Thèse n° 7779

Discourse Phenomena in Machine Translation

Présentée le 1 février 2021

Faculté des sciences et techniques de l'ingénieur
Laboratoire de l'IDIAP
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Lesly Sadiht MICULICICH WERLEN

Acceptée sur proposition du jury

Prof. J.-Ph. Thiran, président du jury
Prof. H. Bourlard, Dr J. Henderson, directeurs de thèse
Prof. R. Sennrich, rapporteur
Prof. L. Specia, rapporteuse
Prof. M. Jaggi, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

Abstract

Machine Translation (MT) has made considerable progress in the past two decades, particularly after the introduction of neural network models (NMT). During this time, the research community has mostly focused on modeling and evaluating MT systems at the sentence level. MT models learn to translate from large amounts of parallel sentences in different languages. The focus on sentences brings a practical simplification for the task that favors efficiency but has the disadvantage of missing relevant contextual information. Several studies showed that the negative impact of this simplification is significant. One key point is that the discourse dependencies among distant words are ignored, resulting in a lack of coherence and cohesion in the text.

The main objective of this thesis is to improve MT by including discourse-level constraints. In particular, we focus on the translation of the entity mentions. We summarize our contributions in four points. First, we define the evaluation process to assess entity translations (i.e., nouns and pronouns) and propose an automatic metric to measure this phenomenon. Second, we perform a proof-of-concept and analyze how effective it is to include entity coreference resolution (CR) in translation. We conclude that CR significantly helps pronoun translation and boosts the whole translation quality according to human judgment. Third, we focus on the discourse connections at the sentence level. We propose enhancing the sequential model to infer long-term connections by incorporating a 'self-attention' mechanism. This mechanism gives direct and selective access to the context. Experiments in different language pairs show that our method outperforms various baselines, and the analysis confirms that the model emphasizes a broader context and captures syntactic-like structures. Fourth, we formulate the problem of document-level NMT and model inter-sentential connections among words with a hierarchical attention mechanism. Experiments on multiple data sets show significant improvement over two strong baselines and conclude that the source and target sides' contexts are mutually complementary. This set of results confirms that discourse significantly enhances translation quality, verifying our main thesis objective.

Our secondary objective is to improve the CR task by modeling the underlying connections among entities at the document-level. This task is particularly challenging for current neural network models because it requires understanding and reasoning. First, we propose a method to detect entity mentions from partially annotated data. We then proposed to model coreference with a graph of entities encoded in a pre-trained language model as an internal structure. The experiments show that these methods outperform various baselines. CR has the potential to help MT and other text generation tasks by maintaining coherence between the entity mentions.

Keywords– Neural machine translation; NMT; coreference resolution; mention detection; document-level translation; discourse phenomenon; pronoun translation

Zusammenfassung

In der maschinellen Übersetzung (MÜ) wurden in den letzten zwei Jahrzehnten erhebliche Fortschritte erzielt, insbesondere seit der Verwendung neuronaler Netzwerkmodelle (NNM). In dieser Zeit konzentrierte sich die Forschungsgemeinschaft hauptsächlich auf die Modellierung und Bewertung von MÜ-Systemen auf Satzebene. MÜ-Modelle lernen von großen Mengen paralleler Sätze in verschiedenen Sprachen. Die Fokussierung auf Sätze bringt eine praktische Vereinfachung mit sich, die zwar Effizienz fördert aber den Nachteil hat, dass relevante Kontextinformationen fehlen. Mehrere Studien haben gezeigt, dass die negativen Auswirkungen dieser Vereinfachung erheblich sind. Ein elementarer Punkt ist, dass die Kontextabhängigkeiten zwischen weiter entfernten Wörtern ignoriert werden, was zu einem Mangel an Kohärenz und Zusammenhang im Text führt.

Das Hauptziel dieser Arbeit ist die Verbesserung der MÜ durch Einbeziehung von Einschränkungen auf Dokumentebene. Insbesondere konzentrieren wir uns auf die Übersetzung von Entitätserwähnungen. Wir fassen unsere Beiträge in vier Punkten zusammen. Zunächst definieren wir den Bewertungsprozess zur Bewertung von Entitätsübersetzungen (d.h. Nomen und Pronomen) und schlagen eine automatische Metrik vor, um dieses Phänomen zu quantifizieren. Zweitens führen wir ein Proof-of-Concept durch und analysieren, wie effektiv es ist, Co-Reference Resolution (CR) von Entitäten in die Übersetzung einzubeziehen. Wir schließen daraus, dass CR die Pronomenübersetzung erheblich unterstützt und die gesamte Übersetzungsqualität nach menschlichen Maßstäben verbessert. Drittens konzentrieren wir uns auf die Kontextverbindungen auf Satzebene. Wir schlagen vor, das sequentielle Modell zu verbessern um auf langfristige Verbindungen zu schließen, indem ein ßelf-attentionMechanismus integriert wird. Dieser Mechanismus ermöglicht den direkten und selektiven Zugriff auf den Kontext. Experimente in verschiedenen Sprachpaaren zeigen, dass unsere Methode verschiedene Ausgangsergebnisse verbessert und die Analyse bestätigt, dass das Modell einen breiteren Kontext betont und syntaktisch ähnliche Strukturen erfasst. Viertens formulieren wir das Problem der NMT auf Dokumentebene und modellieren Verbindungen zwischen Wörtern auf Satzebene mit einem hierarchischen ättention-Mechanismus. Experimente mit mehreren Datensätzen zeigen eine signifikante Verbesserung gegenüber zwei starken Ausgangsergebnissen und kommen zu dem Schluss, dass der Kontext der Ausgangs- und Zielseite komplementär ist. Diese Ergebnisse bestätigen, dass der Kontexts die Übersetzungsqualität erheblich verbessert und unser Hauptziel bestätigt.

Unser sekundäres Ziel ist die Verbesserung der CR-Aufgabe durch Modellierung der zugrunde liegenden Verbindungen zwischen Entitäten auf Dokumentebene. Diese Aufgabe ist für aktuelle neuronale Netzwerkmodelle besonders anspruchsvoll, da sie Verstehen und Argumentation des Dokuments erfordert. Zunächst schlagen wir eine Methode vor, um Erwähnungen von Entitäten aus teilweise kommentierten Daten zu erkennen. Dann kann die Co-Referenz mit einem Graph von Entitäten modelliert werdem, die in einem vorab trainierten Sprachmodell als interne Struktur codiert sind. Experimente zeigen, dass diese Methoden verschiedene Ausgangsergebnisse übertreffen. CR hat das Potenzial, MÜ und anderen Aufgaben bei der Texterzeugung zu helfen, indem die Kohärenz zwischen den Entitätenerwähnungen aufrechterhalten wird.

Schlüsselwörter– Neuronale maschinelle Übersetzung; MÜ; Koreferenzauflösung; Erwähnung der Erkennung; Übersetzung auf Dokumentebene; Diskursphänomen; Pronomenübersetzung

Résumé

La traduction automatique (TA) a fait des progrès considérables au cours des deux dernières décennies, en particulier après l'introduction des modèles de réseaux neuronaux (TAN). Pendant ce temps, la communauté scientifique a principalement concentré ses efforts sur la modélisation et l'évaluation de systèmes de TA à l'échelle de la phrase, c'est-à-dire de modèles qui apprennent à traduire de grandes quantités de phrases en parallèle dans différentes langues. L'accent mis sur les phrases permet une simplification pratique de la tâche qui améliore l'efficacité de la TA mais présente l'inconvénient d'ignorer des informations contextuelles pertinentes. Plusieurs études ont montré que l'impact négatif de cette simplification est significatif. Un point clé est la négligence de la dépendance entre des mots distants dans le discours, ce qui entraîne un manque de cohérence et de cohésion dans le texte.

L'objectif principal de cette thèse est d'améliorer la TA en incluant des contraintes au niveau du discours. Nous nous concentrons en particulier sur la traduction des mentions d'entité. Nos contributions peuvent être résumées en quatre points. Tout d'abord, nous définissons un processus d'évaluation et une mesure de performance automatique pour les traductions d'entités (c'est-à-dire noms et pronoms).

Deuxièmement, nous proposons une preuve de concept et analysons l'efficacité d'une traduction qui inclut la résolution de la coréférence d'entité (RC). Nous concluons que la RC améliore la qualité de la traduction en général, et celle des pronoms en particulier, selon le jugement humain.

Troisièmement, nous nous concentrons sur les connexions dans le discours au niveau de la phrase. Nous proposons d'améliorer le modèle séquentiel pour déduire des connexions à long terme en incorporant un mécanisme d' "auto-attention" qui donne un accès direct et sélectif au contexte. Des expériences dans différentes paires de langues montrent que notre méthode surpasse diverses méthodes de référence, et l'analyse confirme que le modèle proposé met l'accent sur un contexte plus large et capture des structures de type syntaxique.

Quatrièmement, nous formulons le problème de la TAN au niveau du document et modélisons les connexions entre les mots dans différentes phrases avec un mécanisme d'attention hiérarchique. Des expériences sur plusieurs ensembles de données montrent une amélioration significative par rapport à deux solides méthodes de référence et nous en concluons que le contexte de la source et de la cible sont complémentaires. Cet ensemble de résultats confirme que la prise en compte du discours améliore considérablement la qualité de la traduction, vérifiant ainsi notre objectif principal de thèse.

Notre objectif secondaire est d'améliorer la RC en modélisant les connexions sous-jacentes entre les entités au niveau du document. Cette tâche est particulièrement difficile pour les modèles de réseaux de neurones actuels car elle nécessite un certain niveau de compréhension et de raisonnement. Tout d'abord, nous proposons une méthode pour détecter les mentions d'entités à partir de données partiellement annotées. Ensuite, nous proposons de modéliser la coréférence avec un graphe d'entités encodées dans un modèle de langage pré-entraîné en tant que structure interne. Les expériences montrent que ces méthodes surpassent diverses méthodes de référence. La RC a le potentiel d'aider la TA et d'autres tâches de génération de textes en maintenant la cohérence entre les mentions d'entité.

Mots clés– Traduction automatique neuronale ; TA ; résolution de coréférence ; détection des mentions ; traduction au niveau du document ; phénomène de discours ; traduction du pronom

Contents

1 Introduction

Machine Translation (MT) is the task of translating text from one natural language to another automatically. Like any other natural language processing task, it presents several challenges. First, there are numerous languages with a distinct alphabet, grammar, lexicon, and syntactic structures. Second, it requires a correct interpretation and understanding of the source text, which is especially difficult for machines. Third, there is not a unique correct answer, as there are multiple ways to convey the same information. Fourth, natural language is ambiguous, and its meaning depends on context (e.g., shared knowledge, environment, time), which sometimes is not part of the system input. This last challenge is of particular interest for the present thesis.

MT is a decades-old problem (Weaver, 1955), and the approaches to solve it have changed over time. We can group them in three main paradigms: Rule-based Machine Translation (RBMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT). RBMT (Nirenburg, 1989) performs translations following a set of rules formulated by experts. These rules are typically based on syntactical, semantic, and morphological knowledge, but they could not be exhaustive enough to contain all natural-language phenomena. On the contrary, data-driven methods are shown to be more robust and flexible to adapt to language. These methods exploit parallel data in a source and target languages to extract translation patterns. SMT (Koehn, 2009), is a data-driven approach that uses statistical methods to learn these translation patterns. In particular, the phrase-based SMT model demonstrates to be relatively efficient and practical. However, the quality of MT reached to the state-of-the-art with NMT (Cho et al., 2014a; Bahdanau et al., 2015). NMT is also a data-driven approach based on neural networks. NMT achieved high translation quality due to its ability to exploit a larger quantity of data. Despite these advantages, data-driven MT has one significant limitation: they are defined at sentence-level, so contextual information at the document-level is dismissed. In the present chapter, we describe in more detail this problem, our objectives and motivations, and the main contributions of this thesis.

1.1 Problem Description

A document is a complete unit of text; thus, corpora for MT is usually created based on parallel documents in a target and source languages (Resnik, 1998; Resnik and Smith, 2003; Koehn, 2005; Ziemski et al., 2016). Nevertheless, data-driven MT was defined since the beginning as the translation of individual sentences (Brown et al., 1990) due to SMT decoders need smaller units of text to be efficient. Therefore, sentence alignment became an active research field (Brown et al., 1991) and the corpora were transformed into parallel sentences. Following MT models, including NMT ones, continue approaching the translation problem at sentence-level. This tendency can be seen in the main events organized by the MT research community. The first Workshop of Machine Translation (WMT) (Koehn et al., 2005) aimed at encouraging data collection for MT, where the parallel alignment was diverse: words, sentences, paragraphs, and documents. Later WMT events aimed at creating MT models based mostly on the sentence-level parallel corpus, so most of the MT research over the past years was under these conditions. Only recently, due to the growing interest of the research community on expanding the context in MT, WMT starting offering parallel data with document alignment annotation (Bojar et al., 2019) again.

The traditional focus on sentences of MT models brings a practical simplification for the task that favors efficiency but also has the disadvantage of missing contextual information while translating a document. This issue cannot be minimized, as MT is commonly used for translating entire documents rather than individual sentences. One key point is that the dependencies among distant words, related to discourse-level constraints, are being ignored. This practice often results in a lack of coherence and cohesion in the translated text. Efforts to approach these important issues started with SMT, the majority of them were focus on specific linguistic phenomena such as: coreference and anaphora (Le Nagard and Koehn, 2010) particularly for pronoun translations (Bojar et al., 2015; Guillou et al., 2016), lexical cohesion (Xiong et al., 2013a), word sense disambiguation (Vickrey et al., 2005), topic adaptation (Ruiz and Federico, 2011), and discourse connectives (Meyer and Popescu-Belis, 2012). The definition of SMT as a document-level task (Xiao et al., 2011; Hardmeier et al., 2012) instead of sentence-level was an important change of paradigm; however, systems were difficult to optimize. When NMT became the state-of-the-art MT, there was yet no effort to use contextual information. Nevertheless, over time, the interest in document-level NMT was growing, specialty because the sentence-level systems reached high quality, so the issues of missing context were more evident and difficult to solve. The present thesis attempts to solve this issue and advance MT towards a more complete and coherent text translation by modeling the underlying coherent-structures at the document level.

1.2 Motivation

1.2.1 *What aspects of discourse phenomena influence translation?*

Discourse refers to the linguistic elements and constructions that give a coherent structure to the text. It covers a variety of linguistic phenomena that potentially impact translation. Here, we analyze three aspects to illustrate the problem.

Coreferences and anaphora

One main issue of sentence-level MT is the translations of coreferences and anaphora. Coreference occurs when two or more expressions in a text refer to the same entity (e.g., person, object). We refer to entity expressions as mentions. Typically the first mention of an entity is fleshed out (*The president of the U.S.*). Later mentions may be expressed with abbreviated descriptions (*the president*) or pronouns (*she*). When two mentions are coreferential, the one in full form is called antecedent, and the one in short form is called anaphor. Coreferent mentions may be situated in different sentences, so translating individual sentences containing anaphora without antecedent can cause ambiguity. The following example of English-to-French illustrates this issue:

Source (EN) :	–	*it was small.*
Target (FR) :	–	*il/elle était petit.*

Here, the translation of the English pronoun *it* to French is ambiguous to either *il* or *elle* depending on the gender of the antecedent object.

Source (EN) : – *The left slipper remained stuck.*
– *it was small.*

Target (FR) : – *La pantoufle gauche est restée coincée.*
– *elle était petit.*

In this example, the antecedent *slipper* is mentioned in a previous sentence; without this information, the MT system can not translate it correctly.

Lexical cohesion

Another problem of the lack of document-level discourse in translation is the lexical cohesion. Cohesion refers to the various linguistic means by which sentences are linked into higher text units. Repetition, synonyms, collocations, and other devices to express relations between parts of text produce cohesion. The impact of cohesion in translation is related to word sense disambiguation and lexical choice. When a word has several senses, those can have different translations, interpreting the correct sense depends on the context. The following example expresses this idea:

Source (EN) : – *Still some bugs . . .*

Target (FR) : – *Encore quelques insects . . .*

Here, the English word *bugs* have at least two different senses: insect or coding error. Each sense is translated differently into French. Context is required to make the correct translation choice:

Source (EN) : – *And the code?*
– *Still some bugs . . .*

Target (FR) : – *Et le code?*
– *Encore quelques bugs . . .*

Coherence

Coherence is the semantic structure that interconnects sentences in a text. This process is defined in terms of the reader experience and understanding. Incorrect translations like the ones observed in previous examples not only express incorrect information to the reader for a particular sentence, but they can also be a source of misunderstanding of the complete text. Here we show an example:

Translation (EN) : – *I bought a very nice set of plates. The next morning, my game was broken. He hadn't even taken it out of the box.*

Reference (EN) : – *I bought a very nice set of plates. The next morning, my set was broken. I hadn't even taken it out of the box.*

Source (ES) : – *Compré un juego de platos muy lindo. A la mañana siguiente, mi juego estaba roto. Ni siquiera lo había sacado de la caja.*

When comparing the MT translation versus a human reference translation, we can see that the translation errors on only two words create a complete sense of misunderstanding of the text.

1.2.2 *How much discourse phenomena impact translation?*

Quantifying the context's impact in translation is difficult because the phenomena are not clearly defined and depend on the text's language, domain, form, and style. Different studies about discourse in translation analytically portray the issues, as in the previous section. However, few studies perform a quantitative analysis of the problem. They focus on a particular corpus and measure some aspects of discourse to show if its effect is significant in translation. Li et al. (2014) analyzes the discourse connectives and their senses in two corpora from Chinese and Arabic to English. They use human-targeted translation error rate (HTER) (Snover et al., 2006) for their analysis (i.e., a measure of edit distance with respect to a human translation reference). They found that the effect of discourse connectives measure higher in Chinese-to-English than Arabic-to-English, but in both cases, translation quality is adversely affected by translations of discourse relations expressed implicitly in one of the languages. Scarton and Specia (2015) analyzes several discourse phenomena in a corpus of English and French in both directions, using HTER. They found that all evaluated discourse features have a high correlation with HTER, meaning that better translation quality is associated with stronger discourse markers. More specifically, correlation with pronouns, connectives, and elementary discourse units are statistically significant. Recent studies suggest that NMT reached human parity in the case of English-to-Chinese (Hassan et al., 2018). However, Läubli et al. (2018) shows that when changing the evaluation conditions from evaluating single sentences to evaluating documents, the human judges changed their assessments and strongly prefer human translation rather than MT translation. The fact that evaluators have access to context makes evident the translation errors related to coherence, which were not visible on single sentence evaluation.

1.3 Challenges

In the previous sections, we have shown that incorporating discourse and document-level context has a huge potential to improve the translation. However, there exist several challenges, as discussed in the following:

Discourse phenomena is difficult to model Previously, we described some aspects of discourse that affect translation, but those are not exhaustive. Research in this area has been active for a while, but its effectiveness is limited. For example, the fact that text coherence depends on the reader's experience makes it difficult to model for an automatic system. The underlying process is not clearly understood or self-contained in the text.

Evaluation of discourse phenomena is challenging Standard automatic metrics to evaluate MT like BLEU score are not designed to assess this aspect. There is research on quality estimation (QE) metrics (Scarton and Specia, 2015; Fonseca et al., 2019) to measure discourse aspects. However, none has been adopted as standard. On the other hand, human evaluation for translation has been designed at sentence-level. Although recently researches noted the necessity to evaluate translation at document level (Läubli et al., 2018), the methods are just starting to be defined and implemented (Bojar et al., 2019). Nevertheless, human evaluation is expensive and non-flexible.

There is a trade-off between the quantity of context and efficiency Including as much context as needed should improve the translation, but this comes at the cost of higher computation and memory requirements. Recently, specialized hardware for neural networks has reached a higher capacity. Hence, including broader contextual information is more feasible now, but it is still limited.

Language understanding is required Current NMT models have a higher capacity to process complicated patterns in the natural language. However, it is still not clear that reasoning and text understanding can be achieved at the human-level. Moreover, document-level information does not necessarily contain all contextual knowledge required to understand the text. Text comprehension usually demands previous world knowledge and environmental context like place, time, objects, and people. This kind of information is not contained in the document.

