2011), « Hospital discharge data can be used for monitoring procedures and intensive care related to severe maternal morbidity », in : *Journal of clinical epidemiology* 64.9, p. 1014-1022.

CHAPMAN, W. W. et al. (oct. 2001), « A simple algorithm for identifying negated findings and diseases in discharge summaries », in : *Journal of Biomedical Informatics* 34.5, ISSN : 1532-0464, DOI : 10.1006/jbin.2001.1029.

CHARNOCK, Ross (1999), « Les langues de spécialité et le langage technique : considérations didactiques », in : *ASp. la revue du GERAS* 23-26, p. 281-302.

CHO, Kyunghyun et al. (2014), « Learning phrase representations using RNN encoder-decoder for statistical machine translation », in : *arXiv preprint arXiv :1406.1078*, URL : https://arxiv.org/abs/1406.1078.

CHO, Paul S, Ricky K TAIRA et Hooshang KANGARLOO (2003), « Automatic section segmentation of medical reports », in : *AMIA Annual Symposium Proceedings*, t. 2003, American Medical Informatics Association, p. 155.

CHOWDHURY, Md. Faisal Mahbub (2012), « FBK : Exploiting phrasal and contextual clues for negation scope detection », in : *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, p. 340-346.

COHEN, Jacob (1960), « A coefficient of agreement for nominal scales », in : *Educational and psychological measurement* 20.1, p. 37-46.

COHEN, K Bretonnel, Thomas CHRISTIANSEN et Lawrence E HUNTER (2011), « Parenthetically speaking : Classifying the contents of parentheses for text mining », in : *AMIA annual symposium proceedings*, t. 2011, American Medical Informatics Association, p. 267.

COHEN, Kevin Bretonnel et Dina DEMNER-FUSHMAN (2014), *Biomedical natural language processing*, t. 11, John Benjamins Publishing Company.

COLLINS, Michael (2002), « Discriminative training methods for hidden Markov models : Theory and experiments with perceptron algorithms », in : *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, p. 1-8.

COLLOBERT, Ronan et al. (2011), « Natural language processing (almost) from scratch », in : *Journal of machine learning research* 12.Aug, p. 2493-2537.

CONNEAU, Alexis, Kartikay KHANDELWAL et al. (2020), « Unsupervised Cross-lingual Representation Learning at Scale », in : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440-8451.

CONNEAU, Alexis, Holger SCHWENK et al. (avr. 2017), « Very Deep Convolutional Networks for Text Classification », in : *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, Valencia, Spain : Association for Computational Linguistics, p. 1107-1116, URL : https://www.aclweb.org/anthology/E17-1104.

COSSIN, Sébastien et al. (2018), « IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates », in : *CLEF 2018 Online Working Notes. CEUR-WS.*

CUGGIA, Marc, Paolo BESANA et David GLASSPOOL (2011), « Comparing semi-automatic systems for recruitment of patients to clinical trials », in : *International journal of medical informatics* 80.6, p. 371-388.

CULIOLI, Antoine (1988), « La négation : marqueurs et opérations », in : *Travaux du Centre de Recherches sémiologiques* 56, p. 17-38.

CURNS, Aaron T et al. (2010), « Reduction in acute gastroenteritis hospitalizations among US children after introduction of rotavirus vaccine : analysis of hospital discharge data from 18 US states », in : *The Journal of infectious diseases* 201.11, p. 1617-1624.

CUTTING, Douglass et al. (1992), « A practical part-of-speech tagger », in : *Third Conference on Applied Natural Language Processing*, p. 133-140.

CZARNECKI, Jan et al. (2012), « A text-mining system for extracting metabolic reactions from full-text articles », in : *BMC bioinformatics* 13.1, p. 172.

DEERWESTER, Scott et al. (1990), « Indexing by latent semantic analysis », in : *Journal of the American society for information science* 41.6, p. 391-407.

DELÉGER, Louise et Cyril GROUIN (2012), « Detecting negation of medical problems in French clinical notes », in : *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*.

DENNY, JC et JF PETERSON (2007), « Identifying QT prolongation from ECG impressions using Natural Language Processing and negation detection », in : *Medinfo*, p. 1283-8.

DEVLIN, Jacob et al. (2019), « Bert : Pre-training of deep bidirectional transformers for language understanding », in : *Proceedings of NAACL-HLT 2019*.

DONNELLY, Kevin (2006), « SNOMED-CT : The advanced terminology and coding system for eHealth », in : *Studies in health technology and informatics* 121, p. 279.

EL EMAM, Khaled et al. (2009), « Evaluating the risk of re-identification of patients from hospital prescription records », in : *The Canadian journal of hospital pharmacy* 62.4, p. 307.

ELKIN, PL et al. (2005), « A controlled trial of automated classification of negation from clinical notes », in : *BMC Med Inform Decis Mak.* 5.13.

FANCELLU, Federico, Adam LOPEZ et Bonnie WEBBER (2016), « Neural networks for negation scope detection », in : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, t. 1, p. 495-504, URL : `http://www.research.ed.ac.uk/portal/files/25518265/neural_networks_negation_1.pdf`.

