THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Informatique
Unité de recherche : Laboratoire d'Informatique de Grenoble

Résolution de coréférence dans les traductions machines

Coreference REsolution into MAchine Translation

Présentée par :

Fabien LOPEZ

Direction de thèse :

Didier SCHWAB PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Directeur de thèse
Marco DINARELLI CHARGEE DE RECHERCHE, CNRS DELEGATION ALPES	Co-encadrant de thèse
Emmanuelle ESPERANCA-RODIER MAITRESSE DE CONFERENCES, UNIVERSITE GRENOBLE ALPES	Co-encadrante de thèse

Rapporteurs :

Massimo POESIO
FULL PROFESSOR, Queen Mary University of London
Mathieu LAFOURCADE
MAITRE DE CONFERENCES, Université de Montpellier

Thèse soutenue publiquement le **15 juin 2026**, devant le jury composé de :

Didier SCHWAB, PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Directeur de thèse
Massimo POESIO, FULL PROFESSOR, Queen Mary University of London	Rapporteur
Mathieu LAFOURCADE, MAITRE DE CONFERENCES, Université de Montpellier	Rapporteur
Caroline ROSSI, PROFESSEURE DES UNIVERSITES, Université Grenoble Alpes	Examinatrice

Invités :

Marco DINARELLI
CHARGE DE RECHERCHE, CNRS Délégation Alpes
Gabriela-Nicole GONZALEZ-SAEZ
MAITRESSE DE CONFERENCES, Université Grenoble Alpes

Contents

I State-of-the-Art **17**

1 Coreference and Anaphora **20**
- 1.1 What is a coreference? . . . 21
- 1.2 What is a coreference model? . . . 22
- 1.3 Datasets . . . 23
- 1.4 Evaluation Measures . . . 25
- 1.5 Architectures . . . 35

2 Machine Translation **43**
- 2.1 Sentence-Level Machine Translation . . . 44
  - 2.1.1 What is a neural translation model? . . . 44
  - 2.1.2 Datasets . . . 46
  - 2.1.3 Evaluation Measures . . . 50
  - 2.1.4 Architecture . . . 53
- 2.2 Context-Aware Machine Translation . . . 56
  - 2.2.1 Motivations and Discussions . . . 57
  - 2.2.2 What is a Context-Aware Neural Machine Translation? . . . 59
  - 2.2.3 Evaluation Measures . . . 60
  - 2.2.4 Architectures . . . 61
- 2.3 Integrating Coreference into Machine Translation . . . 64

3 Explanation **66**
- 3.1 Introduction . . . 67
  - 3.1.1 Explainability: methods and models . . . 69
  - 3.1.2 Attention is not (not) Explanation . . . 74

II Contributions **76**

4 Methodology **79**
4.1 The Attention Mechanism: Why Focus Here? . . . 80
4.2 Experimental Setup . . . 81
4.2.1 Models and Experiences . . . 81
4.2.2 Post-processing . . . 85
4.2.3 Discussions . . . 89

5 Experiments **92**
5.1 Attention please! . . . 93
5.2 Heads up: Corefs Matters . . . 95
5.3 Cerberus or which head is giving direction? . . . 98

6 Conclusions & perspectives **102**
6.1 Summary of contributions . . . 102
6.2 Perspectives . . . 103

List of Figures

1.1 Coreference gold example 27
1.2 MUC example with a merge between small entities. 29
1.3 MUC example with a merge between big entities. 30
1.4 B-cubed example with one mention per entity. 32
1.5 B-cubed example with one entity for all mentions. 33
1.6 BLANC example with two small entities. 34
1.7 BLANC example with three small entities. 34
1.8 Example of false link. 37
1.9 Example of false new. 37
1.10 Example of wrong link. 38
1.11 Processing pipeline presented by Soon et al. (2001) to identify markables, i.e.: elements likely to be coreferential mentions. 38
1.12 First step on the process to evaluate the antecedent of the mention *"the company"*. 39
1.13 Second step on the process to evaluate the antecedent of the mention *"the company"*. 40

2.1 Architecture Transformer proposed by Vaswani et al. (2017). This figure comes from Vaswani et al. (2017). 55
2.2 Outside and inside context integration representation from Li et al. (2020). 63
2.3 Hierarchical attention network encoder architecture from Miculicich et al. (2018) 64

4.1 Simplified illustration of the attention mechanisms in the HAN model within the multi-encoder architecture proposed by Miculicich et al. (2018). 83
4.2 Processing pipeline for extracting and analyzing attention weights. . . 86

4.3 Examples of modifications applied to attention matrices for improved readability. 87
4.4 Example of the issue related to registers. 91
4.5 Corrected example of the issue related to registers. 91

5.1 An example of heatmap from the concatenation model encoder from sentence *221* to *k2* (sentence *219*). 96
5.2 Heatmap visualization of the values obtained using the metrics proposed in Section 4, showing the specialization of attention heads by layer and by head for the concatenation model and by context-level head and sentence-level head for the multi-encoder model. 100

1 Concat model Average weight metric as a percentage (↑). 107

List of Tables

1.1 Numbers of entities, links and mentions by language and data split in the OntoNotes5.0 corpus version release for the CoNLL-2012 shared task (Pradhan et al., 2012) . . . 24
1.2 Numbers of tokens and mentions in the ARRAU corpus (Poesio and Artstein, 2008) . . . 25
1.3 Numbers of tokens and mentions in the WikiCoref corpus (Ghaddar and Langlais, 2016) . . . 25
1.4 Numbers of tokens and mentions in the Ancor corpus (Poesio and Artstein, 2008) . . . 26
1.5 Table of $B^3_{precision}$, B^3_{recall}, and $B^3_{F-measure}$ results for the $B-cubed$ metric . . . 31

2.1 Table of the different language pairs included in the WMT17 corpus[1] 47
2.2 Information about data in en-de corpora from WMT17. . . . 48
2.3 Information about data in en-de corpora from IWSLT17. . . . 49
2.4 Corpus statistics for text, sentence and token from TED Talks and Newswire texts in ParCorFull 2.0. . . . 50
2.5 Statistics on the annotated mentions and their subcategories: pronouns (pron), nominal phrases (np), verbal phrases (vp), clauses. . . . 50

4.3 Quantitative results in terms of accuracy on the ContraPro test suite, obtained with the CA-NMT models. We show a comparison to a baseline context-agnostic model, and the best models from the literature. Models marked with * use both source and target context. . . . 84
4.4 Caption for table 4 . . . 85

5.1 Automatic results for different models on the 2,280 sentences forming the sub-corpus of ParCorFull2.0 proposed in Section 4 . . . 93
5.2 Automatic results for different models on the 74 sentences forming the sub-corpus of ParCorFull2.0 proposed in Section 4 . . . 93

5.3 Human (manual) evaluation statistics on the 74 selected examples for the multi-encoder (multi-enc) and concatenation (concat) models. . . 97

Abstract

Neural machine translation systems have reached near-expert performance at the sentence level, producing fluent and locally accurate translations. However, this performance degrades when moving to document-level translation, where discourse-level phenomena come into account, such as referential coherence, terminological consistency, and context-dependent disambiguation. In particular, specific linguistic ambiguities (such as those related to coreference) involve only a very small number of tokens (e.g.: pronouns), yet have a decisive impact on the overall coherence of the translated text.

A major limitation of current evaluation practices comes from the fact that standard metrics are largely insensitive to these localised errors or do not require the model to actually generate the translation. As a result, a model may achieve high scores while still failing in practice to correctly translate context-dependent ambiguities. This highlights a gap between measured performance and actual discourse quality and underscores the need for more fine-grained evaluation methods that aim to better understand the model's translation process.

This thesis addresses this issue by investigating how context-aware Transformer models leverage surrounding sentences through their attention mechanisms. Rather than focusing solely on output evaluation, this work adopts an interpretability-orientated perspective, aiming to determine whether models effectively rely on relevant contextual elements, such as the correct antecedents, when producing translations.

To this end, the thesis proposes a methodological framework combining attention weight analysis and controlled input context. It introduces a set of complementary metrics designed to quantify the extent to which attention distributions align with coreference relations across sentences. These metrics capture different types of attention behaviour, ranging from sharp alignment peaks to more subtle contextual activations.

Ultimately, this work contributes to bridging the gap between translation performance and model interpretability by providing tools to assess not only whether a model produces a correct translation but also to questions whether it does so for the right reasons: by effectively leveraging contextual information.

Résumé

Les systèmes de traduction automatique neuronale ont atteint un niveau de performance proche de celui d'un expert au niveau de la phrase, produisant des traductions fluides et localement correctes. Cependant, cette performance s'estompe dès lors que l'on considère la traduction au niveau du document. Il intervient alors des phénomènes discursifs tels que la cohérence référentielle, la consistance terminologique et une désambiguïsation dépendante du contexte. En particulier, certaines ambiguïtés linguistiques (comme celles liées à la coréférence) concernent un nombre très restreint de jetons (e.g.: des pronoms), mais ont un impact déterminant sur la cohérence globale du texte traduit.

Une limite majeure des pratiques actuelles d'évaluation réside dans le fait que les métriques standards sont peu sensibles à ces erreurs localisées ou alors ne demande pas au modèle de générer la traduction. Ainsi, un modèle peut obtenir de bons scores tout en échouant en pratique à traduire correctement des ambiguïtés dépendantes du contexte. Cela met en évidence un décalage entre la performance mesurée et la qualité discursive réelle, et souligne la nécessité de méthodes d'évaluation plus fines cherchant à comprendre le processus de traduction du modèle.

Cette thèse s'inscrit dans cette problématique en étudiant la manière dont les modèles Transformer contextualisés exploitent les phrases environnantes via leurs mécanismes d'attention. Plutôt que de se limiter à l'évaluation des sorties, ce travail adopte une approche orientée vers l'interprétabilité, visant à déterminer si les modèles s'appuient effectivement sur des éléments contextuels pertinents, tels que les bons antécédents, pour produire leurs traductions.

Dans cette optique, la thèse propose un cadre méthodologique combinant l'analyse des poids d'attention et le contrôle du contexte d'entrée. Elle introduit un ensemble de métriques complémentaires permettant de quantifier dans quelle mesure les distributions d'attention s'alignent avec les relations de coréférence entre phrases. Ces métriques capturent différents types de comportement de l'attention, allant de pics d'alignement nets à des activations contextuelles plus discrètes.

En définitive, ce travail contribue à réduire l'écart entre la performance en traduction et l'interprétabilité des modèles, en cherchant à fournir des outils permettant d'évaluer non seulement si un modèle produit une traduction correcte, mais aussi s'il le fait pour les bonnes raisons, c'est-à-dire en mobilisant efficacement le contexte.

Introduction

Neural machine translation (NMT) has become one of the most visible successes of artificial intelligence (AI) in the domain of modern natural language processing (NLP). It is now embedded in everyday tools, professional workflows and multilingual production pipelines in which users increasingly expect outputs that are both fluent and faithful. Much of the recent progress is particularly striking when translation quality is assessed sentence by sentence (i.e.: *sentence-level translation*): for short and relatively self-contained segments. Sentence-level neural machine translation (SL-NMT) often generates grammatically correct, lexically plausible, and stylistically natural translations.

However, this impression of maturity weakens as soon as translation is considered at the level of discourse (i.e.: *document-level translation*). A coherent document is not merely the sum of well-translated sentences. It requires continuity of reference, stable terminological and stylistic choices, and correct handling of information that is not locally available in a single sentence. In practice, many translation errors emerge precisely when meaning depends on links that span multiple sentences: implicit relations, temporal and causal dependencies, shifts in perspective, and, crucially, referential continuity across a text.

Coreference (See Chapter 1for more details) occupies a central position among the discourse phenomena that most clearly expose the gap between strong sentence-level performance and weaker document-level coherence. Coreference concerns the relation between multiple linguistic mentions (for example: pronouns, definite descriptions, proper names, or even larger constituents) that refer to the same underlying entity. Correctly resolving those links is a prerequisite for robust interpretation. Without those links, pronouns may become ambiguous, nominal chains might lose their targets and the thread of the discourse should break. In translation, coreference becomes especially demanding because the target language may require grammatical distinctions that are under-specified or absent in the source language (for instance, gender agreement with the pronoun "*it*" when translating from English to French). As a result, producing a correct translation may require identifying an antecedent that

appears earlier in the document and transferring this decision into morphosyntactic choices in the target language.

This thesis is situated at the intersection of coreference resolution (See Chapter 1 for more details), context-aware neural machine translation, also known as CA-NMT (See Section 2.2 for more details), and explainable IA (see Chapter 3). Its overarching goal is to study how ambiguous phenomena such as coreference affect translation and, more importantly, to understand *how* context-aware translation models use surrounding sentences to generate coherent output. The work is not limited to verifying whether the system produces the correct translation; it seeks to determine whether a model arrives at that translation *for the right reasons* (i.e.: by actually leveraging the appropriate contextual antecedent rather than relying on superficial correlations or spurious cues).

From performance to auditability: the "right reasons" question

As NMT systems improve, a persistent methodological challenge becomes more salient: the difficulty of auditing the internal mechanisms that lead to a prediction. A translation may be correct by coincidence or through shortcuts that do not reflect genuine contextual understanding. This thesis, therefore, adopts an explicitly diagnostic stance. In the specific case of coreference, "right reasons" can be operationalised as follows: when a model must disambiguate a mention (often a pronoun) in the current sentence, does it retrieve and exploit the correct antecedent from the context?

Addressing this question requires tools that go beyond standard end-to-end evaluation. Global metrics and human judgments of fluency remain essential, but they may fail to reveal whether contextual information was truly used. This thesis thus emphasizes *auditability*: developing methods that can provide evidence, at least partially, about which part of the context sentences influence translation decisions (See Section 3.1 for more details).

Why attention: a native tracer of contextual information flow

Transformers (See Subsection 2.1.4 for more details) access contextual information through the attention mechanism. This architectural property motivates the central

methodological choice of the thesis: to study coreference in CA-NMT by analysing attention patterns linking the current sentence to preceding context sentences. Attention is not assumed to be a universal or perfectly faithful explanation of model behaviour. Nonetheless, it is a privileged observational window into information flow in Transformer-based systems, and it can serve as a practical tracer of whether and where the model consider it similar to the ambiguity in the current sentence.

The research question is therefore formulated in operational terms:

> In a Transformer-type translation model, is it possible to find evidence of coreference-resolution links to the context within the attention mechanism, in order to provide an explanation for the translation?

This framing deliberately shifts the focus from output quality alone to the interpretability of contextual utilisation: whether internal attention signals can be connected, in a measurable way, to discourse-level phenomena.

Architectural comparison: concatenation vs. multi-encoder context modelling

To answer this question, the thesis compares two representative families of CA-NMT approaches:

- **Concatenation-based CA-NMT**: the context sentences and the current sentence are combined into a single longer sequence. In this setting, the model can, in principle, attend across sentence boundaries within the same encoder stream, but it may also suffer from diluted signals and difficulty in distinguishing context from the main sentence (See Subsection 4.2.1 for more details).
- **Multi-encoder CA-NMT with hierarchical attention**: the current sentence and each context sentence are encoded separately and then integrated through dedicated mechanisms (e.g., hierarchical attention operating at word and sentence levels). This design explicitly structures the context and may encourage more targeted retrieval of relevant information (See Subsection 4.2.1 for more details).

Comparing these approaches is not merely an architectural exercise. It also reflects two distinct hypotheses about contextual access: whether treating context as a flattened extension of the input sequence is sufficient or whether explicitly modelling context as structured information yields clearer and more reliable contextual signals.

Data: leveraging an annotated parallel corpus for coreference

A major obstacle in studying coreference in translation is data availability. Ideally, one needs texts annotated for coreference and aligned across languages so that referential phenomena can be analysed systematically. This thesis adopts a conservative and reliability-oriented strategy: rather than relying primarily on automatic coreference annotation (which can introduce cascading errors), it leverages **ParCorFull2.0**, a multilingual parallel corpus annotated for coreference and aligned at the sentence level. The empirical analyses in this work focus in particular on the English–German language pair.

ParCorFull2.0 provides richly annotated coreference chains that go beyond pronouns alone and include a variety of mention types. While the corpus is not designed to train large translation models from scratch, it is highly valuable for analysis and fine-grained evaluation because it supplies reference annotations (*gold* coreference links) that can be used to test whether internal model signals align with linguistically meaningful relations (See Subsection 2.1.2 for more details).

Experimental setup: controlled context and attention extraction

To ensure a consistent and interpretable experimental design, inputs are organised as a bundle of sentences in which the current sentence is denoted as $k0$ and the preceding context sentences as $k1$, $k2$, and $k3$ (when available). This notation makes it possible to control context depth and to study how attention is distributed between the current sentence and the context.

The thesis combines complementary experimental lenses:

- A **quantitative analysis** of how attention mass is distributed across current and context sentences, providing a global picture of contextual focus.
- A **qualitative analysis** on a smaller, carefully inspected subset of examples, aimed at verifying whether observed attention patterns correspond to plausible coreference-related behaviour.
- A **head-level analysis** exploring whether specific attention heads specialize in linking mentions to their antecedents, and how such specialization differs across architectures.

Making attention usable: preprocessing, normalization, and metrics

Raw attention matrices are not directly comparable to linguistic annotations. They are defined at the token level (often subword units), constrained by probability normalisation, and sensitive to sentence length and formatting artefacts. This thesis therefore proposes a processing pipeline to improve interpretability and comparability:

- Handling padding and non-informative tokens when present.
- Aggregating subword units into word-level representations, with aggregation choices designed to preserve salient peaks rather than dilute informative signals.
- Define an attention threshold using a uniform baseline to distinguish non-trivial attention from near or below-uniform distributions.
- Applying normalization procedures to mitigate length effects and improve cross-sentence comparability.

On top of this pipeline, the thesis introduces a set of complementary quantitative measures to link attention to coreference relations:

- **Max-weight**: capturing whether the strongest attention from a mention token targets an antecedent token.
- **Non-zero weight**: capturing whether attention above a baseline threshold is assigned to antecedent tokens.
- **Mean value**: capturing the average normalised attention assigned along mention to antecedent links.

Together, these metrics are intended to characterise different regimes of attention behaviour (sharp peaks, broader activation, and average intensity) and to enable systematic comparisons across models and experimental conditions.

Contributions

This thesis makes the following contributions:

- **A diagnostic perspective on context use in NMT**: the thesis reframes evaluation beyond output quality by asking whether a model produces correct context-dependent translations *for the right reasons*, i.e., by retrieving and exploiting the appropriate antecedent in the discourse context.
- **A comparative analysis of CA-NMT architectures**: it empirically contrasts two representative context-aware approaches(*concatenation-based* context modelling and a *multi-encoder* architecture with hierarchical attention) to characterize how their internal mechanisms differ when handling coreference-related ambiguities.
- **A corpus-based study grounded in gold coreference annotations**: the work leverages the ParCorFull2.0 parallel corpus with expert coreference annotation (focusing on English–German) to support controlled, linguistically grounded analyses that avoid the pitfalls of purely automatic annotation.
- **A practical pipeline for attention interpretability in this setting**: the thesis proposes processing steps (e.g., subword-to-word aggregation, application of a threshold relative to a uniform reference, normalization to mitigate length-related effects) that make attention matrices more comparable and usable for discourse-oriented diagnostics.
- **Quantitative metrics linking attention to coreference**: it introduces complementary measures (e.g., *max-weight*, *non-zero weight*, and *mean value*) to operationalize the presence and strength of attention along mention–antecedent links and to enable systematic comparisons across models.
- **An integrated quantitative and qualitative evaluation protocol**: the thesis combines corpus-level statistics with manual qualitative inspection and head-level analyses to assess whether observed attention patterns plausibly reflect coreference resolution behaviour rather than artifacts.

Structure of the thesis

The remainder of this thesis is organised as follows:

- **Part I State-of-the-art**: this first part focuses on background and related work, it is composed of three main chapters. First, Chapter 1 reviews coreference and anaphora. Second, Chapter 2 introduces neural machine translation

and context-aware neural machine translation. Third, Chapter 3 discusses interpretability considerations relevant to attention-based analyses.

- **Part II Contributions**: this second part focuses on the contributions of this thesis. First, Chapter 4 presents the overall framework: model selection, corpus choice and preparation, the experimental setup with controlled context ($k0$–$k3$), attention extraction, preprocessing/normalisation steps, and the definition of quantitative and qualitative evaluation procedures. Second, Chapter 5 is about experiments and analysis. This chapter reports the empirical results, combining corpus-level quantitative analyses, qualitative case studies, and head-level investigations to assess whether attention provides evidence of coreference-related links to context. Lastly, Chapter 6 summarises the main findings, discusses limitations, and outlines future directions, including extensions to additional architectures and discourse phenomena beyond coreference.

Publications

This thesis led to the publication of five articles as first author or contributor. These articles are:

- Fabien Lopez (June 2023). "Approches neuronales pour la détection des chaînes de coréférences : un état de l'art". fra. In: *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*. ed. by Marie Candito et al. Paris, France: ATALA, pp. 101–113. URL: https://aclanthology.org/2023.jeptalnrecital-rjc.7/

- Fabien Lopez et al. (Dec. 2023). "The MAKE-NMTVIZ System Description for the WMT23 Literary Task". In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Philipp Koehn et al. Singapore: Association for Computational Linguistics, pp. 287–295. DOI: 10.18653/v1/2023.wmt-1.30. URL: https://aclanthology.org/2023.wmt-1.30/

- Marco Dinarelli et al. (2024). "Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention". In: *Traitement Automatique des Langues* 64.3. Ed. by Guillaume Wisniewski, pp. 67–91. DOI: 10.57896/2023-tal-64_3_3. URL: https://aclanthology.org/2023.tal-3.4/

- Gabriela Gonzalez-Saez et al. (June 2024b). "The MAKE-NMTViz Project: Meaningful, Accurate and Knowledge-limited Explanations of NMT Systems for Translators". In: *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*. Ed. by Carolina Scarton et al. Sheffield, UK: European Association for Machine Translation (EAMT), pp. 12–13. URL: https://aclanthology.org/2024.eamt-2.7/
- Gabriela Gonzalez-Saez et al. (June 2024a). "Exploring NMT Explainability for Translators Using NMT Visualising Tools". In: *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*. Ed. by Carolina Scarton et al. Sheffield, UK: European Association for Machine Translation (EAMT), pp. 396–410. URL: https://aclanthology.org/2024.eamt-1.34/

Part I

State-of-the-Art

Introduction

This first part is dedicated to establishing the theoretical and technical framework necessary for understanding the work presented in this thesis. The objective is to provide a critical survey of the state-of-the-art in three complementary fields whose intersection lies at the heart of our research: coreference resolution, context-aware neural machine translation, and language model interpretability methods.

In Chapter 1, we define the fundamental concepts of coreference and anaphora. We detail the evolution of models in this field, moving from statistical approaches to "end-to-end" neural architectures while analysing the evaluation metrics that attempt to capture the complexity of this linguistic phenomenon.

Then, Chapter 2 explores the field of Machine Translation. After a review of baseline models, we focus on the crucial transition from sentence-level neural machine translation to document-level translation and context-aware neural machine translation. We specifically examine what changes from an architectural point of view. Then we examine how coreference information has been integrated into these systems to improve discursive coherence.

Finally, Chapter 3 addresses the question of explanation and interpretability in black-box models. We discuss methods for auditing model decisions, with a particular focus on the attention mechanism. This chapter lays the methodological foundation that will allow us, in the following part, to analyse whether translation models rely on "the right reasons" to resolve context-related ambiguities.

Contents

1 Coreference and Anaphora **20**
- 1.1 What is a coreference? . . . 21
- 1.2 What is a coreference model? . . . 22
- 1.3 Datasets . . . 23
- 1.4 Evaluation Measures . . . 25
- 1.5 Architectures . . . 35

2 Machine Translation **43**
- 2.1 Sentence-Level Machine Translation . . . 44
  - 2.1.1 What is a neural translation model? . . . 44
  - 2.1.2 Datasets . . . 46
  - 2.1.3 Evaluation Measures . . . 50
  - 2.1.4 Architecture . . . 53
- 2.2 Context-Aware Machine Translation . . . 56
  - 2.2.1 Motivations and Discussions . . . 57
  - 2.2.2 What is a Context-Aware Neural Machine Translation? . . . 59
  - 2.2.3 Evaluation Measures . . . 60
  - 2.2.4 Architectures . . . 61
- 2.3 Integrating Coreference into Machine Translation . . . 64

3 Explanation **66**
- 3.1 Introduction . . . 67
  - 3.1.1 Explainability: methods and models . . . 69
  - 3.1.2 Attention is not (not) Explanation . . . 74

Chapter 1

Coreference and Anaphora

In this Chapter 1, we will introduce notions related to the concept of coreference link resolution, focusing specifically on coreference within the task of Natural Language Processing: **automatic coreference link resolution**.

Automatic coreference link resolution is a fundamental task in Natural Language Processing that lies at the intersection of linguistic theory and computational modelling. It is the process of identifying all linguistic expressions, or mentions, that refer to the same real-world entity within a document. Achieving referential continuity is not only essential for human text comprehension but also serves as a backbone for higher-level downstream tasks, such as document-level machine translation, summarization, and information extraction.

In the context of this thesis, it is crucial to first establish a rigorous understanding of coreference phenomena. A model's ability to maintain a coherent "thread" of discourse depends on its capacity to link mention like pronouns, nouns, and proper names to their correct antecedents across sentence boundaries.

This chapter provides a comprehensive review of the state of the art in coreference resolution. We begin by clarifying the linguistic distinctions between coreference and anaphora (Section 1.1). We then transition to the computational perspective, defining what constitutes a coreference model and exploring the historical shift from pipeline-based systems to modern end-to-end neural architectures (Section 1.2). Following this, we survey the datasets that have shaped the field (Section 1.3) and critically analyse the complex landscape of evaluation measures such as MUC, B-cubed, and the CoNLL score which are used to quantify model performance (Section 1.4). Finally, we detail the evolution of model architectures, from early statistical heuristics to recent Transformer-based approaches (Section 1.5).

Contents

1.1 What is a coreference? 21
1.2 What is a coreference model? 22
1.3 Datasets 23
1.4 Evaluation Measures 25
1.5 Architectures 35

1.1 What is a coreference?

The coreference resolution task is a task that involves linking mentions in a sentence or text that refer to the same entity in a real or imaginary world. To define the core components:

- A **mention**, from the Latin *mentio* means "recall to memory", is a linguistic expression used to refer to an entity.
- An **entity** from the Latin *entis* means "being", "that which is" on the physical plane or as an object of thought.

A mention that is not linked to any other mention is called a **singleton** which means it is the only mention referring to the entity within the discourse. Conversely, two mentions referring to the same entity are linked together by a **coreference link**. A **coreference chain** (also called a cluster) symbolise an entity. It is represented as a set in which each mention is connected to each other with a coreference link. While standard computational definitions consider a chain to exist with at least two coreferent mentions, some linguistic studies (Corblin, 1995, Schnedecker, 1997) suggest a higher limit increased to a minimum of three elements to constitute a proper coreference chain. Coreference is integral to text comprehension processes (Winograd, 1972), and the specific task of coreference resolution has gained significant spotlight in computer science since the start of the coreference resolution task proposed in MUC6 (Message Understanding Conferences) (Vilain et al., 1995).

It is crucial to distinguish coreference from **anaphora**. An anaphora (from Greek *ana*, "back", and *phorein*, "to carry") creates a dependency where a linguistic element requires an antecedent to be understood. While anaphora and coreference frequently overlap, they are not the same concepts. Example 1.1.1 illustrates the three possible configurations:

- The mention "$[\text{It}]_2$" refers to the same entity as "$[\text{turtle}]_1$" but "$[\text{It}]_2$" needs the mention "$[\text{turtle}]_1$" to be understood. Here, both concepts overlap.

- The mentions "Tortuga$_4$" and "Tortuga$_5$" are coreferential as both refer to the same entity. However, since proper names function independently and do not structurally require an antecedent to be interpreted, they are not considered anaphoric in this context.
- Finally, "[The shell]$_6$" requires the context of "[a turtle]$_1$" to be understood. Yet, they do not refer to the exact same entity: the shell is just a part of the turtle. In that way, there is no coreference but what is called *associative anaphora* (or *bridging anaphora*).

> I saw [a turtle]$_1$. [It]$_2$ was swimming slowly. [Its]$_3$ name is [Tortuga]$_4$. [Tortuga]$_5$ is 100 years old. [The shell]$_6$ was dark green.

Example 1.1.1. Example of coreference and anaphora, coreference without anaphora and anaphora without coreference. Inspired from Kleiber and Vassiliadou (2007).

Although anaphora and coreference are two different concepts in linguistics, in computational linguistics tasks, these distinctions are sometimes merged for practical purposes. Similarly, concepts such as singletons are usually considered one mention coreference chains in NLP, while they are individual mentions in linguistics.

1.2 What is a coreference model?

At its core, a coreference resolution model is a function mapping a document D to a set of entities. The primary goal is to determine which mentions in the text refer to the same real-world entity.