1.4 Objectives and Contributions

The main objective of this thesis is to improve the automatic translation of documents by including discourse-level constraints. In particular, we focus on the translation of entity mentions, including aspects of coreference/anaphora resolution, lexical cohesion, and coherence. We address this problem in four stages: First, we define the evaluation process to assess entity translations (i.e., nouns and pronouns). Second, we performed a proof-of-concept and analyze how effective it is to include coreference resolution in translation. This preliminary work is performed utilizing SMT models, but the outcomes are applicable for NMT models as well. We concluded that coreference resolution significantly helps the pronoun translation even though it is limited by the coreference resolution models' low performance. Thus, for the next stages, we implicitly model discourse and inter-sentential connections in a generic manner. In the third stage, we focus on the discourse connections at the sentence level. Commonly, sentences are modeled as sequences. When generating text, each produced word is conditioned on the previously predicted information, which is contained in a vector with limited memory capacity. Thus, this modeling approach has a bias for recent information and does not always capture the long-range dependencies between words, e.g., those connecting antecedents and anaphors. We propose to enhance the sequential model to be able to infer long-term connections. For this purpose, we incorporate a "self-attention" mechanism. This mechanism gives direct and selective access to context. It helps the network to focus on specific parts of the sentence useful to predict a particular word. In the fourth stage, we expand the previously described idea to model connections among different sentences. We formulate the problem of document-level NMT and model inter-sentential connections among words with a hierarchical attention mechanism.

The secondary objective of this thesis is to improve the coreference resolution task. This task is particularly challenging for current neural network models because it requires understanding and reasoning. We propose to model coreferences with a graph of entities encoded in the language model as an internal structure. We believe it has the potential to help MT and other NLP tasks that involve text generation by maintaining coherence between the entity mentions.

In summary, the contributions of this thesis are the following:

i) We proposed a metric for automatic evaluation of the accuracy of pronoun translation (APT) and show that it strongly correlates with the human scores. The correlation is considerably higher than general-purpose metrics such as BLEU and METEOR, and other similar metrics. It has been adapted to measure the accuracy of noun translation as well.

ii) We proposed two methods to evaluate the impact of coreference resolution in machine translation: re-ranking and post-editing. Post-editing significantly ameliorate pronoun translation; and human judges prefer post-edit translations over the baseline. We empirically show that coreference resolution can improve translation quality.

iii) We proposed a novel decoder to model the underlying language structures at the sentence level. The decoder uses self-attentive residual learning. The residual connections facilitate the flow of contextual information on the target language side and show consistent improvement over an NMT baseline. The analysis of the self-attention confirms that it emphasizes a broader context and captures syntactic-like structures.

iv) We proposed a hierarchical framework to model discourse in MT. The model captures context and inter-sentence connections in a structured and dynamic manner. We show significant improvement over two strong baselines on multiple data sets and conclude that the source and target sides' contextual information are complementary.

v) We model coreference resolution as a graph problem. The model is non-autoregressive, but it refines the predictions iteratively. We improve the performance of this task in comparison to various baselines.

Each chapter of this thesis is based on a publication. The contributions described here are attributed to the thesis' author. Co-authors of different publications provided insightful advice and discussions but did not contribute directly to the work presented here.

1.5 Thesis Outline

The present thesis is organized into nine chapters. **Chapter 1** (current chapter) presents the problem introduction, motivations, challenges, and thesis contributions. **Chapter 2** contains the technical background describing the SMT, and NMT approaches, relevant discourse phenomena, our evaluation method, and datasets used for the experiments. Our work is divided into two parts:

Part I: Machine Translation This part contains our main work related to discourse phenomena in translation. It is divided into four chapters:

Chapter 3 This chapter describes the definition of the accuracy of pronoun translation (APT) metric. APT is a reference-based metric that works with word alignments. We assess APT with a dataset from a workshop of pronoun translations. We compare the Pearson and Spearman correlation of APT and other metrics with human judgment and found that APT presents a higher correlation.

Chapter 4 This chapter shows how coreference can be useful for translation. We present two alternatives to optimize translations based on coreference scores: re-raking and post-editing. We evaluate our approaches in a Spanish-to-English dataset using APT. We also perform a manual evaluation of the general translation quality. The post-editing approach improves pronoun translation, and human judges prefer it over the baseline.

Chapter 5 This chapter presents our approach to model discourse phenomena at sentence-level with a self-attention mechanism. First, we detail the state-of-the-art NMT baseline. Then, we present our approach and our adaption of similar approaches to translation. The experiments are performed in different language pairs, and we show that our methods outperform all baselines. We present a result analysis explaining possible reasons.

Chapter 6 This chapter presents our approach to model discourse phenomena at the document-level with a hierarchical attention network. We describe the state-of-the-art NMT baseline and our approach. We experiment with three different datasets and present improvements over the baselines. We perform evaluations of lexical cohesion, coherence, and APT.

Part II: Entity Detection and Linking This part describes our complementary work related to the particular discourse phenomena of entity linking.

Chapter 7 This chapter presents our approach for mentions detection. We present two alternative methods for mention detection: sequence tagging model and span scoring model. Afterwards, we define the problem of partially annotated data and present our solution. We evaluate our methods for coreference resolution.

Chapter 8 This chapter presents our approach for coreference resolution. We proposed to model coreference as a graph of entity mentions and links between them. A sequence-to-graph model predict the coreference in an iterative manner. The experiments show improvement over various baselines.

Finally, **Chapter 9** contains the general thesis conclusions and possible future directions.

2 Background

This chapter provides a brief background on machine translation and specific aspects of discourse phenomena that we deem relevant for this thesis. Additional background is given in the context of the individual chapters. Although state-of-the-art machine translation approaches have constantly changed during the development of this thesis, the task definition has not changed. Machine translation has two mayor approaches: statistical machine translation and neural machine translation. In both cases, most of the advances have not considered discourse phenomena at document-level, with important exceptions that are described here.

2.1 Statistical Machine Translation

Statistical machine translation (SMT) (Brown et al., 1990) exploits statistical properties of the text to estimate its translation, in contrast to previous rule-based approaches that require expert knowledge to determine the translation rules. The SMT models' parameters are learned from a corpus of parallel sentences in a given source and target languages. The main idea is to estimate translations based on the frequency of word co-occurrences in the source and target sentences. Intuitively, if two words in different languages co-occur with high frequency in several corresponding parallel source-target sentences, they are presumed to be translations of each other.

Here, a sentence is represented as a sequence of words, more specifically, tokens (i.e., words including punctuations, and special treatment for compound words), and the objective is to find the most probable sequence of tokens $\hat{Y}$ in the target language given a sequence of tokens in the source language X:

$$\hat{Y} = \underset{Y}{\arg\max}\, p(Y|X) \approx \underset{Y}{\arg\max}\, p(X|Y)p(Y) \tag{2.1}$$

The translation is expressed as an optimization problem, but the search space increases exponentially with the length of the sequence. Therefore, by applying Bayes Theorem, the problem is divided into two parts: the translation model $p(X|Y)$, and the language model $p(Y)$. This transformation permits the reduction of the search space by giving priority to explore only well-formed sentences, i.e., sentences with higher probabilities according to the language model. In practice, however, Equation 2.1 is replaced by the following equation using log linear models:

$$\begin{aligned}\hat{Y} &= \underset{Y}{\arg\max}\, \exp(\lambda_1 \log(p(X|Y)) + \lambda_2 \log(p(Y))) \\ &= \underset{Y}{\arg\max} \sum_i \lambda_i f_i(X, Y) \qquad (2.2)\end{aligned}$$

where λ_i is the weight of function f_i, called 'feature'. Under this definition, it is possible to add an arbitrary number of features to the model. Detailed descriptions of SMT models can be found at (Koehn, 2009; Cancedda et al., 2009). Notice that Equation 2.1 is an special case of Equation 2.2 when weights are equal to one.

Research on SMT started in the late 1980s with the word-by-word translation model proposed by IBM (Brown et al., 1990), establishing the foundations of SMT. However, this design makes a strong independence assumption by translating each word disregarding its context (i.e., surrounding words). Phrase-based translation model (Koehn et al., 2003) improved this issue by extending the translation units from words to "phrases". In this case, "phrases" refer to sequences of words that frequently co-occur in a dataset of parallel sentences.

Although SMT was state-of-the-art for several years, neural machine translation showed to be a more powerful approach (Bahdanau et al., 2015). One advantage is that the independence assumption between phrases was removed using recurrent models' ability to encode whole sentences. Additionally, SMT models are generative and make assumptions of independence between its components (e.g., translation model and language model). Thus, the joint optimization of different feature components is problematic, limiting the overall accuracy. On the other hand, neural machine translation models are discriminative and set the problem in an end-to-end fashion; they directly optimize the target objective. Another advantage of NMT is that it incorporates flexibility at translating language pairs with different grammatical order, a point at which SMT struggles. A comparison between phrase-based SMT and NMT in terms of quality of translation (Bentivogli et al., 2016), concludes that NMT generates fewer morphological, lexical, and word order errors. For instance, in the latter case, it produces 70% fewer errors than SMT. In general, neural networks have shown superior performance than other machine learning methods in different NLP tasks, especially when large amounts of data are available.

2.2 Neural Machine Translation

Advances in specialized hardware (Dean et al., 2012) for deep learning and artificial neural networks (Nielsen, 2015; Goodfellow et al., 2016; Aggarwal et al., 2018) make its application to large scale tasks such as machine translation feasible. On the other hand, the later research on machine translation helped deep-learning advance towards better architectures for sequence modeling. The earliest attempts to use neural networks in machine translation were to estimate different features used by SMT models. First, the frequency-based *n-gram* language model (Jelinek, 1980) was replaced by a more accurate neural language model (Bengio et al., 2003; Schwenk et al., 2006). Then, feed-forward neural networks were used to re-score the translation probability of phrases (Devlin et al., 2014) though they required the phrases to have a fixed length. This issue was addressed later on by using recurrent neural networks instead (Cho et al., 2014b). Even though these approaches demonstrated to be effective at improving translation performance, the most important contribution by far was the new paradigm of end-to-end neural machine translation (NMT) (Forcada and Ñeco, 1997; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014b; Liepins et al., 2017). Like SMT, the objective is to find the most probable sequence in the target language $\hat{Y}$ given a sequence in the source language X, but instead of applying chain rule, NMT calculates the direct conditional distribution:

$$\hat{Y} = \underset{Y}{\arg\max}\, p(Y|X) \tag{2.3}$$

In the following sections, we describe the different concepts and elements needed to define NMT that will be useful for developing the following chapters. First, we describe the mapping of words into vector representations, followed by the sequential modeling for constructing sentence representations, and finally, the sequence-to-sequence model setup for NMT. The descriptions here aim to be generic, so specific NMT architectures are detailed in each corresponding chapter according to their use.

2.2.1 Word Representation

Previously, we mentioned that in Natural language processing (NLP) sentences are commonly represented as sequences of words. They result from a word segmentation process that can be, among other algorithms, simple tokenization or a subword segmentation (Sennrich et al., 2016b). A vocabulary V is created based on the word segments obtained from a given dataset. Then, each word segment is encoded as one-hot vector w_t i.e., a vector of size of the vocabulary length, with zero values and a single one at the position i of the vocabulary corresponding to that particular word segment. Given that the one-hot encoding requires a finite vocabulary, tokenization results in omitting infrequent out-of-vocabulary (OOV) tokens. Subword segmentation (Sennrich et al., 2016b) process alleviates this issue because the vocabulary is composed of subword units obtained with an algorithm based on *byte pair encoding* (BPE), so OOV words can be represented through those subunits. Finally, the word representation x_t is obtained by projecting w_t to a different vector space of much smaller dimension:

$$x_t = Ew_t \quad s.t \quad E \in d \times |V|, \quad d << |V| \tag{2.4}$$

where E is called embedding matrix. In later models, the matrix embedding weights are tied with the output layer (Pappas et al., 2018).

2.2.2 Sequence Modeling

Recurrent neural networks (RNNs) offer a principled way to manage sequences. They operate over an input sequence of vectors $\mathbf{x} = (x_1, ..., x_t, .., x_T)$ where T is the sequence length, and t is the time step-index. This network is referred to as recurrent because at a given step time t refers back to a previous state at time $t-1$:

$$h_t = f(h_{t-1}, x_t, \Theta) \tag{2.5}$$

where h is the network's hidden state, f is a non-linear function, and Θ are the parameters of the network. It is important to note that RNNs share parameters across various time steps. The RNN creates a directed cycle that allows the network to have a dynamic temporal behavior, so it can manage variable sequence lengths.

Vanilla RNN The simplest network architecture of RNN is called vanilla RNN or sigmoidal

RNN. Here, the hidden state at time t is defined as follow:

$$h_t = \tanh(W_h[h_{t-1}, x_t]) \tag{2.6}$$

where $W_h \in d_h \times (d_h + d_x)$ is a matrix of parameters. In our notation, all parameter matrices contain both weights and biases. d_h and d_x are the dimensions of the hidden state and input vector respectively. The brackets indicate matrix concatenation. The hyperbolic tangent function tanh is commonly used in RNN's, but other popular non-linear functions are sigmoid σ, and rectified linear unit *RELU* (Glorot et al., 2011).

RNNs suffer from the vanishing gradient problem (Pascanu et al., 2013). During the backward pass, neurons in early time steps are updated with smaller gradients than neurons in later time steps. This issue ultimately leads to having very small or even zero gradients with longer sequences. That means that RNNs have limited learning and memory capacity over time steps. Long short term memory networks were proposed to overcome this problem.

Long short term memory Long short term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) are RNNs with a specialized memory vector c_t and gates to control the reading and writing process over it. The input gate i_t filters useful information from the current time-step input into the memory; the forget gate f_t filters out information from the previous memory state; and the output gate o_t serves to select relevant information for the current time step emission. First, the gates are calculated based on the current input and previous hidden states:

$$i_t = \sigma(W_i[c_{t-1}, h_{t-1}, x_t]) \tag{2.7}$$

$$o_t = \sigma(W_o[c_{t-1}, h_{t-1}, x_t]) \tag{2.8}$$

$$f_t = \sigma(W_f[c_{t-1}, h_{t-1}, x_t]) \tag{2.9}$$

Then, the memory c_t, and recurrent hidden state h_t are updated as follows:

$$\hat{c}_t = \tanh(W_c[h_{t-1}, x_t]) \tag{2.10}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \tag{2.11}$$

$$h_t = o_t \odot \tanh c_t \tag{2.12}$$

LSTMs are the state-of-the-art models for sequence. However, they require more hardware resources i.e., memory and processing power than a vanilla RNN. This issue escalates with the number of stacked layers limiting the practical implementation of very deep LSTM networks. Some studies proposed modifications of LSTM to make them less resource consuming. A popular one is the gated recurrent unit.

Gated recurrent unit (GRU) This is a variant of the LSTM introduced by (Cho et al., 2014b). It simplifies the recurrent layer to make it more efficient. It uses an update gate u_t to control the input and output information in the current time step, and a reset gate r_t to filter out information from previous time steps. They are defined as follows:

$$u_t = \sigma(W_u[h_{t-1}, x_t]) \tag{2.13}$$

$$r_t = \sigma(W_f[h_{t-1}, x_t]) \tag{2.14}$$

The memory is included directly into the hidden state as follows:

$$\hat{h}_t = \tanh(W_c[r_t \odot h_{t-1}, x_t]) \tag{2.15}$$

$$h_t = (1 - u_t) \odot h_{t-1} + u_t \hat{h}_t \tag{2.16}$$

For a while, GRU based networks were a feasible alternative to LSTMs. Later on, an empirically better modeling approach, the transformer networks (Vaswani et al., 2017), was proposed.

Bidirectional RNN The sequential modeling approaches presented so far follows a temporal direction exploiting only information from the past. Nevertheless, it is possible to utilize information from future as well. The objective is to represent each time step input with the complete contextual information. The directed cycle of RNNs runs from left-to-right, so by including a second cycle from right-to-left we get a bidirectional RNN (Schuster and Paliwal, 1997). The bidirectional RNN architecture is composed of a *forward* $\overrightarrow{h}_t$ and a *backward* $\overleftarrow{h}_t$ hidden states. At the end, this two vectors are concatenated to obtain a final hidden state representation as follows:

$$\overrightarrow{h}_t = f(\overrightarrow{h}_{t-1}, x_t, \overrightarrow{\Theta}) \tag{2.17}$$

$$\overleftarrow{h}_t = f(\overleftarrow{h}_{t+1}, x_t, \overleftarrow{\Theta}) \tag{2.18}$$

$$h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t] \tag{2.19}$$

Attention and Transformer Alternatively to the RNN based models discussed earlier based on RNN, self-attentive models capture past or future time step representations with an attention mechanism. The attention mechanism calculates a weighted sum of a set of vectors h as:

$$c_t = \sum_i \alpha_i^t h_i \tag{2.20}$$

$$\alpha_i^t = \frac{exp(f(h_t, h_i))}{\sum_j exp(f(h_t, h_j))} \tag{2.21}$$

where f is a scoring function. This mechanism is used in different architectures to model sequences as explained in Chapter 5. More recently, the transformer (Vaswani et al., 2017) architecture was proposed which is completely based on attention mechanism to model the sequence. This is explained in detail on Chapter 6.

2.2.3 Sequence-to-Sequence Model

The sequence-to-sequence model represents a conditional distribution of emitting a sequence given another $p(Y|X)$. It is typically modeled with an encoder-decoder architecture. The encoder builds representations of the source sequence by projecting the input vectors in context into a continuous space. The decoder then uses those representations as input to project them into the target language vector space. In autoregressive decoding, the output probability of a token in the target language is computed using chain rule:

$$p(Y|X) = \prod_{t=0}^{T} p(y_t|y_0,...,y_{t-1};X) \tag{2.22}$$

Note that there is research on non-autoregressive decoding (Gu et al., 2018) as well, where the whole output sequence is estimated at once. However, this is beyond the scope of the present thesis.

The initial models applied for machine translation (Sutskever et al., 2014; Cho et al., 2014b), used RNNs for the encoder and decoder. The encoder emits an output c_T only in the last time step of the sequence, it can be at the last layer as (Cho et al., 2014b) or at each layer as (Sutskever et al., 2014). This output c_T can be understood as the vector representation of the source sentence in a continuous space, where the assumption is that "similar" sentences are close to each other. Based on this representation, the decoder generates a sequence of tokens in the target language as follows:

$$p(y_t|y_1,...y_{t-1},X) = g(h_t,c_T) \quad s.t. \quad h_t = f(h_{t-1},y_{t-1},c_T) \tag{2.23}$$

where h_t is the hidden state of the decoder at time t, g is a nonlinear function that outputs the probability of y_t, and f is a sequence model. Figure 2.1 shows a basic sequence-to-sequence architecture. One drawback of this model is that while short sentences can be represented in a vector successfully, longer ones may not because the performance of translation degrades with the length of the sentences due to the recency bias of RNNs (Pouget-Abadie et al., 2014).