FARKAS, Richárd et al. (2010), « The CoNLL-2010 shared task : learning to detect hedges and their scope in natural language text », in : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, Association for Computational Linguistics, p. 1-12.

FENG, Mengling et al. (2018), « Transthoracic echocardiography and mortality in sepsis : analysis of the MIMIC-III database », in : *Intensive care medicine* 44.6, p. 884-892.

FLEISS, Joseph L, Bruce LEVIN et Myunghee Cho PAIK (1981), *Statistical methods for rates and proportions*, John Wiley & Sons.

FLORIAN, Radu et al. (2003), « Named entity recognition through classifier combination », in : *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, Association for Computational Linguistics, p. 168-171.

FREUND, Yoav et Robert E SCHAPIRE (1995), « A desicion-theoretic generalization of on-line learning and an application to boosting », in : *European conference on computational learning theory*, Springer, p. 23-37.

FUKUDA, Ken-ichiro et al. (1998), « Toward information extraction : identifying protein names from biological papers », in : *Pac symp biocomput*, t. 707, 18, p. 707-718.

GAGE, Philip (1994), « A new algorithm for data compression », in : *C Users Journal* 12.2, p. 23-38.

GENTIMIS, Thanos et al. (2017), « Predicting hospital length of stay using neural networks on MIMIC III data », in : *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, IEEE, p. 1194-1201.

GINDL, S, K KAISER et S MIKSCH (2008), « Syntactical negation detection in clinical practice guidelines », in : *Stud Health Technol Inform*, p. 187-92.

GOBEILL, Julien, Patrick RUCH et Rodolphe MEYER (2020), « Machine Learning for Automatic Encoding of French Electronic Medical Records : Is More Data Better ? », in : *Digital Personalized Health and Medicine : Proceedings of MIE 2020* 270, p. 312.

GOLDBERG, Yoav (2017), « Neural network methods for natural language processing », in : *Synthesis Lectures on Human Language Technologies* 10.1, p. 1-309.

GOODFELLOW, Ian, Yoshua BENGIO et Aaron COURVILLE (2016), *Deep Learning*, `http://www.deeplearningbook.org`, MIT Press.

GRABAR, Natalia et Rémi CARDON (2018), « CLEAR–Simple Corpus for Medical French », in : *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, p. 3-9.

HABIBI, Maryam et al. (2017), « Deep learning with word embeddings improves biomedical named entity recognition », in : *Bioinformatics* 33.14, p. i37-i48.

HARKEMA, Henk et al. (2009), « ConText : an algorithm for determining negation, experiencer, and temporal status from clinical reports », in : *Journal of biomedical informatics* 42.5, p. 839-851.

HARTMANN, Nathan et al. (2017), « Portuguese Word Embeddings : Evaluating on Word Analogies and Natural Language Tasks », in : *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, p. 122-131.

HENRY, Sam et al. (2020), « 2018 N2C2 shared task on adverse drug events and medication extraction in electronic health records », in : *Journal of the American Medical Informatics Association* 27.1, p. 3-12.

HINTON, Geoffrey E et al. (2012), « Improving neural networks by preventing co-adaptation of feature detectors », in : *arXiv preprint arXiv :1207.0580*.

HOBBS, FD Richard et al. (2016), « Clinical workload in UK primary care : a retrospective analysis of 100 million consultations in England, 2007–14 », in : *The Lancet* 387.10035, p. 2323-2330.

HOCHREITER, Sepp et Jürgen SCHMIDHUBER (1997), « Long short-term memory », in : *Neural computation* 9.8, p. 1735-1780.

HORN, Laurence R. (2001), *A Natural History of Negation*. CSLI PUBLICATIONS.

HOWARD, Jeremy et Sebastian RUDER (2018), « Universal language model fine-tuning for text classification », in : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, p. 328-339.

HUANG, Yang et Henry J LOWE (2007), « A novel hybrid approach to automated negation detection in clinical radiology reports », in : *Journal of the American medical informatics association* 14.3, p. 304-311.

HUANG, Zhiheng, Wei XU et Kai YU (2015), « Bidirectional LSTM-CRF models for sequence tagging », in : *arXiv preprint arXiv :1508.01991*.

JAIN, Suvir et al. (2016), « Weakly supervised learning of biomedical information extraction from curated data », in : *BMC bioinformatics*, t. 17, S1, Springer, S1.

JELINEK, Frederick (1985), « Markov source modeling of text generation », in : *The impact of processing techniques on communications*, Springer, p. 569-591.

JIANG, Yufan et al. (2019), « Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition », in : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3576-3581.

JOHNSON, Alistair EW et al. (2016), « MIMIC-III, a freely accessible critical care database », in : *Scientific data* 3.1, p. 1-9.

JOHNSON, Rie et Tong ZHANG (2016), « Supervised and Semi-Supervised Text Categorization Using LSTM for Region Embeddings », in : *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, New York, NY, USA : JMLR.org, p. 526-534.

JOUSSELME, Anne-Laure, Patrick MAUPIN et Éloi BOSSÉ (2003), « Uncertainty in a situation analysis perspective », in : *Proceedings of the Sixth International Conference of Information Fusion*, t. 2, p. 1207-1214.

JOZEFOWICZ, Rafal, Wojciech ZAREMBA et Ilya SUTSKEVER (2015), « An empirical exploration of recurrent network architectures », in : *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, p. 2342-2350.