Formally, coreference resolution is a clustering problem. Let $M = \{m_1, m_2, ..., m_N\}$ be the set of all mentions in a document. The objective of the model is to find the optimal partition $\mathcal{C} = \{c_1, c_2, ..., c_K\}$ of M, where each subset c_k represents a distinct entity identified as a coreference chain containing one or more mentions. Since the number of possible partitions grows exponentially with the number of mentions, the model cannot simply enumerate them. Instead, it must learn a scoring function to estimate the likelihood of a link or a cluster.

Historically, this problem was approached with rule-based systems (e.g., heuristics on gender and number agreement). With the advent of machine learning, and later

deep learning, the focus shifted to learning a compatibility score between mentions or clusters from annotated data.

A generic coreference model aims to be as close as possible to the gold cluster structure Y that maximises a conditional probability given the document D, as formulated in recent neural approaches (Wiseman et al., 2016, Lee et al., 2017):

$$\hat{Y} = \underset{Y \in \mathcal{Y}(D)}{\arg\max}\, P(Y|D) \tag{1.1}$$

Where $\mathcal{Y}(D)$ represents the space of valid coreference assignments for document D. The probability $P(Y|D)$ is usually parameterized by a neural network that scores the compatibility between two mentions independently of the others or the coherence of a whole entity cluster.

Pipeline vs. End-to-End We can categorise coreference models based on how they handle the input:

- **Pipeline approaches**: these models assume that mentions have been identified beforehand or by a separate module in a pipeline. The coreference model strictly focuses on clustering these pre-identified mentions.
- **End-to-End approaches**: as proposed by Lee et al. (2017), these models operate directly on the raw text. They jointly perform mention detection and coreference resolution, treating them as dependent tasks to optimise the global objective function.

Specific architectures and their detailed objective functions (such as the latent structure perception used in Wiseman et al. (2016) or the span-ranking loss in Lee et al. (2017)) will be detailed in Section 1.5.

1.3 Datasets

One of the issues concerning neural learning methods is the quantity and quality of the examples contained in the corpora used, as they directly influence the model's quality. In this section, we will detail the main annotated corpora for coreference in French, English and multilingual ones.

For coreference resolution in English, the main and widely used dataset is OntoNotes5.0. Introduced by Pradhan et al. (2012) as the official corpus of the CoNLL-2012 conference, it is the biggest corpus annotated on coreference with 2.9 million

Language	Type	Train	Development	Test	All
English	Entities	35,143	4,546	4,532	44,221
	links	120,417	14,610	15,232	150,259
	Mentions	155,560	19,156	19,764	194,480
Chinese	Entities	28,257	3,875	3,559	35,691
	Links	74,597	10,308	9,242	94,147
	Mentions	102 854	14 183	12 801	129 838
Arabic	Entities	8,330	936	980	10,246
	Links	19,260	2,381	2,255	23,896
	Mentions	27,590	3,313	3,235	34,138

Table 1.1: Numbers of entities, links and mentions by language and data split in the OntoNotes5.0 corpus version release for the CoNLL-2012 shared task (Pradhan et al., 2012)

tokens distributed across three languages: English, Chinese, and Arabic. 1.7 million tokens are dedicated to English, while 0.9 million are allocated to Chinese, and 0.3 million are allocated to Arabic with entities. The links and mentions are distributed as in Table 1.1. However, some of the choices made when designing this dataset have been discussed since the first version in 2007 (Pradhan et al., 2007). As an example, this dataset is not annotated with singletons.

In 2008, Poesio and Artstein (2008) created the ARRAU corpus based on 4 sub-corpura :

- RST corpus (Carlson et al., 2003): excerpts from the wall street journal.
- TRAINS-93 corpus (Heeman and Allen, 1995): dialogue between a user and an assistant giving him all information needed.
- PEAR corpus (Chafe, 1980): collection of spoken narratives from the Pear Stories.
- GNOME corpus (Poesio, 2004): texts from both medical and art history genres.

Although much smaller than the Ontonotes5.0 corpus, the ARRAU corpus has the advantage of containing more information. For example, annotated singletons can be found in the corpus, as well as discourse deixis and expletives.

Ghaddar and Langlais (2016) created the WikiCoref corpus mainly using the OntoNotes5.0 annotation schema (i.e. : singletons are not annotated, only noun phrases are eligible to be mentions). It's a relatively small corpus annotated by 2 annotators with a Kappa coefficient (Carletta, 1996) of 0.78.

Language	RST	GNOME	PEAR	TRAINS	All
Tokens	228,901	21,458	14,059	83,654	354,072
Mentions	72,013	6,562	4,008	16,999	99,582

Table 1.2: Numbers of tokens and mentions in the ARRAU corpus (Poesio and Artstein, 2008)

Language	WikiCoref
Tokens	59,652
Mentions	7,296
Entities	1,469

Table 1.3: Numbers of tokens and mentions in the WikiCoref corpus (Ghaddar and Langlais, 2016)

For other languages there are more substantial datasets. For example in French there is ANCOR (Muzerelle et al., 2011, 2014) and DEMOCRAT (Lattice et al., 2019). The Ancor corpus is the first substantial one in French, it is based on 4 subcorpora :

- ESLO_CO2 and ESLO_ANCOR (Eshkol-Taravella et al., 2011). This corpus is based on a transcribed section of the ESLO oral corpus. This section corresponds to spontaneous speech and, more specifically, sociolinguistic interviews. ESLO_CO2 consists of 35,000 tokens for 2.5 hours, while ESLO_ANCOR comprises 417,000 words for 25 hours.
- OTG. This corpus also corresponds to spontaneous speech, particularly in face-to-face dialogues between individuals and the staff of the Grenoble Tourist Office, spanning a duration of 2 hours and consisting of 26,000 words.
- Accueil_UBS. This corpus corresponds to spontaneous speech as well, specifically telephone dialogues collected from *Université Bretagne Sud*'s telephone switchboard, with a duration of one hour and 10,000 words.

1.4 Evaluation Measures

The automatic resolution of coreference chains depends on the data provided to the model during training, the specific architecture used, and the evaluation protocol proposed by the research community. This means comparing the output of the

Language	ESLO_CO2	ESLO_ANCOR	OTG	Accueil_UBS	All
Tokens	35,000	417,000	26,000	10,000	488,000
Length (hours)	2.5	25	2	1	30.5

Table 1.4: Numbers of tokens and mentions in the Ancor corpus (Poesio and Artstein, 2008)

model, called the **prediction**, to the reference values, called the **reference**. An objective comparison between two models requires them to be evaluated under similar conditions while also employing appropriate and relevant evaluation methods. In this chapter, we will study the different evaluation methods for the coreference resolution task by discussing their strengths and weaknesses.

Settings: Let us revisit example 1.4.1, where the mentions are: $[I]_1$, $[I]_2$, $[every\ woman]_3$, $[her]_4$, $[Bahr]_5$, $[Men]_6$, $[they]_7$, $[they]_8$, $[their]_9$, $[their\ families]_{10}$, $[their]_{11}$, $[them]_{12}$, $[them]_{13}$, $[she]_{14}$, $[women]_{15}$, $[women]_{16}$, $[they]_{17}$, $[their]_{18}$, and $[they]_{19}$, and could be organised as shown in Figure 1.1. Squares represent the entities, circles represent mentions and lines between two mentions represent a coreference or anaphora link between both mentions. In each entity, each word is linked to every other word of the entity.

- [...]
- "$[I]_1$'m not exaggerating when $[I]_2$ say that $[every\ woman]_3$ who arrives here has dealt with sexual violence on $[her]_4$ way to find refuge." said $[Bahr]_5$.
- "$[Men]_6$ flee because of war, because $[they]_7$ are politically persecuted, because $[they]_8$ are threatened with torture or death, or because $[[their]_9$ families$]_{10}$ pin $[their]_{11}$ hopes on $[them]_{12}$ and send $[them]_{13}$ to Europe." $[she]_{14}$ explains.
- When $[women]_{15}$ flee, other reasons are in the foreground: "Many $[women]_{16}$ sneak away secretly, because $[they]_{17}$ see no other way out."
- In $[their]_{18}$ home countries, $[they]_{19}$ experienced sexual violence, forced marriage, honour killings, slavery or forced prostitution.
- [...]

Example 1.4.1.

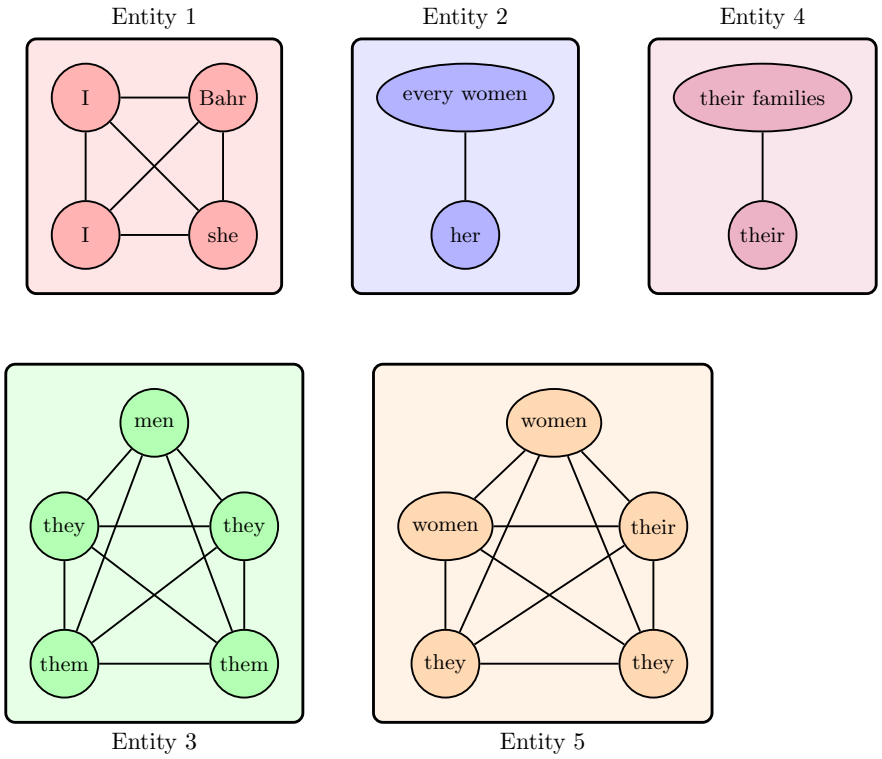


Figure 1.1: Coreference gold example

For all following measures, *precision* will be computed as the number of correct links predicted by the model divided by the number of predicted links. *Recall* will be computed as the number of correct links predicted divided by the number of links in the reference. A high precision means the model doesn't produce many links between non-coreferent mentions. On the other hand, a high recall means the model returns most of the reference links.

F-measure is calculated as Equation 1.2.

$$F\text{-}measure = 2 * \frac{precision * recall}{precision + recall} \tag{1.2}$$

MUC : A first way to assess the model's ability to resolve coreference links could be to verify whether every link is correct. However, because it may be inefficient to check every possible link between every mention of the document, a solution

may be to check, pair by pair, whether the links between the mentions are correct. This is what Vilain et al. (1995) proposes with the MUC (Message Understanding Conference) measure, the first coreference resolution measure ever proposed. MUC is based on the minimal number of links used to describe the entities and penalises extra and/or missing links. $MUC_{precision}$ and MUC_{Recall} are computed as follows:

$$MUC_{precision}(\mathcal{T}, \mathcal{R}) = \sum_{r \in \mathcal{R}} \frac{|r| - |partition(r, \mathcal{T})|}{|r| - 1} \tag{1.3}$$

$$MUC_{recall}(\mathcal{T}, \mathcal{R}) = \sum_{t \in \mathcal{T}} \frac{|t| - |partition(t, \mathcal{R})|}{|t| - 1} \tag{1.4}$$

Let $\mathcal{T}$ denote the set of reference mentions in the dataset, and $\mathcal{R}$ the set of mentions predicted by the model. Individual mentions from these sets are noted as $t \in \mathcal{T}$ and $r \in \mathcal{R}$, respectively. The term $|partition(r, \mathcal{T})|$ refers to the number of elements in $\mathcal{T}$ that have a non-empty intersection with r.

However, this metric has several flaws, the most controversial one is that the links are not weighted based on the size of the entities considered. To illustrate this, let us look back at the example 1.4.1 consider a system output such as the one shown in Figure 1.2 and in Figure 1.3. In both cases, we obtain a $MUC_{precision} = 0.93$, as in both cases only one link is incorrect with a minimal representation of the links. This result is counter-intuitive because we would expect to get a higher precision for a model merging two small entities rather than two large ones.

Moreover, since MUC is based on links, the metric also has the drawback of not taking into account singletons, which are not linked to any other mention.

B-Cubed : A way to account for both the size of the entities and the issue of singletons would be to no longer rely on links but on the mention's entities themselves. By calculating precision and recall for each entity, it might be possible to obtain a more relevant score than MUC. Bagga and Baldwin (1998) thus propose the **B-cubed** measure (also denoted as B^3). $B^3_{precision}$ and B^3_{recall} are then defined as:

$$B^3_{precision} = \frac{1}{\sum\limits_{r_j \in \mathcal{R}} |r_j|} \sum_{r_j \in \mathcal{R}} \sum_{t_i \in \mathcal{T}} \frac{|r_j \cap t_i|^2}{|r_j|} \tag{1.5}$$

$$B^3_{recall} = \frac{1}{\sum\limits_{t_i \in \mathcal{T}} |t_i|} \sum_{t_i \in \mathcal{T}} \sum_{r_j \in \mathcal{R}} \frac{|t_i \cap r_j|^2}{|t_i|} \tag{1.6}$$

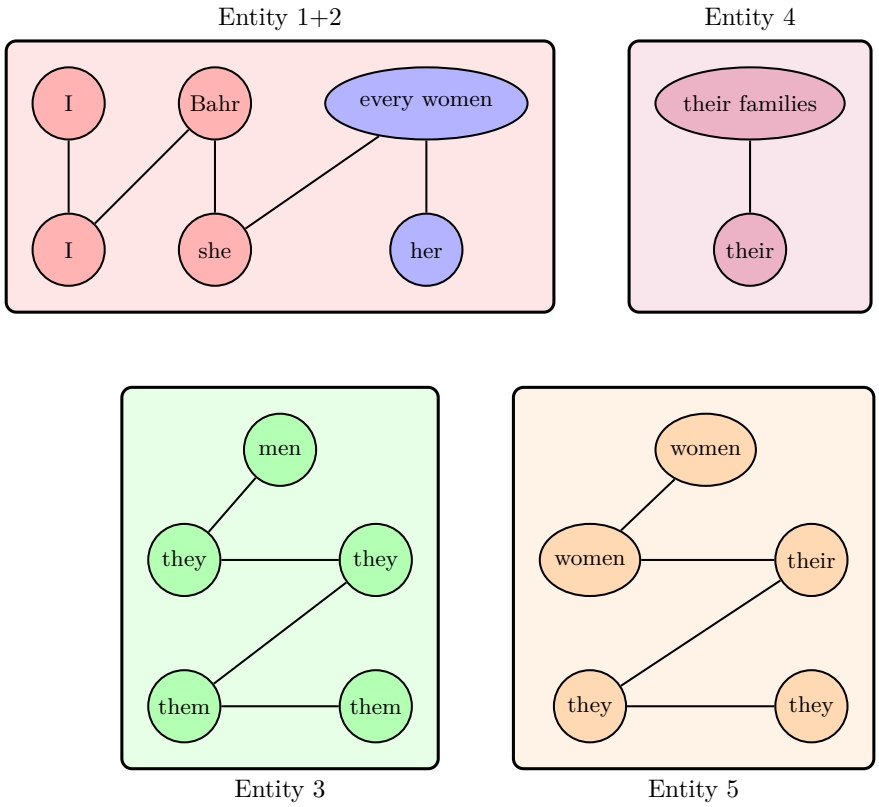


Figure 1.2: MUC example with a merge between small entities.

Since B^3 relies only on mentions to measure the quality of the prediction, the metric suffers from the "mention identification effect" discussed by Luo and Pradhan (2016). An illustrative example proposed by Luo and Pradhan (2016) highlights the issue that arises when key mentions are repeated in the system output. In this example, the gold standard contains a single entity $\{a, b, c\}$, whereas the system predicts three entities: $\{a, b\}$, $\{a, c\}$, and $\{b, c\}$. When the gold entity is compared to the first predicted cluster $\{a, b\}$, it is divided into two subsets, leading to a partial credit of $\frac{2}{3}$ for mentions a and b. Repeating this process with the second predicted cluster $\{a, c\}$ again assigns a credit of $\frac{2}{3}$ to mention a. Consequently, mention a accumulates a total credit of $\frac{4}{3}$; mentions b and c each receive the same amount. This results in an overall recall of $\frac{4/3+4/3+4/3}{3} = \frac{4}{3}$ which is, a value greater than 1, and is clearly inconsistent.

Additionally, as singletons are the most represented elements in documents, B^3

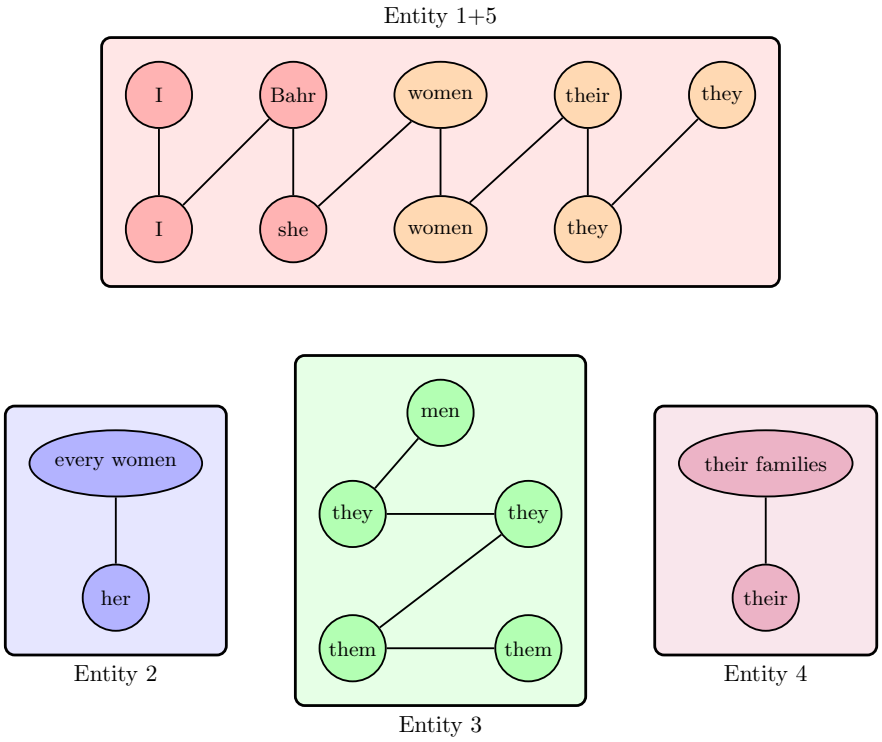


Figure 1.3: MUC example with a merge between big entities.

tends to be easily inflated and leaves little room for comparison between coreference chains. Example 1.4 represent an output with one mention per entity while example 1.5 represent the output with one entity for all mentions. Both examples have $B^3_{precision}, B^3_{recall}$ et $B^3_{F-mesure}$ in table 1.5.

Example 1.4 will have a $B^3_{precision}$ of 1, as there are no non-coreferent mentions within the same entity. However, it will have a $B^3_{recall} = 0.23$, resulting in an F-measure of $B^3_{F-measure} = 0.37$. Conversely, an output where the model places all mentions within a single entity will achieve a $B^3_{recall} = 1$ of 1, as no links are missing. However, the $B^3_{precision} = 0.28$, resulting in an F-measure of 0.43. In both cases mentioned above, the B-cubed metric yields an F-measure that is considered too high relative to the actual quality of the output.

CEAF : Following the proposals of MUC and B^3 with their respective advantages and disadvantages, Luo (2005) introduced **CEAF** (Constrained Entity-Alignment F-Measure), a metric based on an alignment function g^* between the set of reference

	Example multiple singletons	Example one matrice
$B^3_{precision}$	1	0.28
B^3_{recall}	0.23	1
$B^3_{F-measure}$	0.37	0.43

Table 1.5: Table of $B^3_{precision}$, B^3_{recall}, and $B^3_{F-measure}$ results for the $B-cubed$ metric

mention dataset $\mathcal{T}$ and the set of predicted mention $\mathcal{R}$. Furthermore, $CEAF$ uses a similarity metric denoted as ϕ to calculate $CEAF_{precision}$ and $CEAF_{recall}$ as follows:

$$CEAF_{precision} = \frac{\sum_{t_i \in \mathcal{T}^*} \phi(t_i, g^*(t_i))}{\sum_{r_j \in \mathcal{R}} \phi(r_j, r_j)} \tag{1.7}$$

$$CEAF_{recall} = \frac{\sum_{t_i \in \mathcal{T}^*} \phi(t_i, g^*(t_i))}{\sum_{t_i \in \mathcal{T}} \phi(t_i, t_i)} \tag{1.8}$$

Thus, a different ϕ function will yield different precision and recall values. We refer to CEAF$_m$ for $\phi(t_i, r_j) = |t_i \cap r_j|$ and CEAF$_e$ for $\phi(t_i, r_j) = \frac{2*|t_i \cap r_j|}{|t_i|+|r_j|}$.

The CEAF variants are all subject to the mention identification issue noted for B-cubed, as well as the lack of consideration for entity size. Lastly, due to the use of optimal alignment, if the alignment does not correctly match a mention from $\mathcal{T}$ to its counterpart in $\mathcal{R}$, that mention will not count as a correct mention.

MELA : Since each of the three metrics has its own strengths and weaknesses, Denis and Baldridge (2009) proposed **MELA** (Mention, Entity and Link Average). This metric gained popularity during the CoNLL-2011 and CoNLL-2012 events, from which it derives its most common name: the "CoNLL-score". MELA is calculated as the average of the F1-measure from the MUC, B^3, and CEAF$_e$ measures. It is calculated using the formula from equation 1.9.

$$CoNLL\text{-}score = \frac{MUC_{F_1} + B^3_{F_1} + CEAF_{F_1}}{3} \tag{1.9}$$

Although CoNNL-score is the most widely used to date, Moosavi and Strube (2016) argue that averaging three biased metrics cannot provide reliable results.

BLANC : Recasens and Hovy (2011) sought to minimise the impact of a mention detected as coreferent but placed in the wrong coreference chain. To address

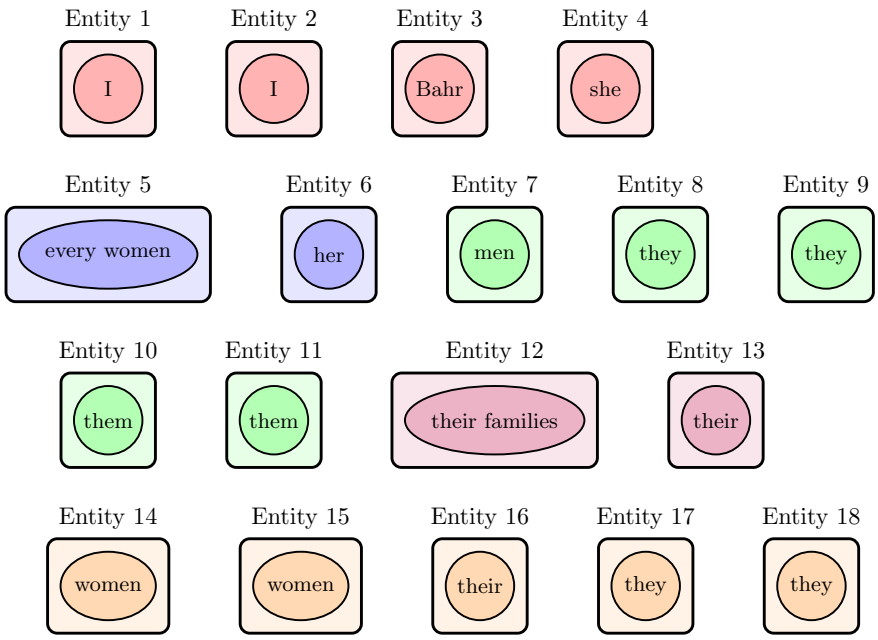


Figure 1.4: B-cubed example with one mention per entity.

this issue, they introduced **BLANC**, based on the Rand Index (Rand, 1971), by adding the concept of non-coreference links that connect two non-coreferent mentions. Recasens and Hovy (2011) defines rc, wc, rn, and wn as shown in equation 1.10. $BLANC_{precision}$ and $BLANC_{Recall}$ are defined as in equation 1.11.

$$\begin{cases} rc\text{: the correct coreference links} \\ wc\text{: the wrong coreference links} \\ rn\text{: the correct non-coreference links} \\ wn\text{: the wrong non-coreference links} \end{cases} \tag{1.10}$$

$$\begin{aligned} P_c &= \tfrac{rc}{rc+wc} \qquad P_n = \tfrac{rn}{rn+wn} \qquad BLANC_{precision} = \tfrac{P_c+P_n}{2} \\ R_c &= \tfrac{rc}{rc+wn} \qquad R_n = \tfrac{rn}{rn+wc} \qquad BLANC_{Recall} = \tfrac{R_c+R_n}{2} \\ F_c &= \tfrac{2P_cR_c}{P_c+R_c} \qquad F_n = \tfrac{2P_nR_n}{P_n+R_n} \qquad BLANC_{F\text{-}measure} = \tfrac{F_c+F_n}{2} \end{aligned} \tag{1.11}$$

However, as illustrated in Figure 1.6, which contains 2 coreference links and 4 non-coreference links, adding an entity with 2 mentions in Figure 1.7 results in 3 coreference links but an increase to 14 non-coreference links. This rapid growth in

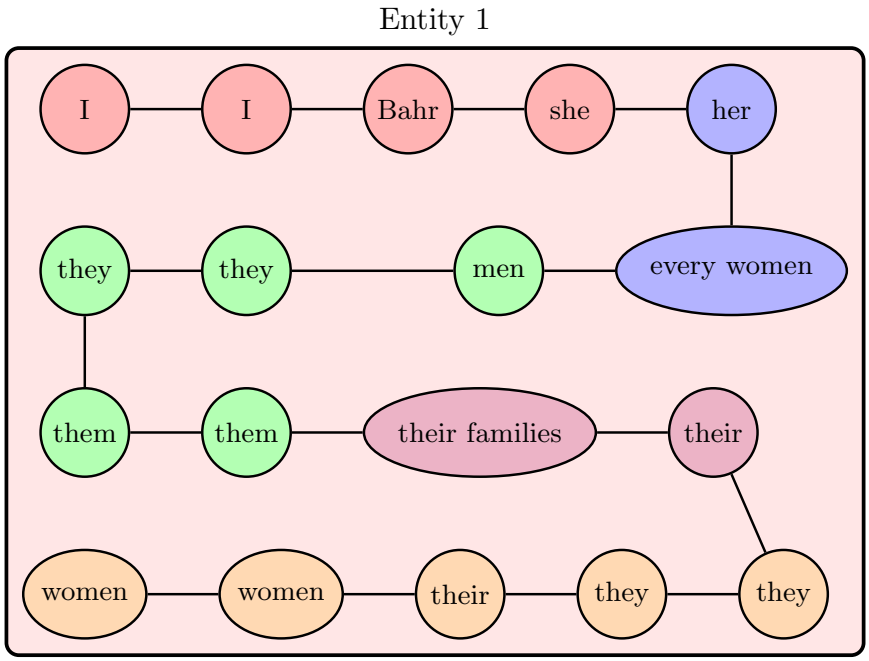


Figure 1.5: B-cubed example with one entity for all mentions.

non-coreference links makes this aspect of the evaluation less sensitive to the quality of the model. Singleton are the worst case scenario where only non-coreferent links are created. The use of non-coreferent links makes BLANC more sensitive than other existing metrics to the mention identification problem (Moosavi and Strube, 2016).

LEA : Taking a step back from the various issues raised by the previous measure, Moosavi and Strube (2016) propose LEA (Link-Based Entity-Aware). LEA metric is computed as a link based resolution score similar to the MUC measure introduced in Subsection 1.4. This score is weighted by an importance coefficient based on entity size to counteract the disadvantages of the MUC metric, allowing a better consideration of long coreference chains. Moreover, since LEA is based on coreference links, the metric is not affected by the mention identification effect, unlike B^3, $CEAF$, and $BLANC$. Furthermore, Moosavi and Strube (2016) propose LEA as a measure capable of handling singletons, despite its coreference link-based approach, by representing singletons as a coreference link to themselves (i.e., a *self-link*). Precision and Recall takes the form of equation 1.12 and is defined as in equations 1.13 and

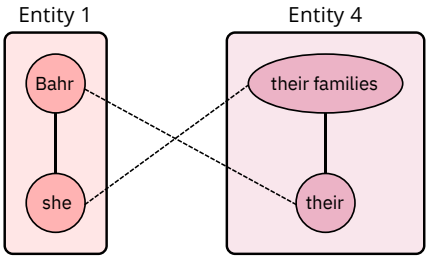


Figure 1.6: BLANC example with two small entities.