To avoid the degradation of the translation of long sentences, Bahdanau et al. (2015) introduced an attention mechanism as alignment function between encoder and decoder, allowing the decoder to select at each step which part of the source sentence is more useful to predict the next output symbol. Therefore, instead of a unique sentence representation, the output depends on a context vector c_t that changes at each time step. Equation 2.23 is modified as follows:

$$p(y_t|y_1,...y_{t-1},c_t) = g(h_t,c_t); \quad s.t \quad h_t = f(h_{t-1},y_{t-1},c_t) \tag{2.24}$$

Figure 2.2 shows a sequence-to-sequence with attention architecture

Note that, here, we are concerned only on supervised learning models given that data is available for our purposes, though significant research has been developed on unsupervised NMT (Lample et al., 2017; Artetxe et al., 2018). In the same manner, we describe the models assuming to have only a pair of source-target languages. Nevertheless, there is an interesting development

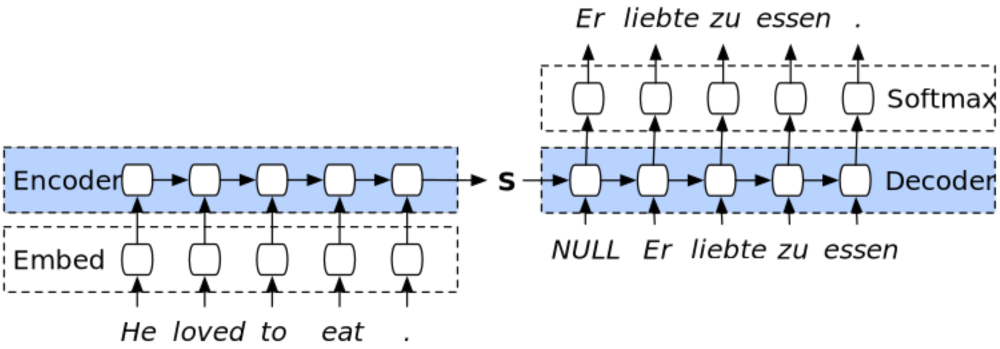


Figure 2.1 – Encoder-Decoder architecture for NMT. *Image taken from* (Merity, 2016)

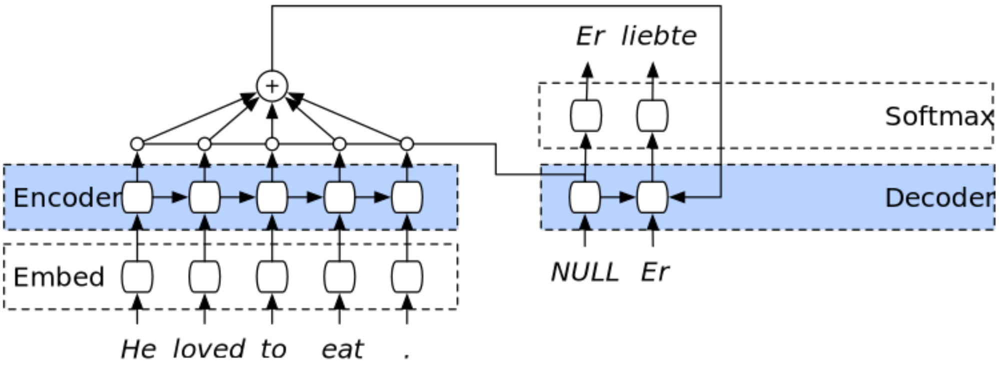


Figure 2.2 – Encoder-Decoder with attention architecture for NMT, *Image taken from* (Merity, 2016)

on multilingual NMT (Johnson et al., 2017; Blackwood et al., 2018) that generalizes the models or some components among different languages, allowing to improve the translation for low resource languages and zero-shot translation.

The machine translation problem that concerns this thesis is that sentences in a text are treated independently, and discourse phenomena are ignored. This problem was pointed out by several scholars (Nakaiwa and Ikehara, 1995; Hardmeier, 2014; Scarton and Specia, 2015) since SMT, and different studies show that omitting discourse phenomena results in lack of lexical cohesion, terminological inconsistency, and sometimes poor word choices (Scarton and Specia, 2015; Läubli et al., 2018).

2.3 Discourse phenomena

Discourse phenomena refer to the linguistic elements and constructions that give a coherent structure to the text. These constructions cross sentence boundaries and are often complex, subtle, and subjective, making it challenging to analyze. Here, we focus on three aspects of discourse relevant for translation: coherence, cohesion, and the particular case of coreferences.

2.3.1 Structure and coherence

In linguistics, coherence is defined as the semantic structure that unifies and interconnects sentences in a text (Bussmann, 2006). It is formed through the interpretations of each sentence in relation to the context, so it involves the reader's mental process and knowledge. Several computational linguistics theories aim to model and interpret coherence in a systematic manner. In particular, the rhetorical structure theory (RST) (Mann and Thompson, 1988) addresses text as rhetoric relations between spans and explains coherence as a hierarchical connected structure that can be modeled in a tree structure. Based on this, discourse parsing (Marcu, 1996, 1997) task was proposed, and later on, shallow discourse parsing (Miltsakaki et al., 2004; Xue et al., 2015). Discourse parsing aims at predicting a tree or graph structure of discourse connections from the text, while shallow discourse parsing predicts relations between pair of sentences independently of other sentences.

Document-level machine translation presents an integrated solution for modeling the discourse structure. Hardmeier et al. (2013) proposed a statistical document-level decoder based on phrase-based MT. It employs a local search approach that scores the translation of an entire document at any time. The optimization is made with a hill-climbing strategy. The document represents a sequence of sentences and a sentence as a sequence of anchored phrases. The initialization is made with a phrase-based MT. At each step, anchored phrases are modified, and the document score is recalculated.

2.3.2 Lexical Cohesion

Cohesion refers to the various linguistics means e.g., grammatical, lexical, phonological, by which sentences are linked into higher text units. Cohesion is produced by 1. the repetition of elements in a text 2. the compacting of the text by relying on context (e.g ellipsis) 3. and linguistic devices to express relations between parts of text (e.g. tense, aspect) (Bussmann, 2006).

In particular, lexical cohesion has been extensively studied to improve SMT (Mascarell, 2017a). One way to do it is by enforcing topic adaptation, for example, with Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA), either in the language model (Woosung Kim and Khudanpur, 2004; Ruiz and Federico, 2011), in the word alignment (Zhao and Xing, 2008), or in the phase table (Tam et al., 2007; Gong et al., 2010; Hasler et al., 2014). Another way is directly modeling lexical changes in document-level SMT (Xiong et al., 2013a,b).

2.3.3 Coreference Resolution

Coreference resolution is the task of grouping together the expressions that refer to the same entity in a text. This task includes two stages: mention identification and coreference resolution. The first stage is usually based on part-of-speech annotation and named-entity recognition. Candidate mentions are usually noun phrases, pronouns, and named entities (Lee et al., 2011). Coreference resolvers follow three main approaches: pair-wise, re-ranking, and clustering. Pair-wise resolvers perform a binary classification, predicting if two mentions refer to the same entity or not. It assumes strong independence of mentions and does not utilize features of the entire entity (Bengtson and Roth, 2008). The second approach lists a set of candidate antecedents for each mention that are simultaneously considered to find the best match. Interpolation between the best and worse candidate is considered (Wiseman et al., 2015; Bengtson and Roth, 2008). Finally, the clustering approach considers the features of a complete cluster of mentions to decide whether a mention belongs or not to a cluster (Clark and Manning, 2015; Fernandes et al., 2012).

The interest on using coreference systems to improve translation emerged while ago (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012). The limited accuracy of coreference resolution may explain its restricted use in MT, although, it is well known that some pronouns require knowledge of the antecedent for correct translation. For instance, Le Nagard and Koehn (2010) trained an English-French translation model where some English pronouns were manually annotated with the gender of their antecedent on the target side. The system translates well 70% of these pronouns, but it did not beat the baseline MT. A model for MT decoding proposed by Luong and Popescu-Belis (2016) used in a probabilistic way several features of the antecedent candidates (e.g., gender, number and humanness values), and demonstrated some improvement in pronouns. Two shared tasks on pronoun-focused translation have been organized (Bojar et al., 2015; Guillou et al., 2016). The results show only marginal improvement of pronoun translation with respect to a baseline SMT system. The best performing systems employed deep neural networks: Luotolahti et al. (2016) and Dabre et al. (2016) summarizing the preceding and following contexts of the pronoun to predict it.

2.4 Evaluation

In this section, we describe automatic evaluation metrics for machine translation and discourse phenomena. Automatic evaluation is more accessible and significantly less costly than human evaluation. However, it is sometimes less precise and extensive. Whenever possible, we have performed a human evaluation to measure the specific objectives of our approaches. They are described in each of the corresponding chapters.

2.4.1 Machine Translation

Automatic metrics for MT compare the translation hypothesis with one or more human reference translations. If the hypothesis is closer to the human references, it is assumed to be a better translation. Here we describe the main ones.

BLEU score (Papineni et al., 2002) is the standard metric for MT. It compares hypothesis and reference translations by counting *n-grams*, with *n* typically from 1 to 4. This yields the calculation of *n-gram* precision scores for each *n*, and the final value is the average logarithm of those scores. The average logarithm is employed because there is an exponential decay with increasing *n*. The recall is not considered, but shorter sentences are penalized. An alternative version of BLEU is NIST (Doddington, 2002) that instead implements an arithmetic average, and weights the *n-grams* according to the frequency of use in general language. A more recent implementation is SacreBLEU (Post, 2018), which standardizes tokenization and normalization schemes to enable stable output scores. BLEU score has several issues, some of them technical (Banerjee and Lavie, 2005; Callison-Burch et al., 2006), e.g., if the precision for one *n* is zero, then the whole score becomes zero. It is also incomplete because it does not calculate recall, and it has no sense of word order. Other issues are methodological (Callison-Burch et al., 2006), exposing the limitations of the BLEU score to correlate with human judgments.

METEOR (Banerjee and Lavie, 2005) was proposed to deal with some weaknesses of the BLEU score. In this case, the comparison between hypothesis and references is made applying unigram matching with word-alignments. Here, precision and recall are calculated, and there is a sense of word order. Moreover, it supports non-exact matching when language-specific resources are available (e.g., stem, Wordnet synonyms, paraphrase table).

Other metrics There has been extensive research on automatic metrics for machine translation (Ma et al., 2019). Some approaches are based on edit distance such as the translation edit rate (TER) from (Snover et al., 2006) or CDER from (Leusch et al., 2006) that edit blocks of words, and the extended edit distance (EDD) from (Stanchev et al., 2019). Other idea is to use character *n-grams* (Popović, 2015; Wang et al., 2016b). More recently, metrics are trained with neural networks to find the best combination of features i.e. word *n-grams*, character *n-grams*, etc. e.g. BEER (Stanojević and Sima'an, 2015) and BLEURT (Sellam et al., 2020). Although these metrics show a higher correlation with human judgment than BLEU, the differences are not significant enough to drop the standard metric, and the fact that researchers have reported the BLEU score for several years perpetuates its utilization for backward comparison.

Test suites Test suites are an alternative way for evaluating translation. In general, they are design to evaluate particular aspects of translation such as discourse. In recent years, the organizers of WMT (Barrault et al., 2019) compile different test suites to evaluate aspects such as document-level translation, gender bias, and domain agreement.

A critical limitation of reference-based metrics is that it is assumed to be wrong if the hypothesis is different from the reference. However, this is not always true because different expressions can transmit the same information. Another limitation is that human effort is required to produce the references, and one can not estimate the quality of translation without a reference. Therefore, researchers have work on Quality Estimation (QE) metrics (Scarton and Specia, 2015; Fonseca et al., 2019) whose objective is to measure translation quality without a human reference and several of them focus on discourse phenomena.

2.4.2 Coreference Resolution

Coreference resolution is typically evaluated in comparison with a gold-standard annotation (Popescu-Belis, 1999; Recasens and Hovy, 2011). The main metrics used for evaluation are: (i) MUC (Vilain et al., 1995), which counts the minimum number of links between mentions to be inserted or deleted in order to map the evaluated document to the gold-standard. (ii) B^3 measure (Bagga and Baldwin, 1998) computes precision and recall for all mentions of a document, while (iii) CEAF (Luo, 2005) computes them at the entity level. (iv) BLANC (Recasens and Hovy, 2011) makes use of the Rand Index, an algorithm for the evaluation of clustering. These metrics are implemented in the scorer for CoNLL 2012 (Pradhan et al., 2014) and the SemEval 2013 one (Màrquez et al., 2013).

2.5 Datasets

NLP community has produced standardized datasets for established tasks such as machine translation and coreference resolution. They are functional in comparing different systems under similar experimental conditions. In this section, we describe the data sources we used for our experiments.

2.5.1 Machine Translation

Datasets for MT consist of parallel sentences on two or more languages, and the test-sets can have more than one human reference to compare. However, multi-parallel datasets are few, and if not mentioned otherwise, the data we use is bilingual with one single human reference for evaluation. We obtain data for our experiments from the following sources:

WMT The Conference on Machine Translation previously called Workshop on Machine Translation (WMT) [1], started in 2005 intending to build multilingual parallel text (Koehn et al., 2005). Until now, this is the most commonly utilized dataset to build and evaluate machine translation models (Bojar et al., 2019). Each year the datasets increase, and different domains are evaluated. However, we only use data from the news translation task. The testing and development sets are created from a sample of online newspapers, while data for training includes different sources such as: (a) Europarl (Koehn, 2005) corpus extracted from the proceedings of the European Parliament which includes different European languages, (b) Common Crawl [2] corpus crawled from web pages, (c) News Commentary [3] (Tiedemann, 2012) corpus with political and economic commentary crawled from the web site Project Syndicate, (d) UN Parallel Corpus (Ziemski et al., 2016) composed of official records and other parliamentary documents of the United Nations, and some others. Since

[1] http://www.statmt.org/

[2] https://commoncrawl.org

[3] http://www.casmacat.eu/corpus/news-commentary.html

2019, WMT (Barrault et al., 2019) offers datasets marked at the document-level for some language pairs such as English-German and English-Czech.

IWSLT The evaluation campaign of the International Workshop on Spoken Language Translation (IWSLT) built a variety of multilingual speech corpora. In particular, we have used the TED and TEDx Talks corpus from 2014 (Cettolo et al., 2014), and 2015 (Cettolo et al., 2015) campaigns. The organizers provide train, development, and test sets.

Subtitles Additionally, we worked with movies and TV programs' subtitles. The Chinese-English subtitles corpus [4] is a compilation of TV program subtitles designed for research on context by (Wang et al., 2018b). The authors proposed splits for testing, development, and training; contrary to other sets, it has three reference translations for evaluation. OpenSubtitles2018 [5] (Lison and Tiedemann, 2016) offers a set of movie subtitles in a variety of languages. However, it is an open corpus, so standard splits for testing, development, and training are not available.

Specific datasets' versions, splits, language, and statistics are described in the corresponding chapter.

2.5.2 Coreference Resolution

Initial datasets for coreference resolution are composed of text documents with annotations of entity clusters, while newer sets have shorter texts and focus only on pair-wise coreference links. In this thesis, we have worked with the English part of the traditional CoNLL 2012 set.

CoNLL 2012 The shared task of the Conference on Computational Natural Language Learning (CoNLL) 2012 (Pradhan et al., 2012) focused on coreference resolution. The organizers built standard test, development, and training sets of the OntoNotes corpus (Hovy et al., 2006; Marcus et al., 2011), which contains newswire data in English, Arabic, and Chinese. Each text document has word-level annotations of part-of-speech, parsing, lemma, word sense, entity name, and speaker; also, document-level annotations for mentions and entity clusters, without considering singleton mentions. The English set has 2802 documents for training, 343 for development, and 348 for testing.

[4] https://github.com/longyuewangdcu/tvsub

[5] http://www.opensubtitles.org

Machine Translation Part I

3 Measuring Accuracy of Pronoun Translation (APT)

This chapter is based on the following paper:

> Miculicich Werlen, L. and Popescu-Belis, A. (2017b). Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics

3.1 Introduction

One of the first steps to include discourse in machine translation is to define an evaluation method. A substantial amount of research has been conducted on the analysis and evaluation of different discourse aspects in translation, but in particular, we are interested in the pronoun translation. This problem has attracted interest of the research community (Hardmeier, 2014; Guillou, 2016), leading to the organization of three shared tasks on the Discourse in Machine Translation (DiscoMT) workshops at 2015 (Hardmeier et al., 2015), 2016 (Guillou et al., 2016), and 2017 (Loáiciga et al., 2017).

Evaluating pronoun translation is challenging due to the interplay between pronouns and antecedents, especially for pro-drop languages (Nakaiwa and Ikehara, 1995) and for the existence of non-referential pronouns. Besides, correct translations often require non-local information, which is a limitation for sentence-level machine translation. It is common to resort to human evaluation, but this type of evaluation comes at a high cost, and in principle, it does not allow repeated evaluations with new candidate sentences. On the other hand, it is considered that automatic evaluation based on reference translations is too restrictive because the amount of legitimate variation is too high. For instance, a particular pronoun can have different valid translations depending on the context or can even be omitted if the target language's grammatical rules allow it. General automatic metrics like BLEU score or METEOR are not sensitive enough to capture improvement in a particular set of POS markers, such as pronouns.

This chapter shows that a simple reference-based metric that estimates the accuracy of pronoun translation (APT) reaches a high correlation with human judgment of quality. The metric includes frequent variations of the pronoun translation obtained from statistics of source-target pronoun occurrences in a parallel corpus. The definition of metric is generic for any set of words in any language pair, but in this work, we define and assess the pronoun translation of English-French using the data from DiscoMT 2015 shared task. This task targets the translation of third-person English pronouns *it* and *they* into French. These pronouns have many possible translations, depending on the referential status of each occurrence, and on the gender and number of its antecedent. The metric compares the candidate translation of each occurrence of *it* and *they* with the reference one. This operation requires a precise alignment of pronouns between these texts. The metric counts the number of identical, equivalent, or different translations in the candidate vs. the reference, and cases when one of the translations is absent or cannot be identified. Several combinations of counts are considered – the most straightforward one gives credit for identical matches and discards all other ones.

Our contributions are the following:

(i) We proposed a metric for automatic evaluation of the accuracy of pronoun translation APT.

(ii) We show that the APT scores correlate strongly with the human scores (0.993–0.999 Pearson and 1.000 Spearman rank correlation), which is considerably higher than general-purpose metrics such as BLEU and METEOR, and other similar metrics.

The rest of the chapter is organized as follows. We present the related work in Section 3.2. In Section 3.3, we define the APT metric, including the alignment procedure and the scoring method. Then, we present the dataset used to validate APT, along with the other metrics and the correlation measures in Section 3.4. Finally, we present the results showing that APT has a higher correlation with human judgment than the other existing metrics in Section 3.5, and our conclusions in Section 3.6.

3.2 Related Work

Initial evaluations for pronoun translation were done by manual inspection. For example, Le Nagard and Koehn (2010) chooses at random a subsample of sentences with occurrences of particular pronouns and reports percentages of correct translations. In order to improve the evaluation efficiency, the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015) created a test suit [1] exclusively to evaluate pronouns from English to French, and compared the candidate translations of different systems with the options deemed correct by human judges previously, so the judges did not see those candidates. The organizers report on a series of metrics, including AutoR, AutoP, and AutoF1 proposed by Hardmeier and Federico (2010). These metrics refer to the recall, precision, and F1 score, respectively, compared to a reference translation. The intuition for using them was that word alignments are not one-to-one so that each pronoun can be linked to multiple elements in the target language. However, Guillou and Hardmeier (2018) later reported that a close examination of this assumption did not show valid samples to support it. We use this metric as a baseline for comparison and describe it in more detail in Section 3.4.2. Later workshops (Guillou et al., 2016; Loáiciga et al., 2017) included additional language pairs for testing i.e., German and Spanish versus English in both directions. They continue reporting on the same automatic metrics, just varying macro-averaged F1 to micro-averaged F1.

After the publication of the present work (Miculicich Werlen and Popescu-Belis, 2017b), additional test suite were created to evaluate particularly challenging cases. Isabelle et al. (2017) proposed a suite of 108 sentences for English-to-French to measure different linguistic aspects, including pronoun translation. Bawden et al. (2018) proposed a contrastive suite from Englis-to-French to evaluate coreference, coherence, and cohesion. A contrastive translation is a translation variant where the correct pronoun is swapped with an incorrect one. Following this line of work, Müller et al. (2018) presented a more extensive test suite of contrastive translations explicitly focused on the translation of pronouns from English-to-German. Compared to automatic metrics, test suites are more accurate at evaluating particular challenging cases where state-of-the-art machine translation fails, as found in (Guillou and Hardmeier, 2018), so wherever possible, they recommend employing them. Nevertheless, the test suites are still limited to a few language pairs.

[1] This test suite was published on (Guillou and Hardmeier, 2016)

3.3 Definition of the APT Metric

In this section, we describe in detail our proposed APT metric. First, we define the used terminology and give an overview of the approach. Then, we explain the pronoun alignments for the triples source-reference-candidate, the scoring process, and finally, the treatment of non-exact matching pronouns.

3.3.1 Terminology

To clarify our terminology, we distinguish *referential* pronouns from non-referential ones, which are also called pleonastic or impersonal. Referential pronouns are also called *anaphoric*, as they point back to a previous item in the discourse, typically but not necessarily a noun phrase, which is called their *antecedent*. An anaphoric pronoun and its antecedent refer to the same (discourse) entity and are therefore *coreferent*. Guillou (2016) argues that a correct translation of pronouns when several options are plausible (i.e., in the case of translation divergences) requires the identification of their function and antecedent (if they are referential). Pronoun and antecedent usually agree in gender and number. The automatic identification of the antecedent of a referential pronoun is called anaphora resolution (Mitkov, 2002).