KALTENBACH, Sophie et Ludovic LHERMITTE (2019), « Apport des examens de biologie dans la prise en charge des hémopathies lymphoïdes B matures. L'intérêt de l'intégration des données clinicobiologiques », in : *La Presse Médicale* 48.7-8, p. 816-824.

KANDULA, Sasikiran, Dorothy CURTIS et Qing ZENG-TREITLER (2010), « A semantic and syntactic text simplification tool for health content », in : *AMIA annual symposium proceedings*, t. 2010, American Medical Informatics Association, p. 366.

KELLY, Liadh et al. (2019), « Overview of the CLEF eHealth evaluation lab 2019 », in : *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, p. 322-339.

KHANDELWAL, Aditya et Suraj SAWANT (mai 2020), « NegBERT : A Transfer Learning Approach for Negation Detection and Scope Resolution », in : *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France : European Language Resources Association, p. 5739-5748.

KILICOGLU, Halil et Sabine BERGLER (2010), « A high-precision approach to detecting hedges and their scopes », in : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, Association for Computational Linguistics, p. 70-77.

KIM, Jin-Dong, Tomoko OHTA et Jun'ichi TSUJII (2008), « Corpus annotation for mining biomedical events from literature », in : *BMC bioinformatics* 9.1, p. 10.

KIM, Yoon (oct. 2014), « Convolutional Neural Networks for Sentence Classification », in : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar : Association for Computational Linguistics, p. 1746-1751, DOI : 10.3115/v1/D14-1181, URL : https://www.aclweb.org/anthology/D14-1181.

KING, Jennifer et al. (2014), « Clinical benefits of electronic health record use : national findings », in : *Health services research* 49.1pt2, p. 392-404.

KLEIN, Sheldon et Robert F SIMMONS (1963), « A computational approach to grammatical coding of English words », in : *Journal of the ACM (JACM)* 10.3, p. 334-347.

KONSTANTINOVA, Natalia et al. (mai 2012), « A review corpus annotated for negation, speculation and their scope », in : *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA), p. 3190-3195, URL : http://www.lrec-conf.org/proceedings/lrec2012/pdf/533_Paper.pdf.

KÖPCKE, Felix et Hans-Ulrich PROKOSCH (2014), « Employing computers for the recruitment into clinical trials : a comprehensive systematic review », in : *Journal of medical Internet research* 16.7, e161.

KOPTIENT, Anaïs, Rémi CARDON et Natalia GRABAR (août 2019), « Simplification-induced transformations : typology and some characteristics », in : *BioNLP 2019*, Florence, Italy, DOI : 10.18653/v1/W19-5033, URL : https://hal.archives-ouvertes.fr/hal-02430514.

KRAUSE, Paul et Dominic CLARK (1993), *Representing uncertain knowledge : an artificial intelligence approach*, Kluwer Academic Publishers.

KUDO, Taku (juill. 2018), « Subword Regularization : Improving Neural Network Translation Models with Multiple Subword Candidates », in : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Melbourne, Australia : Association for Computational Linguistics, p. 66-75, DOI : 10.18653/v1/P18-1007, URL : https://www.aclweb.org/anthology/P18-1007.

KUDO, Taku et John RICHARDSON (nov. 2018), « SentencePiece : A simple and language independent subword tokenizer and detokenizer for Neural Text Processing », in : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, Brussels, Belgium : Association for Computational Linguistics, p. 66-71, DOI : 10.18653/v1/D18-2012, URL : https://www.aclweb.org/anthology/D18-2012.

LAFFERTY, John, Andrew MCCALLUM, Fernando PEREIRA et al. (2001), « Conditional random fields : Probabilistic models for segmenting and labeling sequence data », in : *Proceedings of the eighteenth international conference on machine learning, ICML*, t. 1, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., p. 282-289.

LAMONTAGNE, Pamela et al. (juill. 2018), « OASIS-3 : LONGITUDINAL NEUROIMAGING, CLINICAL, AND COGNITIVE DATASET FOR NORMAL AGING AND ALZHEIMER'S DISEASE », in : *Alzheimer's & Dementia* 14, P138, DOI : 10.1016/j.jalz.2018.06.2231.

LAMPLE, Guillaume et al. (juin 2016), « Neural Architectures for Named Entity Recognition », in : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, San Diego, California : Association for Computational Linguistics, p. 260-270, DOI : 10.18653/v1/N16-1030, URL : https://www.aclweb.org/anthology/N16-1030.

LANDIS, J Richard et Gary G KOCH (1977), « The measurement of observer agreement for categorical data », in : *Biometrics*, p. 159-174.

LAPPONI, Emanuele et al. (2012), « UIO 2 : sequence-labeling negation using dependency features », in : *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, p. 319-327, URL : http://dl.acm.org/citation.cfm?id=2387687.

LE, Hang et al. (2020), « FlauBERT : Unsupervised Language Model Pre-training for French », in : *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 2479-2490.

LECUN, Yann et al. (1989), « Generalization and network design strategies », in : *Connectionism in perspective* 19, p. 143-155.

LERAT, Pierre (1995), *Les langues spécialisées*, FeniXX.