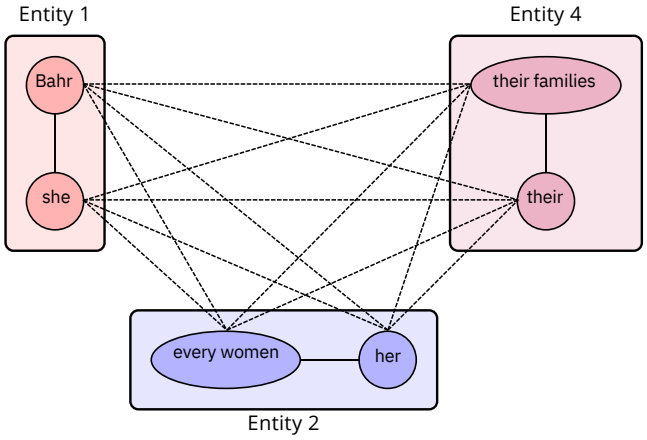


Figure 1.7: BLANC example with three small entities.

1.14, respectively.

$$\frac{\sum_{e_i \in \mathcal{E}}(importance(e_i) \times resolution\text{-}score(e_i))}{\sum_{e_k \in \mathcal{E}} importance(e_k)} \tag{1.12}$$

$$precision = \frac{\sum_{r_j \in \mathcal{R}}(|r_j| * \sum_{t_i \in \mathcal{T}} \frac{link(r_j \cap t_i)}{link(r_j)}}{\sum_{r_z \in \mathcal{R}} |r_z|} \tag{1.13}$$

$$recall = \frac{\sum_{t_i \in \mathcal{T}}(|t_i| * \sum_{r_j \in \mathcal{R}} \frac{link(t_i \cap r_j)}{link(t_i)}}{\sum_{t_z \in \mathcal{T}} |t_z|} \tag{1.14}$$

Lmetrics : Up to this point, we have treated coreference link resolution as a simple clustering problem. However, Chen and Ng (2013) propose the **Lmetrics**, a variant of MUC, B-cubed, and CEAF that is capable of incorporating the linguistic aspects overlooked in the original versions of these evaluation metrics.

To achieve this, they highlight three components: the elements in the response subset, the elements in the gold subset, and the elements common to both subsets. They assign a weight to these elements based on the type of coreference link, considering, for instance, whether the coreference link is a name, nominal, or pronominal.

Conclusion : Currently, the most widely used metric remains the CoNLL score. Its role in coreference resolution is predominant, despite various criticisms (Moosavi and Strube, 2016; Recasens and Hovy, 2011). In particular, Lion-Bouton et al. (2020) investigated which of the above-mentioned measures met the criteria necessary to be considered genuine measures of similarity. Their findings show that the MUC, B^3, $CEAF_e$, CoNLL-score, BLANC, and LEA metrics do not satisfy the triangle inequality, meaning that two very similar predictions can receive significantly different scores. Moreover, MUC, and by extension, the CoNLL score, does not achieve an identity score of 1 in extreme cases, such as when only singleton mentions are present. Finally, BLANC fails to meet the principle of symmetry. Only $CEAF_m$ appears to meet all the necessary criteria to be considered a normalized similarity metric.

1.5 Architectures

We will detail the different automatic coreference resolution systems along three main axes:

- First will be focused on resolving coreference links between pairs of mentions: **mention oriented coreference model**,
- Second will focus on systems seeking to represent the entity to which the mentions refer: **entity oriented coreference model**,
- The third will group together the different architectures that don't fit into the first two categories: **different approach types**.

Automatic coreference resolution systems were initially statistical (Ng and Cardie, 2002; Soon et al., 2001) before shifting towards a deep learning approach, which, thanks to neural network-based models, has better generalisation capabilities (Lee et al., 2018; Miculicich and Henderson, 2022; Wu et al., 2020). We will introduce brief information about statistical model then go deeper into models with neural network approach.

Mention Oriented Coreference Model

As mention oriented approach solves coreference by considering pairs of mentions, we call it **pairwise mention coreference resolution**. This approach use a similarity score based on features of the both mentions. These features can be as explicit as the gender or number of people involved, or implicit, captured through features learned

by training. The idea of the mention oriented approach is to look sequentially at every mention of the document and then solve the coreference with a previous antecedent if it exists.

To find this potential antecedent, different algorithms were implemented like *closest first algorithm* (Soon et al., 2001) or *best first algorithm* (Ng and Cardie, 2002). On one hand, the *closest-first algorithm* applies a binary classifier to every mention pair between the given mention and every preceding mention. For any given mention, it assigns as antecedent the nearest preceding mention that the classifier labels as COREF (i.e.: labeled as a coreferent mention), which explains the term closest-first. On another hand, the *best first antecedent* algorithm computes a similarity score on every mention pair between the given mention and every preceding mention. For any given mention, it assigns as antecedent the mention that achieves the highest similarity score.

Adopting the approach of Ng and Cardie (2002), Rahman and Ng (2009) introduced ϵ, a new potential default antecedent (also called a dummy antecedent) that is used to symbolise an absence of antecedents. Although providing the foundations of what is now used for coreference resolution, this approach remains prone to errors primarily because it only considers the two current mentions without taking into account the context of both current mentions or links that have already been solved by the model. Once the entire document has been processed, coreference chains are built by transitivity.

The various systems proposed up to now were part of a processing chain using neural networks only for the score function (Wiseman et al., 2015).

Wiseman et al. (2015) propose to compute the score between two mentions x and y using equation 1.15, where $\mathbf{u}$, $\mathbf{v}$, g, h_a, and h_p are learnt parameters.

$$s(x,y) \triangleq \begin{cases} \mathbf{u}^\top g\left(\begin{bmatrix} h_a(x) \\ h_p(x,y) \end{bmatrix}\right) + u_0 & \text{if } y \neq \epsilon \\ \mathbf{v}^\top h_a(x) + v_0 & \text{if } y = \epsilon \end{cases} \tag{1.15}$$

This neural approach needs an objective function to be trained. The one used by Wiseman et al. (2015) is shown in equation 1.16.

$$L(\theta) = \sum_{n=1}^{N} \max_{\hat{y} \in \mathcal{Y}(x_n)} \Delta(x_n, \hat{y}) \left(1 + s(x_n, \hat{y}) - s(x_n, y_n^\ell)\right) + \lambda \|\theta\|_1 \tag{1.16}$$

In equation 1.16, $\Delta(x_n, \hat{y})$ represents a coefficient that adjusts based on the type of error made by the model among the three possible error types:

- False link (FL): model links x to the former coreference chain when it should create a new one. An example is illustrated in Figure 1.8, where the mention "every women" is incorrectly linked to the mention "I" when a new cluster should have been created.

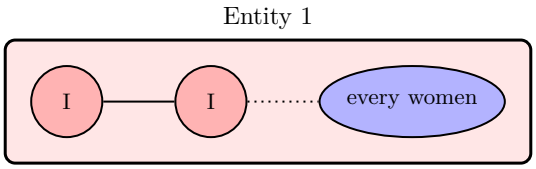


Figure 1.8: Example of false link.

- False new (FN): model wrongly predicts x as a new coreference chain when it should be linked to a former one. An example is provided in Figure 1.9, where the mention "her" incorrectly forms a new entity instead of being linked to the mention "every women" of Entity 2 as it should have been.

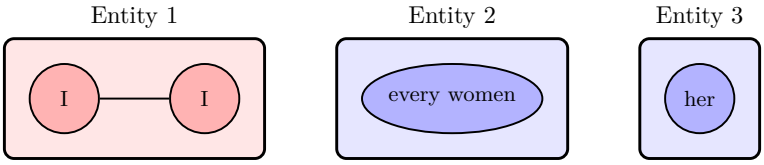


Figure 1.9: Example of false new.

- Wrong link (WL): model links x to the wrong former coreference chain. An example is provided in Figure 1.10, where the mention "her" is incorrectly linked to the mention "I" from Entity 1, whereas it should have been linked to the mention "every women" from Entity 2.

It is important to note that coreference resolution in this type of model is part of a processing pipeline aimed at detecting mention boundaries as well as various information required for the features. An example of such a processing pipeline is shown in Figure 1.11. This pipeline implies that errors made prior to the coreference resolution process will negatively impact the model. Such errors, known as cascading errors, propagate through the pipeline and ultimately hinder the performance of the coreference resolution model.

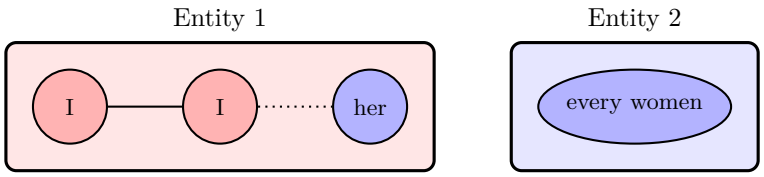


Figure 1.10: Example of wrong link.

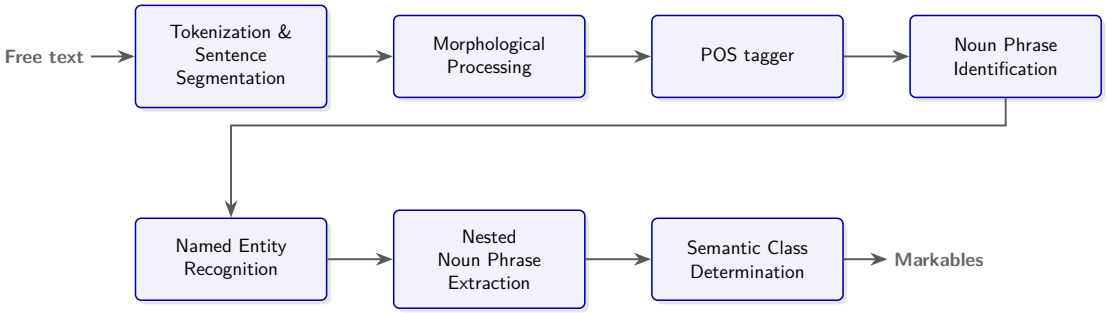


Figure 1.11: Processing pipeline presented by Soon et al. (2001) to identify markables, i.e.: elements likely to be coreferential mentions.

To limit any cascading error behaviour, Lee et al. (2017) proposed the first end-to-end coreference resolution neural model. The model detects mentions and then resolves coreference in an end-to-end process. This method arises several issues. First, in order to detect the maximum number of mentions in the document, the model checks all possible combinations of tokens creating a document of size $\mathcal{O}(T^4)$ with T the length of the document. The number of possibilities is partially addressed through aggressive pruning of potential mentions by limiting the length of mentions and incorporating a mention score s_m, in addition to the antecedent score s_a, which evaluates the probability that y is the antecedent of x. A pairwise coreference score is computed as in Equation 1.17.

$$s(x,y) = \begin{cases} 0 & \text{if } y = \epsilon \\ s_m(x) + s_m(y) + s_a(x,y) & \text{if } y \neq \epsilon \end{cases} \tag{1.17}$$

A second issue raised with the model proposed by Lee et al. (2017) is the use of head words, which were previously obtained through the processing chain. Lee et al. (2017) propose calculating what they call a soft-head, i.e., a mention head obtained

through an attention mechanism (Bahdanau et al., 2014) applied to the potential mention $\hat{x}$.

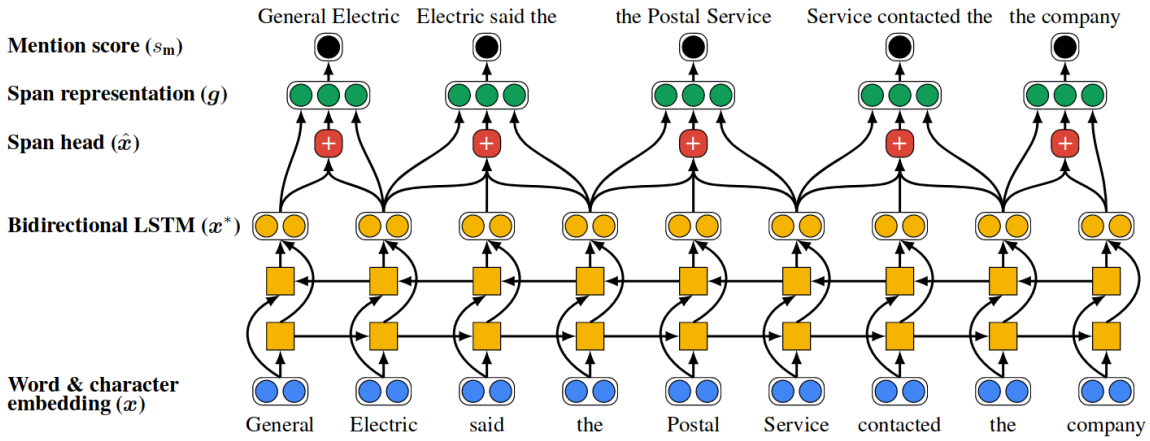


Figure 1.12: First step on the process to evaluate the antecedent of the mention *"the company"*.

An example of the process is provided in Figures 1.12 and 1.13, where the task is to evaluate whether the mention $x =$ "the company" has as its antecedent either $\hat{y}_1 =$ General Electric or $\hat{y}_2 =$ the Postal Service. Starting from the word and character embedding, the model create a span representation and a span.

Entity Oriented Coreference Model

Because the mention-based approach has the drawback of considering only the two current mentions, Clark and Manning (2015) suggest a way to incorporate information about entire coreference chains into their decision-making process. This type of approach has been widely adopted afterward because it is more intuitive and yields better results.

A first type of entity-based approach is proposed by Clark and Manning (2015), which uses an approach based on all pairs of mentions in the coreference chain to represent the characteristics of that chain.

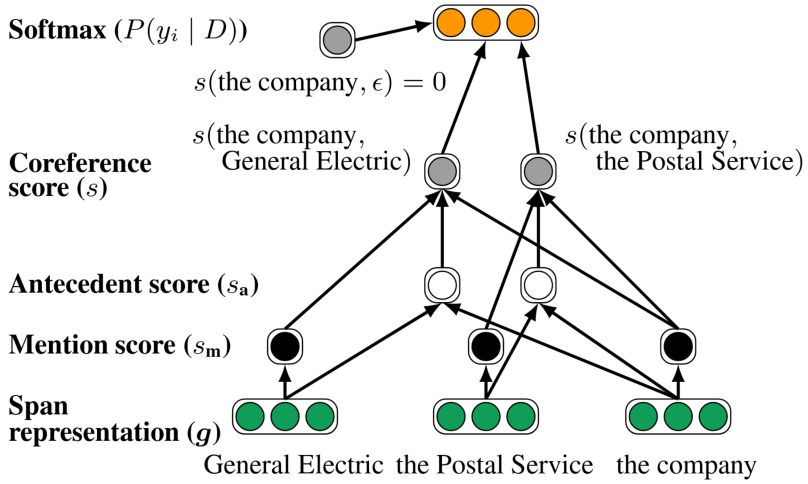


Figure 1.13: Second step on the process to evaluate the antecedent of the mention "*the company*".

> [Bill Clinton]$_1$ and [Hillary Clinton]$_2$ left the White House. [Clinton]$_3$ said that [he]$_4$ did not want to talk about his trip.

Example 1.5.1. Example of sentences where pairwise coreference resolution may be an issue.

With example 1.5.1, if the model links "*[Clinton]*$_3$" with "*[Hillary Clinton]*$_2$" as well as "*[Bill Clinton]*$_1$" with "*[he]*$_4$", since these are, respectively, the closest possible antecedents when only considering these two mentions and the only available antecedent, then the model could prevent the merging of the entities {"*[Clinton]*$_3$", "*[Hillary Clinton]*$_2$"} and {"*[Bill Clinton]*$_1$", "*[he]*$_4$"}. Although there may be a high chance of linking "*[Bill Clinton]*$_1$" with "*[Clinton]*$_3$", the model's features would highlight the incompatibility between "*[Bill Clinton]*$_1$" and "*[Hillary Clinton]*$_2$."

Wiseman et al. (2016) argue for the importance of incorporating global features into coreference resolution models as well as the challenges of implementing such features. To address this issue, they introduced a way to compute the features of coreference chains by encoding the entities as a sequence using recurrent neural networks. At each step, the current hidden state is used as a representation of the entire entity. Clark and Manning (2016) proposed a new method to compute similarity scores between coreference chains through the stacking of multiple encoders and

models. They extract the mention-pair encoder output from mention-oriented model to produce a cluster-level encoder to generate cluster-level features and then train a cluster-ranking model. The mention-oriented model is then used to prune the search space of the cluster ranking model. With $r_m(x, y)$ as the output of the mention-pair encoder, Clark and Manning (2016) compute a combination R_m of the entities' mention hidden state with $R_m(c^x, c^y) = [r_m(c_1^x, c_1^y), r_m(c_1^x, c_2^y)], ..., r_m(c_{|c^x|}^x, c_{|c^y|}^y)$ and then $r_c(c^x, c^y)$ the distributed representation of the cluster pair computed as equation 1.18.

$$r_c(c^x, c^y)_k = \begin{cases} \max\{R_m(c^x, c^y)_{k,\cdot}\} & \text{for } 0 \leq k < d \\ \text{avg}\{R_m(c^x, c^y)_{k-d,\cdot}\} & \text{for } d \leq k < 2d \end{cases} \tag{1.18}$$

Lee et al. (2017) consider only the features between the two current mentions, whereas Lee et al. (2018) propose a variant capable of incorporating cluster information. They do this by adding a cluster-aware parameter $s_c(x, y)$ to the pairwise coreference score, which take into accounts the information of clusters. Furthermore, Lee et al. (2018) introduced a new way to select potential mentions, which is more permissive and balance it by adding an additional layer of control to the system to better limit computational cost overall while achieving better results. Joshi et al. (2019) proposed to replace the GloVe embeddings used by Lee et al. (2018) with Bidirectional Encoder Representations from Transformers (BERT) embeddings (Devlin et al., 2019), highlighting the superior ability of BERT-large to distinguish between related but non-coreferent entities. Joshi et al. (2020) extended this work by introducing a SpanBERT embeddings model trained to predict spans rather than individual words, enabling the model to outperform BERT.

Different approach types

Other types of approaches have emerged. For instance, Miculicich and Henderson (2022) proposed a Graph2Graph Transformer to process an entire document at once, arguing that such a method is essential to effectively capture coreference relations across a document, despite the practical limitation on the maximum length of a document that can be processed in one pass.

On the other hand, Wu et al. (2020) suggested an alternative approach that does not process the entire document at once. Instead, their method deliberately ignores certain mentions initially and later employs a question-answering technique to recover the omitted mentions. However, this method is challenging to compare with other works due to the increased data requirements associated with leveraging question-answering datasets.

Other methods attempting to incorporate more linguistically grounded theories have also been proposed. For example, Jeon and Strube (2020) and Chai and Strube (2022) leveraged **Centering Theory** (Grosz et al., 1983, 1995; Joshi and Weinstein, 1981), which identifies the most relevant element in each sentence to account for shifts in focus across document segments.

Finally, Zhang et al. (2023), inspired by Paolini et al. (2021), proposed Seq2Seq models based on T5 (Raffel et al., 2020) or T0 (Sanh et al., 2022). Those Seq2Seq models take raw text as input and return it annotated with tags that delimit the mentions in the text. Each mention is assigned an identifier to indicate its associated entity. For example, the output for Example 1.5.1 would be formatted as: $<$ m $>$ Bill Clinton$|1$ $<$ $/m$ $>$ and $<$ m $>$ Hillary Clinton$|2$ $<$ $/m$ $>$ left the White House. $<$ m $>$ Clinton$|1$ $<$ $/m$ $>$ stated that $<$ m $>$ he$|1$ $<$ $/m$ $>$ did not want to discuss his trip.

Chapter 2

Machine Translation

Machine Translation has undergone a radical evolution in recent decades, moving from statistical models to deep learning architectures that now achieve near-human fluency in many language pairs. However, as Neural Machine Translation has matured, the focus of the research community has increasingly shifted from simple sentence-by-sentence mapping to the more complex challenge of document-level coherence. This chapter explores the technical and theoretical foundations of neural machine translation, with a specific emphasis on the transition toward context-aware models.

The chapter is structured as follows: Section 2.1 establishes the fundamentals of sentence-level neural machine translation, detailing the objective functions, the pivotal role of subword tokenization (BPE), and the Transformer architecture which serves as the current standard. In Section 2.2, we move beyond the sentence boundary to discuss Context-Aware Machine Translation. We categorize the various types of linguistic ambiguities as lexical, structural, and referential ambiguity that necessitate a context based approach and review the primary architectures used to incorporate context: concatenation-based and multi-encoder models. Finally, Section 2.3 focuses on the specific challenge of integrating coreference information into translation workflows, highlighting the sparsity of referential cues and the current strategies used to ensure discursive consistency across translated texts.

Contents

2.1 Sentence-Level Machine Translation 44
2.1.1 What is a neural translation model? 44
2.1.2 Datasets 46
2.1.3 Evaluation Measures 50
2.1.4 Architecture 53
2.2 Context-Aware Machine Translation 56
2.2.1 Motivations and Discussions 57
2.2.2 What is a Context-Aware Neural Machine Translation? 59
2.2.3 Evaluation Measures 60
2.2.4 Architectures 61
2.3 Integrating Coreference into Machine Translation 64

2.1 Sentence-Level Machine Translation

A second main topic in this thesis on coreference resolution into machine translation is machine translation. In section 2.1, we will detail the case of a sentence-level machine translation task. Context-Aware and Document-Level machine translation tasks will be discussed in section 2.2. While closely related, Sections 2.1 and 2.2 are substantial and specialised enough to be addressed separately.

We will begin by explaining the fundamentals of sentence-level machine translation in subsection 2.1.1, followed by an overview of the main available datasets in subsection 2.1.2. Next, we will explore different evaluation methods highlighting their strengths and weaknesses in subsection 2.1.3. Finally, subsection 2.1.4 will focus on the mainly used architecture for sentence-level translation models: the *Transformer* (Vaswani et al., 2017).

2.1.1 What is a neural translation model?

A neural translation model aims to maximise the probability $P_{\hat{\theta}}(y|x)$, where x is a sentence in the source language and y is a sentence in the target language. $\hat{\Theta}$ represents the model parameters, which are trained end-to-end.

To do this, the model incrementally predicts each token in the sentence y, based on the tokens already predicted and the source sentence. Let $\{x_1, x_2, ..., x_{|x|}\}$ and $\{y_1, y_2, ..., y_{|y|}\}$ represent the tokens of sentences x and y, respectively. Thus, $P_{\hat{\theta}}(y|x)$

can be expressed as in equation 2.1, where y_t denotes the current token to be predicted, while $y_{<t}$ and x represent respectively the target sentence tokens already predicted and the entire source sentence.

$$P_{\hat{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{|\boldsymbol{y}|} P_{\hat{\theta}}(\boldsymbol{y}_t|\boldsymbol{y}_{<t}, \boldsymbol{x}) \tag{2.1}$$

To train the model parameters $\hat{\theta}$, a parallel dataset consisting of source sentences $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^{|\boldsymbol{X}|}\}$ and their corresponding translations $\boldsymbol{Y} = \{\boldsymbol{y}^1, \boldsymbol{y}^2, ..., \boldsymbol{y}^{|\boldsymbol{Y}|}\}$ is used. This parallel corpus is denoted as $\mathcal{C} = \{\boldsymbol{X}, \boldsymbol{Y}\}$, with $|\boldsymbol{X}| = |\boldsymbol{Y}|$, where $\boldsymbol{y}^i$ represents the best proposed translation for $\boldsymbol{x}^i$, along with $1 \leq i \leq |\boldsymbol{X}|$. The ultimate objective is to maximise the log-likelihood $\mathcal{L}$ of the probability of translation $\boldsymbol{y}^i$ for the source sentence $\boldsymbol{x}^i$, which can be written as in equation 2.2.

$$\mathcal{L}(\mathcal{C}, \hat{\theta}) = \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \log P_{\hat{\theta}}(y^i|x^i) \tag{2.2}$$

A neural translation model predicts future tokens from a set called the *dictionary*, which consists of the tokens previously observed in the training corpus $\mathcal{C}$. Using characters as tokens has the main drawback of producing overly long sentences, while using words as tokens is not exhaustive and leads to the issue of *out-of-vocabulary* words. To optimise this problem, Sennrich et al. (2016) proposed BPEs (also known as *subwords*), which help minimise the entropy between the number of tokens in a sentence and the possibility of out-of-vocabulary words. The BPE mechanism works by identifying recurring groups of characters in a dataset and creating a new character that represents the group, allowing for optimal segmentation of the words in the dataset. For example, given the dictionary's words {['low'], ['lowest'], ['newer'], ['wider']}, the BPE learning mechanism could extract the following subwords: {['low'], ['est'], ['new'], ['wid'], ['er']}. This allows the model to generate words like ['lower'], ['newest'], and ['widest'], which are not present in the original dictionary.

Source sentence: "Le président a visité un pays en crise."
Translation 1: "The president visited a country in crisis."
Translation 2: "The leader toured a crisis-stricken nation."
Translation 3: "A head of state paid a visit to a troubled nation."

Example 2.1.1.

Finally, since a neural translation model generates only one word at each step, it can sometimes produce awkward or sub-optimal phrasing. In the example 2.1.1, Translation 1 may represent a typical NMT prediction, while Translations 2 and 3 introduce more diversity, thanks to the *beam search* mechanism proposed by Graves (2012) for speech and adapted by Sutskever et al. (2014) for translation. With the beam search, "*President*" can be replaced with "*leader*" or "*head of state*," and "*country in crisis*" can be swapped for "*crisis-stricken nation*" or "*troubled nation*" allowing for greater translation diversity, with quality assessed based on a more advanced view of the hypothesis.

To generate such alternatives, instead of retaining only the best hypothesis for the token at step t, the model keeps track of the top n hypotheses. It then considers the n^2 possible combinations between the hypotheses from step $t-1$ and the n proposed tokens, selecting the top n hypotheses with the highest probabilities. This process is repeated for the prediction at step $t+1$. Although beam search improves the quality of the hypotheses, it remains computationally expensive to implement, as the model must generate n hypotheses at each step and evaluate n^2 combinations to select the top n candidates.

2.1.2 Datasets

Many translation corpora have been created over the years. What makes them unique is that they are aligned with at least two languages, meaning that each sentence in one language corresponds to a sentence in another language. With a method for aligning these sentences, we can teach an NMT model to produce similar translations.

WMT: WMT (Workshop on Machine Translation) is an annual event that focuses on machine translation and related topics. Every year, they release some standardized datasets for many language pairs. The version of the dataset used

in 2017 (also referred to as WMT17) notably includes data aligned between 7 language pairs, as shown in Table 2.1 (Czech-English, German-English, Finnish-English, Latvian-English, Russian-English, Turkish-English, and Chinese-English), as well as monolingual data.

File	Size	CS-EN	DE-EN	FI-EN	LV-EN	RU-EN	TR-EN	ZH-EN
Europarl v7	628MB	✓	✓					
Europarl v8	238MB			✓	✓			
Common Crawl corpus	876MB	✓	✓			✓		
News Commentary v12	162MB	✓	✓			✓		✓
CzEng 1.6	3.1GB	✓						
Yandex Corpus	121MB					✓		
Wiki Headlines	9.1MB					✓		
SETIMES2	44 MB						✓	
UN Parallel Corpus V1.0	3.6 GB					✓		✓
Rapid corpus EU press	156 MB		✓	✓	✓			
LETA translated news	2 MB				✓			
Digital Corpus EP	123 MB				✓			
Online Books	309 kB				✓			
CWMT Corpus							✓	

Table 2.1: Table of the different language pairs included in the WMT17 corpus[1]

In this thesis, we will focus particularly on the English-German data, which includes the following corpora:

- Europarl V10: data collected from the European Parliament, organized into documents, each representing a discussed topic. Version 10, disclosed in 2023, consists of the raw data from version 7, which was released in 2012, with additional metadata included.
- ParaCrawl V9: data crawled from web pages, organized into sentences and don't have any document-level context.
- Common Crawl: data crawled from web pages, organized into sentences and don't have any document-level context.

[1]Information from https://www.statmt.org/wmt17/translation-task.html.

- News Commentary: data collected from the Project Syndicate website. It contains political and economic commentaries with each commentary represents a document.

Further information on the content present in the four corpora mentioned above is provided in Table 2.2.

Corpus	# Sentences	#Words **de**	#Words **en**
Europarl v7	1 920 209	50 957 968	53 576 202
Common Crawl corpus	2 399 123	55 956 928	53 576 202
News Commentary v12	270 769	6 981 067	59 894 034
Rapid corpus EU press	1 329 041	25 578 080	6 906 193
Total	5 919 142	139 474 043	146 585 551

Table 2.2: Information about data in en-de corpora from WMT17.

IWSLT17: IWSLT (International Conference on Spoken Language Translation)[2] is focus on spoken languages translation. This conference provide standardized training, development and testing dataset. The corpus primarily consists of subtitles from TED and TEDxTalks conferences, aligned across Arabic, German, English, French, Japanese, Chinese, and Korean. The English-German subset is divided into eight files, each comprising multiple documents, thereby facilitating context-aware training. After the removal of XML tags, the English and German data are structured as outlined in Table 2.3. Despite its smaller size compared to the WMT17 corpus, the IWSLT17 corpus is characterized by the presence of explicit document boundaries, enabling training models on document-level texts. IWSLT17 is frequently used as a benchmark in shared tasks, particularly for evaluating context-aware and discourse-level models.