3.3.2 Overview of the Approach

The APT metric relies on a referential human-translation. It compares a candidate translation (i.e., produced by the MT system) with the human reference to compute the evaluation scores. It uses both source and reference pronouns to consider equivalent and dropped pronouns as valid translation alternatives. This process is performed as follows: Given a word-level alignment of the source, reference, and candidate translations, APT first identifies triples of pronouns: (*source pronoun*, *reference pronoun*, *candidate pronoun*). It then compares each candidate against the corresponding reference, assuming that a pronoun is well translated when it is identical to the reference. (This assumption is validated below by comparing APT scores with human ones, averaged over multiple instances.) Partial matches defined using equivalence classes can also contribute to the score, but these classes depend on the target language and must be defined *a priori*.

'Equivalent' pronouns are those that can be exchanged in most contexts without affecting the meaning of the sentence. For example in Italian, the formal use of the pronouns *egli* and *lui* is analogous to *he* and *him* in English. However, the use of *egli* becomes less common and is being replaced by *lui*. Therefore, these pronouns can be considered as equivalents, though they are not always exchangeable. (e.g. *egli* can never translate *him*). Also, in some languages, one should consider the possibility of identical pronouns with different forms. For example, French has pronoun contractions such as *c'* for *ce*, in the expletive construction *c'est* (meaning *it is*). Or, the German pronoun *es*, translating the English *it*, is sometimes contracted to *'s*. Moreover, the correspondence between source, reference, and candidate pronouns is not always guaranteed.

Pronouns can be missing either from the reference, candidate or from both of them. Plausible reasons for missing pronouns are: human translation choices (omit to translate a source pronoun which the system will most likely try to render), or mismatches in the alignment of pronouns. The percentage of missing pronouns should be small if the alignment is correct. As this condition is not always real, we also propose a simple algorithm to fix missing alignments of pronouns.

3.3.3 Pronoun Alignment

Given the list of source pronouns considered for evaluation, the first step is to obtain their corresponding alignments in the target language texts. In the candidate translation case, the alignment can be directly obtained from the MT system when available. However, in the case of the reference, it is necessary to perform automatic word alignment. We use here the GIZA++ system (Och and Ney, 2003). We attach to the processed sentences a larger corpus to ensure an acceptable accuracy, since GIZA++ has no separate training vs. testing stages. The alignment is made in both directions: source-to-target and target-to-source. The results are then merged using the *grow-diag-final* heuristic from Moses (Koehn et al., 2007).

Accurate pronoun alignment is essential to APT. To estimate its accuracy, we manually evaluated 100 randomly selected sentences from the WIT3 parallel corpus of English-French TED Talks (Cettolo et al., 2012), containing the pronouns *it* and *they*. We found that the alignments of 19 out of 100 pronouns were missing and that 4 pronouns were incorrectly aligned. As expected, the majority of misalignments involved infrequently-used target pronouns.

We defined several pronoun-specific heuristics to improve the alignment. Our four-step procedure is exemplified in Table 3.1 where the alignment between the pronouns *it* and *il* was not identified by GIZA++. First, we identify possible misalignments: source pronouns which are not aligned to any word, or which are aligned to a non-pronoun, or multiple target words. This task can be performed by using a predefined list of pronouns or a POS tagger. If among the multiply-aligned target words, there is a pronoun, then it is considered the alignment. If not, we identify the corresponding alignments (called markers) of the words preceding and following the pronoun (position -1 and +1). Second, we define a range in the target-side neighborhood by considering one word before the first marker and one after the second one, to expand the range of options. Third, we test whether this range includes any likely translations of the source pronoun. Finally, we choose as the aligned word the closest word to the center of the range. The proposed procedure helped address 22 out of the 23 misalignments found in the WIT3 test data described above. The heuristics here requires either POS tagger or a list of predefined pronouns on source and target sides.

3.3.4 Computing APT Scores

The first step of the evaluation is to compare each pair of candidate and reference translations of each source pronoun. We define six cases based on those from a similar metric for discourse

Step	Example
0	E: *The system is so healthy that* ***it*** *purifies the water.* F: *Le système est si sain qu'* ***il*** *purifie l' eau.*
1	E: *The system is so healthy that it purifies the water.* F: *Le système est si sain qu' il purifie l' eau.*
2	F: *Le système est si [sain qu' il purifie l'] eau.*
3	F: *Le système est si [sain*$_2$ *qu'* ***il***$_1$ *purifie* ***l'***$_2$ *] eau.*
4	From the list {*il*, *l'*}, the closest to the center: *il*.

Table 3.1 – Example of applying the heuristics to improve pronoun alignment: *it* in the English source.

connectives (Hajlaoui and Popescu-Belis, 2013): (1) Identical pronouns, (2) Equivalent pronouns, (3) Different or incompatible pronouns, (4) Candidate translation not found, (5) Reference translation not found, (6) Both translations not found. We associate a score or weight to each case. It reflects certainty about the translation in that case. For instance, the first case (candidate identical to reference) is likely a correct translation, so its weight is 1.

Let $C = c_1, .., c_m$ be the set of $m = 6$ cases defined above, n_{c_i} the number of pronoun translation pairs that belong to case c_i, and $w_i \in [0, 1]$ the weight or score associated with case c_i. We denote the subset of discarded cases as $C_d \subseteq C$. The APT score is computed as the number of correctly translated pronouns over the total number of pronouns, formally expressed as:

$$APT = \left(\sum_{i=1, c_i \notin C_d}^{m} w_i n_{c_i} \right) / \left(\sum_{i=1, c_i \notin C_d}^{m} n_{c_i} \right).$$

The APT metric's input parameters are the weights, the discarded cases if any, and the lists of equivalent and identical pronouns in the target language. The case and weights for our experiments on evaluating English to French pronoun translation are set as follows (The setup summary is shown in Table 3.3):

Case 1: Candidate pronouns identical to the reference are considered correct, $w_1 = 1$.

Case 2: In this case, the candidate pronoun is only deemed 'equivalent' to the reference one according to a predefined list (see Section 3.3.5). Counting them always as correct may lead to an indulgent metric, while the contrary might unduly penalize the candidate. We experiment with three options: counted as incorrect ($w_2 = 0$), as partially correct ($w_2 = 0.5$), or as correct ($w_2 = 1$).

Case 3: Candidate pronouns different from the reference are considered as incorrect ($w_3 = 0$).

Case 4: When the reference pronoun is found but not the candidate one, which is then likely absent, the pair is counted as incorrect ($w_4 = 0$), although in some cases omitting a pronoun may still be correct.

Case 5: This is a special scenario because there is no reference pronoun to compare with, therefore we assume two possibilities: either discard these cases or consider them for evaluation. With the second option, case 5 is necessarily considered as incorrect ($w_5 = 0$), but contributes to the denominator in the definition of APT above.

Case 6: Like case 5, we have two possibilities: discard these cases entirely, or evaluate them. If we evaluate them, there are situations when neither the reference nor the candidate translation of a source pronoun could be found, which can often be supposed to be correct, but sometimes reflect complex configurations with wrong candidate translations. Due to this uncertainty, we experiment with three possibilities: counted as incorrect ($w_6 = 0$), as partially correct ($w_6 = 0.5$), or as correct ($w_6 = 1$).

3.3.5 Equivalent Pronouns

The pronouns considered identical were defined based on insights from a French grammar book (Grevisse and Goosse, 2007), which were verified and optimized based on the following quantitative study of observed equivalents.

We built a baseline MT system using Moses (Koehn et al., 2007), and then performed a manual evaluation with 100 randomly selected sentences from the parallel dataset of English-French TED Talks WIT3 (Cettolo et al., 2012), containing the pronouns *it* and *they*. Each pronoun translation was marked as correct or incorrect. The probability of a correct equivalence of different pronouns is defined as $p(c = 1|t, r)$ where t and r are the candidate and reference pronouns, $r \neq t$, and $c \in \{0, 1\}$ corresponds to the manual evaluation (0 incorrect, 1 correct). First we filtered all pairs (t, r) with a frequency of appearance smaller than 5% of the total sample. Then, we calculated the probability by counting the number of correct samples given a particular pair (t, r). Finally, we selected all pairs where $p(c = 1|t, r) > 0.5$, which indicates that the two pronouns are more likely to be correct translation alternatives than not. The final lists found for French are shown in Table 3.2. Two examples of pronoun equivalence in English/French translation are: *"it is difficult ..."* translated to *"il / c' est difficile ..."*, and *"it would be nice ..."* to *"ce / ça serait beau ..."*. Equivalent pronouns require language-specific annotated resources. It is possible to obtain the list of equivalent pronouns from a parallel corpus to obtain each pronoun's most frequent translations. In both cases, the equivalence is context-dependent and not always correct. Thus, we add a weight to this count to balance the cases.

3.4 Experimental Settings

In this section, we describe our experimental setup for assessing APT. We present the DiscoMT dataset, comparable metrics, and the evaluation method.

Identical	Equivalent
ce, c'	*ce, il* ($p = 0.6$)
ça, ç', cela	*ce, ça* ($p = 0.6$)

Table 3.2 – APT lists of identical and equivalent pronouns in French, constructed from a data set where the translation options for *it* and *they* were limited to *il*, *elle*, *ils*, *elles*, *ce*, *on*, *ça*, and *cela*.

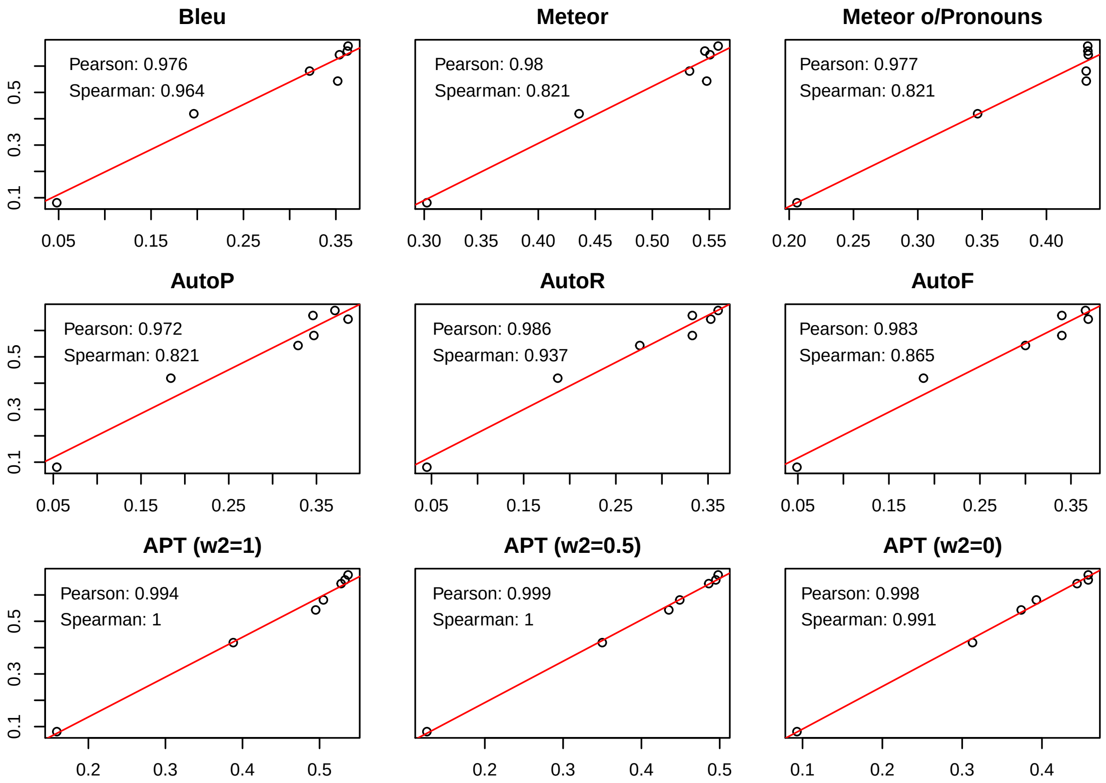


Figure 3.1 – Correlation between the manual evaluation (vertical axis) and different automatic metrics (horizontal axis). The red line is the linear regression model. Pearson's and Spearman's correlations values are showed. The values of APT correspond to the setting: $w_6 = 0$ and $C_d = \{\emptyset\}$ i.e. all cases are counted in the APT score.

Case	Weight	Assumption
c_1 *(identical)*	$w_1 = 1$	candidate = reference $\implies$ correct translation
c_2 *(equivalent)*	$w_2 = \{0, 0.5, 1\}$	candidate ~ reference $\implies$ correct translation
c_3 *(different)*	$w_3 = 0$	candidate $\neq$ reference $\implies$ incorrect translation
c_4 *(not in candidate)*	$w_4 = 0$	incomplete translation $\implies$ incorrect translation
c_5 *(not in reference)*	$w_5 = \varnothing$ $w_5 = 0$	$C_d = \{c_5, c_6\}$: no reference $\implies$ discarded $C_d = \{\emptyset\}$: candidate $\neq$ reference $\implies$ incorrect translation
c_6 *(not in both)*	$w_6 = \varnothing$ $w_6 = \{0, 0.5, 1\}$	$C_d = \{c_5, c_6\}$: no reference $\implies$ discarded $C_d = \{\emptyset\}$: candidate = reference $\implies$ correct translation

Table 3.3 – Settings of the APT metric in our experiments.

3.4.1 DiscoMT Dataset

The data set we use for our experiments was generated during the shared task on pronoun-focused translation at the DiscoMT 2015 workshop (Hardmeier et al., 2015). The systems participating in this task were given 2,093 English sentences to translate into French. The evaluation was focused on the correctness of the translation of the English pronouns *it* and *they* into French. Only a sample of 210 pronouns was manually evaluated for each of the six submitted systems plus a baseline one. The methodology of evaluation was gap-filling annotation: instead of correcting the translation, the annotators were asked to fill the gaps of hidden French candidate sentences with one or more of the following options: *il*, *elle*, *ils*, *elles*, *ce*, *on*, *ça/cela*, *other* or *bad translation*. The accuracy of each submitted translation was calculated with respect to the human annotations using several metrics: accuracy with or without the *other* category, pronoun-specific F-scores (harmonic mean of precision and a lenient version of recall), and general F-score (based on micro-averages of pronoun-specific recall and precision). Additional possible metrics are presented hereafter.

3.4.2 Other Metrics for Comparison

We compare the results of APT with two well-known automatic metrics for MT: BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). Additionally, we include the METEOR score restricted to the French pronouns present in the manual annotation. For this purpose, we set the *function words list* of METEOR to the list of French pronouns defined in DiscoMT (listed above), and its δ parameter to 0 to give preference to the evaluation of the function words (in our case, pronouns).

Additionally, we include the *AutoP*, *AutoR* and *AutoF* metrics proposed by Hardmeier and Federico (2010) for automatic evaluation of pronoun translation. These metrics were inspired by BLEU score. First, they extracts a list C of all words aligned to the source pronouns from the candidate text, and similarly a list R from the reference text. Then, they compute a clipped count of a candidate word w, defined as the minimum value between the number of times it occurs in C and R:

$$c_{clip}(w) = min(c_{C(w)}, c_{R(w)})$$

Finally, all the clipped counts from the words in C are summed up, in order to calculate the precision and recall as follows:

$$AutoP = \sum_{w \in C} c_{clip}(w)/|C|$$

$$AutoR = \sum_{w \in C} c_{clip}(w)/|R|$$

3.4.3 Assessment Method

We use for the assessment of the correlation between each automatic metric and the human judgments the Pearson and Spearman correlation coefficients. Pearson's correlation coefficient r measures the linear dependency between two variables. The formulation we use for our data is:

$$r = \frac{\sum_{i=1}^{n}(h_i - \bar{h})(a_i - \bar{a})}{\sqrt{\sum_{i=1}^{n}(h_i - \bar{h})^2}\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}}$$

where $\{h_1, .., h_n\}$ and $\{a_1, .., a_n\}$ represent the human and automatic scores for the $n = 7$ systems, and $\bar{h}$ and $\bar{a}$ are the means of those scores.

Spearman's rank correlation coefficient is a non-parametric measure of the possibility to express the relation between two variables as a monotonic function. In contrast to Pearson's correlation coefficient, it does not measure the extent to which the metrics are linearly dependent but compares only the rankings resulting from each metric. The formulation we use is the same as for r where we replaced $\{h_1, .., h_n\}$, $\{a_1, .., a_n\}$, $\bar{h}$ and $\bar{a}$ with the rankings given by the human and automatic metrics and their means.

In the pronoun-focused translation shared task at DiscoMT 2015 (Hardmeier et al., 2015), three different human evaluation metrics were used: accuracy including the category *others*, accuracy without *others*, and precision. The organizers selected the first one for the official ranking of the systems, because it allows evaluating the whole sample, and penalizes MT systems that tend to classify many difficult cases as *others*. Therefore, we also use this metric in our correlation experiments hereafter.

3.5 Results Analysis

Here, we show the experimental results and analysis. We compare the correlation coefficients with respect to human evaluation. Then, we analyze the role of the weights and the percentage of errors per case.

3.5.1 Correlation Coefficients

Figure 3.1 shows the correlations of several automatic metrics with the human evaluation scores (i.e. accuracy with *other*, the official DiscoMT 2015 shared task metric): three versions of APT (at the bottom, with $w_2 \in \{0, 0.5, 1\}$), and six previous metrics: BLEU, METEOR (general and restricted to pronouns), and recall/precision/F-score from Hardmeier and Federico (2010). The plots display the values of Pearson's and Spearman's correlation coefficients and the linear regression model fitted for the first coefficient. For all automatic metrics, Pearson's correlation is over 0.97, which is a relatively high value. METEOR has the lowest Spearman correlation, and contrary to what we expected, METEOR evaluated only over pronouns does not perform better than its generic version. Although BLEU and METEOR are not specialized for the evaluation of pronouns, there Pearson's correlation with human judgments is relatively high. These values should be considered as lower bounds when studying metrics dedicated to pronouns. Another interpretation of BLEU and METEOR's high correlations with human judgments of pronouns is that MT systems that are good at translation, in general, are also good at translating pronouns. The performance of the metric proposed by Hardmeier and Federico (2010) is better than generic metrics, especially for recall *AutoR*. Therefore, this specific metric appears to model better the human evaluation for this particular task.

As shown in the lowest row of Figure 3.1, the three tested versions of APT have the best performance, regardless of the weight w_2 given to case 2 occurrences, namely 'equivalent' pronouns. If data for metric tuning were available, we could tune w_2 to reach optimal scores.

Finally, one can argue that the linear correlation between the manual evaluation and the different metrics is inflated because we included a prominent outlier system. This system, coded 'A3-108' in Hardmeier et al. (2015), shows a markedly poor performance at predicting pronouns compared to the other systems. Thus, we also present the correlation values without the outlier, in Table 3.4, and observe that in comparison with the values shown in Figure 3.1, APT remains almost the same while the correlation of the other metrics have a small degradation. Therefore, our conclusions hold regardless of the outlier system.

3.5.2 Role of Weights

Table 3.5 shows the correlation values between APT and other metrics for different values of the weights of cases 2 and 6, with two alignment options. When applying APT with the basic alignment method, always considering equivalent pronouns (case 2) as incorrect translations $w_2 = 0$ has better performance than considering them as partially incorrect $w_2 = 0.5$ or totally correct $w_2 = 1$. The same observation can be made for the weight of case 6, i.e., when considering missing pronoun pairs as correct or not.

	BLEU	**METEOR**	**METEOR o/Pron.**
Pearson	0.902	0.893	0.863
Spearman	0.943	0.714	0.714
	AutoP	**AutoR**	**AutoF**
Pearson	0.923	0.965	0.955
Spearman	0.714	0.919	0.804
	APT $(w_2 = 1)$	**APT** $(w_2 = 0.5)$	**APT** $(w_2 = 0)$
Pearson	0.994	0.999	0.998
Spearman	1.000	1.000	0.989

Table 3.4 – Correlation between the manual evaluation and different automatic metrics without the outlier system. The values of APT are obtained with $w_6 = 0$ and $C_d = \{\emptyset\}$, i.e. all cases are counted in the APT score.