LEVENSHTEIN, Vladimir I (1966), « Binary codes capable of correcting deletions, insertions, and reversals », in : *Soviet physics doklady*, t. 10, 8, p. 707-710.

LI, Hao et Wei LU (2018), « Learning with structured representations for negation scope extraction », in : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 533-539.

LI, Junhui et al. (2010), « Learning the scope of negation via shallow semantic parsing », in : *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, p. 671-679.

LI, Li et al. (2008), « Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening : a case study », in : *AMIA Annual Symposium Proceedings*, t. 2008, American Medical Informatics Association, p. 404.

LI, Xinxin et al. (2010), « Exploiting rich features for detecting hedges and their scope », in : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, Association for Computational Linguistics.

LIGHT, Marc, Xin Ying QIU et Padmini SRINIVASAN (2004), « The language of bioscience : Facts, speculations, and statements in between », in : *HLT-NAACL 2004 workshop : linking biological literature, ontologies and databases*, p. 17-24.

LIMA, Luciano RS de, Alberto HF LAENDER et Berthier A RIBEIRO-NETO (1998), « A hierarchical approach to the automatic categorization of medical documents », in : *Proceedings of the seventh international conference on Information and knowledge management*, p. 132-139.

LIN, Yu-Kai, Hsinchun CHEN et Randall A BROWN (2013), « MedTime : A temporal information extraction system for clinical narratives », in : *Journal of biomedical informatics* 46, S20-S28.

LIU, Fei et al. (2011), « Insertion, deletion, or substitution ? : normalizing text messages without pre-categorization nor supervision », in : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, Association for Computational Linguistics, p. 71-76.

LIU, Hanxiao, Karen SIMONYAN et Yiming YANG (2019), « Darts : Differentiable architecture search », in : *Proceedings of the 2019 International Conference on Learning Representations (ICLR 2019)*, p. 1-13.

LIU, Qianchu, Federico FANCELLU et Bonnie WEBBER (mai 2018), « NegPar : A parallel corpus annotated for negation », in : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA), URL : https://www.aclweb.org/anthology/L18-1547.

LIU, Yinhan, Jiatao GU et al. (2020), « Multilingual denoising pre-training for neural machine translation », in : *arXiv preprint arXiv :2001.08210*.

LIU, Yinhan, Myle OTT et al. (2019), « Roberta : A robustly optimized bert pretraining approach », in : *arXiv preprint arXiv :1907.11692*.

MA, Xuezhe et Eduard HOVY (août 2016), « End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF », in : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Berlin, Germany : Association for Computational Linguistics, p. 1064-1074, DOI : 10.18653/v1/P16-1101, URL : https://www.aclweb.org/anthology/P16-1101.

MADEC, Julia et al. (2019), « eHOP Clinical Data Warehouse : From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. », in : *Studies in health technology and informatics* 264, p. 1536-1537.

MALDONADO, José A et al. (2009), « LinkEHR-Ed : A multi-reference model archetype editor based on formal semantics », in : *International journal of medical informatics* 78.8, p. 559-570.

MARTIN, Louis et al. (juill. 2020), « CamemBERT : a Tasty French Language Model », in : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 7203-7219.

MCCALLUM, Andrew et Wei LI (2003), « Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons », in : *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 188-191, URL : https://www.aclweb.org/anthology/W03-0430.

MEYSTRE, SM et al. (2017), « Clinical data reuse or secondary use : current status and potential future progress », in : *Yearbook of medical informatics* 26.1, p. 38.

MIFTAHUTDINOV, Zulfat et Elena TUTUBALINA (2017), « KFU at CLEF eHealth 2017 Task 1 : ICD-10 Coding of English Death Certificates with Recurrent Neural Networks », in : *CLEF 2017 Online Working Notes. CEUR-WS.*

MILOJEVIC, Ai et al. (2014), « Short-term effects of air pollution on a range of cardiovascular events in England and Wales : case-crossover analysis of the MINAP database, hospital admissions and mortality », in : *Heart* 100.14, p. 1093-1098.

MORANTE, Roser et Eduardo BLANCO (2012), « * SEM 2012 shared task : Resolving the scope and focus of negation », in : *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, p. 265-274, URL : http://dl.acm.org/citation.cfm?id=2387679.

MORANTE, Roser et Walter DAELEMANS (2009a), « A metalearning approach to processing the scope of negation », in : *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, p. 21-29.

— (2009b), « Learning the scope of hedge cues in biomedical texts », in : *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Association for Computational Linguistics, p. 28-36.

MORANTE, Roser, Anthony LIEKENS et Walter DAELEMANS (2008), « Learning the scope of negation in biomedical texts », in : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 715-724.

MORANTE, Roser, Sarah SCHRAUWEN et Walter DAELEMANS (2011), *Annotation of Negation Cues and their Scope. Guidelines v1.0.*

MORANTE, Roser, Vincent VAN ASCH et Walter DAELEMANS (2010), « Memory-based resolution of in-sentence scopes of hedge cues », in : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, Association for Computational Linguistics, p. 40-47.

MULLER, Claude (1991), *La négation en français : syntaxe, sémantique et éléments de comparaison avec les autres langues romanes*, Librairie Droz.