ParCorFull2.0: ParCorFull2.0 (Lapshinova-Koltunski et al., 2022) is a corpus that builds upon and extends the ParCorFull corpus (Lapshinova-Koltunski et al., 2018), which itself is based on the ParCor corpus (Guillou et al., 2014). The corpus consists of transcriptions of planned speech, particularly TEDxTalks, and news texts. Although smaller in size compared to the corpora mentioned earlier, it has the specific characteristic of being both aligned across four languages (English, German, French,

[2]https://huggingface.co/datasets/IWSLT/iwslt2017

File	# Sentences	#Words **de**	#Words **en**
train	201 131	4 147 634	3 906 663
dev2010	896	20 130	19 049
tst2010	1 579	32 049	30 422
tst2011	1 449	26 929	26 266
tst2012	1 715	30 712	29 233
tst2013	1 009	20 869	19 657
tst2014	1 320	24 793	23 840
tst2015	1 092	20 850	19 740
Total	210 191	4 323 966	4 074 870

Table 2.3: Information about data in en-de corpora from IWSLT17.

and Portuguese) and annotated for coreference in each language. While its size is not sufficient for training a machine translation model from scratch, the authors describe it as being large enough for fine-tuning on the task of Machine Translation. In terms of coreference-related content, its size is around one third of the ARRAU corpus (Poesio and Artstein, 2008), which offers richer coreference annotation and covers a broader variety of coreferential relations.

The coreference annotations include coreference chains involving **pronouns**, **nouns**, **noun phrases**, **verbal phrases**, and **clauses**. The following examples illustrate the types of annotations found in the corpus:

- **Indirect pronouns : *and* et *en*:**
  - **fr**: "Jean est allé à [Paris]. Il [y] a trouvé le bonheur."
  - **en**: "Jean has gone to [Paris]. He's found happiness [there]."
- **Relative pronoun as a referring expression:**
  - **pt**: "Mas [a ideia] [de que] não devemos permitir que a ciência faça o seu trabalho porque temos medo, é de facto muito sufocante."
  - **en**: "But [the idea] [that] we should not allow science to do its job because we're afraid, is really very deadening."
  - **fr**: "Le docteur m'a dit que j'étais guéri, [ce qui] m'a surpris."
  - **en**: "The doctor told me I was cured, [which] surprised me."

- **Pronominal reduplication:**
  - **pt**: "A possibilidade de [um indivíduo] [se] ver a [si] próprio como capaz."
  - **en**: "The possibility of [an individual] (themselves) to see [themselves] as capable."

A detailed breakdown of the sentences and tokens elements contained in ParCorFull2.0 can be found in Table 2.4 meanwhile Table 2.5 give coreference information contained in ParCorFull2.0.

Language	**TED Talks**			**News**			**Total**		
	#txt	**#snt**	**#token**	**#txt**	**#snt**	**#token**	**#txt**	**#snt**	**#token**
English	20	3,277	70,736	19	464	10,798	39	3,741	81,534
German	20	2,829	66,783	19	281	10,602	39	3,110	77,385
French	20	1,959	76,229	–	–	–	20	1,959	76,229
Portuguese	9	1,488	27,898	11	309	6,522	20	1,797	34,420
Total	69	9,553	241,646	49	1,054	27,922	118	10,607	269,568

Table 2.4: Corpus statistics for text, sentence and token from TED Talks and Newswire texts in ParCorFull 2.0.

	English			**German**			**French**			**Portuguese**			**total**
	news	TED	total	news	TED	total	TED	news	total	TED	news	total	total
pron	400	3,772	4,172	477	3,840	4,317	5,140	329	5,469	1,772	329	2,101	15,730
np	434	2,206	2,640	446	2,401	2,847	3,327	410	3,737	1,501	410	1,911	10,725
vp	6	126	132	9	126	135	182	15	197	104	15	119	568
clause	12	323	335	18	317	335	360	11	371	127	11	138	1,168
all	852	6,427	7,279	950	6,684	7,634	9,009	765	9,774	3,504	765	4,269	28,191

Table 2.5: Statistics on the annotated mentions and their subcategories: pronouns (pron), nominal phrases (np), verbal phrases (vp), clauses.

2.1.3 Evaluation Measures

Unlike the task of coreference resolution, where metrics aim to verify the links established or not between mentions to evaluate how closely the model has approached the desired result, machine translation metrics will attempt to assess the model's ability to preserve the meaning of the source sentence and the model's ability to generate an

appropriate and meaningful hypothesis in the target language. These, respectively, correspond to the concepts of adequacy and fluency.

Among the various possibilities, the most relevant and reliable method is human evaluation. However, since human evaluation is time-consuming and expensive to implement, efforts are made to find a faster and more cost-effective alternative, albeit one that is more prone to criticism: automatic evaluation.

Human evaluation is considered the gold standard, and the goal is to determine whether other metrics produce assessments that closely align with human judgments.

Human based evaluation

Several types of human evaluation methods are proposed to consider a wide range of factors while minimising evaluation flaws. Thus, the following methods are suggested:

- **Direct evaluation**: a human have to rate a sentence. Usually with a Likert's scale from 1 to 5, where 1 is the worse sentence and 5 the best possible one (Bojar et al., 2016), or a continuous scale (Graham et al., 2013). Although inter-annotator agreement values are typically low, this is the evaluation method used every year at the Workshop on Machine Translation (WMT) between 2017 and 2021.

- **Relative ranking**: the objective is to compare at least two candidate translations and ask evaluators to rank the different outputs from best to worst. In addition to the potentially large number of sentences to be translated, this method allows for ranking outputs in a relative, rather than absolute, manner.

- **Multidimensional Quality Metrics (MQM)**: the goal of this approach is to assign a weight to each error made in the target sentence based on the type and severity of the error. This makes the evaluation more specific to the translation task under study. It was employed for both WMT 2021 and WMT 2022.

While human evaluation is regarded as the most reliable, it is not a guarantee of quality. It is often subjective and can be unreliable when protocols for evaluation are flawed.

Automatic evaluation

On the other hand, there is automatic evaluation, which is divided into two parts. The first part relies solely on the n-grams used, while the second part is based on contextual word embeddings in the sentence.

- **BLEU** (Papineni et al., 2001): the evaluation score computes machine translation quality by comparing the tokens between the predicted sentence to a reference sentence. Because short predictions tend to achieve high tokens matches and therefore achieve an overvalued quality, a coefficient called brevity penalty BP (or length penalty) is applied to discourage the generation of overly short translation. BP is defined as :

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}$$

Where c is the length of the predicted sentence meanwhile r is the length or the reference sentence. In order to captures local fluency, BLEU score is divided in four different precision p_n for $n = 1, ..., 4$. Each precision p_n counts n consecutive tokens matches between predicted sentences and reference sentence thus p_1 considers individual tokens, p_2 considers consecutive token pairs, p_3 involves consecutive token triplets and p_4 compares consecutive token quadruplets. The overall BLEU score is computed as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right)$$

Where $w = 1/4$ if $n = 1, ..., 4$. However, BLEU score suffer of a major drawback: it is insensitive to semantic similarity. BLEU score will assign the same score to a token that is semantically close but different from the reference token to a token that is semantically different. For example, an English translation of "*Le chat mange la souris*" is "*The cat eat the mouse*", a first translation "*The cat devour the mouse*" will be scored same as "*The cat is the mouse*" which is obviously incorrect.

- **METEOR** (Denkowski and Lavie, 2011): the METEOR score considers the precision of correct words in the predicted sentence compared to the reference sentence, as well as the recall of words in the predicted sentence compared to the source sentence. To avoid a problem similar to the one mentioned for BLEU (not considering the meaning of tokens), METEOR uses tables listing different synonyms or acronyms of words, thus giving a higher METEOR score for the translation "The cat devours the mouse" than for the translation "The cat is the mouse." METEOR score relies on these tables and can only be used for certain languages such as English, French, German, Spanish, and Czech.

There are many other token-based evaluation scores such as NIST, ChrF, WER, PER, TER, and CharacTER. However, all of these are limited to comparing the tokens in the sentence, thus not taking into account the meaning of the predicted sentence and the meaning of the reference sentence. Although token-based metrics like BLEU have long been used for automatic evaluation, the field has witnessed advancements towards learnt metrics such as BERTscore and COMET score that can better capture the semantic nuances of translations.

- **BERTScore** (Zhang et al., 2020): BERTScore relies on cosine similarity of word embeddings to consider the meaning of words, thereby mitigating potential errors stemming from synonym usage or acronyms. A score is added because some works(Banerjee and Lavie, 2005) show that rare words can be more indicative for sentence similarity than common words.

- **COMET** (Rei et al., 2020): a supervised metric trained on WMT workshop judgments to predict human judgments based on the source sentence, reference translation, and predicted sentence. It is currently emerging as the new standard after the BLEU score. Freitag et al. (2021) demonstrated that COMET has a strong correlation with human judgment.

Despite its strong performance, the use of learnt metrics has a black-box effect that deters complete abandonment of the BLEU score, despite its shortcomings. Additionally, learnt metrics highlight the issue of *universal translation* (Yan et al., 2023; Zouhar et al., 2024), where specific translations consistently receive high scores regardless of their semantic similarity to the source sentence and the target gold sentence.

2.1.4 Architecture

Now that we have discussed various methods of evaluating translations in section 2.1.3, let's turn our focus towards the task of generating these translations. While there are different approaches to automatically produce translations, such as rules-based translation models or statistical translation models, we will primarily delve into neural machine translation models quickly introduced in section 2.1.1. As mentioned, NMT models require parameters represented by θ, which need to be trained to approximate an unknown function. In our case, a function for translating from a source language to a target language. The model will aim to optimise its $\hat{\theta}$ parameters using a large set of examples with the goal of minimising the differences between its examples and its translated predictions.

The NMT architecture that we will primarily discuss in subsection 2.1.4 is the Transformer architecture, proposed by Vaswani et al. (2017). While other architectures for machine translation exist, such as recurrent neural networks like models based on long short-term memory, also called LSTM (Hochreiter and Schmidhuber, 1997), or Gated Recurrent Units, also called GRU (Cho et al. (2014)), the Transformer architecture offers the best performance in terms of translation quality Chen et al. (2018), as well as its ability to effectively handle contextual information.

The Transformer architecture is primarily built upon the attention mechanism, which we will introduce in subsection 2.1.4. Following that, we will delve deeper into the specifics of the Transformer architecture in section 2.1.4.

Attention Mechanism

The attention mechanism was introduced by Bahdanau et al. (2014). It relies on three learned matrices, denoted as $\boldsymbol{W}_q \in \mathbb{R}^{d_k \times d}$, $\boldsymbol{W}_k \in \mathbb{R}^{d_k \times d}$, and $\boldsymbol{W}_v \in \mathbb{R}^{d_v \times d}$, where d represents the embedding size, d_k is an unconstrained size, and d_v is the output dimension. Using these three matrices, the attention matrix can be computed as follows: $Q = \boldsymbol{W}_q \times x$, $K = \boldsymbol{W}_k \times x$, and $V = \boldsymbol{W}_v \times x$, with x being the input data. The dot-product attention used by Vaswani et al. (2017) is then calculated as shown in equation 2.3.

$$Attention(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}(\frac{\mathrm{Q}.\mathrm{K}^T}{\sqrt{d_k}}).\mathrm{V} \tag{2.3}$$

Where $softmax(\frac{Q.K^T}{\sqrt{d_k}})$ is a way to compute a normalised similarity between the query Q and the key K thanks to the softmax function (Bridle, 1989).

Transformer

The Transformer architecture (see Figure 2.1), introduced by Vaswani et al. (2017), is an encoder-decoder model that relies on the attention mechanism. The model uses pre-learned embeddings of dimension d_{model}, however due to the nature of the attention mechanism, it lacks the ability to encode the positional information of words in a sequence. To address this, Vaswani et al. (2017) proposed adding a positional encoding of dimension d, which is summed with the pre-trained embeddings. This positional encoding is defined by equations 2.4, where pos is the token's position in the sequence, and i is the encoding dimension ($1 \leq i \leq d$).

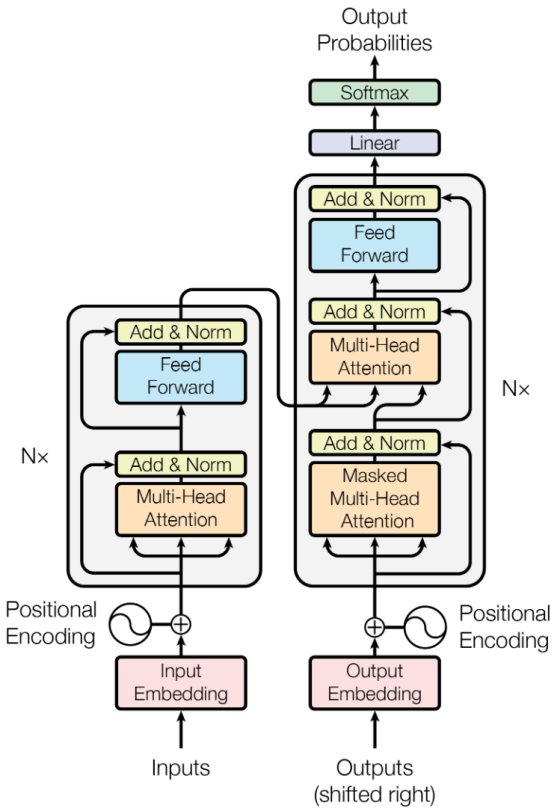


Figure 2.1: Architecture Transformer proposed by Vaswani et al. (2017). This figure comes from Vaswani et al. (2017).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.4}$$

Vaswani et al. (2017) found that the attention mechanism alone was not fully optimal and introduced multi-head attention to improve it. Multi-head attention involves concatenating several attention mechanisms, each based on different linear transformations of d_q, d_k and d_v, allowing the model to capture various aspects of the input's representations. The equation used for multi-head attention is defined in equation 2.5, where each $head_i$ represents an attention mechanism as per equation 2.3: specifically, $head_i = Attention(Q\boldsymbol{W}_Q^i, K\boldsymbol{W}_K^i, V\boldsymbol{W}_V^i)$.

$$MultiHead = \text{Concat}(head_1, head_2, ..., head_h)\boldsymbol{W}_O \tag{2.5}$$

Following this multi-head attention block, Vaswani et al. (2017) introduce a *Feed Forward Neural Network* (FFNN) layer, consisting of two linear transformations separated by a ReLU activation function (Nair and Hinton, 2010), as defined in Equation 2.6.

$$FFNN(x) = max(0, x\boldsymbol{W}^1 + b^1)\boldsymbol{W}^2 + b^2 \tag{2.6}$$

The output from the multi-head attention mechanism and the FFNN layer is summed with their input to create a residual connection.

The encoder in the model proposed by Vaswani et al. (2017) is composed of six identical blocks, as shown on the left side of Figure 2.1. The decoder (on the right side of Figure 2.1) also consists of six layers with two main modifications compared to the encoder block. The first modification is in the self-attention mechanism, where a mask is added to ensure that the model only relies on elements that have already been produced. The second modification adds cross multi-head attention: a new multi-head attention layer in which the K and V values come from the encoder output, while the Q value is from the output of the masked multi-head attention layer. Finally, a linear layer and a softmax function are applied to the decoder output to compute probabilities over the target vocabulary for predicting the next token.

2.2 Context-Aware Machine Translation

In Section 2.2, we will detail models that use context to predict the translation of a current sentence. Subsection 2.2.1 explores the motivations behind context-aware

translation models, highlighting the various ambiguities that can arise in a document. We then explain, in Subsection 2.2.2, the distinction between translation models that focus solely on the current sentence and those that incorporate contextual information. Finally, Subsection 2.2.3 addresses the challenges of evaluating context-aware models, followed by Subsection 2.2.4, where we review the main architectures of such models.

2.2.1 Motivations and Discussions

With advancements in neural models over recent years, sentence-level translation models as seen in section 2.1 have achieved a performance comparable to bilingual human translators (Hassan et al., 2018; Läubli et al., 2018; Wu et al., 2016). However, this claim holds primarily for sentence-level translation task and is no longer valid for document-level translation task (Toral et al., 2018). Indeed, there are instances where *ambiguity* arises and can only be resolved using information beyond the sentence (if resolution is possible). Ambiguity is defined as a situation or statement that is unclear because it can be understood in more than one way. This unclearness arises from various factors and multiple form. We describe here a non-exhaustive list of ambiguity:

- **Lexical ambiguity** arises when a word is polysemous. For instance, in the sentence *"She walked along the bank"*, the term *bank* can refer to either the edge of a river or a financial institution.
- **Structural (or syntactic) ambiguity** occurs when a sentence can be parsed into multiple syntactic structures. Consider the example: *"I saw the man with a telescope."*. This could imply that the speaker used a telescope to observe the man, or that the man being observed possessed a telescope.
- **Semantic ambiguity** stems from a combination of words conveying multiple meanings, even when the syntax is clear. In *"Pastor John married my sister"*, the verb *married* could mean he performed the wedding ceremony or that he became her husband.
- **Pragmatic ambiguity** is related to the speaker's intent. The question *"Can you close the window?"* can be interpreted literally as an inquiry into physical ability ("Are you able to?"), or pragmatically as a polite request to close the window.

- **Reference ambiguity** concerns the identification of the entity to which a pronoun refers. In the sentence *"Nathan says to Gaël that he has to leave"*, the pronoun *he* is ambiguous and could refer to either Nathan or Gaël.

- **Scope ambiguity** relates to the logical scope of an expression, typically involving negation or quantifiers. For example, *"I didn't eat all the cookies"* can signify that no cookies were eaten at all, or that some were eaten but not the entire batch.

In addition to the linguistic ambiguities mentioned above, the task of translation involving multiple languages introduces specific challenges inherent to the translation process between source and target languages. These can be categorized as follows:

- **Honorific ambiguity** occurs when the target language distinguishes between levels of formality that do not exist in the source. For instance, the English term *"you"* can be translated into French as *"tu"* (informal and singular) or *"vous"* (formal and/or plural). Resolving this ambiguity requires context regarding to the relationship between speakers.

- **Morphological ambiguity** poses a significant challenge when translating from a language with neutral nouns to a gendered language. For instance, translating *"the friend"* from English to French requires context to determine whether output *"l'ami"* (masculine) or *"l'amie"* (feminine) is the most suitable translation.

- **Ellipsis (or Zero Anaphora)** arises in "pro-drop" languages (such as Japanese, Chinese, or Spanish) where subjects are often omitted. When translating to a non-pro-drop language like English, the model must infer the missing pronoun from the surrounding context to generate a grammatically correct sentence.

- **Cross-lingual lexical ambiguity** refers to cases where a single word in the source language corresponds to multiple distinct words in the target language depending on the context. While similar to monolingual lexical ambiguity, the translation model is forced to make a concrete choice (e.g., translating *"river"* to French *"fleuve"* or *"rivière"* depending on whether it flows into the sea).

Although these cases may be rare within a document, they remain essential to maintaining the coherence and cohesion of a translation. The concept of **coherence** refers to the harmony within a document, the absence of contradiction in the flow of its parts, while **cohesion** pertains to the structural aspect of the text, the way its

elements are interconnected (Charolles, 1995). During translation, the model must decide how to address ambiguities, and providing a broader and more comprehensive context generally facilitates this resolution:

- **Target-side context** helps maintain coherence with elements already translated. It is crucial for *lexical consistency* which is the act of ensuring that specific terminologies remain uniform (e.g., consistently translating "device" as "appareil" rather than switching between "appareil" and "dispositif") and for *anaphora resolution* (linking pronouns to previously mentioned entities).

- **Source-side context** adds another layer of information, enriching the understanding of nuances. It is particularly indispensable for *cataphora resolution*, allowing the model to access information that appears later in the source text but has not yet been generated in the target translation.

However, there are still instances where resolving ambiguity remains impossible (e.g.: a completely isolated ambiguous sentence but same issue can occur with more context). In such cases, the model is forced to make a decision that may seem arbitrary, yet is influenced by its inherent biases. These biases often defaulting to the most frequent translation observed in the training data, which can lead to stereotypical errors. In addition, expanded context windows may introduce spurious signals (noise), resulting in a degradation of the model's overall performance (Lopez et al., 2023).

2.2.2 What is a Context-Aware Neural Machine Translation?

In the case of context-aware neural translation models, multiple sentences are provided as input. We denote the set of source sentences as $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^{|X|}\}$ and the set of target sentences as $\boldsymbol{Y} = \{\boldsymbol{y}^1, \boldsymbol{y}^2, ..., \boldsymbol{y}^{|Y|}\}$. The equation 2.1 is then modified to reflect this context:

$$P_{\hat{\theta}}(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{i=1}^{|\boldsymbol{X}|} \prod_{t=1}^{|\boldsymbol{y}|} P_{\hat{\theta}}(y_t^i | y_{<t}^i, \boldsymbol{Y}_{<j}, \boldsymbol{X}) \tag{2.7}$$

We distinguish between **document-level models** and **context-aware models** based on their input scope. A document-level model processes the entire document at once, while context-aware models only have access to a limited number of sentences from the document at each step. This latter approach is often preferred because the

elements necessary to disambiguate a sentence are typically found in close proximity to the current sentence. Furthermore, the elements needed for disambiguation are often infrequent within a document, and adding excessive contextual information can introduce noise into the translation.

Moreover, the various context-aware translation models can be categorized into three main families: *Jumping K-to-K*, *Sliding K-to-K*, and *Sliding K-to-1*, where K represents the number of context sentences.

Jumping K-to-K (Junczys-Dowmunt, 2019) involves dividing the document into segments of K sentences and translating these K sentences as a single block before concatenating the resulting blocks. However, this approach only attempts to sidestep the issue without effectively addressing it. In fact, it does not resolve the problem, as the first sentence will only have access to subsequent sentences, while the last sentence will only be able to refer to preceding ones.

The Sliding K-to-1 solution proposed by Tiedemann and Scherrer (2017) involves providing the model with K context sentences in addition to the current sentence, allowing it to make predictions solely based on the current sentence. In contrast, the Sliding K-to-K approach (Tiedemann and Scherrer, 2017) enables the model to predict all K context sentences alongside the current one, but ultimately retains only the current sentence. While Sliding K-to-K facilitates a better consideration of the context in the target language, it requires generating multiple translations for the same sentence without retaining these outputs for the final prediction.

These new approaches to think translation by incorporating context sentences have several implications. In particular, 2 key questions arise regarding how to evaluate the ability of context-aware models to leverage context to enhance translation quality. We will discuss this in detail in subsection 2.2.3, along with the new architectures proposed, which will be addressed in subsection 2.2.4.

2.2.3 Evaluation Measures

Although the various improvements in the metrics presented earlier, a major drawback still persists among them: all of them compare translation at the sentence level. However, it has been demonstrated that in the field of machine translation, context plays a crucial role in addressing certain linguistic phenomena, such as pronoun translation or polysemous nouns. This kind of issue is encountered in Example 2.2.1 where the gold translation is the best among the ones offered, while the possible translation using the word “pierre” (cobble) and the article “la” (feminine one) remains understandable though suboptimal. In contrast, the false translation presented is clearly incorrect due to gender disagreement.

Gold source sentence: "*Adèle picks up [a pebble]. She throws [it] into the pond.*"
Gold target sentence: "*Adèle ramasse [un galet]. Elle [le] jette dans la mare.*"
Possible translation: "*Adèle ramasse [une pierre]. Elle [la] jette dans la mare.*"
Wrong translation: "*Adèle ramasse [un galet]. Elle [la] jette dans la mare.*"

Example 2.2.1. Example of cohesion issue in a translated sentence.

As a result, a branch of the automatic machine translation evaluation task has thus introduced certain solutions called 'test suites' with the aim of assessing the model's ability to utilize context for generating better translations. Some of them (Guillou and Hardmeier, 2016; Wong et al., 2020) target the evaluation of anaphoric pronoun translation. However, they suffer from the limitation of comparing the model's translation to a reference translation. As seen in Example 4, a translation can be correct but suboptimal. Other test suites (Bawden et al., 2018; Lopes et al., 2020; Muller et al., 2018) aim to measure the model's ability to assign a higher probability to a correct translation compared to an incorrect one, both of which have already been produced. This second approach offers more flexibility in evaluation but doesn't directly assess the model's ability to generate the translation. Typically, it needs to be complemented with a previously mentioned metric, like BLEU score, to be meaningful. While there are numerous automatic or manual means of evaluating translation, these methods primarily focus on sentence-level evaluation. Document-level translation evaluation methods are more oriented towards the model's ability to ensure that context is used correctly without necessarily guaranteeing the model's capabilities. Although Miculicich Werlen and Popescu-Belis (2017), Jwalapuram et al. (2019) and Wong and Kit (2012) proposed metrics based on aligning source sentences with predicted sentences, the errors introduced by the tools used make them poorly correlated with human judgment (Guillou and Hardmeier, 2018; Wong and Kit, 2012).

2.2.4 Architectures

Now that we have covered the new methods for evaluating the performance of context-aware models, we will explore in this section the main approaches proposed

for contextualizing the current sentence.

To incorporate context into the input sentence, approaches generally fall into two main categories (Kim et al., 2019): the concatenation method (Kim et al., 2019; Lupo et al., 2022b) and the multi-encoder method (Lupo et al., 2022a; Miculicich et al., 2018). The first approach focuses primarily on modifying the input sentences, while the second modifies the model's encoder-decoder architecture.

The first approach, the concatenation method, involves processing multiple sentences by adding specific tokens to indicate sentence boundaries to the model. This approach maintains an objective similar to that defined in subsection 2.1.1. For example, in a sliding K-to-1 model, equation 2.1 becomes:

$$P_{\hat{\theta}}(\boldsymbol{y}|\boldsymbol{x}^K) = \prod_{t=1}^{|\boldsymbol{y}|} P_{\hat{\theta}}(\boldsymbol{y}_t|\boldsymbol{y}_{<t}, \boldsymbol{x}^K) \tag{2.8}$$

With K representing the set of context sentences provided as input to the model along with the current sentence.

While a concatenation approach might initially appear suitable, Lupo et al. (2022b) emphasize the need for caution when training context-aware models and propose certain modifications:

The first modification addresses the observation that elements requiring contextualization for accurate translation are relatively rare. Consequently, training a model based on a scoring function similar to equation 2.2 may be sub-optimal. A proposed solution to address this issue involves adjusting the loss function by introducing a coefficient, termed the context-discount CD, to lessen the model's focus on contextual elements. The modified loss function is then given by equation 2.9.

$$\mathcal{L}(\mathcal{C}, \hat{\theta}) = CD * \mathcal{L}_{context}(\mathcal{C}, \hat{\theta}) + \mathcal{L}_{current}(\mathcal{C}, \hat{\theta}) \tag{2.9}$$

Another proposal suggests that a sentence x^j within a context x^K, usable for a model such as the one described in Equation 2.8, would take the form $x^K = x^{j-K+1} < s > x^{j-K+2} < s > ... < s > x^j < end >$. For instance, in a sentence like "Hey there! $< s >$ How are you? $< end >$", the token "How" from the next phrase ends up as close to the token "!" as to the token "Hey," despite the latter being part of a different question. To address this, Lupo et al. (2022b) propose modifying the positional embeddings by artificially increasing the distance between sentences. This corresponds to introducing a jump in the values of the positional embeddings.

Thus, it is suggested that concatenation-based models, while relatively simple to implement, should not be used in the same straightforward manner as sentence-level

models. Various techniques to refine the use of context are essential to leverage it effectively without compromising the quality of the full translation.

The second category of context-aware models relies on a multi-encoder architecture. This approach introduces a module designed to incorporate the meaning of the context into the hidden states of the encoder outside (Outside Context Integration) or inside (Inside Context Integration) the decoder, as depicted in the left and right sections of Figure 2.2 (Li et al., 2020).

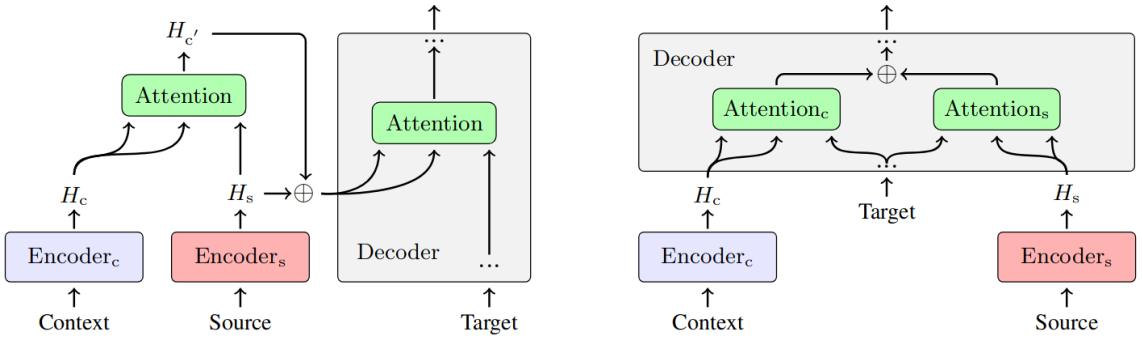


Figure 2.2: Outside and inside context integration representation from Li et al. (2020).