Nevertheless, the situation changes when applying APT with the heuristics for the pronoun alignment described above. Here, the partially correct scenarios present better performance than the others. There is a balanced percentage of correct and incorrect samples for case 2 (as seen in Table 3.6, with heuristic-based alignment), which could explain why $w_2 = 0.5$ leads to a slightly better correlation than other values. On the contrary, all occurrences in case 6 are found to be incorrect according to the manual evaluation. Although this could lead us to set $w_6 = 0$, this does not lead to the best correlation value; a possible explanation is the fact that all MT systems are compared against the same reference. In general, each configuration's differences are too small to lead to firm conclusions about the weights. If more data with human judgments were available, the weights could be optimized on such a set.

3.5.3 Analysis of Scores

Figure 3.2 shows the distribution of cases identified by APT. Most of the samples are identified as case 1 (equal to reference) or case 3 (different from it). This indicates that most candidate translations are either correct or incorrect and that the number of missing pronouns (on either side) is much smaller. Moreover, the heuristics for pronoun alignment help to reduce the number of reference misaligned pronouns (mainly cases 5 and 6, but not exclusively). When comparing the reference and the manual annotation, the proportion of perfect matches increases from 61% to 66% after applying the heuristics.

Table 3.6 shows a breakdown of the comparison between APT scores and manual evaluation into six different cases. The result of the comparison is: *Correct* when the manual annotator's choice of pronoun coincides with the system's translation; *Incorrect* when it does not coincide; and *Bad Translation* when the annotator indicated that the entire sentence is poorly translated and the pronoun cannot be scored. Table 3.6 provides the total number of judgments for the six

	$\mathbf{w_2}$	$\mathbf{w_6}$	**Pearson**	**Spearman**
Basic alignment	0	0	**0.999**	**1.000**
	1	0	0.992	0.987
	0.5	0	0.998	1.000
	1	1	0.994	0.964
	0.5	0.5	0.999	0.987
Alignment with heuristics	0	0	0.998	0.989
	1	0	0.994	1.000
	0.5	0	**0.999**	**1.000**
	1	1	0.995	0.964
	0.5	0.5	**0.999**	**1.000**

Table 3.5 – Correlation between the manual evaluation and APT scores for different values of the parameters of APT, namely the w_2 and w_6 weights of cases 2 and 6.

	Manual Evaluation			
Case	Correct	Incorr.	Bad Tr.	**Total**
$c1$ (same)	**84%**	13%	3%	534
$c2$ (similar)	**43%**	47%	10%	135
$c3$ (different)	26%	**60%**	14%	581
$c4$ (not in candidate)	0%	**76%**	24%	129
$c5$ (not in reference)	53%	**36%**	11%	81
$c6$ (not in both)	0%	**76%**	24%	38
Total	47%	43%	10%	1498

Table 3.6 – Comparison between APT and the manual evaluation for each case identified by APT.

systems and the baseline.

We observe that 84% of the instances in case 1 (candidate identical to reference) are considered correct. Conversely, for case 3 (different pronouns) and case 4 (candidate translation not found), the vast majority of occurrences were indeed judged as incorrect. However, a sizable 26% of case 3 occurrences were considered correct translations by the annotator – presumably due to legitimate variations that cannot be captured by a reference-based metric such as APT. As for case 2 (‘equivalent’ translations), the percentages of correct vs. incorrect translations are nearly balanced. That indicates that the definition of equivalent pronouns is quite problematic, as there are equal chances that ‘equivalent’ pronouns are substitutable or not. Another direction for improvement is the cases with no reference pronoun to compare a candidate: 53% of occurrences in case 5 are considered correct by humans, but APT cannot evaluate them correctly for lack of a comparison term. These cases could be discarded for APT evaluation, but if the goal is to compare several systems with the same reference, they will all be equally penalized by these cases.

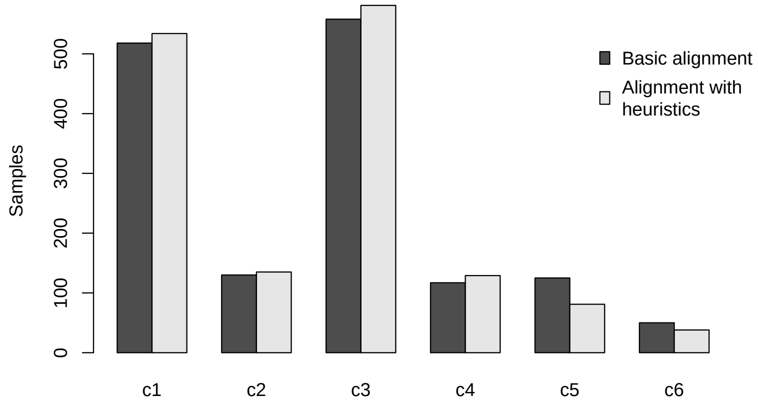


Figure 3.2 – Distribution of pronoun occurrences in each of APT's six cases, with and without heuristics for alignment.

3.6 Conclusion

In this chapter, we have shown that a simple reference-based metric for the accuracy of pronoun translation (APT) has a high correlation with human judgment of correctness, over the scores of seven systems submitted to the DiscoMT 2015 shared task on pronoun-focused translation. While intrinsically, the APT metric seems to set strong constraints on the pronouns' correctness, when averaged over a large number of translations, it appears that improved APT scores reflect quite accurately an improvement in the human perception of pronoun translation quality. A precise alignment of the source and target pronouns, for the reference and the candidate translations, appears to be an essential requirement for APT's accuracy and should be improved in the future. Similarly, a better understanding of 'equivalent' pronouns and their proper weighing in the APT score should improve the quality of the metric and better models of omitting pronouns in translation.

APT has been used for evaluating Spanish-to-English pronoun translation (Rios Gonzales and Tuggener, 2017; Luong et al., 2017; Miculicich Werlen and Popescu-Belis, 2017a), showing that it can be adapted to other language pairs. Despite the correlation between APT and human judgments, (Guillou and Hardmeier, 2018) found that automatic evaluation of pronoun translation misses essential parts of the problem. They point out that APT identifies good translations with relatively high precision, but fails to reward spatially challenging cases. This is a general issue for automatic evaluation. However, the metric serves its propose whenever another kind of evaluation (e.i. human or test suite based) is not feasible.

4 Coreference for Machine Translation

This chapter is based on the following paper:

> Miculicich Werlen, L. and Popescu-Belis, A. (2017a). Using coreference links to improve spanish-to-english machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain. Association for Computational Linguistics

4.1 Introduction

Coreference resolution has the potential to guide machine translation towards a consistent translation of entities. Although several attempts to integrate coreference and translation, they have shown few or none significant improvements over the baselines (Section 4.2). The integration is challenging because translation is defined at the sentence-level while coreference at the document-level. To deal with this issue, we do not modify the translation model but rather rerank or post-edit its output based on coreference scores defined at the document-level.

As good translations should provide the reader the same understanding of entities as the source texts, we propose to use the similarity of coreference links between a source text and its translation as a criterion to improve translation hypotheses. This information should be beneficial to the translation of pronouns, which often depends on their antecedent properties, but should also ensure lexical consistency in the translation of coreferent nouns, pronouns, and noun phrases.

With this work, we provide the first proof-of-concept showing that the coreference criterion can lead to measurable improvements in the translation of referring expressions in Spanish-to-English machine translation (MT). We consider that better translations should have coreference links that are closer to those in the source text, and implement it in two ways. First, we define a similarity measure between source and target coreference structures by projecting the target ones onto the source ones and reusing existing monolingual coreference metrics. Based on this similarity measure, we rerank the translation hypotheses of a baseline MT system for each sentence. Alternatively, to address the lack of diversity of mentions among the MT hypotheses, we focus on mention pairs and integrate their coreference scores with MT ones, resulting in post-editing decisions. Experiments with Spanish-to-English MT on the AnCora-ES corpus show that our second approach yields a substantial increase in pronoun translation accuracy, while BLEU scores remain constant.

Our contributions are the following:

(i) We proposed two methods to evaluate the impact of coreference resolution in machine translation: reranking and post-editing.

(ii) The post-editing is shown to be preferable for human judges with respect to the baseline, and it significantly improves pronoun translation.

Our work is organized as follows. In Section 4.2, we present an overview of related work on coreference and anaphora resolution in translation. In Section 4.3, we define a measure of entity cluster similarity between source and target-side at document-level, and using this measure, we propose to re-rank the translation hypothesis in Section 4.4.3. Alternatively, we propose a post-editing method in Section 4.5. The experiments and results are presented in Section 4.6 and Section 4.7 respectively. Finally, the conclusion is draw in Section 4.8.

4.2 Related Work

Numerous anaphora resolution systems were designed in the past decades (Mitkov, 2002; Ng, 2010), as a consequence, the interest in using them to improve pronoun translation emerged (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012). However, the limited accuracy of anaphora resolution caused those attempts to implement a coreference-aware machine translation model had limited or no impact. Le Nagard and Koehn (2010) trained an English-French translation model on an annotated corpus in which each occurrence of the English pronouns *it* and *they* was annotated with the gender of its antecedent on the target side. Their system correctly translated 40 pronouns out of the 59 that they examined but did not outperform the MT baseline. A model for MT decoding proposed by Luong and Popescu-Belis (2016, 2017) combined several features of the antecedent candidates (gender, number, and humanness) with an MT decoder, in a probabilistic way, and demonstrated small improvement on pronouns. Furthermore, shared tasks on pronoun-focused translation were organized. In the first one (Hardmeier et al., 2015), the improvement of pronoun translation was only marginal with respect to a baseline SMT. In the second one (Guillou et al., 2016), some of the best systems avoided the direct use of anaphora resolution (except Luong et al. (2015a)). For example, (Callin et al., 2015) designed a classifier based on a feed-forward neural network, which is considered a feature all preceding nouns and determiners along with their part-of-speech tags. The winning systems (Luotolahti et al., 2016; Dabre et al., 2016) summarized the contexts of the pronoun with a recurrent neural network, so they assumed that anaphora resolution could be made in an unsupervised manner. Later to the publication of this chapter, Rios Gonzales and Tuggener (2017) integrated coreference resolution with phrase-based SMT to solve the problems of elided subjects and drop-pronouns in Spanish-to-English. At the same time, Luong et al. (2017) showed that pronoun translations improved Spanish-to-English when anaphora probabilities were included as features in an SMT system.

4.3 Coreference Resolution for MT

A principle of translation is that the information conveyed in a document should be preserved in its translation. Here, we focus on the referential information, i.e., the coreference links between mentions. If we apply coreference resolution to a source text and a faithful translation of it, then the mentions' grouping should be identical. We thus formulate the following criterion for MT: *better translations should have coreference links that are more similar to the source.* Table 4.1 illustrates the above criterion on an example of Spanish-to-English translation, extracted from the AnCora-ES corpus (Recasens and Martí, 2010),[1] with source coreference chains coming from the AnCora-ES annotations. The automatic translation comes from a commercial online MT system. The Stanford Statistical Coreference Resolution system (Clark and Manning, 2015)[2] was applied to both translations, and the resulting coreference chains are indicated in the table with

[1] http://clic.ub.edu/corpus/

[2] http://stanfordnlp.github.io/CoreNLP/coref.html

numbers and colors. We observe that the chains in the human translation match well those in the source, but this is less the case for the automatic translation, in particular, due to wrong pronoun translations. Although the MT output is still understandable, this requires more time than human translation, due to the wrong set of coreference links inferred by the reader.

In what follows, we will implement a proof-of-concept coreference-aware MT system for Spanish-to-English translation. This pair is particularly challenging because Spanish is a pro-drop language so that an MT system must not only select the correct translation of pronouns, but it must also generate English pronouns from Spanish null ones. In this study, in order to avoid introducing errors made by the coreference resolution system, we will always use on the source side the gold-standard coreference annotation from AnCora-ES (Recasens and Martí, 2010), which was used in the SemEval-2010 Task 1 on coreference resolution in multiple languages (Recasens et al., 2010).[3] As our proposal does not require specific training on coreference-annotated data, AnCora-ES will be used for testing only. On the target side, as coreference resolution must be performed for each translation hypothesis, we must use an automatic system. One advantage of the Spanish-to-English direction is that English coreference resolution systems have been studied and developed for a long time, more than any other language, thus keeping coreference errors minimum. We use again the Stanford Statistical Coreference Resolution system proposed by Clark and Manning (2015). Moreover, to obtain pairwise mention scores, needed in Section 4.5, we use the code of the pairwise classifier available with the source code of the Stanford CoreNLP toolkit (Manning et al., 2014)[4].

4.4 Reranking Approach

We propose to use the document-level coreference similarity score to rerank the n-best translation hypotheses of each sentence in the text. The n-best translation hypotheses can be obtained with any MT system.

4.4.1 Coreference Similarity Score

To obtain a coreference similarity score, we first apply coreference resolution to the source and candidate translation and compare both with the source playing the role of the ground-truth or gold-standard. Then, we used traditional metrics for evaluating coreference resolution: MUC, B^3, and CEAF-m (see Section 2.4.2). Traditional metrics were designed to compare texts in the same language and not across different languages. This issue raises difficulties for matching the referring expressions (i.e., mentions, or markables). So we propose to project the mentions of the target text back to the source text so that each word in the source is aligned with its corresponding translation (one or more words). This alignment can be obtained directly from the Moses MT system (see the start of Section 4.4.3). There is not always a one-to-one word correspondence

[3] http://stel.ub.edu/semeval2010-coref/

[4] Source class 'edu.stanford.nlp.scoref.PairwiseModel' at http://stanfordnlp.github.io/CoreNLP/.

Source	Human Translation	Machine Translation
La película narra la historia de [un joven parisiense]$_{c_1}$ que marcha a Rumanía en busca de [una cantante zíngara]$_{c_2}$, ya que [su]$_{c_1}$ fallecido padre escuchaba siempre [sus]$_{c_2}$ canciones.	The film tells the story of [a young Parisian]$_{c_1}$ who goes to Romania in search of [a gypsy singer]$_{c_2}$, as [his]$_{c_1}$ deceased father used to listen to [her]$_{c_2}$ songs.	The film tells the story of [a young Parisian]$_{c_1}$ who goes to Romania in search of [a gypsy singer]$_{c_2}$, as [his]$_{c_2}$ deceased father always listened to [his]$_{c_2}$ songs.
Pudiera considerarse un viaje fallido, porque [$\emptyset$]$_{c_1}$ no encuentra [su]$_{c_1}$ objetivo, pero el azar [le]$_{c_1}$ conduce a una pequeña comunidad...	It could be considered a failed journey, because [he]$_{c_1}$ does not find [his]$_{c_1}$ objective, but the fate leads [him]$_{c_1}$ to a small community...	It could be considered [a failed trip]$_{c_3}$, because [it]$_{c_3}$ does not find [its]$_{c_3}$ objective, but the chance leads $\emptyset$ to a small community...

Table 4.1 – Comparison of coreference clusters in the Spanish source vs. English human and machine translations. English clusters were obtained with the Stanford coreference resolver (Manning et al., 2014). The clusters are numbed $c_1, \ldots, c_n$, and are also color-coded. The void symbol $\emptyset$ indicates a null subject pronoun in Spanish. The third coreference clusters (c_3) in the MT output is erroneous.

between the words in the source and target sentences, and word order also differs. Thus, we apply the following heuristic to improve the cross-language mapping of the mentions. Through word-alignment, the words that comprise the mentions may have changed its original order in the translation. Thus, we take the first and last words on the target side that align with any word of the mention in the source and assume that all words in between are also part of the mention. The null pronouns are transferred to the next immediate verb, and we refine the alignment to be sure these verbs are aligned to the generated pronoun in the target. Once the target mentions are mapped to the source, we apply the MUC, B^3, and CEAF-m coreference similarity metrics from the CoNLL 2012 scorer between the source document d_s and the projected target one d_t. To mitigate individual variations, we use the average of the three scores at the similarity criterion and note it $C_{sim}(d_t, d_s)$.

4.4.2 Coreference and Translation Quality

To validate the insight that better translations correlate with better coreference similarity scores, we present in Table 4.2 the MUC, B^3 and CEAF scores of a human translation vs. two systems: the Moses baseline phrase-based MT system used below and an online commercial MT system. The source is a set of documents with ca. 3.5 thousand words with gold-standard coreference annotation from AnCora-ES. The English translation was done manually. On the target side, we applied the Stanford automatic coreference resolution system (Manning et al., 2014). By definition, the best translation is the human one. Then, according to the BLEU score measured

Metric	Translation	Recall	Precision	F1
MUC	Human	31	46	37
	Commercial MT	21	38	28
	Baseline MT	18	33	23
B^3	Human	24	49	32
	Commercial MT	20	38	26
	Baseline MT	17	40	24
CEAF	Human	41	40	41
	Commercial MT	34	39	36
	Baseline MT	32	35	33

Table 4.2 – Coreference similarity scores (%) between source and target texts for different translations. The scores increase with the quality of translations.

on the same set of documents, the second-best translation is done by the commercial MT with 49.4, and the last one by the baseline MT with 43.7. We observe that the coreference scores also decrease in this order, and they decrease consistently for the three evaluation metrics. These results thus support the principle that translation quality and coreference similarity are correlated. We will now show how to use this principle to improve translation quality.

4.4.3 Reranking Hypotheses

Our goal is to find a combination of sentence translations that optimize the document-level coreference similarity score defined before. First, we use the Moses toolkit to build a phrase-based statistical MT system (Koehn et al., 2007), with training data from the translation task of the WMT 2013 workshop (Bojar et al., 2013). The English-Spanish training set consists of 14 million sentences, with approximately 340 million tokens. The tuning set is the *News Test 2010-2011* one, with ca. 5,500 sentences and almost 120k tokens. We built a 4-gram language model from the same training data augmented by ca. 5,500 sentences monolingual data from *News Test 2015*. Our baseline system has a BLEU score of 30.8 on the *News Test 2013* with 3,000 sentences.

A translated document d_t is represented as an array of translations $d_t = (s^1, s^2, ..., s^M)$, where each sentence can be selected from a list of n-best translation hypotheses $s^i \in \{s^i_1, s^i_2, ..., s^i_N\}$. The objective is to select the best combination of hypotheses based on their coreference similarity C_{sim} with the source, i.e.:

$$\underset{h_1,h_2,..,h_M}{\arg\max}\, C_{sim}((s^1_{h_1}, s^2_{h_2}, ..., s^m_{h_M}), d_s)$$

To limit the decrease of sentence-level translation scores when optimizing the document-level objective, we keep track of the former and select the sentences with the best translation scores if they lead to the same C_{sim}. This combinatorial problem is expensive, so we try to reduce the search space to allow reasonable performance. First, we filter out candidate sentences. In this approach, the essential variations in the translation are entity mentions. Thus sentences are modeled as sets of mentions, and duplicate sets are filtered out. Second, we apply beam search optimization. Based on the fact that the first mentions of entities usually contain more information than the next ones, the beam search starts from the first sentence and aggregates at each step the translation hypothesis with the highest similarity scores with the preceding ones. We foresee several limitations of this approach. First, with a sentence containing several mentions, there is no guarantee that the n-best hypotheses include a combination of mention translations that optimize all mentions simultaneously. What is worse, the correct translation of a given mention may not be present at all among the n-best hypotheses because the differences among the top hypotheses are often minimal, especially when sentences are long. In order to solve these problems, we present a second approach.

4.5 Post-editing Approach

This approach differs from the previous one in two aspects. First, it uses hypotheses of translation of individual coreferent mentions rather than complete sentences. It allows us to optimize the translation of each mention independently and increase the variety of hypotheses of each mention. Second, coreference resolution is applied only on the source side. So, instead of searching for similar clustering on the target side, we try to induce it. The selection of the best translation hypothesis of a mention is based on a cluster-level coreference score. First, we apply automatic coreference resolution at source side to obtain the mentions and clusters, and translate the source document. Second, we find the aligned words of each source mention in the target translation. Then, we create a new document by coping the translated text and replacing the mentions with the source version. Thus, the new document is in the target language except for the mentions that are written in the source language. Then, we make a second pass translation using the new document so we obtain translation candidates only for the mentions (other words are already in the target language). Finally, we choose the translations hypothesis that correlates better with other mentions in the same cluster (obtained in the first step). This automatic method improves the performance because it uses coreference resolution only once instead of multiple times, and as shown in the experimental section, it is more effective at improving the translation of mentions.