MURPHY, Daniel R et al. (2014), « Electronic health record-based triggers to detect potential delays in cancer diagnosis », in : *BMJ quality & safety* 23.1, p. 8-16.

MUTALIK, P. G., A. DESHPANDE et P. M. NADKARNI (déc. 2001), « Use of general-purpose negation detection to augment concept indexing of medical documents : a quantitative study using the UMLS », in : *Journal of the American Medical Informatics Association : JAMIA* 8.6, ISSN : 1067-5027.

NÉVÉOL, Aurélie, K Bretonnel COHEN et al. (2016), « Clinical information extraction at the CLEF eHealth evaluation lab 2016 », in : *CEUR workshop proceedings*, t. 1609, NIH Public Access, p. 28.

NÉVÉOL, Aurélie, Cyril GROUIN et al. (2014), « The Quaero French medical corpus : A ressource for medical entity recognition and normalization », in : *In Proc BioTextM, Reykjavik*, Citeseer.

NÉVÉOL, Aurélie, Aude ROBERT, Robert ANDERSON et al. (2017), « CLEF eHealth 2017 Multilingual Information Extraction task Overview : ICD10 Coding of Death Certificates in English and French. », in : *CLEF 2017 Online Working Notes. CEUR-WS.*

NÉVÉOL, Aurélie, Aude ROBERT, Francesco GRIPPO et al. (2018), « CLEF eHealth 2018 Multilingual Information Extraction Task Overview : ICD10 Coding of

Death Certificates in French, Hungarian and Italian. », in : *CLEF 2018 Online Working Notes. CEUR-WS.*

NGUYEN, Dat Quoc et al. (avr. 2014), « RDRPOSTagger : A Ripple Down Rules-based Part-Of-Speech Tagger », in : *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden : Association for Computational Linguistics, p. 17-20, URL : http://www.aclweb.org/anthology/E14-2005.

OLEYNIK, Michel et al. (2019), « Evaluating shallow and deep learning strategies for the 2018 N2C2 shared task on clinical text classification », in : *Journal of the American Medical Informatics Association* 26.11, p. 1247-1254.

OLIVEIRA, Lucas Emanuel Silva e et al. (2020), *SemClinBr – a multi institutional and multi specialty semantically annotated corpus for Portuguese clinical NLP tasks*, arXiv : 2001.10071 [cs.CL].

OROSZ, György, Attila NOVÁK et Gábor PRÓSZÉKY (2013), « Hybrid text segmentation for Hungarian clinical records », in : *Mexican International Conference on Artificial Intelligence*, Springer, p. 306-317.

ØVRELID, Lilja, Erik VELLDAL et Stephan OEPEN (2010), « Syntactic scope resolution in uncertainty analysis », in : *Proceedings of the 23rd international conference on computational linguistics*, Association for Computational Linguistics, p. 1379-1387.

PACKARD, Woodley et al. (2014), « Simple Negation Scope Resolution through Deep Parsing : A Semantic Solution to a Semantic Problem. », in : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Baltimore, Maryland : Association for Computational Linguistics, p. 69-78, URL : https://www.aclweb.org/anthology/P/P14/P14-1007.pdf.

PAKHOMOV, Serguei V, James BUNTROCK et Christopher G CHUTE (2005), « Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier », in : *Journal of biomedical informatics* 38.2, p. 145-153.

PENG, Yifan, Catalina TUDOR et al. (oct. 2012), « iSimp : A sentence simplification system for biomedical text », in : t. 1, p. 1-6, ISBN : 978-1-4673-2559-2, DOI : 10.1109/BIBM.2012.6392671.

PENG, Yifan, Xiaosong WANG et al. (2018), « NegBio : a high-performance tool for negation and uncertainty detection in radiology reports », in : *AMIA 2018 Informatics Summit*, URL : http://arxiv.org/abs/1712.05898.

PESTIAN, John P et al. (2007), « A shared task involving multi-label classification of clinical free text », in : *Proceedings of the Workshop on BioNLP 2007 : Biological, Translational, and Clinical Language Processing*, Association for Computational Linguistics, p. 97-104.

PETERS, Matthew E et al. (2018), « Deep contextualized word representations », in : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana : Association for Computational Linguistics, p. 2227-2237.

PHILIPS, Lawrence (juin 2000), « The Double Metaphone Search Algorithm », in : *C/C++ Users J.* 18.6, p. 38-43, ISSN : 1075-2838.

PURUSHOTHAM, Sanjay et al. (2018), « Benchmarking deep learning models on large healthcare datasets », in : *Journal of Biomedical Informatics* 83, p. 112-134, ISSN : 1532-0464, DOI : https://doi.org/10.1016/j.jbi.2018.04.007, URL : http://www.sciencedirect.com/science/article/pii/S1532046418300716.

QIAN, Zhong et al. (2016), « Speculation and negation scope detection via convolutional neural networks », in : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 815-825.

QUERLER, N.L. (1996), *Typologie des modalités*, Presses universitaires de Caen, ISBN : 9782841330546, URL : https://books.google.fr/books?id=Xj5cAAAAMAAJ.

RAU, Lisa F (1991), « Extracting company names from text », in : *[1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application*, t. 1, IEEE, p. 29-32.