In the outside context integration method (Lupo et al., 2022a; Maruf et al., 2018, 2019; Miculicich et al., 2018; Voita et al., 2018; Zhang et al., 2018; Zheng et al., 2020), the hidden state of the current sentence H_s serves as the query, while the hidden state of the context sentence H_c is used as the key and value to produce a contextually enriched hidden state of the context sentence $H_{c'}$. This enriched state is then gated with the current sentence's hidden state H_s and used into the decoder, as explained in Section 2.1.4.

A popular approach that we will use in this work is the method proposed by Miculicich et al. (2018), where they propose an hierarchical attention model based on attention mecanism (see Figure 2.3 for an encoder example) to contextualize the current sentence t using k context sentences, in contrast to Yang et al. (2016) who introduced it with LSTM-based models. The hidden state h_t of the current sentence t is used to compute q_w and q_s.

q_w serves as the query for word-level attention mechanisms, where the values/keys are the hidden states h_j^i of the context sentences, allowing the computation of a vector s_j. Here, $i = \{1, ...|x_j|\}$ and $j = \{t-k, ..., t-1\}$, with k being the number of context sentence considered.

A sentence-level attention mechanism is then applied, using q_s as the query and the vectors s_j as values/keys. The resulting output is merged with h_t using a gate function (Tu et al., 2018), thus producing a contextualized representation $\tilde{h}_t$.

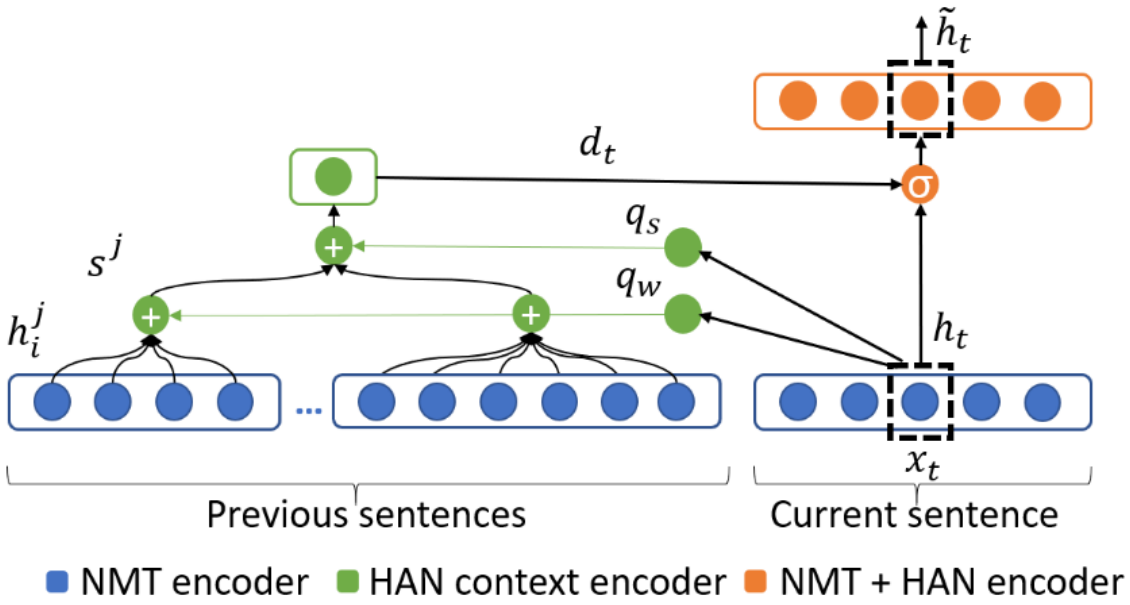


Figure 2.3: Hierarchical attention network encoder architecture from Miculicich et al. (2018)

The inside context integration method (Bawden et al., 2018; Kuang et al., 2018; Tan et al., 2019; Tu et al., 2018; Voita et al., 2019), on the other hand, separately applies twice the attention mechanism as described in Section 2.1.4 by using the embedding of the target sentence as the query and the hidden states H_c and H_s of the context and current sentences as keys/values. The resulting information is then gated before being used.

2.3 Integrating Coreference into Machine Translation

As discussed in Section 2.2, the challenges associated with context-aware translation are more subtle than those encountered in sentence-level translation. The primary issue lies in the sparsity of ambiguous terms: few tokens within a sentence are ambiguous, and resolving this ambiguity typically requires access to a limited number

of cues, often located outside the current sentence.

Coreferent or anaphoric elements serve as the most prominent examples of this phenomenon. Consequently, few studies have attempted to incorporate these elements into translation models (Le Nagard and Koehn, 2010; Miculicich and Popescu-Belis, 2017).

One approach proposed by Miculicich and Popescu-Belis (2017) is premised on the assumption that the higher the similarity between the coreference links in the source and target sentences, the greater the translation quality. Their method involves resolving alignments between mentions in the source and target sentences. A similarity score is then calculated based on the previously generated document content and the n-best hypotheses for the current sentence. Finally, this score is utilized to re-rank the translation hypotheses.

Chapter 3

Explanation

As established in the previous chapters, while context-aware neural machine translation models have achieved remarkable predictive performance, they operate largely as black boxes. This opacity creates a significant challenge: a model may produce a linguistically correct translation by relying on spurious correlations rather than a genuine understanding of discourse-level dependencies. In the context of coreference resolution, this leads to a fundamental question: does the model resolve a pronoun by correctly identifying its contextual antecedent, or is it merely producing a guess based on frequency biases?

This chapter explores the theoretical and methodological landscape of explainable AI as it applies to natural language processing. The goal is to move beyond the evaluation of model outputs to the auditing of model internals, seeking to determine if a system achieve the right translation for the right reasons.

The chapter is divided into two main sections. Section 3.1.1 introduces the core concepts of explainable AI, distinguishing between predictive performance and model faithfulness.

Finally, Section 3.1.2 addresses the central methodological debate of this thesis: the role of the attention mechanism as an explanatory tool. We navigate the "Attention is not (not) Explanation" (Wiegreffe and Pinter, 2019a) controversy, reviewing the empirical critiques and defences that have shaped the field (Méloux et al., 2025). By the end of this chapter, we establish why, despite the ongoing debate, attention weights remain a privileged tracer of information flow for studying coreference in context-aware NMT, providing the necessary justification for the experimental framework presented in the second part of this thesis.

Contents

3.1 Introduction 67
3.1.1 Explainability: methods and models 69
3.1.2 Attention is not (not) Explanation 74

3.1 Introduction

As observed in Section 2.2, the model requires additional context to achieve optimal translation, particularly when dealing with anaphoric or coreferential out-of-sentence elements. This fact raises two key questions:

- *(Q1)*: How does the model resolve out-of-sentence coreference links?
- *(Q2)*: Does the model predict the right translation for the right reasons?

To elaborate on the second question (Q2), evaluating a translation model solely on the basis of its outputs can provide information about the quality of the translation but does not guarantee that the model achieves this quality by taking the right elements into account. This discrepancy highlights a fundamental distinction in machine learning: the difference between predictive power and model faithfulness. Predictive power measures how well an explanation can serve as a proxy to predict model output, whereas faithfulness evaluates the alignment between a provided explanation and the model's actual logic (Jacovi and Goldberg, 2020).

A well-known illustration of this issue comes from Ribeiro et al. (2016), involving an image recognition model designed to classify photos of huskies and wolves. While the model achieved high accuracy, an investigation using explanation methods revealed that it relied primarily on the pixels in the background (the presence of snow) rather than the anatomical features of the animals. This "shortcut" occurred because, in the training set, wolves were mainly photographed in snowy wild environments, whereas huskies were shown in domestic settings. This phenomenon demonstrates how models can exploit spurious correlations in data rather than learn the causal aspects of the task.

Such errors are particularly prevalent and subtle in NLP, especially concerning context-aware tasks. Consider example 3.1.1 from ParCorFull2.0 dataset translated from English to French. While the translation appears correct, the English pronoun "[them]$_4$" could refer to either "[every woman]$_1$" which appear first in the text, "[Men]$_2$" or either "[woman]$_3$". If the model correctly generates the token "*les*", it might be doing so because it successfully resolved the coreference link or simply

because, in this case, gender does not affect the prediction. Without explainability, there is no guarantee that the model is not relying on a linguistic shortcut rather than a proper analysis of the discourse context.

English version: "I'm not exaggerating when I say that [every woman]$_1$ who arrives here has dealt with sexual violence on her way to find refuge," said Bahr. "[Men]$_2$ flee because of war, because they are politically persecuted, because they are threatened with torture or death, or because their families pin their hopes on them and send them to Europe," she explains. When [women]$_3$ flee, other reasons are in the foreground: "Many women sneak away secretly, because they see no other way out". In their home countries, they experienced sexual violence, forced marriage, honour killings, slavery or forced prostitution. That propels [them]$_4$ forward.
French version: "Je n'exagère pas quand je dis que [toutes les femmes]$_1$ qui arrivent ici ont été victimes de violences sexuelles avant de trouver refuge", explique Bahr. "[Les hommes]$_2$ fuient à cause de la guerre, parce qu'ils sont persécutés politiquement, parce qu'ils sont menacés de torture ou de mort, ou parce que leurs familles placent leurs espoirs en eux et les envoient en Europe", explique-t-elle. Lorsque [les femmes]$_3$ fuient, d'autres raisons sont au premier plan : "Beaucoup de femmes s'enfuient en secret, car elles ne voient pas d'autre issue". Dans leur pays d'origine, elles ont été victimes de violences sexuelles, de mariages forcés, de crimes d'honneur, d'esclavage ou de prostitution forcée. C'est ce qui [les]$_4$ pousse à partir.

Example 3.1.1. Example of misleading ambiguity for the right antecedent of "them" translated into "les". Multiple reasons give the right answer. Example come from the Parcorfull2.0 dataset

To address these concerns, contrastive test suites (Bawden et al., 2018, Muller et al., 2018, Lopes et al., 2020) were introduced, as described in Section 2.2.3. These tools measure a model's ability to assign a higher probability to a linguistically correct variant over an incorrect one (e.g., matching a pronoun's gender with its antecedent in the context). However, as noted by Belinkov and Glass (2019), these techniques only evaluate the outcome of the decision process. They act as a "black-box" evaluation: they confirm *that* a model can distinguish the right answer but fail to explain *how* the

model leveraged the context to reach that conclusion. Moreover, they do not evaluate the model's capacity to generate the right answer. To understand if a model is truly "context-aware", we must delve into the mechanics of its internal representations.

The need for explanation became critical as NLP architectures evolved from statistical machine translation to deep learning models. Statistical machine translation models were inherently more transparent: translation was based on explicit phrase tables and alignment parameters that could be directly inspected. On the other hand, modern architectures like LSTMs or Transformers rely on high-dimensional latent spaces and millions of non-linear parameters. While these models have significantly improved translation quality, they have done so at the cost of interpretability, functioning as "black-boxes" where the decision logic is distributed across complex weight matrices.

In the literature, explainability methods can be categorised according to two dimensions:

- **Ante-hoc vs. Post-hoc:** *ante-hoc* interpretability refers to models that are transparent by design (like short decision trees or models built with this specific aspect in mind). Since modern NMT models are too complex for this, research focuses mainly on *post-hoc* explanation methods applied to a trained model to extract insights into its behaviour.
- **Local vs. Global:** *local* explanations aim to justify a single prediction, while *global* explanations seek to understand the overall logic of the model.

Since *ante-hoc* methods are relatively uncommon today, we will focus more specifically on *post-hoc* methods. These can be categorised into the main families described in Subsection 3.1.1.

3.1.1 Explainability: methods and models

In this section, we will discuss the principles behind the main methods implemented for AI explainability (also noted XIA). We will begin by discussing feature attribution methods, then move on to probing methods before discussing sparse autoencoder models. Next, we will briefly discuss concept-based explanation methods, causal approaches, and circuit methods before concluding with natural language explanation methods.

Feature Attribution Methods Feature attribution methods aim to quantify the contribution of each input feature (e.g.: individual tokens, words, or sentence segments) to a model's specific output through the assignment of importance scores.

These methods are generally categorised by their underlying paradigms: those relying on gradient computation and those based on systematic perturbations of the input data.

Gradient-based methods operate on the principle that the partial derivative of the output with respect to the input is linked to the model's sensitivity to that specific feature. These methods are called *model-specific* methods, as they require internal access to the architecture and weights of deep neural networks.

- **Saliency Maps**: introduced by Simonyan et al. (2014) primarily for computer vision, this technique visualises the magnitude of the raw gradient at a local point to highlight salient features. However, saliency maps suffer from the "gradient saturation" problem, where the gradient becomes near-zero in regions where the model's prediction is confident, despite the feature's importance. Furthermore, Adebayo et al. (2018) demonstrated that saliency maps often fail "sanity checks," appearing visually plausible even when model weights are randomised.

- **Integrated Gradients**: to address the gradient saturation issue, Sundararajan et al. (2017) proposed Integrated Gradients. This method computes the integral of the gradients along a straight-line path from a baseline input to the actual input. Integrated gradient methods are designed to satisfy three fundamental axioms:

  - *Sensitivity*: If an input and a baseline differ in one feature and result in different predictions, that feature must receive a non-zero attribution.
  - *Completeness*: The sum of all attribution scores must exactly equal the difference between the prediction for the actual input and the prediction for the baseline.
  - *Implementation Invariance*: Functionally equivalent models (producing identical outputs for all inputs) must have identical attributions, regardless of their internal implementation.

Conversely, *model-agnostic* methods treat the network as a "black box," observing how variations in the output occur because of systematic occlusions or modifications of the input. A prominent example is SHAP (Lundberg and Lee, 2017) or LIME (Ribeiro et al., 2016).

Despite their widespread adoption, feature attribution methods face significant criticism regarding their reliability and robustness. As noted by Méloux et al. (2025), AI interpretability is increasingly susceptible to "false positives". Ghorbani et al.

(2019) demonstrated that gradient-based attributions are highly fragile to minor changes in the input, which can radically alter the resulting explanation without changing the model's prediction. These findings raise the critical issue of *identifiability*: if multiple conflicting explanations can be generated for the same prediction with equal fidelity, determining which explanation truly reflects the model's internal reasoning remains an open challenge.

Probing methods Probing methods aim to identify what linguistic or conceptual information is encoded within the latent representations of a pre-trained model. The core principle involves training a simple supervised model (e.g.: a linear or structural classifier referred to as a "probe") on the hidden states of the network, while keeping the original model's weights frozen (Conneau et al., 2018, Hewitt and Manning, 2019). If the classifier can predict a specific property (e.g.: number, gender or even part-of-speech tagging) with high accuracy, it is inferred that the information is encoded within the model's latent space. However, interpreting these results requires caution: high probing performance does not necessarily imply that the original model actively utilizes this information to perform its final task, highlighting the ongoing debate between information encoding and causal utility (Ravichander et al., 2021, Méloux et al., 2025).

Sparse AutoEncoders In standard neural architectures, individual neurons are frequently polysemantic, meaning they activate in response to a diverse set of unrelated features (Christopher et al., 2020). This phenomenon, known as *superposition*, arises when a model learns a greater number of distinct features than the available dimensions in a given layer (Elhage et al., 2022). To address this, Sharkey et al. (2022) utilize sparse autoencoders to learn compact and structured latent representations of data, orthogonal directions for feature representation by imposing a *sparsity constraint* (i.e.: a mechanism that restricts the number of active neurons for any specific input). In practice, this is achieved through the inclusion of an L1 regularization term in the loss function, forcing the model to represent information using only a small subset of its neurons. This process encourages the emergence of distinct, "monosemantic" features, thereby enhancing the interpretability of individual units (Bricken et al., 2023, Cunningham et al., 2023). However, recent work has shown that sparse autoencoders can extract seemingly interpretable components even from randomly initialized Transformers (Heap et al., 2025). Furthermore, Li et al. (2026) demonstrated that sparse autoencoders are highly vulnerable to adversarial input perturbations, suggesting that their robustness as reliable explanatory tools remains an open question.

Concept-based Explanation Concept-based methods (Bau et al., 2017, Kim et al., 2018) aim to introduce high-level concepts (i.e.: theoretically more human-interpretable) in order to align them with a model's internal representations. Rather than explaining a prediction through raw features such as pixels, tokens, or latent dimensions, these methods aim to link the model's decision-making process to explicit semantic notions, such as shapes, patterns, linguistic properties, or abstract attributes defined in advance. Various approaches have been proposed to handle these concepts. Notably, Kim et al. (2018) introduced Concept Activation Vectors (CAV), which extract the hidden states of a neural model (as a CNN, RNN, or Transformer) from data samples containing or lacking a specific target concept. A linear classifier is then trained on these states to identify the directional vector in the latent space that best represents the concept. While this approach seeks to identify the concepts leveraged by the model during inference, it faces several limitations. Specifically, assigning human-readable names to identified directions can be challenging. Furthermore, as demonstrated for BERT (Devlin et al., 2019), certain neurone activations may appear to encode a specific concept in a spurious manner, leading to what Bolukbasi et al. (2021) terms an "interpretability illusion."

Subsequent research has sought to adapt these methodologies to Natural Language Processing. For instance, LACOAT (Yu et al., 2024) aims to generate clusters of embeddings representing concepts, following the methodology proposed by Dalvi et al. (2022). In this framework, word embeddings from a corpus are projected into an embedding space and aggregated into clusters using agglomerative hierarchical clustering (Gowda and Krishna, 1978). However, several methodological concerns have been raised. As highlighted by Méloux et al. (2025), LACOAT effectively identifies concepts correlated with the model's predictions but fails to demonstrate that these concepts play a necessary or causal role in the model's underlying reasoning process.

Causal Approaches While traditional attribution methods rely primarily on statistical correlations, causal approaches aim to establish direct causal links between a model's internal components and its final predictions. Grounded in the theoretical framework of causal inference (Pearl, 2012), these techniques utilize systematic interventions, often referred to as *interchange interventions* (Geiger et al., 2021). The core principle involves modifying or replacing the activations of a specific component (i.e.: a neuron, an attention head, or a full layer) with those derived from a different input to observe the resulting effect on the model's output. A prominent methodology in this area is *causal mediation analysis*, which quantifies the signal flow through specific pathways within the network to explain particular behaviors, such as gender

bias (Vig et al., 2020) or factual knowledge retrieval (Meng et al., 2023). However, these approaches face the challenge of *causal overdetermination*: due to the inherent redundancy in Transformer architectures called *overdetermination* or the *Hydra effect* (McGrath et al., 2023). As alternative circuits may compensate for the loss, the lack of a measurable effect following a component's ablation does not necessarily imply its irrelevance. This lack of identifiability emphasizes the importance of integrating causal analysis with rigorous uncertainty quantification (Méloux et al., 2025).

Circuits As part of causal approach methods, Mechanistic interpretability seeks to "reverse-engineer" neural networks by treating them as complex programmes composed of human understandable computational circuits (Olah et al., 2020) (i.e.: subnetwork of the main neural network). This approach aims to uncover the precise internal algorithms used by the model to transform those features into predictions. A landmark success in this field was the discovery of induction heads Elhage et al. (2021, 2022), which are specific attention heads capable of implementing a "copy-and-paste" logic. A mechanism highly relevant to coreference resolution and context-aware tasks. In the context of Natural Language Processing (NLP), mechanistic interpretability has been used to identify circuits dedicated to indirect object identification (Wang et al., 2022) or factual knowledge retrieval. However, as noted by Méloux et al. (2025), this "bottom-up" approach faces significant scalability and identifiability challenges. Identifying a single, unique circuit for a linguistic behaviour is often complicated by the model's internal redundancy. Moreover, Sutter et al. (2025) demonstrate that, in general, current causal abstraction methods may produce mechanistic-style explanations even when applied to randomly initialised networks. This work highlights potential limitations of these methods as indicators of meaningful internal structure.

Natural Language Explanation A more recent and human-centric paradigm in explainability involves generating Natural Language Explanations, where a model (usually a Large Language Model) is tasked with producing a textual reasoning to justify its own predictions (Marasovic et al., 2022; Wiegreffe et al., 2022, 2021). This often takes the form of self-rationalization, where the model externalizes its "reasoning" steps before providing a final answer (Wei et al., 2023). While these explanations significantly enhance the *plausibility* of model outputs for human end-users, they are frequently criticized for their lack of *faithfulness* (Lanham et al., 2023; Turpin et al., 2023). Natural language explanation are particularly susceptible to the "dead salmon" artifact of AI interpretability (Méloux et al., 2025): because the hypothesis space of natural language is virtually unconstrained, a model can easily construct a

logically coherent story to justify a prediction made for entirely different—or even spurious—reasons. This phenomenon, referred to as *post-hoc confabulation*, suggests that the generated rationale may be a reflection of the model's linguistic capabilities rather than its actual decision-making process.

3.1.2 Attention is not (not) Explanation

In this section, we will go deeper into the explainability power of the attention mechanism. Since the introduction of the Transformer architecture, attention weights have been widely used as a primary tool for interpreting model behaviour, particularly in tasks involving distant dependencies such as in-context translation. However, the validity of attention as a faithful explanation of a model's decision-making process has been the subject of debate in the scientific community, primarily sparked by the conflicting findings of Jain and Wallace (2019) and Wiegreffe and Pinter (2019b).

The critique formulated by Jain and Wallace (2019) is based on two main empirical observations. First, they found that attention weights often exhibit low correlation with other established feature importance measures, such as gradient-based saliency maps. This suggests that the elements the model "seems to look at" are not necessarily the same as those that most significantly influence the output. Second, they demonstrated the existence of "counterfactual" attention distributions. In particular, Jain and Wallace (2019) showed a way to find a completely different set of attention weights that would result in the same model prediction. Based on this, they concluded that attention weights are not unique and therefore cannot be considered a faithful explanation of the model's internal reasoning.

However, the existence of a potential alternative explanation does not invalidate the one actually used by the model (Wiegreffe and Pinter, 2019b). They noted that attention is an integral component of the model's forward pass, not an external post-hoc explanation. As such, it represents the actual routing of information through the network. Furthermore, they conducted adversarial training experiments and found that models forced to use "meaningless" attention distributions usually suffered from a decrease in performance when used with new data, suggesting that the original attention patterns do indeed capture necessary linguistic signals.

In the context of this thesis, this debate invites a cautious but productive use of attention analysis. While Serrano and Smith (2019) also found that attention does not always provide a clear causal link to predictions, they acknowledged that it remains a valuable indicator of the model's intent regarding information gathering. For coreference resolution in context-aware NMT, we argue that attention weights are particularly relevant not as a measure of absolute feature importance, but as a

way to evaluate how much the model is able to use the context sentences in the right way. When a model translates a gender-ambiguous pronoun in the current sentence by attending to a specific noun in the context, it indicates that the model is actively seeking to incorporate contextual information into the current representation. Even if this attention is not the sole cause of the final prediction, it serves as a trace of the mechanism used to overcome the sparsity of information within the isolated sentence. Consequently, the analysis of attention patterns presented in Chapter 5 aims to characterise how different NMT model variants (multi-encoder vs. concatenation) handle the flow of contextual cues while acknowledging the limitations raised by Jain and Wallace (2019) through the use of additional cleaning and normalisation techniques. The review of explainability methods discussed in Section 3.1.1 highlights a paradox inherent to language models: while their predictive performance increases, their faithfulness becomes increasingly difficult to audit. Within the scope of this thesis, this challenge is compounded by the nature of the phenomenon under study: context-aware coreference resolution.

As we highlighted with question (Q2) (*"Does the model predict the right translation for the right reasons?"*), it is not enough for a model to simply produce the correct pronoun. It is imperative to determine whether it has derived this information from the correct antecedent within the context. While methods such as Saliency Maps or Integrated Gradients are powerful, they suffer from significant noise and instability, which complicates the fine-grained interpretation of precise linguistic links like coreference. Although the scientific literature debates its value as a universal explanation, the attention mechanism remains the only native component of the model that has explicit access to the information in the context sentences necessary for the disambiguation of ambiguous tokens in the current sentence. By choosing to focus on the model's attention weights, we do not seek a singular, intrinsic, and actionable explainability, but rather to determine whether the model resolves coreference by observing "where it looks" to resolve it. Ultimately, this approach aims to transform the alignment mechanism that is attention into a metric akin to the score produced by test suites.

Part II

Contributions

Building on the theoretical and technical foundations established in Part I, this second part presents the main research and findings of this thesis. Having reviewed the topic of automatic coreference link resolution, the evolution of context-aware neural machine translation, and the challenges of model explainability, we now focus on a diagnostic analysis of model behaviours.

The central motivation of these contributions is to move beyond the assessment of translation quality alone based on either metrics such as BLEU (Papineni et al., 2001) or any learnt metrics. Then, to audit the internal mechanisms of neural machine translation systems. As established in the "right reasons" question, we seek to determine if models that produce correct context-dependent translations effectively leverage referential links. This motivation is condensed in the main research question of this work:

"In a Transformer-type translation model, is it possible to find evidence of coreference resolution links to the context within the attention mechanism to provide an explanation for the translation?"

To answer this question, Part II is organised into three chapters:

First, Chapter 4, we define our experimental framework. We introduce a processing pipeline and a set of quantitative metrics designed to transform raw, noisy attention matrices into more interpretable signals of contextual retrieval. We also detail the behaviour of two representative families of context-aware architectures: concatenation-based models and multi-encoder hierarchical models.

Second, Chapter 5 presents the results of three complementary experiments. We perform a large-scale quantitative study, followed by a fine-grained qualitative analysis of 74 expert-annotated sentences, and finally a quantitative head-level investigation. This analysis reveals how different architectures "look" at the context, highlighting a contrast between the specialised "coreference heads" found in concatenation models and the more diffuse integration strategies of multi-encoder systems.

Finally, Chapter 6 summarises the contributions of this thesis, discusses the limitations of our diagnostic approach, and presents future research avenues.

Contents

4 Methodology **79**
4.1 The Attention Mechanism: Why Focus Here? 80
4.2 Experimental Setup 81
4.2.1 Models and Experiences 81
4.2.2 Post-processing 85
4.2.3 Discussions 89

5 Experiments **92**
5.1 Attention please! 93
5.2 Heads up: Corefs Matters 95
5.3 Cerberus or which head is giving direction? 98

6 Conclusions & perspectives **102**
6.1 Summary of contributions 102
6.2 Perspectives 103

Chapter 4

Methodology

In Part I, we provided a description of the Natural Language Processing (NLP) task of machine translation, detailing the Transformer model architectures (Vaswani et al., 2017) and the Hierarchical Attention Network (HAN) variant (Miculicich et al., 2018). We further elaborated on the functioning of automatic coreference resolution and concluded with the question: "Does a translation model predict the correct translation for the right reasons in ambiguous sentences?" As this question is too broad and opens multiple research avenues, we chose to focus on a more specific aspect:

"In a Transformer-type translation model, is it possible to find evidence of coreference resolution links to the context within the attention mechanism to provide an explanation for the translation?"

Contents

4.1 The Attention Mechanism: Why Focus Here? 80
4.2 Experimental Setup 81
4.2.1 Models and Experiences 81
4.2.2 Post-processing 85
4.2.3 Discussions 89

4.1 The Attention Mechanism: Why Focus Here?

Transformer models, as described in Section 2.1.4, access contextual information exclusively through the attention mechanism. To answer our research question, it is therefore pertinent to observe the behavior of this mechanism with respect to coreferent elements in the current sentence and their relationship to context sentences. Moreover, with the attention mechanism acting as a similarity measure on one hand, and elements referring to the same entity on the other, it is intuitive that the attention mechanism should be able to capture this similarity. Clark et al. (2019) demonstrated that BERT's attention mechanism (Devlin et al., 2019) is capable of at least partially capturing coreference phenomena, while Voita et al. (2018) showed similar results on a variant of the Transformer model.

Other methods for analyzing attention exist, such as perturbation techniques or gradient-based explainability methods. However, for this study, we chose to focus on directly examining attention weights because they provide an intuitive and directly exploitable interpretation of the flow of contextual information within the model.

The objective of this thesis is to investigate whether similar behavior emerges in more commonly used context-aware Transformer variants, specifically a concatenation model and a multi-encoder model for CA-NMT.

To achieve this objective, three experiments described below were proposed. For each experiment, a notation was assigned to the involved sentences. The sentences are in the form "<*k3*><*k2*><*k1*><*k0*>" with *k0* symbolizes the current sentence, *k1* represents the sentence preceding *k0*. Similarly, *k2* is the notation for the sentence preceding *k1*, and finally, *k3* corresponds to the sentence preceding *k2*. It should be noted that *k1*, *k2*, and *k3* do not necessarily exist for a sentence *k0*, for example, at the beginning of a document.

Experiment 1: "Quantitative Study of Attention Weights" (**Attention please!**) will consist of a quantitative study on the average attention weights output by the encoder's attention mechanism between the current sentence *k0* and the existing context sentences *k1*, *k2*, and *k3*. This first experiment will be combined with a second

experiment: "Qualitative Study on a Subsample" (**Heads up: Corefs Matters**), consisting of a qualitative study on a subsample of the initial data selected during an internship. Finally, Experiment 3: "Detailed Analysis of Attention Heads" (**Cerberus or which head is giving direction ?**) will delve deeper by detailing the behavior of each available attention head to observe if one or more of them specialize in resolving coreference links between tokens of the current sentence *k0* and tokens of the context sentences *k1*, *k2*, and *k3*.