4.5.1 Candidate Translations

It is essential to include the surrounding context in the translation to obtain the n-best translation hypotheses of the entity mentions. Otherwise, an independent translation could lead to the construction of invalid or erroneous sentences. We would like to have an MT system that brings hypotheses corresponding only to mentions and fix the translations of other words so that we can interchange the hypotheses of one mention in the same text. Building such an MT system would require a significant modification of the baseline. As an alternative solution, we will simply perform two passes of MT. The first pass is a simple translation of the text. Then, the mentions are identified in the target text, and their source-language version replaces them. That results in a mixed language text. Then, this text is passed to the MT system for the second time, so that the system identifies and translates only the words in the source language. Nevertheless, the language and reordering models are still going to evaluate the complete sentence. To avoid any translation of the context words (i.e., not mentions) in the second pass, we filter out from the translation table all words not corresponding to mentions. It is important to note that we consider only the heads of mentions obtained from the parse tree (this annotation is included in AnCora corpus) to avoid long mentions such as the ones with subordinate clauses, and focus on the most crucial part of each mention.

4.5.2 Cluster-level Coreference Score

In this approach, we rely on the coreference resolver applied to the source side to define mentions' clusters. Each cluster is defined as a set of mentions $c_x = \{m^i, m^j, .., m^k\}$, where each mention can be selected from a set of translation hypotheses $m^i \in \{m^i_1, m^i_2, ..., m^i_N\}$. By definition, the mentions in a cluster represent the same entity. Thus, they have to correlate in features such as gender, number, and animation. To achieve this objective on the target side, we define a cluster-level coreference score C_{ss}. It represents the likelihood that all mentions in that cluster belong to the same entity. So, for each given cluster, we select the combination of translation hypotheses of mentions with a higher cluster-level coreference score. This combinatorial problem is expensive. Therefore, it is simplified with a beam search approach. Mentions are processed one at a time. The translation hypotheses of a new upcoming mention are compared with each of the previously selected ones. Then, the combinations with lower C_{ss} are pruned. The algorithm continues in the same manner until it processes the last mention.

In order to compare two mentions, we use the mention pair scorer from (Clark and Manning, 2015). It uses a logistic classifier to assign a probability to a pair of hypotheses, which represents the likelihood that they are coreferent. The pair score is defined as follows:

$$p_{pair}(m^i_{h_i}, m^j_{h_j}) = (1 + e^{\theta^T f(m^i_{h_i}, m^j_{h_j})})^{-1}$$

where $f(m^i_{h_i}, m^j_{h_j})$ is a vector of feature functions of the mentions and θ is the vector of feature weights. Finally, we define the cluster-level coreference score C_{ss} as the product of the individual pairwise probabilities:

$$C_{ss}(c_x) = \prod_{m^i \in c_x} \prod_{m^{i \neq j} \in c_x} p_{pair}(m^i_{h_i}, m^j_{h_j})$$

We illustrate this idea with an example. Here, we have a sentence in Spanish and its translation to English. We show one coreference cluster c_1 formed by three mentions:

Source (es): *La alcaldesa de Málaga y cabeza del* [*partido*]$_{c_1}$ [*que*]$_{c_1}$ *ganó en esta ciudad, pidió a los militantes de* [*este partido político*]$_{c_1}$*...*

Target (en): *The mayor of Malaga and head of the* $[m_1]_{c_1}$ $[m_2]_{c_1}$ *won in this city, asked the militants of this* $[m_3]_{c_1}$ *to...*

In this example, the three marked mentions have the following translation hypotheses: $m_1 \in \{match, party\}$, $m_2 \in \{who, which\}$, and $m_3 \in \{political\ party\}$. We calculate the pairwise score p_{pair} of each combination and show the results in the following table.

m_1, m_2	$(match, who) = 0.03, (match, which) = \mathbf{0.35}$,
	$(party, who) = 0.01, (party, which) = \mathbf{0.26}$
m_1, m_3	$(match, political\ party) = 0.08$,
	$(party, political\ party) = \mathbf{0.53}$
m_2, m_3	$(political\ party, who) = 0.12$,
	$(political\ party, which) = \mathbf{0.27}$

Finally, we find that the set of translation hypotheses with the highest cluster-level coreference C_{ss} score is {'party', 'which', 'political party'}, with a score of 0.04. Intuitively, we can verify that this final combination is the best solution for the example.

4.5.3 Joint Scoring

The proposed score guides the system to select translation hypotheses, which are more likely to refer to the same entity in a cluster. In order to enhance the decision process, we include two sources of additional information: the translation frequency, that can help to decide between synonyms by selecting the most frequently translated one; and information of the entity in the source side, which enriches the knowledge of the entity. The information about the frequency of translation can indicate how well a particular hypothesis translates the mention. Therefore, we define a translation score, T_s, at mention-level. The translation score of a hypothesis is calculated based on its relative frequency of emission by the MT system, as follows:

$$T_s(m^i_{hi}) = count(m^i_{hi}) / \sum_j count(m^i_j)$$

The information about the entity in the source side can indicate how well a particular hypothesis represents it. Thus, we define a simple representation of an entity by setting relevant features such as gender, number, and animation. The features are extracted and summarized from all mentions in the cluster. That is a naive representation and more advanced work on entity-level representations have been performed concerning coreference resolution (Clark and Manning, 2016; Wiseman et al., 2016), which could be applied here in the future. Having an entity representation, we define a simple scoring function which measures how well a candidate represents an entity with respect to other alternatives:

$$E_s(m^i_{hi} = f(m^i_{h_i}, \theta_{e_x}) / \sum_j f(m^i_j, \theta_{e_x})$$

where f is a linear function and θ_{e_x} are the entity features.

Finally, the decision is made by combining the three previous scores: cluster-level coreference, translation, and entity matching. As one additional step, we adjust the coreference score to the same scale as others:

$$C_s = C_{ss}(m^i_{h_i}, m^j_{h_j}, \ldots) / \sum_{x,y,\ldots} C_{ss}(m^i_x, m^j_y, \ldots).$$

The final score is defined as follows:

$$\begin{aligned} C_{score}(m^i_{h_i}, m^j_{h_j}, \ldots) = & C_s(m^i_{h_i}, m^j_{h_j}, \ldots)^{\lambda_1} \times \\ & [T_s(m^i_{h_i}).T_s(m^j_{h_j})\ldots]^{\lambda_2} \times \\ & [E_s(m^i_{h_i}).E_s(m^j_{h_j})\ldots]^{\lambda_3} \end{aligned}$$

where $\sum_i \lambda_i = 1$ are predefined hyper-parameters of the function. The final set is given by:

$$(m^i, m^j, \ldots) = \underset{h_i, h_j, \ldots}{\arg\max} \quad C_{score}(m^i_{h_i}, m^j_{h_j}, \ldots).$$

These three hyper-parameters were optimized on a different subset of AnCora-ES than the one used for evaluation. The optimized values are λ_1=0.5, λ_2=0.1, and λ_3=0.4.

4.6 Experimental Setting

Our initial experiments' objective is to measure how much coreference can improve the correct choices of translation of mentions and the impact of these choices on global translation quality. We translated 10 sample documents from the test set to serve as reference translations for evaluation. The evaluation of global MT quality is made with the well-known BLEU *n*-gram precision metric (Papineni et al., 2002), while the evaluation of mentions, being less standardized, is performed in several ways. We reuse previous insights on pronoun translation and therefore score them with a metric that automatically computes the accuracy of pronoun translation (APT) in terms of the number of identical pronouns vs. different from a human reference translation (Miculicich Werlen and Popescu-Belis, 2017b)[5]. More originally, to provide a complete view of the performance, we compute the "accuracy of noun translation" (ANT) by reusing the same idea as in APT to count the number of exactly matched nouns between MT and the reference translation.

We test the two proposed methods, reranking and post-editing vs. the phrase-based statistical MT (PBSMT) baseline described in Section 4.4.3. We also include a neural machine translation (NMT) baseline (Bahdanau et al., 2015) as a reference for comparison. We chose to build our systems over a PBSMT system for simplicity because the word-alignment can be obtained directly from the system. Additionally, we also present the results obtained with an automatic coreference resolver in the source side, namely the CorZu system (Tuggener, 2016; Rios, 2015), for the

[5] https://github.com/idiap/APT

System	BLEU	APT	ANT
Baseline PBSMT	46.5±4.3	0.35±0.07	0.78±0.08
Baseline NMT	46.9±3.7	0.37±0.07	0.78±0.07
PBSMT + Re-rank	41.7±3.9***	0.40±0.10*	0.74±0.01**
PBSMT + Post-edit	46.4±3.9	0.59±0.13***	0.78±0.07
PBSMT + Post-edit + Automatic coreference	46.1±4.3	0.41±0.07*	0.76±0.09

Table 4.3 – Comparison of baseline MT and our proposals for reranking or post-editing, for three metrics. In addition to the average scores and standard deviation over the ten test documents, we indicate the statistical significance level of the difference between each of our systems and the baseline (* for 95.0%, ** for 99.0% and *** for 99.9%).

post-editing approach.

4.7 Results Analysis

In this section, we describe and analyze the experimental results with both automatic and human evaluation.

4.7.1 Automatic Evaluation

Table 4.3 shows the results of the experiments. We first calculate BLEU, APT, and ANT values at the document-level and show the average and standard deviation values for the three evaluated systems: baseline and our two proposed approaches. Additionally, we show the significance levels (t-test) of the results in comparison to the baseline. The post-editing approach improves the pronoun translation quite significantly, without decreasing the overall quality of translation. This improvement is demonstrated by the rise of the APT score, whereas the BLUE score remains without significant change. However, the quality of the translation of nouns does not change significantly, as shown by the ANT.

The reranking approach shows a significant increase in the quality of pronoun translation. Nevertheless, the overall quality of translation decreases significantly, as well as the quality of noun translation. The limitations of this approach can explain these results. The optimization was done by considering the correlation of mentions, but the changes were made at the sentence-level. The overall quality of translation at the sentence-level was not considered. A combination of coreference similarity and translation probability could be used in the future to address the described problem.

Figure 4.1 shows the distribution of pronouns translated by the three evaluated systems (i.e., baseline, reranking, and post-editing) compared to the reference. The number of pronouns equal to the reference increases for both proposed approaches, especially for the post-editing. The pronouns that improve the most were the third-person personal and possessive ones. Also, the

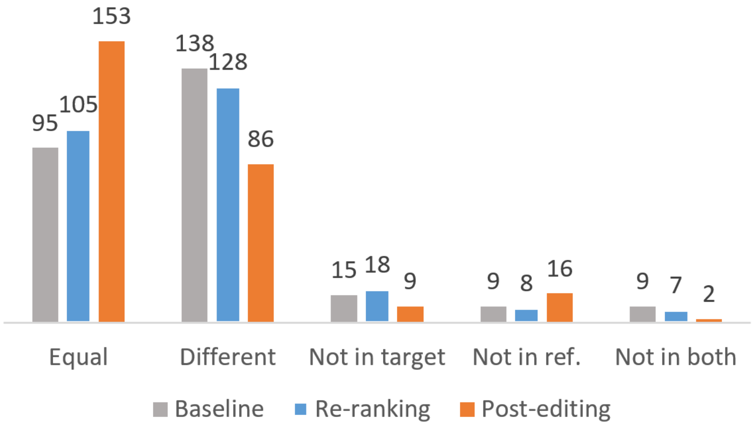


Figure 4.1 – Pronoun translation in comparison with the reference: numbers of equal vs. different pronouns for the three systems, including also missing pronouns in target, reference, and both sides (counts based on source pronouns).

Evaluation	**Scoring points**	**Baseline**		**Re-rank**		**Post-edit**	
		(#)	**(score)**	**(#)**	**(score)**	**(#)**	**(score)**
Incorrect	0	53	0	55	0	21	0
Acceptable	1	21	21	19	19	28	28
Correct	2	115	230	115	230	140	280
Total score			251		249		308

Table 4.4 – Manual evaluation of fourth randomly selected documents. The evaluation was done over nouns and pronouns.

translation of some of the null pronouns in the source was improved. The association with other mentions of the same entity and the entity's representation coming from the source side was important for this improvement.

4.7.2 Human Evaluation

Finally, we perform manual evaluation by examining source mentions, as annotated over AnCora-ES, and evaluating their translations by the baseline MT along with the two approaches presented above (in Sections 4.4 vs. 4.5). When presented to the evaluator, the three translations of each source sentence are provided in a random order, so that the evaluator does not know to which system they belong. The evaluator assigned a score of '2' to a translation identical to the reference, '1' for translation that is different but still good or acceptable, and '0' to a wrong or unacceptable translation. To minimize the time spent on manual evaluation at this stage, one evaluator rated four test documents.

Table 4.4 shows the results of the manual evaluation, scored as explained above, which includes

Example 1: Correct pronoun gender

S: [Barton]$_3$, por [su]$_3$ parte , también dudó de la capacidad de [Megawati]$_2$ en [su]$_3$ [nueva tarea]$_4$.

R: [Barton]$_3$, for [his]$_3$ part , also doubted [Megawati]$_2$'s ability in [her]$_2$ [new task]$_4$.

B: [Barton]$_3$, for [its]$_3$ part , also doubted the capacity of Megawati in [his]$_2$ [new task]$_4$.

P: [Barton]$_3$, for [his]$_3$ part , also doubted the capacity of [Megawati]$_2$ in [her]$_2$ [new task]$_4$.

Example 2: Correct drop-pronoun

S: ... que " [parece estar]$_2$ abrumada ... críticos consideran que [no será]$_2$ capaz de hacerse con el papel de líder .

R: ...that " [she seems]$_2$ overwhelmed ... critics consider [she will not be]$_2$ able to take the lead role .

B: ... that " [appears to be]$_2$ overwhelmed ... critics believe that [it will not be]$_2$ able to take a leading role .

P: ...that " [she seems]$_2$ to be overwhelmed ... critics believe that [she will not be]$_2$ able to take a leading role .

Example 3: Incorrect drop-pronoun

S: - ¿ [Es]$_1$ iconoclasta por valenciano ? - .

R: - [Are you]$_1$ iconoclastic by Valencian ? - .

B: - [Is]$_1$ an iconoclast by Valencian ? - .

P: - [he is]$_1$ an iconoclast by Valencian ? - .

Table 4.5 – Examples of source, reference, baseline and post-edited sentences.

nouns and pronouns together. In general, it supports the results of the automatic evaluation. Here, the post-editing approach has 32 less mentions scored as "wrong" than the baseline, 7 of them were score as "acceptable", and the rest 25 as identical to the reference. The reranking approach, despite the theoretical appeal of its definition, fails to improve noun and pronoun translation. Table 4.5 shows examples of translations obtained with our approaches. The noun translations by the baseline are good. The discrepancies with respect to the reference are, in many cases, due to synonyms and acronyms. Still, some nouns suffer from sense ambiguity, which may be improved by our method. However, this particular test set is too small and does not contain enough instances of this type to evaluate their translations with certainty.

4.8 Conclusion

We presented two methods for evaluating the impact of including coreference scores in translation, based on the coreference similarity of source and translated texts. Both approaches are shown to improve the pronoun translations but not the noun translations. However, the reranking decreases the whole translation quality because the sentence-level optimization makes it difficult to maintain the balance between coreference and translation scores. The post-editing approach brought a significant improvement from Spanish-to-English pronoun translation without degradation of the translation quality. Even though the BLEU scores of the baseline and the post-editing were similar, the human evaluation determined that the post-editing translations were preferable in a significant number of cases. The results presented here were later confirmed on larger data sets (Rios Gonzales and Tuggener, 2017; Luong et al., 2017). However, post-editing has several restrictions. First, the process of obtaining candidate translations requires a second pass of the text to the MT system, and sporadically, the correct translations are not part of the hypotheses. Second, the final probabilities are a combination of independent probabilities, which requires extra hyperparameters. We believe that these problems can be addressed by integrating coreference resolution with MT in an unsupervised or semi-supervised manner. In that way, the access to hypotheses is not restricted, and the optimization can be done by taking into account all affecting probabilities at the same time while decoding.

5 Self-Attention for Neural Machine Translation

This chapter is based on the following paper:

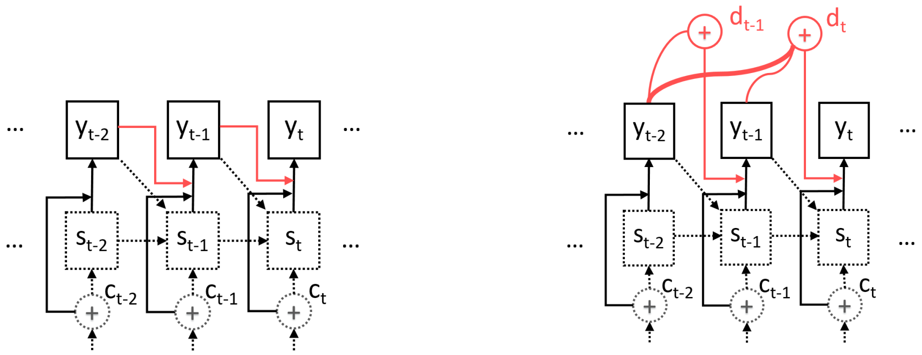
> Miculicich Werlen, L., Pappas, N., Ram, D., and Popescu-Belis, A. (2018). Self-attentive residual decoder for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1366–1379, New Orleans, Louisiana. Association for Computational Linguistics

5.1 Introduction

In the previous chapters, we focused on coreference and anaphora resolution for improving the translation of nouns and pronouns. We defined an evaluation metric and showed that coreference resolution effectively boosts the pronoun translation producing sentences preferred by human judges. Nevertheless, we also concluded that without full integration of discourse phenomena on the translation, the impact is limited. In addition to that, the advances in neural machine translation (NMT) led us to aim at incorporating discourse as an integral part of the neural network. Neural networks are flexible, so we hypothesize that they can learn the text's intrinsic discourse phenomena when adding the correct mechanism for this purpose. In this chapter, we work on the discourse context and semantics present at the sentence-level leaving open to the model to learn any structure necessary, including but not limited to coreferences.

At the time of writing this chapter, the attention-based NMT model designed by Bahdanau et al. (2015) was considered the de-facto baseline, achieving remarkable performance with respect to the ancestor phrase-based SMT. One of the reasons for their effectiveness is their ability to capture relevant source-side contextual information at each time-step prediction through an attention mechanism. However, the target-side context is solely based on the sequence model, which, in practice, is prone to recency bias and cannot capture effectively non-sequential dependencies among words. This architecture comprises two recurrent neural networks (RNNs), an encoder and a decoder, and an attention mechanism between them for modeling a soft word-alignment. First, the model encodes the complete source sentence and then decodes one word at a time. The decoder has access to all the context on the source side through the attention mechanism. However, on the target side, the contextual information is represented only through a fixed-length vector, namely the decoder's hidden state. As observed by Bahdanau et al. (2015), this creates a bottleneck, which hinders the sequential model's ability to learn longer-term information effectively. Additionally, such RNN models do not fully capture the structural composition of language Cheng et al. (2016). Several models have been proposed to address these limitations, namely memory networks (Cheng et al., 2016; Tran et al., 2016; Wang et al., 2016a) and self-attention networks (Daniluk et al., 2016; Liu and Lapata, 2018). We experimented with these methods, applying them to NMT: *memory RNN* (Cheng et al., 2016) and *self-attentive RNN* (Daniluk et al., 2016). However, we observed no significant gains in performance over the baseline architecture.

In this work, we propose a self-attentive residual recurrent decoder, presented in Figure 5.1b, which, if unfolded over time, represents a densely-connected residual network. The self-attentive residual connections focus selectively on previously translated words and propagate useful information to the decoder's output, within an attention-based NMT architecture. The attention paid to the previously predicted words is analogous to a read-only memory operation and enables the learning of syntactic-like structures useful for the translation task. Our evaluation over three language pairs shows that the proposed model improves over several baselines, with only a small increase in computational overhead. In contrast, other similar approaches have lower scores but a higher computational overhead. The decoder's analysis of the attention learned confirms that it



(a) Baseline NMT decoder

(b) Self-attentive residual dec.