READ, Jonathon et al. (2012), « Uio 1 : Constituent-based discriminative ranking for negation resolution », in : *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, Montréal, Canada : Association for Computational Linguistics, p. 310-318, URL : http://dl.acm.org/citation.cfm?id=2387686.

REHUREK, Radim et Petr SOJKA (2010), « Software framework for topic modelling with large corpora », in : *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta : ELRA, p. 45-50.

REI, Marek et Ted BRISCOE (2010), « Combining Manual Rules and Supervised Learning for Hedge Cue and Scope Detection », in : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*, CoNLL '10 : Shared Task, Uppsala, Sweden : Association for Computational Linguistics, p. 56-63, ISBN : 9781932432848.

RIOS, Anthony et Ramakanth KAVULURU (2018), « Few-shot and zero-shot multi-label learning for structured label spaces », in : *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, t. 2018, NIH Public Access, p. 3132.

RUCH, Patrick et al. (2008), « From episodes of care to diagnosis codes : automatic text categorization for medico-economic encoding », in : *AMIA Annual Symposium Proceedings*, t. 2008, American Medical Informatics Association, p. 636.

RUSSO, Elise et al. (2016), « Challenges in patient safety improvement research in the era of electronic health records », in : *Healthcare*, t. 4, 4, Elsevier, p. 285-290.

SALTON, G. (1971), *The SMART Retrieval System—Experiments in Automatic Document Processing*, Upper Saddle River, NJ, USA : Prentice-Hall, Inc.

SANDFORT, Veit et al. (2019), « Prolonged elevated heart rate and 90-Day survival in acutely Ill patients : Data from the MIMIC-III database », in : *Journal of intensive care medicine* 34.8, p. 622-629.

SANG, Erik F et Fien DE MEULDER (2003), « Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition », in : *arXiv preprint cs/0306050*.

SARKER, Abeed et al. (2015), « Utilizing social media data for pharmacovigilance : a review », in : *Journal of biomedical informatics* 54, p. 202-212.

SAURÍ, Roser et James PUSTEJOVSKY (sept. 2009), « FactBank : A corpus annotated with event factuality », in : *Language Resources and Evaluation* 43, p. 227-268, DOI : 10.1007/s10579-009-9089-9.

SAVOVA, Guergana K et al. (2010), « Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) : architecture, component evaluation and applications », in : *Journal of the American Medical Informatics Association* 17.5, p. 507-513.

SCHAPIRE, Robert E et Yoram SINGER (2000), « BoosTexter : A boosting-based system for text categorization », in : *Machine learning* 39.2-3, p. 135-168.

SCHERPF, Matthieu et al. (2019), « Predicting sepsis with a recurrent neural network using the MIMIC III database », in : *Computers in biology and medicine* 113, p. 103395.

SCHMID, Helmut (1994a), « Part-of-speech tagging with neural networks », in : *Proceedings of the 15th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, p. 172-176.

— (1994b), « Probabilistic Part-of-Speech Tagging Using Decision Trees », in : *Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK*. P. 44-49.

SCHUSTER, M. et K. NAKAJIMA (2012), « Japanese and Korean voice search », in : *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5149-5152.

SCHUSTER, Mike et Kuldip K. PALIWAL (1997), « Bidirectional recurrent neural networks », in : *IEEE Transactions on Signal Processing* 45.11.

SEIFERT, Stephan et Werner WELTE (1987), *A basic bibliography on negation in Natural Language*, t. 313, Gunter Narr Verlag.

SEKINE, Satoshi (1998), « Description of the Japanese NE System Used for MET-2 », in : *Seventh Message Understanding Conference (MUC-7) : Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, URL : https://www.aclweb.org/anthology/M98-1019.

SENNRICH, Rico, Barry HADDOW et Alexandra BIRCH (août 2016), « Neural Machine Translation of Rare Words with Subword Units », in : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Berlin, Germany : Association for Computational Linguistics, p. 1715-1725, DOI : 10.18653/v1/P16-1162, URL : https://www.aclweb.org/anthology/P16-1162.

SERGEEVA, Elena et al. (nov. 2019), « Neural Token Representations and Negation and Speculation Scope Detection in Biomedical and General Domain Text », in : *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, Hong Kong : Association for Computational Linguistics, p. 178-187, DOI : 10.18653/v1/D19-6221, URL : https://www.aclweb.org/anthology/D19-6221.

SETTLES, Burr, Mark CRAVEN et Lewis FRIEDLAND (2008), « Active learning with real annotation costs », in : *Proceedings of the NIPS workshop on cost-sensitive learning*, Vancouver, CA : p. 1-10.

SHAH, Parantu K et al. (2003), « Information extraction from full text scientific articles : where are the keywords ? », in : *BMC bioinformatics* 4.1, p. 20.

SINGH, Hardeep et al. (2009), « Improving follow-up of abnormal cancer screens using electronic health records : trust but verify test result communication », in : *BMC medical informatics and decision making* 9.1, p. 1-7.

SINNENBERG, Lauren et al. (2017), « Twitter as a tool for health research : a systematic review », in : *American journal of public health* 107.1, e1-e8.