4.2 Experimental Setup

4.2.1 Models and Experiences

To address coreference in documents written in the source language, working solely with a coreference corpus might have been sufficient. However, with the aim of initially focusing on the encoder part and subsequently addressing the decoder part of translation models, this implies either:

- Using a coreference resolution model on a text in a source language and on the predicted translation of that text, then aligning the coreference predictions between both source and target side.
- Using a coreference-annotated corpus aligned across the languages being processed.

Using a coreference resolution model in the first solution would automate the process so that it could be adapted to different corpora. However, this carries a risk of errors if no verification is performed after automatic annotation. Conversely, the second solution implies that the coreference annotation is correct because it is produced by experts. However, in practice, this also means that the corpus size will potentially be smaller due to a heavier workload.

As detailed in Section 2.1.2, such a corpus exists, albeit a relatively small one. Indeed, the ParCorFull2.0 corpus (Lapshinova-Koltunski et al., 2022) is coreference-annotated and sentence-aligned in English, German, French, and Portuguese. Although the work presented in this thesis is solely on the encoder side, using such a corpus from the outset allows for better continuity for future work mentioned in 6. After cleaning the ParCoFull2.0 corpus for the English-German pair, a set of 2,280 usable sentences, whose information is summarized in Table 4.1, was produced. Given that the corpus size is too large for a qualitative analysis at our scale, and coreference phenomena are relatively rare in a document (Lupo et al., 2022a), a sub-corpus of

100 examples distributed over 74 sentences was selected (Dinarelli et al., 2024) and will be studied in Experiment **Heads up: Corefs Matters**. A summary of the details of this sub-corpus is provided in Table 4.2.

Language	*Sentences*	*Tokens*	*Mentions*	*Coref. chains*
English	2,280	42,798	4,206	425
German	2,280	40,261	3,377	306

Table 4.1: Statistics on the English-German data from the ParCorFull2.0 corpus used for our quantitative analyses.

Language	*Sentences*	*Tokens*	*Mentions*	*Coref. chains*
English	74	557	135	73
German	74	605	132	73

Table 4.2: Statistics on the English-German data from the ParCorFull2.0 sub-corpus used for our qualitative analyses.

Once the corpus is selected, the choice of models is crucial. These models must perform sufficiently well on the measures discussed in Sections 2.1.3 and 2.2.3. This means achieving satisfactory BLEU scores (the primary measure used for comparing translation models despite its flaws) as well as good scores on the test suites described in Section 2.2.3. To this end, we focused on the models from Lupo et al. (2022a) and Lupo et al. (2022b), specifically:

- **A multi-encoder model** by Lupo et al. (2022a) trained with the *divide & rule* (*d&r*) method, which is intended to improve the model's ability to find information in the context when an ambiguous term is present in the current sentence. To achieve this, each sentence is split, and each part is provided to the multi-encoder model as a separate sentence. This mechanism forces the model to retrieve the information necessary for translation from parts of the current sentence considered as context sentences, whose information is only accessible via the context encoders. In the following experiments, as indicated in Figure 4.1, we will use the term *word-level attention mechanism* to refer to the attention mechanisms between *k0* and each context sentence *k3*, *k2*, and *k1* independently. The term *sentence-level attention mechanism* will be used for the higher-level attention between *k0* and word-level attention mechanisms.

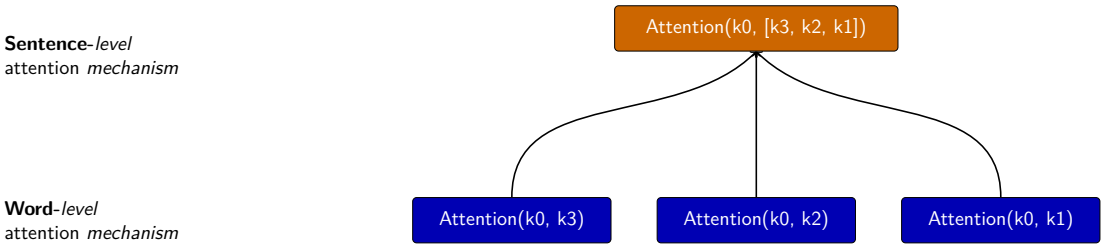


Figure 4.1: Simplified illustration of the attention mechanisms in the HAN model within the multi-encoder architecture proposed by Miculicich et al. (2018).

  - For Experiment **Attention please!**, this involves retrieving the average of the multi-head attention weight matrices from the HAN component between the current sentence and each context sentence independently (i.e., the word-level attention mechanism). The word-level HAN component calculates attention from the current sentence *k0* to *k1*, then attention from *k0* to *k2*, and finally attention from *k0* to *k3*, denoted respectively as *attention(k0, k1)*, *attention(k0, k2)*, and *attention(k0, k3)*.
  - For Experiment **Heads up: Corefs Matters**, this involves defining annotation rules to specify the desired behavior to be annotated, as well as establishing an agreement journal for specific cases under conditions similar to Experiment **Attention please!**, but with a reduced dataset size of 73 sentences.
  - For Experiment **Cerberus or which head is giving direction ?**, since the multi-encoder model only accesses context through word-level and sentence-level attention mechanisms, this implies contextualizing the attention weights of each word-level attention mechanism, i.e., *attention(k0, k3)*, *attention(k0, k2)*, and *attention(k0, k1)*. The sentence-level mechanism assigns a weight between each token of the current sentence *k0* and each word-level attention mechanism (this can be seen as a weight between the token of the current sentence and the context sentence). We decided to use this as a multiplicative coefficient on the word-level mechanism weights to interpret it as allowing contextualization of the word-level attention mechanisms.

- **A concatenation model** by Lupo et al. (2022b) trained with the *shift embeddings* and *focused-concat* methods. These methods are designed to improve the model's ability to distinguish context from the current sentence by increasing

the spacing between sentences at the positional embedding level described in Section 2.1.4.

- For Experiment **Attention please!**, this involves segmenting the average of the multi-head self-attention matrices from the last encoder layer to retain only the elements between the current sentence and each context sentence.
- For Experiment **Heads up: Corefs Matters**, similarly to the multi-encoder model, this involves defining annotation rules to specify the behavior we wish to annotate, as well as establishing an agreement journal for specific cases.
- For Experiment **Cerberus or which head is giving direction ?**, this involves segmenting the self-attention matrices of each layer to retain only the elements between the current sentence and all context sentences.

Both the multi-encoder and concatenation models were chosen due to their performance on ContraPro scores (see Table 4.3), BLEU (see Table 4.4), and COMET (see Table 4.4). These scores indicate their certain capabilities, including for the Machine Translation task.

NMT model	*ContraPro Accuracy*
Baseline	45.00
Zhang et al. (2018)	42.60
Tu et al. (2018) *	45.20
Muller et al. (2018) concat21	48.00
Muller et al. (2018) concat22*	70.80
Maruf et al. (2019) *	39.15
Voita et al. (2018)	42.55
Stojanovski and Fraser (2019)	52.55
Muller et al. (2018) * best	58.13
Multi-encoder	61.09
Concat *	74.39

Table 4.3: Quantitative results in terms of accuracy on the ContraPro test suite, obtained with the CA-NMT models. We show a comparison to a baseline context-agnostic model, and the best models from the literature. Models marked with * use both source and target context.

NMT model / Metric	*BLEU*	*COMET*	*ChrF*	*TER*
Multi-encoder	32.17	0.83	59.04	56.53
Concat *	32.08	0.81	58.62	57.38

Table 4.4: Quantitative results in terms of BLEU[a], COMET[b], ChrF and TER scores, obtained with the multi-encoder and concatenation models on the ParCorFull2 corpus. * means the model uses both source and target context.

a) Using sacrebleu (Post, 2018=, signature: nrefs:1—case:mixed—eff:no—tok:13a—smooth:exp—version:2.3.1.

b) Using model wmt22-comet-da: https://huggingface.co/Unbabel/wmt22-comet-da.

4.2.2 Post-processing

Once the models are selected, it is crucial to establish a method for extracting and analyzing data from their attention mechanisms. To this end, a processing pipeline has been implemented as depicted in Figure 4.2. The elements provided by the corpus are in orange (gold sentence and coreference links between mentions contained in sentences), the processing of these elements is in green, and finally, the analysis of the processed data is in blue. As indicated, sentences from the ParCorFull2.0 corpus are fed as input to the translation models, from which multi-head attention weights are extracted from the encoder output (see example on Figure 4.3a) and at each decoder step during prediction. The obtained data has token-level granularity, with the sum of weights from current sentence tokens to context sentence tokens equaling 1. This data is cleaned to avoid model-specific formatting. Thus, padding tokens are removed if necessary, and BPEs are merged (see example on Figure 4.3b). BPE merging is performed by retaining the maximum of the weight values for tokens belonging to the same word. Previous work, such as Clark et al. (2019), preferred to merge BPE weights using the average. We believe that retaining the maximum of the weight values for tokens of the same word is more relevant because the model processes each element at the token level. Keeping only the average weights could result in a token where the model would place significant attention weight not being taken into account in the calculation of metrics. Conversely, keeping only the sum of the weights could skew the results for long words. Since the model only sees words broken down into tokens, we believe that using the max function is the most appropriate approach. For example, if for a 4 token length word $m = [t_0, t_1, t_2, t_3]$, the model may assigns a very high weight to token t_2 (this BPE being considered highly relevant by the model), but low weights to tokens $[t_0, t_1, t_3$. Then using the average weight risks lowering the attention value between token t_2 and then word m to an imperceptible level. Meanwhile, with the same word, the use of sum function may

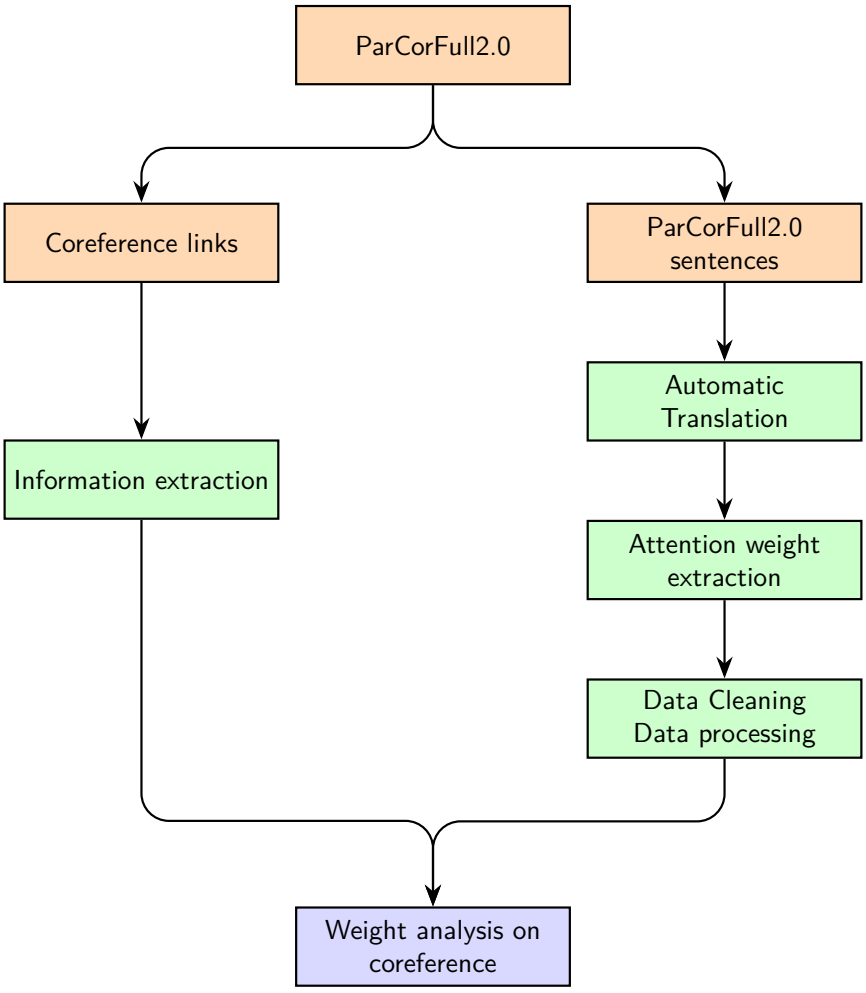
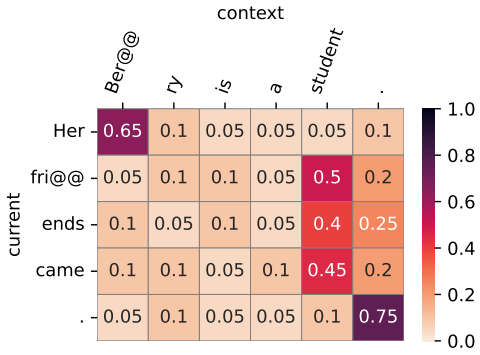


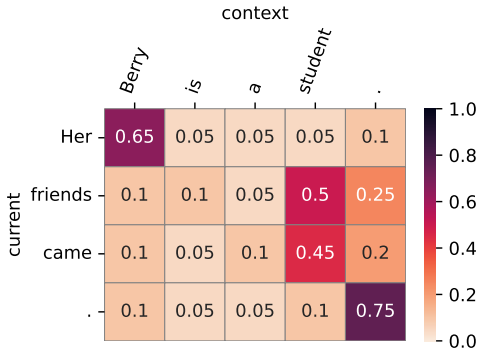
Figure 4.2: Processing pipeline for extracting and analyzing attention weights.

overestimate the attention of the word m. Once the matrices are standardized for each model, the weights are processed to be more interpretable. A "null" attention is assigned to weights below the value $u = \frac{1}{\text{number of tokens in the sentence}}$. The intuition being if the model's attention mechanisms are well-learned, then each weight greater than zero indicates a certain similarity from the current sentence token to the context sentence token (see example on Figure 4.3c). Finally, since the attention mechanism cannot assign an attention weight of 0, the longer the sentence, the closer the weight values will be to 0. To mitigate this behavior, a min-max normalization $norm$ is

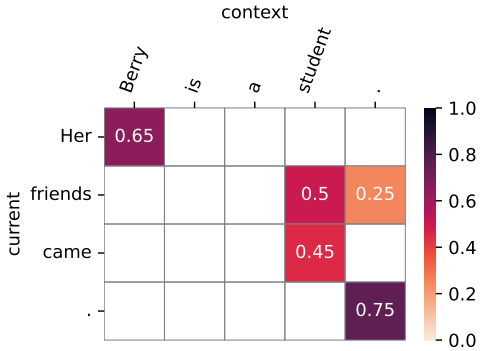
applied. This projects the elements between *min* and *max* into $[0; 1]$ (see example on Figure 4.3d). Information related to coreference links is extracted from the corpus data to be aligned between English and German (the languages used).



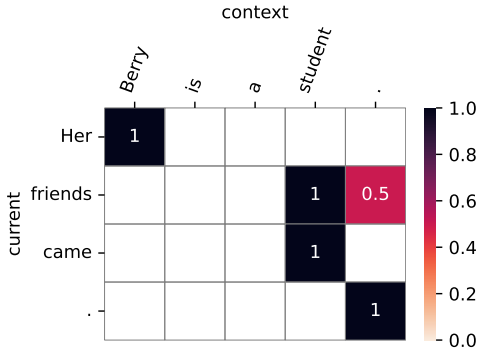
(a) Example of an attention matrix from scratch.



(b) Example of an attention matrix with BPEs merged.



(c) Example of an attention matrix with non-significant values deleted.



(d) Example of an attention matrix with normalized values.

Figure 4.3: Examples of modifications applied to attention matrices for improved readability.

The modifications applied to the attention weights allow for the implementation of different metrics used in Experiment **Attention please!** and reused in Experiment **Cerberus or which head is giving direction ?**. Each metric aims to quantify an aspect of the models' ability to resolve coreference through their attention weights. The proposed metrics are *max-weight*, *non-zero weight* and *mean value*.Each of these

metrics seeks to capture a different element of attention weight behavior.

- *Max weight* corresponds to the number of times the model placed a maximum weight from a token of a mention in the current sentence to a token of an antecedent in the context sentence. This indicates that the similarity value is maximal between elements referring to the same entity.
- *Non-zero weight* corresponds to the number of times the model placed a weight greater than u (i.e., the value of a uniform distribution). This means the model evaluated that the subword of a mention in the current sentence has a positive similarity with a subword referring to the same entity.
- *Mean value* corresponds to the average value attributed after normalization by the model between the current token and the coreferent token.

For the qualitative analysis in Experiment **Heads up: Corefs Matters**, an annotation system was established to specify, for a mention x in the current sentence with at least one antecedent a in a context sentence, whether certain tokens in x have a significant attention weight towards one of the tokens in a. To do this, several pieces of information has been requested from the annotators:

- Information necessary for recognizing the coreference under study. That is:
  - the sentence ID,
  - the ID of the model being studied,
  - the token from the current sentence,
  - the coreferent token from the context sentence.
- Information necessary for understanding the result. That is:
  - the token with the maximum weight,
  - whether the token with the maximum weight is indeed the coreferent token (for cases where the same token appears multiple times in the context sentence but only one of them is coreferent with the token from the current sentence),
  - whether the coreference is "naive" (if the mention in the current sentence is the same as in the context sentence),
  - the value of the attention weight assigned by the attention mechanism from the current token to the coreferent token in the context sentence,

  - whether context was necessary to disambiguate the token in the current sentence.

- Additional information to better understand potential counter-intuitive behaviors. That is:
  - whether the sentence is difficult, meaning if the annotator had trouble understanding it,
  - whether the coreference is difficult, meaning if the coreferent tokens are identical or not.

4.2.3 Discussions

For each experiment, due to the construction of the various models by Lupo et al. (2022a) and Lupo et al. (2022b), different results are intuitively expected in Experiment **Attention please!** and **Cerberus or which head is giving direction ?** if the attention mechanism is capable of capturing coreference:

- For Experiment **Attention please!**:
  - The concatenation model, observing both the current sentence and the context sentences as a single flattened sequence, may have high attention value from a current sentence token to a context sentence coreferent token as the attention mechanism is an alignement mechanism. However, we would expect most of the attention to be on the current sentence mainly because of the *focused-concat* behavior (Lupo et al., 2022b) which trained by focusing on the current sentence.
  - The multi-encoder model, however, computes attention between the current sentence and one context sentence at a time , which may lead to a more accurate attention as there is fewer token to look at. Then because the attention mecanism is an alignment function, the multi-encoder model can be expected to also have a high *non-zero weight* and *max-weight* metric. Since the model processes only one context sentence at a time, there may be less risk of ambiguity in the case of a sentence containing a coreferent element. Conversely, in the case of a potential but non-coreferent antecedent in a sentence, it might still receive a high attention weight (such cases will be highlighted in Experiment **Heads up: Corefs Matters**).

- For Experiment **Cerberus or which head is giving direction ?**:

– The concatenation model uses 6 layers and 8 attention heads per layer (as part of the self-attention module). It is reasonable to expect that at least one of the attention heads will specialize in resolving the coreference phenomenon between the current sentence and the context sentences (potentially including the current sentence itself). This would lead to an increase in the *max-weight* and *non-zero weight* metrics.

– The multi-encoder model, on the other hand, uses only 2 layers in the HAN module. We expect that each head should maintain equivalent *max-weight* and *non-zero weight* metrics score. A decrease of either *non-zero weight* or *max-weight* metrics would imply that the sentence-level mechanism placed a high attention weight on an incorrect context sentence. Since each sentence is analyzed individually, an increase in these metrics for the multi-encoder model is not possible because it would mean that the coefficient applied using the sentence-level attention mechanism weight is not applied in the same way to the entire context sentence.

To facilitate the interpretation of attention matrices, they will be represented as heatmaps. Each row represents the attention of a token/word from the current sentence to a token/word from the context sentence placed in a column. To further aid readability, a color scale has been implemented. Thus, darker colors indicate higher attention weight values, with a value of 1 represented by a black cell, while a value of 0 will be represented by a white color. To maintain a way to differentiate close values, the weight values remain indicated in the cells.

Finally, by analyzing the attention matrices extracted from the multi-encoder and concatenation models, a specific behavior emerged. An example of this behavior is shown in the heatmap in Figure 4.4.

In this example, different highlighted weights are visible, as indicated in Section 4.2.2. In particular, the pronoun *it* has significant attention weights on both *malaria* tokens, but also all tokens of the current sentence have significant attention weights on the tokens “*,*”, “*.*”, and “*<eos>*”. This second behavior is not uncommon in the studied dataset and mainly concerns punctuation tokens such as “*,*”, “*.*”, “*?*”, “*...*”, but also function words which are abundant in the corpus but provide very little linguistic information to the sentence. This behavior has already been observed and described by Darcet et al. (2024). Indeed, Darcet et al. (2024) suggest in the image recognition task that visual transformers use certain pixels of the image that do not contain information about the object to be identified as buffers called ”registers”. The behavior highlighted in this paragraph seems similar in the context of the machine translation task due to the elements involved and the models used (Trans-

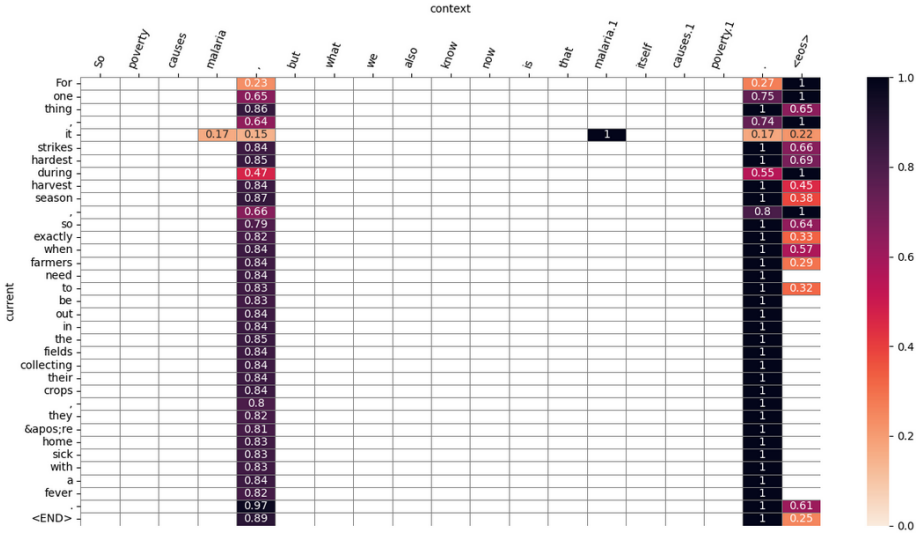


Figure 4.4: Example of the issue related to registers.

former in both cases). The additional suppression of attention weights pertaining to the concerned tokens leads to the evolution of the example from Figure 4.4 to Figure 4.5.

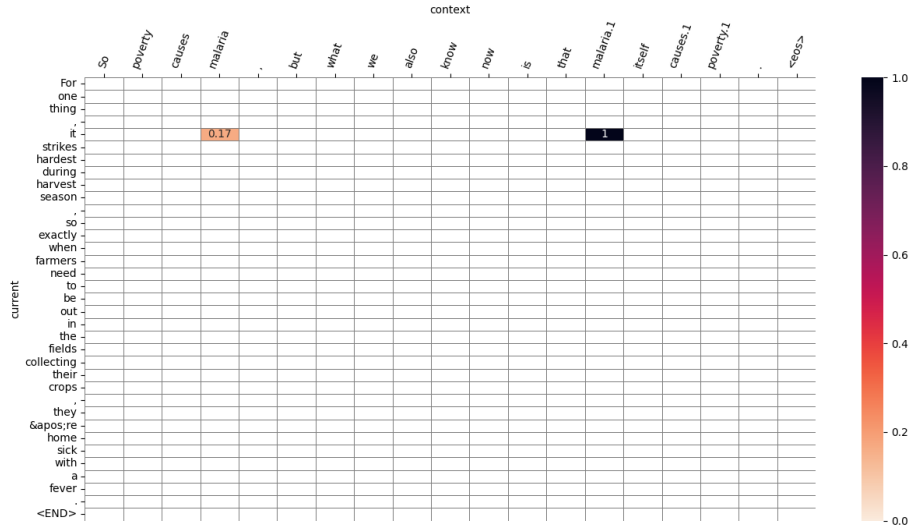


Figure 4.5: Corrected example of the issue related to registers.

Chapter 5

Experiments

The Chapter 4 details the methodologies for the three experiments whose results will be described in this chapter. First, the "Attention Please!" experiment will be presented, where its results will be elaborated upon. Subsequently, we will detail and discuss the outcomes of the "Heads up: Corefs Matters" experiment, then addressing the "Cerberus or which head is giving direction" experiment. Both Experiments "Attention Please!" and "Heads up: Corefs Matters" were published in:

Marco Dinarelli et al. (2024). "Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention". In: *Traitement Automatique des Langues* 64.3. Ed. by Guillaume Wisniewski, pp. 67–91. DOI: 10.57896/2023-tal-64_3_3. URL: https://aclanthology.org/2023.tal-3.4/.

Contents

5.1 Attention please! 93
5.2 Heads up: Corefs Matters 95
5.3 Cerberus or which head is giving direction? 98

5.1 Attention please!

In this section, we will delve into Experiment 1, "Attention Please!", which aims to quantitatively study the distribution of attention weights between the current sentence and context sentences for both the concatenation model and the multi-encoder model. Specifically, we will focus on the average of the attention heads of the last layer for the concatenation model and the average of the first-level HAN attention heads for the multi-encoder model, as discussed in Section 4.

Metric / NMT model	Multi-encoder (src)	Concat (src)	Concat (tgt)
Max-weight	45.91%	10.45%	13.25%
Non-zero weight	88.83%	50.98%	33.22%
Average weight	0.8183	0.2994	0.2136

Table 5.1: Automatic results for different models on the 2,280 sentences forming the sub-corpus of ParCorFull2.0 proposed in Section 4

The results obtained using the three metrics: *max-weight*, *non-zero weight*, and *average weight*, are listed in Table 5.1 for the corpus of 2,280 sentences from the English-German data of ParCorFull2. Similarly, Table 5.2 lists these three metrics applied to the sub-corpus of 74 sentences extracted from the 2,280 English-German sentences of ParCorFull2.

Metric / NMT model	Multi-encoder (src)	Concat (src)	Concat (tgt)
Max-weight	49.31%	9.49%	10.71%
Non-zero weight	92.36%	51.82%	29.46%
Average weight	0.8574	0.3039	0.2117

Table 5.2: Automatic results for different models on the 74 sentences forming the sub-corpus of ParCorFull2.0 proposed in Section 4

We first observe that both tables exhibit similar trends. The **max-weight** value is particularly high for the *multi-encoder (src)* model, at 45.91% and 49.31% respectively, for the 2,280-sentence corpus and the 74-sentence sample extracted from *ParCorFull2.0*. In these instances, the maximum attention weight is attributed to a coreferent element. In comparison, the *Concat (src)* and *Concat (tgt)* models show significantly lower values. Specifically, *Concat (src)* achieves only 10.45% and 9.49% on the same corpora, while *Concat (tgt)* obtains slightly higher values of 13.25% and 10.71%. This aligns with the expectations discussed in Subsection 4.2.3, although it is interesting to note the slight performance increase in the decoder part of the *Concat (tgt)* model compared to the encoder part of the same model (*Concat (src)*). The **non-zero weight** value remains higher for the *multi-encoder (src)* model (88.83% and 92.36% respectively) than for the *Concat (src)* (50.98% and 51.82%) and *Concat (tgt)* (33.22% and 29.46%) models. However, it can be observed that while this value slightly increases on the source side in the concatenation model, it decreases on the target side as examples become more complex. The fact that the max-weight metric is slightly better while the non-zero weight metric is half as high tends to highlight that the decoder is much more accurate, potentially because the benefit of disambiguating ambiguous tokens appears mainly on the decoder side.

The **average weight** value slightly increases for the *multi-encoder (src)* model (0.8183 and 0.8574 respectively), whereas for the *Concat (src)* (0.2994 and 0.3039) and *Concat (tgt)* (0.2136 and 0.2117) models, this value remains relatively stable.

Thus, it appears that the *multi-encoder (src)* model exhibits an attention weight distribution that is not only more concentrated on coreferent elements but also more robust to sentence complexity. This observation tends to confirm that the multi-encoder architecture favors better utilization of coreferent elements within the context, while the concatenation model struggles to effectively capture these coreference relationships, particularly on the target side. This conclusion differs from the results obtained using ContraPro Accuracy, as indicated in Table 4.3, but can be explained by the presence of a decoder that considers the target side for the concatenation model. Further research should be conducted on a concatenation model not utilizing the target side or on a multi-encoder model utilizing the target side.

As a reminder, the objective of this "Attention, Please!" experiment was to demonstrate the relevance of using attention as a complementary evaluation tool for how context sentences are utilized by context-aware translation models, which appears to be the case in this experiment. To further elaborate on this experiment, analyzing these same attention matrices through a qualitative study, in order to observe in more detail the behaviors detected by the proposed metrics, seems pertinent. This constitutes the second part of this section, found in Subsection 5.2.