Figure 5.1 – Comparison between the decoder of the baseline NMT and the proposed decoder with self-attentive residual connections.

emphasizes a broader context and captures syntactic-like structures.

Our contributions are the following:

(i) We propose and compare several options for using self-attentive residual learning within a standard decoder, facilitating the flow of contextual information on the target side.

(ii) We demonstrate consistent improvements over a standard baseline and two advanced variants that use memory and self-attention on three language pairs (English-to-Chinese, Spanish-to-English, and English-to-German).

(iii) We perform an ablation study and analyze the learned attention function, providing additional insights on its actual contributions.

The rest of the chapter is organized as follows: Section 5.2 describes the relevant related work and Section 5.3 the baseline. Our self-attentive approach is detailed in Section 5.4, and alternative self-attentive models are presented in Section 5.5. The experiments and results are discussed in Sections 5.6 and 5.7 respectively. Finally, the conclusions are draw in Sections 5.8.

5.2 Related Work

Several studies have been proposed to enhance sequential models by capturing longer contexts. Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is the most commonly used recurrent neural network (RNN), because its internal memory allows to retain information from a more distant past than a vanilla RNN. Other studies attempt to increase the memory capacity of LSTMs by using memory networks (Weston et al., 2015; Sukhbaatar et al., 2015). For instance, Cheng et al. (2016) incorporate different memory cells for each previous output representation, which are later accessed by an attention mechanism. Tran et al. (2016) include a memory block to access recent input words selectively. Both methods show improvements in language modeling. For NMT, Wang et al. (2016a) presented a decoder enhanced with external shared memory. Memory networks extend the network's capacity and have the potential to read, write, and forget information. Our method, which attends over previously predicted words, can be seen as a read-only memory, which is simpler but computationally more efficient because it does not require additional memory space. Other studies aim to improve the modeling of source-side contextual information, for example through a context-aware encoder using self-attention (Zhang et al., 2017), or a recurrent attention NMT (Yang et al., 2017) that is aware of previously attended words on the source-side to better predict which words will be attended in the future. Additionally, variational NMT (Zhang et al., 2016a) introduces a latent variable to model the underlying semantics of source sentences. In contrast to these studies, we focus instead on the contextual information *on the target side*.

The application of self-attention mechanisms to RNNs have been previously studied, and in general, they seem to capture syntactic dependencies among distant words (Liu and Lapata, 2018; Soltani and Jiang, 2016; Lee et al., 2017b; Lin et al., 2017). Daniluk et al. (2016) explore different approaches to self-attention for language modeling, leading to improvements over a baseline LSTM and memory-augmented methods. However, the methods do not fully utilize a more extended context. Our approach's main difference is that we apply attention to the output embeddings rather than the hidden states. Thus, the connections are independent of the recurrent layer representations, which is beneficial to NMT, as we show below.

Our model relies on residual connections, which have been shown to improve deep neural networks' learning process by addressing the vanishing gradient problem (He et al., 2016). These connections create a direct path from previous layers, helping the transmission of information. Several architectures using residual connections with LSTMs have been proposed for sequence prediction (Zhang et al., 2016b; Kim et al., 2017; Zilly et al., 2017; Wang and Tian, 2016). To our knowledge, our study is the first one to use self-attentive residual connections within residual RNNs for NMT. In parallel to our study, a similar method was recently proposed for sentiment analysis (Wang, 2017).

5.3 Neural Machine Translation with Attention

Neural machine translation aims to compute the conditional distribution of emitting a sentence in a target language given a sentence in a source language, denoted by $p_\Theta(y|x)$, where Θ is the set of parameters of the neural model, and $y = \{y_1, ..., y_n\}$ and $x = \{x_1, ..., x_m\}$ are respectively the representations of source and target sentences as sequences of words. The parameters Θ are learned by training a sequence-to-sequence neural model on a corpus of parallel sentences. In particular, the learning objective is to maximize the following conditional log-likelihood:

$$\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} \log(p_\Theta(y|x)) \tag{5.1}$$

The models typically use gated recurrent units (GRUs) (Cho et al., 2014b) or LSTMs (Hochreiter and Schmidhuber, 1997). Their architecture has three main components: an encoder, a decoder, and an attention mechanism.

The goal of the encoder is to build meaningful representations of the source sentences. It consists of a bidirectional RNN which includes contextual information from past and future words into the vector representation h_i of a particular word vector x_i, formally defined as $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$. Here, $\overrightarrow{h_i} = f(x_i, h_{i-1})$ and $\overleftarrow{h_i} = f(x_i, h_{i+1})$ are the hidden states of the forward and backward passes of the bidirectional RNN respectively, and f is a non-linear function.

The decoder (see Figure 5.1a) is in essence a recurrent language model. At each time step, it predicts a target word y_t conditioned over the previous words and the information from the encoder using the following posterior probability:

$$p(y_t|y_1, ..., y_{t-1}, c_t) \approx g(s_t, y_{t-1}, c_t) \tag{5.2}$$

where g is a non-linear multilayer function. The hidden state of the decoder s_t is defined as:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{5.3}$$

and depends on a *context vector* c_t that is computed by the attention mechanism.

The attention mechanism allows the decoder to select which parts of the source sentence are more useful to predict the next output word. This goal is achieved by considering a weighted sum over all hidden states of the encoder as follows:

$$c_t = \sum_{i=1}^{m} \alpha_i^t h_i \tag{5.4}$$

where α_i^t is a weight calculated using a normalized exponential function a, also known as *alignment function*, which computes how good is the match between the input at position $i \in \{1, ..., n\}$ and the output at position t:

$$\alpha_i^t = softmax(e_i^t) \tag{5.5}$$

$$e_i^t = a(s_{t-1}, h_i) \tag{5.6}$$

Different types of alignment functions have been used for NMT, as investigated by Luong et al. (2015b). Here, we use the one originally defined by Bahdanau et al. (2015).

5.4 Self-Attentive Residual Decoder

The decoder of the attention-based NMT model uses a skip connection from the previously predicted word to the output classifier to enhance translation performance. As we can see in Equation (5.2), the probability of a particular word is calculated by a function g which takes as input the hidden state of the recurrent layer s_t, the representation of the previously predicted word y_{t-1}, and the context vector c_t. Within g, these quantities are typically summed up after going through simple linear transformations. Hence the addition of y_{t-1} is indeed a skip connection as in residual networks (He et al., 2016). In theory, s_t should be sufficient for predicting the next word given that it is dependent on the other two local-context components, according to Equation (5.3). However, the y_{t-1} quantity makes the model emphasize the last predicted word for generating the next word. How can we make the model consider a broader context?

To answer this question, we propose to include into the decoder's formula skip connections not only from the previous time step y_{t-1}, but from all previous time steps from y_0 to y_{t-1}. This defines a residual recurrent network that, unfolded over time, can be seen as a densely connected residual network. These connections are applied to all previously predicted words and reinforce the recurrent layer's memory towards what has been translated so far. At each time step, the model decides which of the previously predicted words should be emphasized to predict the next one. To deal with the dynamic length of this new input, we use a target-side summary vector d_t that can be interpreted as the representation of the decoded sentence until the time t in the word embedding space. We therefore modify Equation (5.2) replacing y_{t-1} with d_t:

$$p(y_t|y_1,...,y_{t-1},c_t) \approx g(s_t,d_t,c_t) \tag{5.7}$$

The replacement of y_{t-1} with d_t means that the number of parameters added to the model is dependent only on the calculation of d_t. Figure 5.1b illustrates the change made to the decoder. We define two methods for summarizing the context into d_t, described in the following sections.

5.4.1 Mean Residual Connections

One simple way to aggregate information from multiple word embeddings is by averaging them. This average can be seen as the sentence representation until time t. We hypothesize that this representation is more informative than using only the embedding of the previous word. Formally:

$$d_t^{avg} = \frac{1}{t-1}\sum_{i=1}^{t-1} y_i \tag{5.8}$$

5.4.2 Self-Attentive Residual Connections

Averaging is a simple and cheap way to aggregate information from multiple words, but may not be sufficient for all kinds of dependencies. Instead, we propose a dynamic way to aggregate information in each sentence, such that different words have different importance according to their relation with the prediction of the next word. We propose to use a shared self-attention mechanism to obtain a summary representation of the translation, i.e., a *weighted average representation* of the words translated from y_0 to y_{t-1}. This mechanism aims to model, in part, important non-sequential dependencies among words, and serves as a complementary memory to the recurrent layer.

$$d_t^{cavg} = \sum_{i=1}^{t-1} \alpha_i^t y_i \tag{5.9}$$

$$\alpha_i^t = softmax(e_i^t) \tag{5.10}$$

The weights of the attention model are computed by a scoring function e_i^t that predicts how important each previous word ($y_0, ...,$ or y_{t-1}) is for the current prediction y_t.

We experiment with two different scoring functions, as follows:

$$e_i^t = v^\intercal tanh(W_y y_i + W_s s_t) \ \ \textit{(content+scope)} \tag{5.11}$$

$$\text{or} \ \ e_i^t = v^\intercal tanh(W_y y_i) \ \ \textit{(content)} \tag{5.12}$$

where $v \in \mathbb{R}^e$, $W_y \in \mathbb{R}^{e \times e}$, and $W_s \in \mathbb{R}^{e \times d}$ are weight matrices, e and d are the dimensions of the embeddings and hidden states respectively. Firstly, we study the scoring function noted *content+scope*, as proposed by Bahdanau et al. (2015) for NMT. Secondly, we explore a scoring function noted as *content*, which is calculated based only on the previous hidden states of the decoder, as proposed by Pappas and Popescu-Belis (2017). In contrast to the first attention function, which makes use of the hidden vector s_t, the second one is based only on the previous word representations, therefore, it is independent of the current prediction representation. However, the normalization of this function still depends on t.

5.5 Other Self-Attentive Networks

To compare our approach with similar studies, we adapted two representative self-attentive networks for application to NMT.

5.5.1 Memory RNN

The *Memory RNN* decoder is based on the proposal by Cheng et al. (2016) to modify an LSTM layer to include a memory with different cells for each previous output representation. Thus at each time step, the hidden layer can select past information dynamically from the memory. To adapt it to our framework, we modify Equation (5.3) as:

$$s_t = f(\tilde{s}_t, y_{t-1}, c_t) \tag{5.13}$$

$$\text{where} \qquad \tilde{s}_t = \sum_{i=1}^{t-1} \alpha_i^t s_i \tag{5.14}$$

$$\alpha_i^t = softmax(e_i^t) \tag{5.15}$$

$$e_i^t = a(h_i, y_{t-1}, \tilde{s}_{t-1}) \tag{5.16}$$

5.5.2 Self-Attentive RNN

The *Self-Attentive RNN* is the simplest one proposed by Daniluk et al. (2016), and incorporates a summary vector from past predictions calculated with an attention mechanism. Here, the attention is applied over previous hidden states. This decoder is formulated as follows:

$$p(y_t|y_1, ..., y_{t-1}, c_t) \approx g(s_t, y_{t-1}, c_t, \tilde{s}_t) \tag{5.17}$$

$$\text{where} \qquad \tilde{s}_t = \sum_{i=1}^{t-1} \alpha_i^t s_i \tag{5.18}$$

$$\alpha_i^t = softmax(e_i^t) \tag{5.19}$$

$$e_i^t = a(s_i, s_t) \tag{5.20}$$

Additional details of the formulations in Sections 5.3, 5.4, and 5.5 are described in the Appendix A.

5.6 Experimental Settings

Here we describe the datasets and general model setup.

5.6.1 Datasets

To evaluate the proposed MT models in different conditions, we select three language pairs with increasing amounts of training data: English-Chinese (0.5M sentence pairs), Spanish-English (2.1M), and English-German (4.5M).

For English-to-Chinese, we use a subset of the UN parallel corpus (Rafalovitch and Dale, 2009)[1], with 0.5M sentence pairs for training, 2K for development, and 2K for testing. For training Spanish-to-English MT, we use a subset of WMT 2013 (Bojar et al., 2013), corresponding to Europarl v7 and News Commentary v11 with ca. 2.1M sentence pairs. Newstest2012 and Newstest2013 were used for development and testing respectively. Finally, we use the complete English-to-German set from WMT 2016 (Bojar et al., 2016) with a total of ca. 4.5M sentence pairs. The development set is Newstest2013, and the testing set is Newstest2014. Additionally,

[1] http://www.uncorpora.org/

we include as testing sets Newstest2015 and Newstest2016, for comparison with the state of the art. We report translation quality using (a) BLEU over *tokenized* and *truecased* texts, and (b) NIST BLEU over *detokenized* and *detruecased* texts[2].

5.6.2 Model Configuration

We use the implementation of the attention-based NMT baseline provided in `team2016theano`[3] developed in Python using Theano (Team et al., 2016). The system implements an attention-based NMT model, described above, using one layer of GRUs (Cho et al., 2014b). The vocabulary size is 25K for English-to-Chinese NMT, and 50K for Spanish-to-English and English-German. We use the byte pair encoding (BPE) strategy for out-of-vocabulary words (Sennrich et al., 2016b). For all cases, the maximum sentence length of the training samples is 50, the dimension of the word embeddings is 500, and the dimension of the hidden layers is 1,024. We use dropout with a probability of 0.5 after each layer. The models' parameters are initialized randomly from a standard normal distribution scaled to a factor of 0.01. The loss function is optimized using Adadelta (Zeiler, 2012) with $\epsilon = 10^{-6}$ and $\rho = 0.95$ as in the original paper. The systems were trained in 7–12 days for each model on a Tesla K40 GPU at the speed of about 1,000 words/sec.

5.7 Results Analysis

Table 5.1 shows the BLEU scores and the number of parameters used by the different NMT models. Along with the NMT baseline, we included a statistical machine translation (SMT) model based on Moses (Koehn et al., 2007) with the same training/tuning/test data as the NMT. The performance of *memory RNN* is similar to the baseline. As confirmed later, it mainly focuses the attention on the prediction at $t-1$. The *self-attentive RNN* method is inferior to the baseline, which can be attributed to the overhead on the hidden vectors that have to learn the recurrent representations and the attention simultaneously. The proposed models outperform the baseline, and the NMT model obtains the best scores with *self-attentive residual connections*. Despite their simplicity, the *mean residual connections* already improve the translation, without increasing the number of parameters.

Tables 5.2 and 5.3 show further experiments with the proposed methods on various English-German test sets, compared to several previous systems. Table 5.2 shows BLEU values calculated by *multi-bleu*, and includes the NMT system proposed by Luong et al. (2015b) which replaces unknown predicted words with the most strongly aligned word on the source sentence. Also, the table includes other systems described in Section 5.2. Additionally, Table 5.3 shows values calculated by the NIST BLEU scorer, as well as results reported by the "Winning WMT" systems for each test set respectively: UEDIN-SYNTAX (Williams et al., 2014), UEDIN-SYNTAX

[2] Scrips from Moses toolkit (Koehn et al., 2007): BLEU *multi-bleu*, NIST BLEU *mteval-v13a.pl*, *tokenizer.perl*, *truecase.perl*.

[3] https://github.com/nyu-dl/dl4mt-tutorial

Model	|Θ|	BLEU En–Zh	BLEU Es–En
SMT baseline	–	21.6	25.2
NMT baseline	108.7M	22.6	25.4
+ Memory RNN	109.7M	22.5	25.5
+ Self-attentive RNN	110.2M	22.0	25.1
+ Mean residual connections	108.7M	23.6	25.7
+ Self-attentive residual connections	108.9M	**24.0**	**26.3**

Table 5.1 – BLEU score (multi-bleu) on *tokenized* text. The highest score per dataset is marked in bold. The self-attentive residual connections make use of the *content* attention function. |Θ| indicates the number of parameters per model.

Model	BLEU NT14	BLEU NT15
NMT (unk. word repl.) (Luong et al., 2015b)	20.9	–
Context-aware NMT (Zhang et al., 2017)	22.57	–
Recurrent attention NMT (Yang et al., 2017)	22.1	25.0
Variational NMT (Zhang et al., 2016a)	–	25.49
NMT baseline	22.3	24.8
+ Memory RNN	22.6	24.9
+ Self-attentive RNN	22.0	24.3
+ Mean residual connections	22.9	24.9
+ Self-attentive residual connections	**23.2**	**25.5**

Table 5.2 – BLEU score (multi-bleu) on *tokenized* text for English-to-German on *Newstest (NT) 2014, and 2015*. The highest score per dataset is marked in bold. The self-attentive residual connections makes use of the *content* attention function.

(Williams et al., 2015), and UEDIN-NMT (Sennrich et al., 2016a). Also, we include the results reported by Sennrich et al. (2016b) for a baseline encoder-decoder NMT with BPE for unknown words similar to our configuration, and finally, the system proposed by Nadejde et al. (2017), an explicit syntax-aware NMT that introduces combinatory categorial grammar (CCG) super tags on the target side by predicting words and tags alternately. The comparison with this work is relevant for the analysis described later in Section 5.7.5. The results confirm that the *self-attentive residual connections* improve the translations significantly. To evaluate the significance of the improvements against the NMT baseline, we performed a one-tailed paired t-test.

5.7.1 Impact of the Attention Function

We now examine the two scoring functions that can be used for the *self-attentive residual connections* model presented in Equation (5.11), considering English-to-Chinese and Spanish-to-English. The BLEU scores are presented in Table 5.4: the best option is the *content* matching

	BLEU (NIST)		
Model	**NT14**	**NT15**	**NT16**
Winning WMT	20.1	24.4	**34.2**
NMT (BPE) (Sennrich et al., 2016b)	–	22.8	–
Syntax NMT (Nadejde et al., 2017)	–	–	29.0
NMT Baseline	21.0	24.4	28.8
+ Mean residual connections*	21.4	24.7	29.6
+ Self-attentive residual connections**	**21.7**	**25.0**	29.7

Table 5.3 – NIST BLEU scores on *detokenized* and *detruecased* text for English-to-German on *Newstest (NT) 2014, 2015, 2016.* Significance test: * $p < 0.05$, ** $p < 0.01$. The Winning WMT systems are listed in the text below.

	BLEU	
Attention function	**En-Zh**	**Es-En**
Content+Scope	23.1	25.6
Content	**24.0**	**26.3**

Table 5.4 – BLEU scores for two scoring variants of the attention function of the proposed decoder.

function, which depends only on the word embeddings. The *content+scope* function, which depends additionally on the hidden representation of the current prediction, is better than the baseline but scores lower than *content*.

The idea that the context's importance depends on the current prediction is appealing because it can be interpreted as learning internal dependencies among words. However, the experimental results show that it does not necessarily lead to the best translation. On the contrary, the *content* attention function may be extracting representations of the whole sentence, which are easier to learn and generalize.

5.7.2 Performance According to Human Evaluation

Manual evaluation on samples of 50 sentences for each language pair helped to corroborate the conclusions obtained from the BLEU scores, and to provide a qualitative understanding of the improvements brought by our model. For each language, we employed one evaluator who was a native speaker of the target language and had good knowledge of the source language. The evaluators ranked three translations of the same source sentence – one from each of our models: *baseline*, *mean residual connections*, and *self-attentive residual connections* – according to their translation quality. The three translations were presented in a random order so that the system that had generated them could not be identified. To integrate the judgments, we proceed in pairs and count the number of times each system was ranked higher, equal to, or lower than

	Ranking (%)								
Models	**En–Zh**			**Es–En**			**En–De**		
	>	=	<	>	=	<	>	=	<
Mean vs. Baseline	26	56	18	20	64	16	28	58	24
Self-attentive vs. Baseline	28	60	12	28	56	16	32	54	14
Self-attentive vs. Mean	24	62	14	28	58	14	32	56	12

Table 5.5 – Human evaluation of sentence-level translation quality on three language pairs. We compare the models in pairs, indicating the percentages of sentences that were ranked higher (>), equal to (=), or lower (<) for the first system with respect to the second one. The values correspond to percentages (%).