SITTIG, Dean F et Hardeep SINGH (2012), « Improving test result follow-up through electronic health records requires more than just an alert », in : *Journal of general internal medicine* 27.10, p. 1235-1237.

SMETS, Philippe et Amihai MOTRO (1997), *Uncertainty Management in Information Systems : From Needs to Solutions*, Kluwer Academic publishers.

SMITHSON, Michael (1989), *Ignorance and Uncertainty.* Springer-Verlag.

SOYSAL, Ergin et al. (2018), « CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines », in : *Journal of the American Medical Informatics Association* 25.3, p. 331-336.

STANFILL, Mary H et al. (2010), « A systematic literature review of automated clinical coding and classification systems », in : *Journal of the American Medical Informatics Association* 17.6, p. 646-651.

STRAKOVÁ, Jana, Milan STRAKA et Jan HAJIC (juill. 2019), « Neural Architectures for Nested NER through Linearization », in : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy : Association for Computational Linguistics, p. 5326-5331, DOI : 10.18653/v1/P19-1527, URL : https://www.aclweb.org/anthology/P19-1527.

STUBBS, Amber, Michele FILANNINO et al. (2019), « Cohort selection for clinical trials : N2C2 2018 shared task track 1 », in : *Journal of the American Medical Informatics Association* 26.11, p. 1163-1171.

STUBBS, Amber, Christopher KOTFILA et Özlem UZUNER (2015), « Automated systems for the de-identification of longitudinal clinical narratives : Overview of 2014 I2B2/UTHealth shared task Track 1 », in : *Journal of biomedical informatics* 58, S11-S19.

SUN, Chi et al. (2019), « How to fine-tune bert for text classification ? », in : *China National Conference on Chinese Computational Linguistics*, Springer, p. 194-206.

SUN, Weiyi, Anna RUMSHISKY et Ozlem UZUNER (2013), « Evaluating temporal relations in clinical text : 2012 i2b2 Challenge », in : *Journal of the American Medical Informatics Association* 20.5, p. 806-813.

SUN, Xu (2014), « Structure regularization for structured prediction », in : *Advances in Neural Information Processing Systems*, p. 2402-2410.

TANABE, Lorraine et W John WILBUR (2002), « Tagging gene and protein names in full text articles », in : *Proceedings of the ACL-02 workshop on Natural Language Processing in the biomedical domain*, p. 9-13.

TANG, Buzhou et al. (2010), « A cascade method for detecting hedges and their scope in natural language text », in : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, Association for Computational Linguistics, p. 13-17.

TEPPER, Michael et al. (2012), « Statistical Section Segmentation in Free-Text Clinical Records. », in : *LREC*, p. 2001-2008.

TOURILLE, Julien et al. (2017), « Neural architecture for temporal relation extraction : a Bi-LSTM approach for detecting narrative containers », in : *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 224-230.

TRIEU, Hai-Long et al. (2019), « Coreference Resolution in Full Text Articles with BERT and Syntax-based Mention Filtering », in : *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, p. 196-205.

TRUSCHNEGG, Astrid et al. (2016), « Epulis : a study of 92 cases with special emphasis on histopathological diagnosis and associated clinical data », in : *Clinical Oral Investigations* 20.7, p. 1757-1764.

UZUNER, Özlem et al. (2011), « 2010 I2B2/VA challenge on concepts, assertions, and relations in clinical text », in : *Journal of the American Medical Informatics Association* 18.5, p. 552-556.

VAN DEN BERCKEN, Laurens, Robert-Jan SIPS et Christoph LOFI (2019), « Evaluating neural text simplification in the medical domain », in : *The World Wide Web Conference*, p. 3286-3292.

VAN MULLIGEN, Erik M et al. (2016), « Erasmus MC at CLEF eHealth 2016 : Concept recognition and coding in French texts », in : *CLEF 2016 Online Working Notes. CEUR-WS.*

VASWANI, Ashish et al. (2017), « Attention is all you need », in : *Advances in neural information processing systems*, p. 5998-6008.

VELLDAL, Erik, Lilja ØVRELID et Stephan OEPEN (juill. 2010), « Resolving Speculation : MaxEnt Cue Classification and Dependency-Based Scope Rules », in : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, Uppsala, Sweden : Association for Computational Linguistics, p. 48-55, URL : https://www.aclweb.org/anthology/W10-3007.

VELLDAL, Erik, Lilja ØVRELID et al. (juin 2012), « Speculation and Negation : Rules, Rankers, and the Role of Syntax », in : *Computational Linguistics* 38.2, ISSN : 0891-2017, 1530-9312, DOI : 10.1162/COLI_a_00126.

VELUPILLAI, Sumithra, Hercules DALIANIS et Maria KVIST (jan. 2011), « Factuality Levels of Diagnoses in Swedish Clinical Text », in : *Studies in health technology and informatics* 169, p. 559-63, DOI : 10.3233/978-1-60750-806-9-559.

VINCENT, Jean-Louis et al. (2018), « Mean arterial pressure and mortality in patients with distributive shock : a retrospective analysis of the MIMIC-III database », in : *Annals of intensive care* 8.1, p. 107.

VINCZE, Veronika (2015), « Uncertainty detection in natural language texts », thèse de doct., szte.