5.2 Heads up: Corefs Matters

In this section, we will discuss Experiment 2, "Heads up: Corefs Matters," which aims to qualitatively study the distribution of attention weights between the current sentence and context sentences for both the concatenation model and the multi-encoder model, as introduced in Section 4.

It is important to note that our qualitative analysis distinguishes between cases where disambiguation is necessary for correct sentence translation and cases where it is not. In Example 5.2.1, we can observe that the German (*de*) translation of the English (*en*) sentence is the same in *Case 1* and *Case 2*, whereas the French (*fr*) translation differs in both cases. Indeed, the pronouns "They" and "Sie" are gender-neutral in the English-German language pair, thus the translation of "They" will not require disambiguation. Conversely, in the English-French language pair, the translation will necessitate disambiguation by resolving the coreference link between "They" and its respective antecedents to be translated as "Ils" (masculine plural) or "Elles" (feminine plural).

- Case 1:
  - **en**: "Pierre-Marie and Nathan were on a trip. They went to Grenoble."
  - **de**: "Pierre-Marie und Nathan waren auf einer Reise. Sie sind nach Grenoble gefahren."
  - **fr**: "Pierre-Marie et Nathan sont partis en voyage. Ils sont allés à Grenoble."
- Case 2:
  - **en**: "Noéma and Louise were on a trip. They went to Grenoble."
  - **de**: "Noéma und Louise waren auf einer Reise. Sie sind nach Grenoble gefahren."
  - **fr**: "Noéma et Louise sont partis en voyage. Elles sont allées à Grenoble."

Example 5.2.1. Example of sentences without issue when translated from English to German but an issue may occur when translated to French.

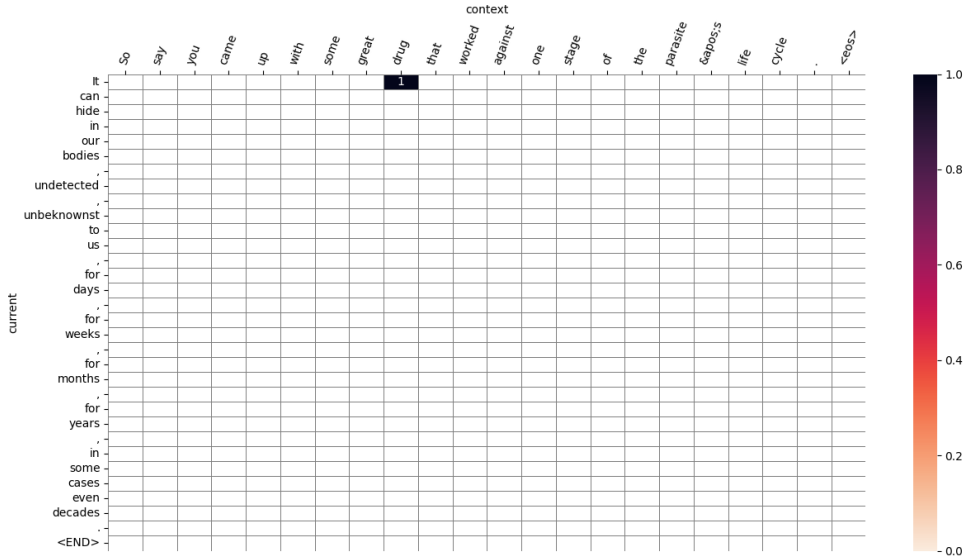


Figure 5.1: An example of heatmap from the concatenation model encoder from sentence *221* to *k2* (sentence *219*).

To facilitate the understanding of images representing attention weight distributions as heatmaps, we reiterate that these heatmaps should be read line by line, with the current sentence tokens on the leftmost column, while each column corresponds to the attention placed on a context sentence token. The specific context sentence (i.e., *k1*, *k2*, or *k3*) is detailed in the image caption. We also recall that the heatmaps represent matrices after the post-processing indicated in Section 4.2.2. Thus, Figure 5.1 illustrates the average attention of the last encoder layer of the concatenation model by Lupo et al. (2022b) between current sentence number *221* in the corpus and context sentence *k2* (context sentence at distance 2, i.e., sentence number *219*). It can be observed that the token *it* has a maximum attention weight on the token *drug* in the context sentence. No other token in the current sentence shows attention to this specific context sentence.

The results of the "Heads up: Corefs Matters" experiment, whose methodology is described in Section 4, are detailed in Table 5.3.

The table presents the category **All cases**, encompassing all annotated mentions, which in turn includes the categories **Ctx needed**, **Ctx needed & hard coref**, and **Positive attention**. **Ctx needed** refers to mentions where context is necessary for disambiguation. **Ctx needed & hard coref** indicates cases where, in addition to requiring context for disambiguation, the coreference was also deemed difficult,

	All cases		Ctx needed		Ctx needed & hard coref		Positive attention	
	Model							
	multi-enc	**concat**	**multi-enc**	**concat**	**multi-enc**	**concat**	**multi-enc**	**concat**
# of mentions	224	224	160 (-64)	160 (-64)	119 (-41)	119 (-41)	217	118
Naive links (%)	12.9%	14.3%	3.1% (-9.8)	5.5% (-8.8)	2.5% (-0.6)	7.6% (+2.1)	13.4%	27.1%
Max-weight	58.5%	41.5%	56.4% (-2.1)	40.5% (-1)	50.4% (-6)	34.5% (-6)	60.4%	78.8%
Non-zero weight	82.6%	43.8%	84% (+1.4)	42.9% (-0.9)	80.7% (-3.3)	37% (-5.9)	84.3%	83.1%
Average-weight	0.884	0.464	0.891 (+0.007)	0.466 (+0.002)	0.88 (-0.011)	0.494 (+0.028)	0.884	0.464
Dispersion	6.43	3.42	6.63	1.11	7.49	1.06	7.53	1.02

Table 5.3: Human (manual) evaluation statistics on the 74 selected examples for the multi-encoder (multi-enc) and concatenation (concat) models.

meaning the tokens of the current mention differ from the tokens of the potential antecedent. Finally, **Positive attention** represents all instances where the model applies a significant attention weight to one or more tokens in the context sentence, regardless of whether it refers to a correct antecedent. The table provides statistics obtained for the multi-encoder and concatenation models, as described in Chapter 4, using the three metrics **Max-weight**, **Non-zero weight**, and **Average-weight**, supplemented by **Naive links** and **Dispersion**. **Naive links** indicates the percentage of occurrences where the model applies the attention weight from the current sentence token to the identical token in the context sentence, while **Dispersion** represents the average number of significant attention weights (i.e., above a uniform distribution value) from current sentence tokens to all context sentence tokens, meaning the average number of context tokens receiving significant attention.

Observing the results in Table 5.3, it emerges that in the **All cases** category, the multi-encoder model places significant attention (**Non-zero weight**) on the correct antecedent in 82.6% of cases, compared to 43.8% for the concatenation model. However, the multi-encoder model applies a maximum weight to an antecedent element in 58.5% of cases, versus 41.5% for the concatenation model. Thus, when the multi-encoder model applies attention, it applies maximum attention to an antecedent in 70.82% of cases ($\frac{\textbf{Max-weight}}{\textbf{Non-zero weight}}$), whereas the concatenation model applies maximum attention in 94.75% of cases when it applies attention. This result is interesting. Indeed, the concatenation-based model proves to be more accurate than the multi-encoder model. This may be explained by the fact that the concatenation-based model processes a larger amount of context simultaneously, thereby benefiting from a more comprehensive view of the global sentence context. Another possible hypothesis is that the HAN module of the multi-encoder model, which consists of only two attention layers, must encode information required for an accurate translation of the current sentence beyond coreferential elements.

In cases where context is necessary (**Ctx needed**, -64 mentions), the behavior of both models remains relatively similar. Both the multi-encoder and concatenation models show a slight decrease in the **Max-weight** metric, dropping to 56.4% (-2.1%) and 40.5% (-1%) respectively. The **Non-zero weight** metric experiences a slight increase for the multi-encoder model, rising to 84% (+1.4). This result suggests that the models implicitly handle coreference, whether or not such resolution is required.

Compared to cases where context is necessary, in cases where context is necessary but coreference is also deemed difficult (**Ctx needed & hard coref**, -41 mentions), the metric values decrease. The **Max-weight** metric values drop to 50.4% (-6%) for the multi-encoder model and to 34.5% (-6%) for the concatenation model. The **Non-zero weight** metric shows a smaller decrease to 80.7% (-3.3%) for the multi-encoder model and a decrease to 37% (-5.9%) for the concatenation model. This result highlights that, while the models are able to implicitly resolve coreference links between a token in the current sentence and a coreferent token in a context sentence, they face increased difficulty when the coreference is noted as difficult by a human annotator.

The results obtained during Experiment "Heads up: Corefs Matters" through the qualitative analysis on the 74-sentence sub-corpus of ParCorFull2.0 follow the same trend as those obtained during Experiment "Attention, Please!" through quantitative analysis, whether performed on the entire ParCorFull2.0 corpus or only on the 74-sentence sub-corpus. Furthermore, these results are consistent with our expectations when designing the automatic metrics. **Max-weight** indeed proves to be the most restrictive metric, while **Non-zero weight**, being more flexible, tends to increase along with the **Dispersion** metric. However, the multi-encoder model exhibits a higher **max-weight** value than the concatenation model, which contradicts the results obtained with the ContraPro test suite. A potential explanation could be that the integration of context using the sentence-level attention mechanism is not optimal, or that the concatenation model primarily utilizes target-side context.

Having established that the automatic metrics **Max-weight**, **Non-zero weight**, and **Average weight** behave similarly to expert annotators, we decided to use them in Experiment "Cerberus or which head is giving direction ?" to analyze in greater depth all attention heads of each layer for these same multi-encoder and concatenation models. This analysis is presented in Section 5.3.

5.3 Cerberus or which head is giving direction?

In this section, we present Experiment 3, "Cerberus, or which head is giving direction?", which aims to quantitatively study the distribution of attention weights

between the current sentence and the context sentences for both the concatenation-based and multi-encoder models. Specifically, this involves analysing each attention head across every layer for the concatenation model and studying the combination of matrices between each sentence-level head of the contextualised HAN module with each context-level head of the same HAN module for the multi-encoder model, as described in Section 4.

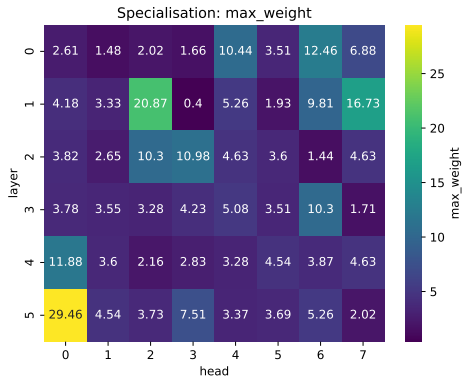
The results obtained focus on the *max-weight* and *non-zero weight* metrics across the 2,308 sentences extracted from the Parcorfull2.0 corpus, detailed in Chapter 4 and used in Experiment 5.1. Unfortunately, technical issues in the implementation prevented us from obtaining the *average weight* metric for the multi-encoder model. Although this metric remains relevant, it is the most sensitive to preprocessing variations (e.g., changes in BPE merging strategies, etc.). Here, we will only discuss the *max-weight* and *non-zero weight* metrics for both the concatenation and multi-encoder models. However, the results for the *average weight* metric on the concatenation model are available in Appendix 6.2.

The results for the *max-weight* and *non-zero weight* metrics are shown in Figure 5.2. The multi-encoder HAN attention module contextualises the current sentence with the context sentences based on two layers:

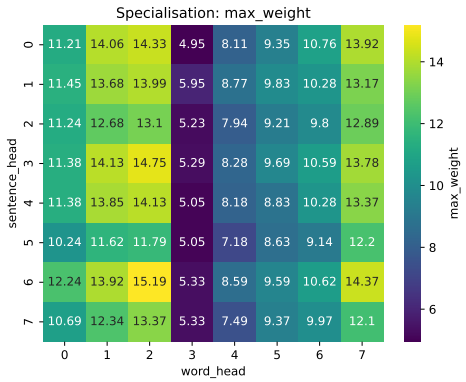
- The first layer, which will be called the "sentence-level" layer because it processes each context sentence independently one after the other.
- The second layer, which will be called the "context-level" layer because it processes the output of the sentence-level layer for each context sentence all together.

Heatmaps of size 8×8 in Figures 5.2b and 5.2d represent the value of respectively the max-weight metric and the non-zero weight when a matrix multiplication between sentence-level and context-level layers is applied. Each row represents the value of those metrics when a full sentence-level attention mechanism is "contextualised" by one head of the context-level of the attention. Conversely, the heatmaps in Figures 5.2a and 5.2c represent the different heads of a single layer in columns, while each row corresponds to a multi-head attention layer, resulting in sparser values.

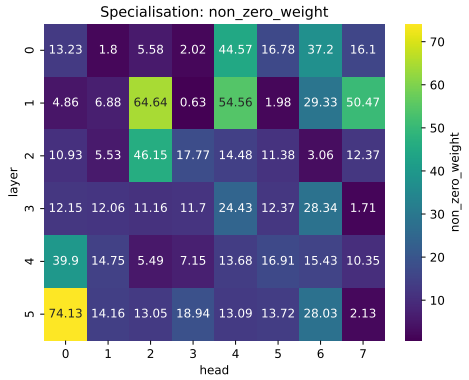
In the multi-encoder model (Figure 5.2d), variations across the *context-level* attention layers are negligible; however, a clear specialisation is visible among the *sentence-level* attention heads. Indeed, four *sentence-level* heads assign significant attention weight to the coreferential element in approximately 80% of cases. Three other *sentence-level* heads assign significant weight in 70% of cases, while the final *sentence-level* head drops to 60% for the *non-zero weight* metric. In Figure 5.2b,



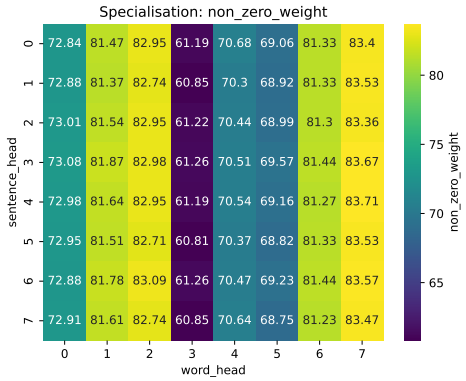
(a) Concat model max-weight metric in percentage (↑)



(b) Multi-encoder model max-weight metric in percentage (↑)



(c) Concat model non-zero weight (value between 0 and 1) (↑)



(d) Multi-encoder model non-zero weight (value between 0 and 1) (↑)

Figure 5.2: Heatmap visualization of the values obtained using the metrics proposed in Section 4, showing the specialization of attention heads by layer and by head for the concatenation model and by context-level head and sentence-level head for the multi-encoder model.

the *context-level* heads have a more pronounced effect on the *sentence-level* heads, leading to a variation of up to 3% in the *max-weight* metric for *sentence-level* head 2. Additionally, we observe a specialisation of certain *sentence-level* heads, with differences of up to 10% between heads. As noted during the "Attention please!" experiment in Section 5.1, the multi-encoder model seems to favour a more diffuse but less precise attention, explaining the high *non-zero weight* but lower *max-weight* (even if 15.19% remains relatively high for a model without specific coreference resolution training).

For the concatenation-based model (Figure 5.2c, *non-zero weight* metric), one attention head in encoder layer 5 stands out, reaching a score of 74.13%. This same head is identified in Figure 5.2a, where it assigns the maximum context weight to the coreferential token in 29.46% of cases. This behaviour is also found to a lesser extent in layer 1, where two heads obtain *max-weight* scores of 20.87% and 16.73%. Unlike the multi-encoder model, the concatenation model appears to have an attention head more specialised in the coreference resolution mechanism, being less diffuse and more precise.

In conclusion, although the concatenation model has a lower *non-zero weight* than the multi-encoder model (74.13% vs. 83.71%), it appears more precise and less dispersed when assigning maximum attention weight to a coreferential token (29.46% vs. 15.19%). This may be due to architectural differences, but also to the fact that only two layers in the multi-encoder model must encode the necessary context information, whereas six layers are utilized in the concatenation model.

As explained in the "Attention please!" experiment in Section 5.1, the *max-weight* and *non-zero weight* metrics are capable of quantifying a model's ability to monitor the context for information disambiguation. Furthermore, the head-by-head analysis across different multi-head attention layers allows for a finer examination of this context-analysis capability. However, these quantitative measures do not account for the semantic nature of the errors. To understand, for instance, why the multi-encoder model "hesitates" despite significant weights, a qualitative analysis of gender or number ambiguity cases is necessary.

The findings presented here highlight the relevance of evaluating model capabilities using the metrics defined in Chapter 4, though they require further investigation, which will be addressed in Chapter 6.

Chapter 6

Conclusions & perspectives

The objective of this thesis was to explore the interface between coreference resolution and context-aware neural machine translation (CA-NMT). Through an analysis of attention mechanisms, we sought to understand not only whether models correctly translate ambiguous coreferential elements, but more importantly, whether they do so for the right reasons.

6.1 Summary of contributions

First, we established an original methodological framework by proposing quantitative metrics based on attention: *max-weight*, *non-zero weight*, and *average weight*. These tools allowed us to approach attention matrices, often perceived as opaque or noisy, as interpretable indicators of contextual information retrieval.

Our experiments on the ParCorFull2.0 (Lapshinova-Koltunski et al., 2022) corpus highlighted fundamental differences between the two major architectures studied:

- The concatenation-based model (Lupo et al., 2022b) revealed a remarkable capacity for specialization. We identified specific "coreference heads" (notably at layer 5) capable of targeting the correct antecedent with high precision (high max-weight).
- The multi-encoder model (Lupo et al., 2022a), while performing well, appears to adopt a more diffuse strategy. Its attention is more dispersed (high non-zero weight), suggesting a more global integration of context that is less focused on the exact linguistic unit of the antecedent.

Furthermore, our work provides answers to the two central questions posed at the beginning of this thesis:

- *(Q1): How does the model resolve out-of-sentence coreference links?*
  We demonstrated that CA-NMT models do not merely capture superficial statistical correlations. There is a tangible trace within the attention mechanisms of an information being process between the current sentence and the context.

- *(Q2): Does the model predict the right translation for the right reasons?*
  The analysis of "registers" nonetheless nuanced this observation. We observed that models sometimes allocate massive attention to non-informative tokens (punctuation, function words), confirming that raw translation performance can mask computational shortcuts. However, the application of our cleaning and normalization methods made it possible to isolate a signal specifically related to coreference.

In conclusion, this thesis demonstrates that attention, although debated as a universal explanation, remains an indispensable element for the explainability of neural models. It highlight that head specialization is not an accident, but an emergent property of well-trained contextual architectures. While the field of Natural Language Processing (and Artificial Intelligence in general) is currently dominated by Large Language Models, the explainability methods and faithfulness metrics developed here aim to pave the way for a new way of context usage evaluation for translation systems. These systems should no longer aim only for fluency, but also for verifiable discursive coherence, thus ensuring more reliable translation. However, these metrics do not seek to provide a single, perfect explanation of model behavior, but rather to highlight the challenges and spark reflections on "What constitutes a good translation." We hope that the preliminary solutions proposed here, as well as the perspectives described in Section 6.2, will inspire other minds within the NLP community.

6.2 Perspectives

The work presented in this thesis has highlighted a specialization of attention heads within context-aware Transformer architectures. In particular, we demonstrated that the concatenation-based model (described in Chapter 4) possesses a head capable of identifying the correct antecedent with remarkable precision: in 29% of cases, it assigns the maximum attention weight to the coreferential element, and it provides a significant weight in 74.13% of situations. Conversely, the multi-encoder model favors a more diffuse strategy: while it provides the maximum weight in only 15.19% of cases, attention is directed toward a coreferential element in 83.71% of instances.

These results, while promising for models that have not undergone specific training for coreference resolution, must be interpreted with caution. Indeed, the analysis was limited to two distinct models. Generalizing these conclusions therefore requires a systematization of the analysis across a broader sample of models. In this regard, complementary work has been initiated via a dedicated web platform, developed for a post-doctoral researcher associated with this project. This tool allows for the application of our metrics to a wider variety of translation models. Preliminary observations from these third-party analyses tend to confirm the relevance and reliability of the metrics proposed in this manuscript. However, as we did not directly produce these additional results, we have chosen not to include them in this document, while noting that they pave the way for a robust, large-scale validation of our methodology.

Several research ideas could not be explored during this doctorate and constitute promising perspectives. A code refactoring is already planned to improve the scalability of the analysis tools. Furthermore, the following areas deserve particular attention:

- **Ablation study and baseline models:** It would be crucial to train baseline models for each architecture to precisely isolate the contribution of *focused concatenation* (Lupo et al., 2022b) and *divide & rule* (Lupo et al., 2022a) techniques to the emergence of specialized heads.

- **Exploration of the decoder:** Our analyses focused on the encoder. However, the decoder plays a critical role in the final token selection. Applying our metrics to modules such as beam search would help verify whether the decoder effectively use the encoder's specialized heads during pronoun generation.

- **Impact of context topology:** Building on the work of Kang et al. (2020) and Kim et al. (2019), a systematic study on antecedent distance (and thus the context window) and sentence length would allow for defining the operational limits of attention head specialization.

- **Validation of the register hypothesis:** The hypothesis of "attention registers," mentioned in Chapter 4, remains to be confirmed. Following the approach of Darcet et al. (2024), the explicit addition of register tokens during training could verify whether it is possible to naturally "clean" the attention signal and thus improve model explainability.

- **Generalization to Large Language Models:** The emergence of large language models is reshaping the NLP landscape. We deliberately chose not to

integrate these architectures into our work to focus on context-aware machine translation models and limit the resources required for training. It would be interesting to observe whether this specialization (a single head in deep layers) is an emergent property of scaling. Unlike BERT (Clark et al., 2019), which is a bidirectional encoder, current LLMs are mainly decoders. It would be necessary to observe whether coreference resolution is equally localized or more diffuse.

- **Contribution of Sparse Autoencoders:** As mentioned in Section 3.1.1, sparse autoencoders allow for the isolation of interpretable features within latent representations. It would be interesting to observe variations in attention behavior in such models to identify even finer coreference circuits.

Conclusion and outlook

This thesis set out from a simple but consequential observation: although neural machine translation has reached impressive sentence-level fluency, document-level coherence remains fragile, especially when correct translation depends on information located outside the current sentence. As argued in the introduction, coreference is a particularly revealing stress test for this limitation, because translating a mention may require retrieving an antecedent from the surrounding discourse and reflecting that choice in morphosyntactic decisions in the target language. Against this backdrop, the central question of this work was not only whether a context-aware system can produce the correct translation, but whether it does so *for the right reasons* by actually exploiting the relevant context.

To address this question, the thesis adopted a diagnostic approach grounded in Transformer attention. By treating attention as a practical tracer of information flow (rather than as a perfect explanation), we developed a methodology that makes attention signals measurable and comparable: controlled context packaging ($k1$–$k3$), extraction of attention matrices, and a series of preprocessing steps designed to mitigate confounds (subword aggregation, baseline thresholding and normalization). On top of this pipeline, we proposed complementary quantitative metrics that operationalize how attention aligns current mentions with contextual antecedents, and we validated these signals through qualitative inspection and finer-grained head-level analyses.

The empirical study, conducted on a parallel corpus with gold coreference annotation, allowed us to compare two representative families of context-aware neural machine translation: concatenation-based and multi-encoder hierarchical models. Be-

yond differences in translation scores, this comparison revealed that the way context is presented to the model shapes the internal patterns through which it is accessed. In particular, the analyses indicate that some architectures exhibit more concentrated or more frequent attention toward contextual material in settings where coreference is relevant, while others distribute contextual influence more diffusely. Importantly, the thesis also highlighted potential pitfalls: attention can be affected by sentence length, tokenization, and other artifacts, which makes rigorous preprocessing and cautious interpretation essential.

Taken together, the findings support the broader claim advanced at the beginning of the manuscript: evaluating context-aware translation should not rely exclusively on output-level metrics. Document-level phenomena such as coreference call for methods that can *audit* whether and how context is used. In this sense, attention-based diagnostics, when paired with appropriate controls and validation, provide a valuable additional lens for understanding and comparing CA-NMT systems.

Appendix

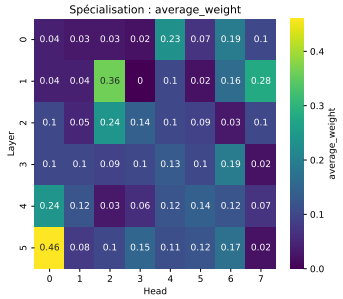


Figure 1: Concat model Average weight metric as a percentage (↑).

Bibliography

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim (2018). *Sanity Checks for Saliency Maps*. arXiv: 1810.03292 [cs.CV]. URL: https://arxiv.org/abs/1810.03292.

Bagga, Amit and Breck Baldwin (1998). "Algorithms for Scoring Coreference Chains". In: URL: https://api.semanticscholar.org/CorpusID:14142164.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Banerjee, Satanjeev and Alon Lavie (June 2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. URL: https://aclanthology.org/W05-0909.

Bau, David, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba (2017). *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. arXiv: 1704.05796 [cs.CV]. URL: https://arxiv.org/abs/1704.05796.

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow (June 2018). "Evaluating Discourse Phenomena in Neural Machine Translation". 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, pp. 1304–1313. DOI: 10.18653/v1/N18-1118. URL: https://aclanthology.org/N18-1118.

Belinkov, Yonatan and James Glass (2019). "Analysis Methods in Neural Language Processing: A Survey". In: *Transactions of the Association for Computational Linguistics* 7. Ed. by Lillian Lee, Mark Johnson, Brian Roark, and Ani Nenkova, pp. 49–72. DOI: 10.1162/tacl_a_00254. URL: https://aclanthology.org/Q19-1004/.

Bojar, Ondrej, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia (2016). "Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"". In: URL: https://api.semanticscholar.org/CorpusID:67046386.

Bolukbasi, Tolga, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg (2021). *An Interpretability Illusion for BERT*. arXiv: 2104.07143 [cs.CL]. URL: https://arxiv.org/abs/2104.07143.

Bricken, Trenton et al. (2023). "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bridle, John Scott (1989). "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition". In: *NATO Neurocomputing*. URL: https://api.semanticscholar.org/CorpusID:59636530.

Carletta, Jean (1996). "Assessing Agreement on Classification Tasks: The Kappa Statistic". In: *Computational Linguistics* 22.2, pp. 249–254. URL: https://aclanthology.org/J96-2004.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski (2003). "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory". In: *Current and New Directions in Discourse and Dialogue*. Ed. by Jan van Kuppevelt and Ronnie W. Smith. Dordrecht: Springer Netherlands, pp. 85–112. ISBN: 978-94-010-0019-2. DOI: 10.1007/978-94-010-0019-2_5. URL: https://doi.org/10.1007/978-94-010-0019-2_5.

Chafe, Wallace L. (1980). "The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production". In: URL: https://api.semanticscholar.org/CorpusID:143332222.

Chai, Haixia and Michael Strube (July 2022). "Incorporating Centering Theory into Neural Coreference Resolution". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 2996–3002. DOI: 10.18653/v1/2022.naacl-main.218. URL: https://aclanthology.org/2022.naacl-main.218.

Charolles, Michel (Jan. 1995). "Cohésion, cohérence et pertinence du discours". In: 29.

Chen, Chen and Vincent Ng (Oct. 2013). "Linguistically Aware Coreference Evaluation Metrics". In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Ed. by Ruslan Mitkov and Jong C. Park. Nagoya,

Japan: Asian Federation of Natural Language Processing, pp. 1366–1374. URL: https://aclanthology.org/I13-1193.

Chen, Mia Xu et al. (2018). *The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation*. arXiv: 1804.09849 [cs.CL]. URL: https://arxiv.org/abs/1804.09849.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio (2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. arXiv: 1409.1259 [cs.CL]. URL: https://arxiv.org/abs/1409.1259.

Christopher, Olah, Cammarata Nick, Schubert Ludwig, Goh Gabriel, Petrov Michael, and Carter Shan (2020). "Zoom In: An Introduction to Circuits". In: URL: https://api.semanticscholar.org/CorpusID:215930358.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning (Aug. 2019). "What Does BERT Look at? An Analysis of BERT's Attention". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes. Florence, Italy: Association for Computational Linguistics, pp. 276–286. DOI: 10.18653/v1/W19-4828. URL: https://aclanthology.org/W19-4828/.

Clark, Kevin and Christopher D. Manning (July 2015). "Entity-Centric Coreference Resolution with Model Stacking". 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, pp. 1405–1415. DOI: 10.3115/v1/P15-1136. URL: https://aclanthology.org/P15-1136.

— (Aug. 2016). "Improving Coreference Resolution by Learning Entity-Level Distributed Representations". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 643–653. DOI: 10.18653/v1/P16-1061. URL: https://aclanthology.org/P16-1061.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni (July 2018). "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 2126–2136. DOI: 10.18653/v1/P18-1198. URL: https://aclanthology.org/P18-1198/.

Corblin, Francis (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes. URL: https://hal.science/ijn_00550962.