VINCZE, Veronika et al. (2008), « The BioScope corpus : biomedical texts annotated for uncertainty, negation and their scopes », in : *BMC Bioinformatics* 9, p. 38-45, ISSN : 1471-2105, DOI : 10.1186/1471-2105-9-S11-S9.

WRENN, Sean M, Peter W CALLAS et Wasef ABU-JAISH (2017), « Histopathological examination of specimen following cholecystectomy : are we accepting resect and discard ? », in : *Surgical endoscopy* 31.2, p. 586-593.

WYDMUCH, Marek et al. (2018), « A no-regret generalization of hierarchical softmax to extreme multi-label classification », in : *Advances in Neural Information Processing Systems*, p. 6355-6366.

XU, Huimin et al. (2015), « Trends and patterns of five antihypertensive drug classes between 2007 and 2012 in China using hospital prescription data », in : *Int J Clin Pharmacol Ther* 53.6, p. 430-437.

XU, Jinghong et al. (2019), « Association of sex with clinical outcome in critically Ill sepsis patients : a retrospective analysis of the large clinical database MIMIC-III », in : *Shock (Augusta, Ga.)* 52.2, p. 146.

YANG, Christopher C et al. (2012), « Social media mining for drug safety signal detection », in : *Proceedings of the 2012 international workshop on Smart health and well-being*, p. 33-40.

YANG, Zhilin et al. (2019), « Xlnet : Generalized autoregressive pretraining for language understanding », in : *Advances in neural information processing systems*, p. 5753-5763.

YANG, Zichao et al. (2016), « Hierarchical attention networks for document classification », in : *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics : human language technologies*, p. 1480-1489.

YOUNG, Tom et al. (2018), « Recent trends in deep learning based Natural Language Processing », in : *IEEE Computational intelligence magazine* 13.3, p. 55-75.

ZAHEER, Manzil et al. (2020), « Big bird : Transformers for longer sequences », in : *arXiv preprint arXiv :2007.14062*.

ZHANG, Jun et al. (2010), « Automatic patient search for breast cancer clinical trials using free-text medical reports », in : *Proceedings of the 1st ACM International Health Informatics Symposium*, p. 405-409.

ZHOU, Huiwei et al. (2010), « Exploiting multi-features to detect hedges and their scope in biomedical texts », in : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, Association for Computational Linguistics, p. 106-113.

ZHU, Qiaoming et al. (oct. 2010), « A Unified Framework for Scope Learning via Simplified Shallow Semantic Parsing », in : *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA : Association for Computational Linguistics, p. 714-724, URL : https://www.aclweb.org/anthology/D10-1070.

ZOU, Bowei, G. ZHOU et Q. ZHU (jan. 2013), « Tree kernel-based negation and speculation scope detection with structured syntactic parse features », in : *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, p. 968-976.

ZOU, Bowei, Qiaoming ZHU et Guodong ZHOU (juill. 2015), « Negation and Speculation Identification in Chinese Language », 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)*, Beijing, China : Association for Computational Linguistics, p. 656-665, DOI : 10.3115/v1/P15-1064, URL : https://www.aclweb.org/anthology/P15-1064.

ZWEIGENBAUM, Pierre et Thomas LAVERGNE (2017), « Multiple Methods for Multi-class, Multi-label ICD-10 Coding of Multi-granularity, Multilingual Death Certificates. », in : *CLEF 2017 Online Working Notes. CEUR-WS.*

UNIVERSITÉ DE RENNES 1

Titre : Fouille de texte et extraction d'informations dans les données cliniques

Mots clés : TALN, étiquetage de séquence, négation, incertitude, classification multi-étiquette, textes cliniques

Résumé : Avec la mise en place d'entrepôts de données cliniques, de plus en plus de données de santé sont disponibles pour la recherche. Si une partie importante de ces données existe sous forme structurée, une grande partie des informations contenues dans les dossiers patients informatisés est disponible sous la forme de texte libre qui peut être exploité pour de nombreuses tâches. Dans ce manuscrit, deux tâches sont explorées : la classification multi-étiquette de textes cliniques et la détection de la négation et de l'incertitude. La première est étudiée en coopération avec le centre hospitalier universitaire de Rennes, propriétaire des textes cliniques que nous exploitons, tandis que, pour la seconde, nous exploitons des textes biomédicaux librement accessibles que nous annotons et diffusons gratuitement. Afin de résoudre ces tâches, nous proposons différentes approches reposant principalement sur des algorithmes d'apprentissage profond, utilisés en situations d'apprentissage supervisé et non-supervisé.

Title: Text mining and information extraction in clinical data

Keywords: NLP, sequence labeling, negation, speculation, multi-label classification, clinical narratives

Abstract: With the introduction of clinical data warehouses, more and more health data are available for research purposes. While a significant part of these data exist in structured form, much of the information contained in electronic health records is available in free text form that can be used for many tasks. In this manuscript, two tasks are explored: the multi-label classification of clinical texts and the detection of negation and uncertainty. The first is studied in cooperation with the Rennes University Hospital, owner of the clinical texts that we use, while, for the second, we use publicly available biomedical texts that we annotate and release free of charge. In order to solve these tasks, we propose several approaches based mainly on deep learning algorithms, used in supervised and unsupervised learning situations.