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey (2023). *Sparse Autoencoders Find Highly Interpretable Features in Language Models*. arXiv: 2309.08600 [cs.LG]. URL: https://arxiv.org/abs/2309.08600.

Dalvi, Fahim, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad (2022). "Discovering Latent Concepts Learned in BERT". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=POTMtpYI1xH.

Darcet, Timothée, Maxime Oquab, Julien Mairal, and Piotr Bojanowski (2024). *Vision Transformers Need Registers*. arXiv: 2309.16588 [cs.CV]. URL: https://arxiv.org/abs/2309.16588.

Denis, Pascal and Jason Baldridge (Jan. 2009). "Global joint models for coreference resolution and named entity classification". In: *Procesamiento del lenguaje natural, ISSN 1135-5948, Nº. 42, 2009, pags. 87-96* 42.

Denkowski, Michael and Alon Lavie (July 2011). "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 85–91. URL: https://aclanthology.org/W11-2107.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

Dinarelli, Marco, Dimitra Niaouri, Fabien Lopez, Gabriela Gonzalez-Saez, Mariam Nakhlé, Emmanuelle Esperança-Rodier, Caroline Rossi, Didier Schwab, and Nicolas Ballier (2024). "Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention". In: *Traitement Automatique des Langues* 64.3. Ed. by Guillaume Wisniewski, pp. 67–91. DOI: 10.57896/2023-tal-64_3_3. URL: https://aclanthology.org/2023.tal-3.4/.

Elhage, Nelson et al. (2021). "A Mathematical Framework for Transformer Circuits". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html.

Elhage, Nelson et al. (2022). *Toy Models of Superposition*. arXiv: 2209.10652 [cs.LG]. URL: https://arxiv.org/abs/2209.10652.

Eshkol-Taravella, Iris, Olivier Baude, Denis Maurel, Linda Hriba, Céline Dugua, and Isabelle Tellier (2011). "Un grand corpus oral " disponible " : le corpus d'Orléans 1 1968-2012". In: *Revue TAL*. Ressources Linguistiques Libres 53.2, pp. 17–46. URL: https://shs.hal.science/halshs-01163053.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar (Nov. 2021). "Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain". In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 733–774. URL: https://aclanthology.org/2021.wmt-1.73.

Geiger, Atticus, Hanson Lu, Thomas Icard, and Christopher Potts (2021). *Causal Abstractions of Neural Networks*. arXiv: 2106.02997 [cs.AI]. URL: https://arxiv.org/abs/2106.02997.

Ghaddar, Abbas and Phillippe Langlais (May 2016). "WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 136–142. URL: https://aclanthology.org/L16-1021.

Ghorbani, Amirata, Abubakar Abid, and James Zou (July 2019). "Interpretation of Neural Networks Is Fragile". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01, pp. 3681–3688. DOI: 10.1609/aaai.v33i01.33013681. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4252.

Gonzalez-Saez, Gabriela et al. (June 2024a). "Exploring NMT Explainability for Translators Using NMT Visualising Tools". In: *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*. Ed. by Carolina Scarton et al. Sheffield, UK: European Association for Machine Translation (EAMT), pp. 396–410. URL: https://aclanthology.org/2024.eamt-1.34/.

Gonzalez-Saez, Gabriela et al. (June 2024b). "The MAKE-NMTViz Project: Meaningful, Accurate and Knowledge-limited Explanations of NMT Systems for Translators". In: *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*. Ed. by Carolina Scarton et al. Sheffield, UK: European Association for Machine Translation (EAMT), pp. 12–13. URL: https://aclanthology.org/2024.eamt-2.7/.

Gowda, K. Chidananda and G. Krishna (1978). "Agglomerative clustering using the concept of mutual nearest neighbourhood". In: *Pattern Recognit.* 10, pp. 105–112. URL: https://api.semanticscholar.org/CorpusID:5186751.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel (Aug. 2013). "Continuous Measurement Scales in Human Evaluation of Machine Translation". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 33–41. URL: https://aclanthology.org/W13-2305.

Graves, Alex (2012). *Sequence Transduction with Recurrent Neural Networks*. arXiv: 1211.3711 [cs.NE]. URL: https://arxiv.org/abs/1211.3711.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein (June 1983). "Providing a Unified Account of Definite Noun Phrases in Discourse". In: *21st Annual Meeting of the Association for Computational Linguistics*. Cambridge, Massachusetts, USA: Association for Computational Linguistics, pp. 44–50. DOI: 10.3115/981311.981320. URL: https://aclanthology.org/P83-1007.

— (1995). "Centering: A Framework for Modeling the Local Coherence of Discourse". In: *Computational Linguistics* 21.2. Ed. by Julia Hirschberg, pp. 203–225. URL: https://aclanthology.org/J95-2003.

Guillou, Liane and Christian Hardmeier (May 2016). "PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 636–643. URL: https://aclanthology.org/L16-1100.

— (Oct. 2018). "Automatic Reference-Based Evaluation of Pronoun Translation Misses the Point". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4797–4802. DOI: 10.18653/v1/D18-1513. URL: https://aclanthology.org/D18-1513.

Guillou, Liane, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber (May 2014). "ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3191–3198. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/298_Paper.pdf.

Hassan, Hany et al. (2018). *Achieving Human Parity on Automatic Chinese to English News Translation*. arXiv: 1803.05567 [cs.CL]. URL: https://arxiv.org/abs/1803.05567.

Heap, Thomas, Tim Lawson, Lucy Farnik, and Laurence Aitchison (2025). *Sparse Autoencoders Can Interpret Randomly Initialized Transformers*. arXiv: 2501.17727 [cs.LG]. URL: https://arxiv.org/abs/2501.17727.

Heeman, Peter A. and James F. Allen (1995). "The TRAINS 93 Dialogues". In: URL: https://api.semanticscholar.org/CorpusID:195602390.

Hewitt, John and Christopher D. Manning (June 2019). "A Structural Probe for Finding Syntax in Word Representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4129–4138. DOI: 10.18653/v1/N19-1419. URL: https://aclanthology.org/N19-1419/.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780.

Jacovi, Alon and Yoav Goldberg (July 2020). "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://aclanthology.org/2020.acl-main.386/.

Jain, Sarthak and Byron C. Wallace (2019). *Attention is not Explanation*. arXiv: 1902.10186 [cs.CL]. URL: https://arxiv.org/abs/1902.10186.

Jeon, Sungho and Michael Strube (Nov. 2020). "Centering-based Neural Coherence Modeling with Hierarchical Discourse Segments". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 7458–7472. DOI: 10.18653/v1/2020.emnlp-main.604. URL: https://aclanthology.org/2020.emnlp-main.604.

Joshi, Aravind K. and Scott Weinstein (1981). "Control of Inference: Role of Some Aspects of Discourse Structure-Centering". In: *International Joint Conference on Artificial Intelligence*. URL: https://api.semanticscholar.org/CorpusID:6853300.

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy (Jan. 2020). "SpanBERT: Improving Pre-training by Representing and Predicting Spans". In: *Transactions of the Association for Computational Linguistics* 8, pp. 64–77. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00300. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a\

_00300/1923170/tacl_a_00300.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00300.

Joshi, Mandar, Omer Levy, Luke Zettlemoyer, and Daniel Weld (Nov. 2019). "BERT for Coreference Resolution: Baselines and Analysis". 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)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 5803–5808. DOI: 10.18653/v1/D19-1588. URL: https://aclanthology.org/D19-1588.

Junczys-Dowmunt, Marcin (Aug. 2019). "Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation". In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Ed. by Ondřej Bojar et al. Florence, Italy: Association for Computational Linguistics, pp. 225–233. DOI: 10.18653/v1/W19-5321. URL: https://aclanthology.org/W19-5321.

Jwalapuram, Prathyusha, Shafiq Joty, Irina Temnikova, and Preslav Nakov (Nov. 2019). "Evaluating Pronominal Anaphora in Machine Translation: An Evaluation Measure and a Test Suite". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 2964–2975. DOI: 10.18653/v1/D19-1294. URL: https://aclanthology.org/D19-1294.

Kang, Xiaomian, Yang Zhao, Jiajun Zhang, and Chengqing Zong (Nov. 2020). "Dynamic Context Selection for Document-level Neural Machine Translation via Reinforcement Learning". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 2242–2254. DOI: 10.18653/v1/2020.emnlp-main.175. URL: https://aclanthology.org/2020.emnlp-main.175.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres (2018). *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)*. arXiv: 1711.11279 [stat.ML]. URL: https://arxiv.org/abs/1711.11279.

Kim, Yunsu, Duc Thanh Tran, and Hermann Ney (Nov. 2019). "When and Why is Document-level Context Useful in Neural Machine Translation?" In: *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*. Ed. by Andrei Popescu-Belis, Sharid Loáiciga, Christian Hardmeier, and Deyi

Xiong. Hong Kong, China: Association for Computational Linguistics, pp. 24–34. DOI: 10.18653/v1/D19-6503. URL: https://aclanthology.org/D19-6503.

Kleiber, Georges and Helene Vassiliadou (June 2007). "Sur les approches intuitives de la relation d'élaboration". In: *Scolia* 22. DOI: 10.3406/scoli.2007.1109.

Kuang, Shaohui, Deyi Xiong, Weihua Luo, and Guodong Zhou (Aug. 2018). "Modeling Coherence for Neural Machine Translation with Dynamic and Topic Caches". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 596–606. URL: https://www.aclweb.org/anthology/C18-1050.

Lanham, Tamera et al. (2023). *Measuring Faithfulness in Chain-of-Thought Reasoning*. arXiv: 2307.13702 [cs.AI]. URL: https://arxiv.org/abs/2307.13702.

Lapshinova-Koltunski, Ekaterina, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier (June 2022). "ParCorFull2.0: a Parallel Corpus Annotated with Full Coreference". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 805–813. URL: https://aclanthology.org/2022.lrec-1.85.

Lapshinova-Koltunski, Ekaterina, Christian Hardmeier, and Pauline Krielke (2018). *ParCorFull: A Parallel Corpus Annotated with Full Coreference*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. URL: http://hdl.handle.net/11372/LRT-2614.

Lattice, LiLPa, ICAR, and IHRIM (2019). *Democrat*. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr. URL: https://hdl.handle.net/11403/democrat/v1.1.

Läubli, Samuel, Rico Sennrich, and Martin Volk (Oct. 2018). "Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, pp. 4791–4796. DOI: 10.18653/v1/D18-1512. URL: https://aclanthology.org/D18-1512.

Le Nagard, Ronan and Philipp Koehn (July 2010). "Aiding Pronoun Translation with Co-Reference Resolution". In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan. Uppsala, Sweden: Association for Computational Linguistics, pp. 252–261. URL: https://aclanthology.org/W10-1737.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (Sept. 2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 188–197. DOI: 10.18653/v1/D17-1018. URL: https://aclanthology.org/D17-1018.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (June 2018). "Higher-Order Coreference Resolution with Coarse-to-Fine Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 687–692. DOI: 10.18653/v1/N18-2108. URL: https://aclanthology.org/N18-2108.

Li, Aaron J., Suraj Srinivas, Usha Bhalla, and Himabindu Lakkaraju (2026). *Evaluating Adversarial Robustness of Concept Representations in Sparse Autoencoders*. arXiv: 2505.16004 [cs.LG]. URL: https://arxiv.org/abs/2505.16004.

Li, Bei, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li (July 2020). "Does Multi-Encoder Help? A Case Study on Context-Aware Neural Machine Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 3512–3518. DOI: 10.18653/v1/2020.acl-main.322. URL: https://aclanthology.org/2020.acl-main.322.

Lion-Bouton, Adam, Loïc Grobol, Jean-Yves Antoine, Sylvie Billot, and Anaïs Lefeuvre-Halftermeyer (June 2020). "Comment arpenter sans mètre : les scores de résolution de chaînes de coréférences sont-ils des métriques ? (Do the standard scores of evaluation of coreference resolution constitute metrics ?)" French. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e Édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*. Ed. by Gilles Adda, Maxime Amblard, and Karën Fort. Nancy, France: ATALA et AFCP, pp. 10–18. URL: https://aclanthology.org/2020.jeptalnrecital-eternal.2.

Lopes, Antonio V, M Amin Farajian, Rachel Bawden, Michael Zhang, and Andre F T Martins (2020). "Document-level Neural MT: A Systematic Comparison". en. In: p. 10.

Lopez, Fabien (June 2023). "Approches neuronales pour la détection des chaînes de coréférences : un état de l'art". fra. In: *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*.

Ed. by Marie Candito, Thomas Gerald, and José G Moreno. Paris, France: ATALA, pp. 101–113. URL: https://aclanthology.org/2023.jeptalnrecital-rjc.7/.

Lopez, Fabien et al. (Dec. 2023). "The MAKE-NMTVIZ System Description for the WMT23 Literary Task". In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz. Singapore: Association for Computational Linguistics, pp. 287–295. DOI: 10.18653/v1/2023.wmt-1.30. URL: https://aclanthology.org/2023.wmt-1.30/.

Lundberg, Scott M. and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964.

Luo, Xiaoqiang (Oct. 2005). "On Coreference Resolution Performance Metrics". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 25–32. URL: https://aclanthology.org/H05-1004.

Luo, Xiaoqiang and Sameer Pradhan (2016). "Evaluation Metrics". In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuckardt, and Yannick Versley. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 141–163. ISBN: 978-3-662-47909-4. DOI: 10.1007/978-3-662-47909-4_5. URL: https://doi.org/10.1007/978-3-662-47909-4_5.

Lupo, Lorenzo, Marco Dinarelli, and Laurent Besacier (May 2022a). "Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 4557–4572. DOI: 10.18653/v1/2022.acl-long.312. URL: https://aclanthology.org/2022.acl-long.312.

— (Dec. 2022b). "Focused Concatenation for Context-Aware Neural Machine Translation". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Ed. by Philipp Koehn et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 830–842. URL: https://aclanthology.org/2022.wmt-1.77.

Marasovic, Ana, Iz Beltagy, Doug Downey, and Matthew Peters (July 2022). "Few-Shot Self-Rationalization with Natural Language Prompts". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Ed. by Marine Carpuat,

Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 410–424. DOI: 10.18653/v1/2022.findings-naacl.31. URL: https://aclanthology.org/2022.findings-naacl.31/.

Maruf, Sameen, André F. T. Martins, and Gholamreza Haffari (Oct. 2018). "Contextual Neural Model for Translating Bilingual Multi-Speaker Conversations". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pp. 101–112. DOI: 10.18653/v1/W18-6311. URL: https://www.aclweb.org/anthology/W18-6311.

— (June 2019). "Selective Attention for Context-aware Neural Machine Translation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3092–3102. DOI: 10.18653/v1/N19-1313. URL: https://www.aclweb.org/anthology/N19-1313.

McGrath, Thomas, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg (2023). *The Hydra Effect: Emergent Self-repair in Language Model Computations*. arXiv: 2307.15771 [cs.LG]. URL: https://arxiv.org/abs/2307.15771.

Méloux, Maxime, Giada Dirupo, François Portet, and Maxime Peyrard (2025). *The Dead Salmons of AI Interpretability*. arXiv: 2512.18792 [cs.AI]. URL: https://arxiv.org/abs/2512.18792.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov (2023). *Locating and Editing Factual Associations in GPT*. arXiv: 2202.05262 [cs.CL]. URL: https://arxiv.org/abs/2202.05262.

Miculicich, Lesly and James Henderson (May 2022). "Graph Refinement for Coreference Resolution". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 2732–2742. DOI: 10.18653/v1/2022.findings-acl.215. URL: https://aclanthology.org/2022.findings-acl.215.

Miculicich, Lesly and Andrei Popescu-Belis (2017). "Using Coreference Links to Improve Spanish-to-English Machine Translation". In: URL: https://api.semanticscholar.org/CorpusID:9732862.

Miculicich, Lesly, Dhananjay Ram, Nikolaos Pappas, and James Henderson (2018). *Document-Level Neural Machine Translation with Hierarchical Attention Networks*. arXiv: 1809.01576 [cs.CL]. URL: https://arxiv.org/abs/1809.01576.

Miculicich Werlen, Lesly and Andrei Popescu-Belis (Sept. 2017). "Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT)". In: *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 17–25. DOI: 10.18653/v1/W17-4802. URL: https://aclanthology.org/W17-4802.

Moosavi, Nafise Sadat and Michael Strube (Aug. 2016). "Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 632–642. DOI: 10.18653/v1/P16-1060. URL: https://aclanthology.org/P16-1060.

Muller, Mathias, Annette Rios, Elena Voita, and Rico Sennrich (Oct. 2018). "A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pp. 61–72. DOI: 10.18653/v1/W18-6307. URL: https://aclanthology.org/W18-6307.

Muzerelle, Judith, Anaïs Lefeuvre, Jean-Yves Antoine, Emmanuel Schang, Denis Maurel, Jeanne Villaneau, and Iris Eshkol (June 2011). "ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement". In: *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles*. Ed. by ATALA. Les Sable d'Olonne, France, pp. 555–563. URL: https://hal.science/hal-01016562.

Muzerelle, Judith, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau (May 2014). "ANCOR_Centre, a Large Free Spoken French Coreference Corpus: description of the Resource and Reliability Measures". In: *LREC'2014, 9th Language Resources and Evaluation Conference*. Ed. by ELRA. Reyjavik, Iceland, pp. 843–847. URL: https://hal.science/hal-01075679.

Nair, Vinod and Geoffrey E. Hinton (2010). "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, pp. 807–814. ISBN: 9781605589077.

Ng, Vincent and Claire Cardie (2002). "Improving Machine Learning Approaches to Coreference Resolution". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, pp. 104–111. DOI: 10.3115/1073083.1073102. URL: https://doi.org/10.3115/1073083.1073102.

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter (2020). "Zoom In: An Introduction to Circuits". In: *Distill*. https://distill.pub/2020/circ in. DOI: 10.23915/distill.00024.001.

Paolini, Giovanni, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto (2021). "Structured Prediction as Translation between Augmented Natural Languages". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=US-TP-xnXI.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2001). "BLEU: a method for automatic evaluation of machine translation". en. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Philadelphia, Pennsylvania: Association for Computational Linguistics, p. 311. DOI: 10.3115/1073083.1073135. URL: http://portal.acm.org/citation.cfm?doid=1073083.1073135 (visited on 08/24/2022).

Pearl, Judea (2012). "The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms". In: *Prevention Science* 13, pp. 426–436. URL: https://api.semanticscholar.org/CorpusID:1895243.

Poesio, Massimo (July 2004). "Discourse Annotation and Semantic Annotation in the GNOME corpus". In: *Proceedings of the Workshop on Discourse Annotation*. Barcelona, Spain: Association for Computational Linguistics, pp. 72–79. URL: https://aclanthology.org/W04-0210.

Poesio, Massimo and Ron Artstein (May 2008). "Anaphoric Annotation in the ARRAU Corpus". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/297_paper.pdf.

Post, Matt (Oct. 2018). "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Ed. by Ondřej Bojar et al. Brussels, Belgium: Association for Computational Linguistics, pp. 186–191. DOI: 10.18653/v1/W18-6319. URL: https://aclanthology.org/W18-6319/.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (July 2012). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Joint Conference on EMNLP and CoNLL - Shared Task*. Ed. by Sameer Pradhan, Alessandro Moschitti, and Nianwen Xue. Jeju Island, Korea: Association for Computational Linguistics, pp. 1–40. URL: https://aclanthology.org/W12-4501/.

Pradhan, Sameer S., Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel (2007). "OntoNotes: A Unified Relational Semantic Representation". In: *International Conference on Semantic Computing (ICSC 2007)*, pp. 517–526. DOI: 10.1109/ICSC.2007.83.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (Jan. 2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *J. Mach. Learn. Res.* 21.1. ISSN: 1532-4435.

Rahman, Altaf and Vincent Ng (Aug. 2009). "Supervised Models for Coreference Resolution". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 968–977. URL: https://aclanthology.org/D09-1101.

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66, pp. 846–850.

Ravichander, Abhilasha, Yonatan Belinkov, and Eduard Hovy (Apr. 2021). "Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance?" In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 3363–3377. DOI: 10.18653/v1/2021.eacl-main.295. URL: https://aclanthology.org/2021.eacl-main.295/.

Recasens, M. and E. Hovy (2011). "BLANC: Implementing the Rand index for coreference evaluation". In: *Natural Language Engineering* 17.4, pp. 485–510. DOI: 10.1017/S135132491000029X.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie (Nov. 2020). "COMET: A Neural Framework for MT Evaluation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213. URL: https://aclanthology.org/2020.emnlp-main.213.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv: 1602.04938 [cs.LG]. URL: https://arxiv.org/abs/1602.04938.

Sanh, Victor et al. (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=9Vrb9D0WI4.

Schnedecker, Catherine (1997). *Nom propre et chaînes de référence*. Ed. by Klincksieck. Recherches linguistiques. Librairie KLINCKSIECK, p. 231. URL: https://hal.science/hal-00808797.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). *Neural Machine Translation of Rare Words with Subword Units*. arXiv: 1508.07909 [cs.CL]. URL: https://arxiv.org/abs/1508.07909.

Serrano, Sofia and Noah A. Smith (July 2019). "Is Attention Interpretable?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 2931–2951. DOI: 10.18653/v1/P19-1282. URL: https://aclanthology.org/P19-1282/.

Sharkey, Lee, Dan Braun, and Beren Millidge (2022). "Taking features out of superposition with sparse autoencoders". In: *AI Alignment Forum*. Vol. 6, pp. 12–13.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. arXiv: 1312.6034 [cs.CV]. URL: https://arxiv.org/abs/1312.6034.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim (Dec. 2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases". In: *Computational Linguistics* 27.4, pp. 521–544. ISSN: 0891-2017. DOI: 10.1162/089120101753342653. eprint: https://direct.mit.edu/coli/article-pdf/27/4/521/1797672/089120101753342653.pdf. URL: https://doi.org/10.1162/089120101753342653.

Stojanovski, Dario and Alexander Fraser (Aug. 2019). "Improving Anaphora Resolution in Neural Machine Translation Using Curriculum Learning". In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. Dublin, Ireland: European Association for Machine Translation, pp. 140–150. URL: https://www.aclweb.org/anthology/W19-6614.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). *Axiomatic Attribution for Deep Networks*. arXiv: 1703.01365 [cs.LG]. URL: https://arxiv.org/abs/1703.01365.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Sutter, Denis, Julian Minder, Thomas Hofmann, and Tiago Pimentel (2025). *The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability?* arXiv: 2507.08802 [cs.LG]. URL: https://arxiv.org/abs/2507.08802.

Tan, Xin, Longyin Zhang, Deyi Xiong, and Guodong Zhou (Nov. 2019). "Hierarchical Modeling of Global Context for Document-Level Neural Machine Translation". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 1576–1585. DOI: 10.18653/v1/D19-1168. URL: https://www.aclweb.org/anthology/D19-1168.

Tiedemann, Jörg and Yves Scherrer (Sept. 2017). "Neural Machine Translation with Extended Context". In: *Proceedings of the Third Workshop on Discourse in Machine Translation*. Ed. by Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann. Copenhagen, Denmark: Association for Computational Linguistics, pp. 82–92. DOI: 10.18653/v1/W17-4811. URL: https://aclanthology.org/W17-4811.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way (Oct. 2018). "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Ed. by Ondřej Bojar et al. Brussels, Belgium: Association for Computational Linguistics, pp. 113–123. DOI: 10.18653/v1/W18-6312. URL: https://aclanthology.org/W18-6312/.

Tu, Zhaopeng, Yang Liu, Shuming Shi, and Tong Zhang (2018). "Learning to Remember Translation History with a Continuous Cache". In: *Transactions of the Association for Computational Linguistics* 6, pp. 407–420. DOI: 10.1162/tacl_a_00029. URL: https://www.aclweb.org/anthology/Q18-1029.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman (2023). *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. arXiv: 2305.04388 [cs.CL]. URL: https://arxiv.org/abs/2305.04388.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber (2020). "Investigating Gender Bias in Language Models Using Causal Mediation Analysis". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 12388–12401. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman (1995). "A Model-Theoretic Coreference Scoring Scheme". In: *Sixth Message Un-*

derstanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995. URL: https://aclanthology.org/M95-1005.

Voita, Elena, Rico Sennrich, and Ivan Titov (July 2019). "When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1198–1212. DOI: 10.18653/v1/P19-1116. URL: https://www.aclweb.org/anthology/P19-1116.

Voita, Elena, Pavel Serdyukov, Rico Sennrich, and Ivan Titov (July 2018). "Context-Aware Neural Machine Translation Learns Anaphora Resolution". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1264–1274. DOI: 10.18653/v1/P18-1117. URL: https://aclanthology.org/P18-1117.

Wang, Kevin, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt (2022). *Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small*. arXiv: 2211.00593 [cs.LG]. URL: https://arxiv.org/abs/2211.00593.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903 [cs.CL]. URL: https://arxiv.org/abs/2201.11903.

Wiegreffe, Sarah, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi (July 2022). "Reframing Human-AI Collaboration for Generating Free-Text Explanations". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 632–658. DOI: 10.18653/v1/2022.naacl-main.47. URL: https://aclanthology.org/2022.naacl-main.47/.

Wiegreffe, Sarah, Ana Marasović, and Noah A. Smith (Nov. 2021). "Measuring Association Between Labels and Free-Text Rationales". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10266–10284. DOI: 10.18653/v1/2021.emnlp-main.804. URL: https://aclanthology.org/2021.emnlp-main.804/.

Wiegreffe, Sarah and Yuval Pinter (2019a). *Attention is not not Explanation*. arXiv: 1908.04626 [cs.CL]. URL: https://arxiv.org/abs/1908.04626.

— (2019b). *Attention is not not Explanation*. arXiv: 1908.04626 [cs.CL]. URL: https://arxiv.org/abs/1908.04626.

Winograd, Terry (1972). "Understanding natural language". In: *Cognitive Psychology* 3.1, pp. 1–191. ISSN: 0010-0285. DOI: https://doi.org/10.1016/0010-0285(72)90002-3. URL: https://www.sciencedirect.com/science/article/pii/0010028572900023.

Wiseman, Sam, Alexander M. Rush, Stuart Shieber, and Jason Weston (2015). "Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution". 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)*. 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, pp. 1416–1426. DOI: 10.3115/v1/P15-1137. URL: http://aclweb.org/anthology/P15-1137 (visited on 03/18/2022).

Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber (June 2016). "Learning Global Features for Coreference Resolution". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, pp. 994–1004. DOI: 10.18653/v1/N16-1114. URL: https://aclanthology.org/N16-1114.

Wong, Billy T. M. and Chunyu Kit (July 2012). "Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1060–1068. URL: https://aclanthology.org/D12-1097.

Wong, KayYen, Sameen Maruf, and Gholamreza Haffari (2020). *Contextual Neural Machine Translation Improves Translation of Cataphoric Pronouns*. arXiv: 2004.09894 [cs.CL].

Wu, Wei, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li (July 2020). "CorefQA: Coreference Resolution as Query-based Span Prediction". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online:

Association for Computational Linguistics, pp. 6953–6963. DOI: 10.18653/v1/2020.acl-main.622. URL: https://aclanthology.org/2020.acl-main.622.

Wu, Yonghui et al. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. arXiv: 1609.08144 [cs.CL]. URL: https://arxiv.org/abs/1609.08144.

Yan, Yiming, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang (July 2023). "BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 5428–5443. DOI: 10.18653/v1/2023.acl-long.297. URL: https://aclanthology.org/2023.acl-long.297.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy (June 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*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, pp. 1480–1489. DOI: 10.18653/v1/N16-1174. URL: https://aclanthology.org/N16-1174/.

Yu, Xuemin, Fahim Dalvi, Nadir Durrani, Marzia Nouri, and Hassan Sajjad (2024). *Latent Concept-based Explanation of NLP Models*. arXiv: 2404.12545 [cs.CL]. URL: https://arxiv.org/abs/2404.12545.

Zhang, Jiacheng, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu (Nov. 2018). "Improving the Transformer Translation Model with Document-Level Context". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 533–542. DOI: 10.18653/v1/D18-1049. URL: https://www.aclweb.org/anthology/D18-1049.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (Feb. 2020). *BERTScore: Evaluating Text Generation with BERT*. en. arXiv:1904.09675 [cs]. URL: http://arxiv.org/abs/1904.09675 (visited on 08/24/2022).

Zhang, Wenzheng, Sam Wiseman, and Karl Stratos (2023). *Seq2seq is All You Need for Coreference Resolution*. arXiv: 2310.13774 [cs.CL]. URL: https://arxiv.org/abs/2310.13774.

Zheng, Zaixiang, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch (Feb. 2020). "Towards Making the Most of Context in Neural Machine Translation". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial*

Intelligence, IJCAI-20. Ed. by Christian Bessiere. International Joint Conferences on Artificial Intelligence Organization, pp. 3983–3989. DOI: 10.24963/ijcai.2020/551. URL: https://doi.org/10.24963/ijcai.2020/551.

Zouhar, Vilém, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow (2024). *Pitfalls and Outlooks in Using COMET*. arXiv: 2408.15366 [cs.CL]. URL: https://arxiv.org/abs/2408.15366.