Model	d	**Perplexity**
LSTM (Daniluk et al., 2016)	300	85.2
LSTM + Attention (Daniluk et al., 2016)	296	82.0
LSTM + 4-gram (Daniluk et al., 2016)	968	75.9
LSTM + Mean residual connections	296	80.2
LSTM + Self-attentive residual connections	296	80.4

Table 5.6 – Evaluation of the proposed methods on language modeling. The number of parameter for all models is 47M.

another competing system. The results shown in Table 5.5 indicate that the *self-attentive residual connections* model outperforms the one with *mean residual connections*, and both outperform the baseline, for all three language pairs. The rankings are thus identical to those obtained using BLEU in Tables 5.1 and 5.3.

5.7.3 Performance on Language Modeling

To examine whether language modeling (LM) can benefit from the proposed method, we incorporate the residual connections into a neural LM. We use the same setting as Daniluk et al. (2016) for a corpus of Wikipedia articles (22.5M words), and we compare it with two methods proposed in the same paper, namely attention LSTM and 4-gram LSTM. As shown in Table 5.6, the proposed models outperform the LSTM baseline as well as the self-attention model, but not the 4-gram LSTM. Experiments using 4-gram LSTM for NMT showed poor performance (13.9 BLEU points for English-Chinese), which can be attributed to the difference between the LM and NMT tasks. Both tasks predict one word at a time conditioned over previous words. However, in NMT the previous target-word-inputs are not given. They have to be generated by the decoder. Thus, the output could be conditioned over previous erroneous predictions affecting the 4-gram LSTM model in higher proportion. This result shows that even if a model improves language modeling, it does not necessarily improve machine translation.

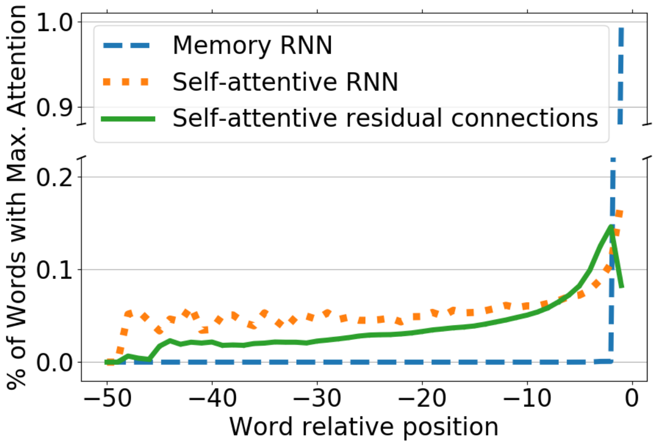


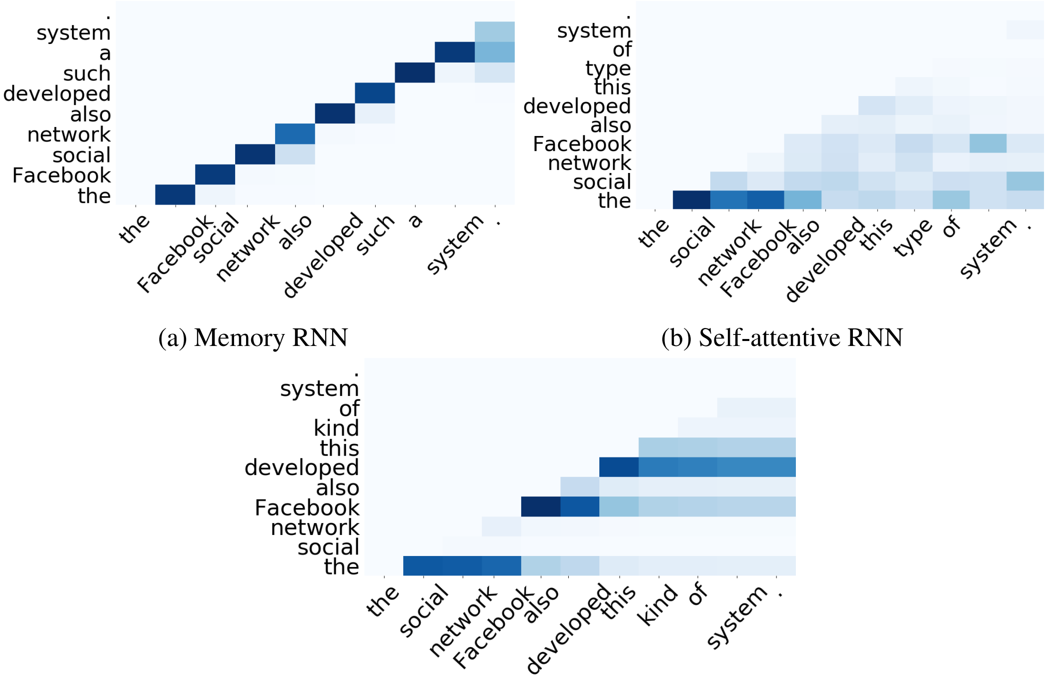
Figure 5.2 – Percentage of words that received maximum attention at a given relative position, ranging from −1 to −50 (maximum length).

5.7.4 Distribution of Attention

Figure 5.2 shows a comparison of the distribution of attention of the different self-attentive models described in this paper, on Spanish-to-English NMT (the other two language pairs exhibit similar distributions). The values correspond to the number of words that received maximal attention for each relative position (x-axis). At each prediction, we selected the preceding word with maximal weight and counted its relative position. We normalized the count by the number of previous words at the time of each prediction.

We observe that the *memory RNN* almost always selects the immediately previous word ($t-1$) and ignores the rest of the context. On the contrary, the other two models distribute attention more evenly among all previous words. In particular, the *self-attentive RNN* uses a longer context than the *self-attentive residual connections*, but, as the performance on BLEU score shows, this fact does not necessarily mean better translation.

Figure 5.3 shows the attention to previous words generated by each model for one sentence translated from Spanish to English. The matrices present the target-side attention weights, with the vertical axis indicating the previous words, and the color shades at each position (cell) representing the attention weights. The weights of the *memory RNN* are concentrated on the diagonal, indicating that the attention is generally located on the previous word, making the model almost equivalent to the baseline. The weights of the *self-attentive RNN* show that attention is more distributed towards the distant past, and they vary for each word because the attention function depends on the current prediction. This model tries to find dependencies among words, although complex relations seem challenging to learn. On the contrary, the proposed *self-attentive residual connections* model strongly focuses on particular words, and we present a more exhaustive analysis of it in the following section.



(a) Memory RNN

(b) Self-attentive RNN

(c) Self-attentive residual connections

Figure 5.3 – Matrix of distribution of the attention weights to previous words. The vertical axis represents the previous words. A darker shade indicates a higher attention weight.

5.7.5 Structures Learned by the Model

When visualizing the matrix of attention weights generated by our model (Figure 5.3c), we observed the formation of sub-phrases that are grouped depending on their attention to previous words. To build the sub-phrases in a deterministic fashion, we implemented Algorithm 1, which iteratively splits the sentence into two sub-phrases every time the focus of attention changes to a new word, from left-to-right. The results are binary tree structures containing the sub-phrases, exemplified in Figure 5.4.

We formally evaluate the syntactic properties of the binary tree structures by comparing them with the results of an automatic constituent parser (Manning et al., 2014), using the ParsEval approach (Black et al., 1991), i.e., by counting the precision and recall of constituents, excluding single words. Our models reach a precision of 0.56, which is better than the precision of 0.45 obtained by a trivial right-branched tree model[4]. Note that these structures were neither optimized for parsing nor learned using part-of-speech tagging as most parsers do. Our interpretation of the results is that they are "syntactic-like" structures. However, given the model's simplicity, they could also be viewed as more limited structures, similar to sentence chunks.

[4] A model constructed by dividing iteratively one word and the rest of the sentence, from left-to-right.

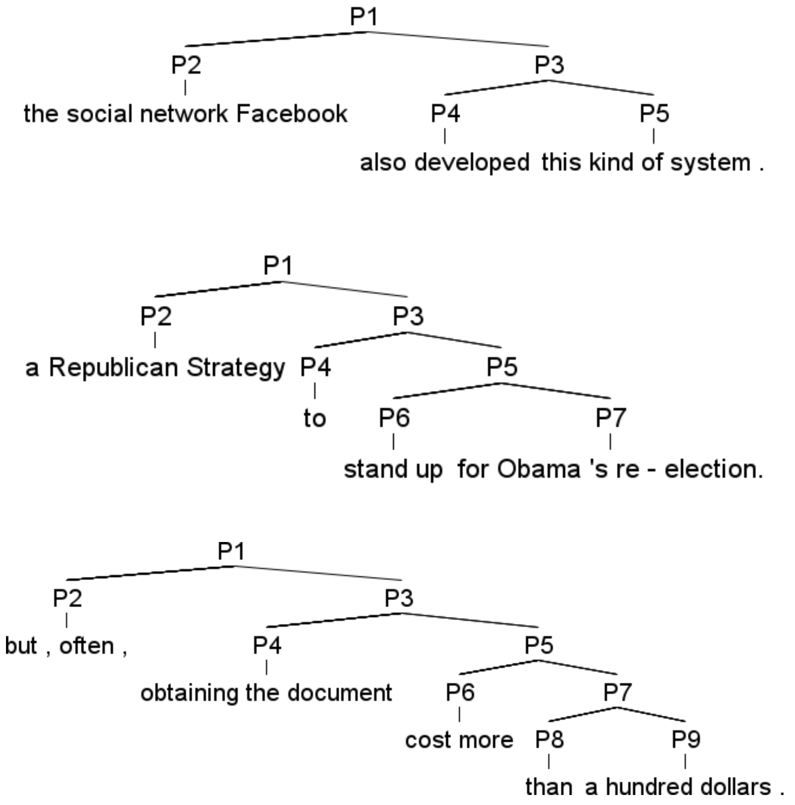


Figure 5.4 – Examples of hypothesized syntactic structures obtained with Algorithm 1.

Algorithm 1 Binary Parse Tree

```
Require: A matrix of attention of size $N \times N$
Require: s sentence as list of words of size $N$
function SPLIT($tree, \mathbf{A}, \mathbf{s}$)
    $n \leftarrow length(s)$
    $i \leftarrow 0$
    while $max(\mathbf{A}[:][i]) = 0$ or $i < n$ do
        $i \leftarrow i + 1$
    end while
    $tree.addChild(\mathbf{s}[0: i])$
    if $i < n$ then
        $subtree \leftarrow newTree()$
        SPLIT($subtree, \mathbf{A}[i: n][i: n], \mathbf{s}[i: n]$)
        $tree.addChild(subtree)$
    end if
end function
$tree \leftarrow newTree()$; SPLIT($tree, \mathbf{A}, \mathbf{s}$)
```

BLUE score: self-attentive > baseline

Example 1: Correct sentence structure

S: Estudiantes y profesores se están tomando a la ligera la fecha.

R: Students and teachers are taking the date lightly.

B: Students and teachers are being taken lightly to the date.

O: Students and teachers are **taking the date lightly**.

Example 2: Correct drop-pronoun translation and sentence structure

S: No porque ∅ compartiera su ideología, sino porque para él los Derechos Humanos son indivisibles.

R: Not because he shared their world view, but because for him, human rights are indivisible.

B: Not because I share his ideology, but because he is indivisible by human rights.

O: Not because **he** shared his ideology, but because **for him human rights are indivisible**.

BLUE score: self-attentive < baseline

Example 3: Incorrect sentence structure

S: El gobierno intenta que no se construyan tantas casas pequeñas.

R: The Government is trying not to build so many small houses.

B: The government is trying **not to build so many small houses**.

O: The government is trying to ensure that so many small houses are not built.

Example 4: Alternative translation

S: Otras personas pueden tener niños .

R: Other people can have children.

B: **Other people can** have children.

O: **Others may** have children.

Table 5.7 – Examples from Spanish to English.

5.7.6 Translation Examples

Table 5.7 shows examples of translations produced with the baseline and the *self-attentive residual connections* model. The first part shows examples for which the proposed model reached a higher BLEU score than the baseline. Here, the structure of the sentences, or at least the word order, are improved. The second part contains examples where the baseline achieved a better BLEU score than our model. In the first example, the sentence structure is different but the content and quality are similar, while in the second one, lexical choices differ from the reference.

5.8 Conclusion

We presented a novel decoder that uses self-attentive residual connections to previously translated words to enrich the target-side contextual information in NMT. To cope with the variable lengths of previous predictions, we proposed two methods for context summarization: *mean residual connections* and *self-attentive residual connections*. Additionally, we showed how similar previous proposals, designed for language modeling, can be adapted to NMT. We evaluated the methods over three language pairs: Chinese-to-English, Spanish-to-English, and English-to-German. In each case, we improved the BLEU score compared to the NMT baseline and two variants with memory-augmented decoders. A manual evaluation over a small set of sentences for each language pair confirmed the improvement. Finally, a qualitative analysis showed that the proposed model distributes weights throughout an entire sentence and learns structures resembling syntactic ones.

6 Document-level Neural Machine Translation

This chapter is based on the following paper:

> Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics

6.1 Introduction

In the previous chapter, we introduced a self-attentive decoder to facilitate modeling discourse and semantic structures useful for translation at the sentence-level. In this chapter, we aim at modeling discourse at the document-level. During the development of this work, research on self-attention networks for sequences was gaining strength, and finally, a network entirely based on attention, the transformer network (Vaswani et al., 2017), showed significant improvement over the previous state-of-the-art models based on LSTMs. The transformer network's attention structure is flexible and allows performing parallel computations for each word in the sequence. Thus, it is efficient at training and makes it easier to increase the network's depth, resulting in better learning capabilities.

However, the main issue of this thesis persists. The NMT models do not consider discourse and context at the document-level. By ignoring discourse connections between sentences and other valuable contextual information, the model potentially degrades the coherence and cohesion of a translated document (Hardmeier, 2012; Meyer and Webber, 2013; Sim Smith, 2017). Some previous studies (Tiedemann and Scherrer, 2017; Jean et al., 2017; Wang et al., 2017; Tu et al., 2018) have demonstrated that adding contextual information to the NMT model improves the general translation performance, and more importantly, improves the coherence and cohesion of the translated text (Bawden et al., 2018; Lapshinova-Koltunski and Hardmeier, 2017). Most of these methods use an additional encoder (Jean et al., 2017; Wang et al., 2017) to extract contextual information from previous source-side sentences. However, this requires additional parameters, and it does not exploit the representations already learned by the NMT encoder. Tu et al. (2018) has shown that a cache-based memory network performs better than the above encoder-based methods. The cache-based memory keeps past context as a set of words, where each cell corresponds to one unique word keeping the hidden representations learned by the NMT while translating it. However, in this method, the word representations are stored irrespective of the sentences where they occur, and the representations are disconnected from the original NMT network.

We propose to use a hierarchical attention network (HAN) (Yang et al., 2016) to model the contextual information in a structured manner using word-level and sentence-level abstractions. In contrast to the hierarchical recurrent neural network (HRNN) used by (Wang et al., 2017), here the attention allows dynamic access to the context by selectively focusing on different sentences and words for each predicted word. Besides, we integrate two HANs in the NMT model to account for target and source context. The HAN encoder helps in the disambiguation of source-word representations, while the HAN decoder improves the target-side lexical cohesion and coherence. The integration is done by (i) re-using the hidden representations from both the encoder and decoder of previous sentence translations and (ii) providing input to both the encoder and decoder for the current translation. This integration method enables it to optimize for multiple-sentences jointly. Furthermore, we extend the original HAN with a multi-head attention (Vaswani et al., 2017) to capture different types of discourse phenomena. The HAN architecture can adapt to any underlying NMT architecture because it works as an additional layer to the encoder and decoder. In this work, we used transformer networks.

Our main contributions are the following:

(i) We propose a HAN framework for translation to capture context and inter-sentence connections in a structured and dynamic manner.

(ii) We integrate the HAN in a very competitive NMT architecture (Vaswani et al., 2017) and show significant improvement over two strong baselines on multiple data sets.

(iii) We perform an ablation study of the contributions of each HAN configuration, showing that contextual information obtained from source and target sides are complementary.

The rest of the chapter is organized as follows: additional work related to document-level machine translation is described in Section 6.2. The baseline transformer architecture for our experiments is detailed in Section 6.3, and our HAN approach in Section 6.4. The experimental setting and result analysis are in Sections 6.5 and 6.6 respectively. Finally, our conclusions are drawn in Section 6.7.

6.2 Related Work

As discussed in Section 2.3 of the Background Chapter, the necessity for introducing discourse in translation was present since statistical machine translation (SMT). Some initial studies were based on cache memories of previous translations (Tiedemann, 2010; Gong et al., 2011), however, most of the work explicitly modeled discourse phenomena (Sim Smith, 2017; Mascarell, 2017a) such as lexical cohesion (Meyer and Popescu-Belis, 2012; Xiong et al., 2013b; Loáiciga and Grisot, 2016; Pu et al., 2017a), word sense disambiguation (Vickrey et al., 2005; Chan et al., 2007; Pu et al., 2017b), coherence (Born et al., 2017), and coreference (Rios Gonzales and Tuggener, 2017; Miculicich Werlen and Popescu-Belis, 2017a). However, these studies were based on sentence-level translations. Some of studies proposed to optimize the translation of the whole text (Xiao et al., 2011). For example, Hardmeier et al. (2013, 2012) implemented a document-level decoder for phrase-based SMT, referred to as "Docent". The decoder allows to incorporate feature functions (i.e. scoring functions for optimizations) at document-level. Some studies used this functionality to incorporate different discourse features (Garcia et al., 2015; Mascarell, 2017b; Garcia et al., 2017; Born et al., 2017).

Previous to our work, several studies proved that context could be effectively used to improve neural machine translation (NMT). Tiedemann and Scherrer (2017) use the concatenation of multiple sentences as input and output of a standard NMT, Jean et al. (2017) adds a context encoder for the previous source sentence, Wang et al. (2017) includes a hierarchical RNN to summarize source-side context, and Tu et al. (2018) use a dynamic cache memory to store representations of previously translated words.

Since the publication of the present work, research on document-level machine translation gained strength, especially since the conference on machine translation (WMT) included evaluations of sentences in context and complete documents (Ma et al., 2019). Although sentence-level NMT reached high performance comparable to human quality under certain evaluation conditions, Läubli et al. (2018) showed that when sentences were evaluated in context, judges still prefer human translation and MT errors related to discourse are spotted. Several studies were done in this direction. Maruf and Haffari (2018) proposed a document-level NMT using memory-networks. Junczys-Dowmunt (2019) proposed a large scale document-level NMT system using a simple encoder-decoder architecture with whole documents as input/output. Voita et al. (2018) showed that context-aware NMT improves, in particular, the anaphoric pronouns. (Voita et al., 2019b) present a complete study of context-aware NMT based on deixis, ellipsis, and lexical cohesion. (Yang et al., 2019b) proposed to use capsule networks to model the hierarchical structure of languages. Similar to the work presented here, Tan et al. (2019) uses a hierarchical modeling approach for global context, and Maruf et al. (2019) improves HAN with a selective attention. Alternative approaches are, for instance, to model discourse at the target side with a monolingual language model (Voita et al., 2019a), or to set document translation as a domain adaptation problem (Kothur et al., 2018).

6.3 Transformer Network

The transformer network proposed by (Vaswani et al., 2017) is an encoder-decoder architecture for sequence-to-sequence modeling. Instated of using LSTMs, the sequence is represented based solely on attention mechanisms, so the intra-sentence contextual information for each word is calculated with a multi-head self-attention. In order to keep the sense of order, a positional embedding p_t is added to the word input embedding x_t at time-step t of the sequence. The position vector is calculated as follows:

$$\begin{aligned} p_t^{2i} &= \sin(t/1000^{2i/d}) \\ P_t^{2i+1} &= \cos(t/1000^{2i/d}) \end{aligned}$$

where i is the index dimension of p_t, and d is the embedding dimension.

The $\mathsf{Attention}$ function is defined over three variables: a *query* vector q_t, a *key* matrix K, and a *value* matrix V. It is calculated as follows:

$$\mathsf{Attention}(q_t, K, V) = \text{softmax}(\frac{q_t \cdot K^{\top}}{\sqrt{d}}) \cdot V \tag{6.1}$$

here *query* q_t refers to the element of interest at time-step t either during encoding or decoding. K and V refer to the context from where we want to obtain information, the names *key* and *value* refer to their functionality in the equation. During encoding, these matrices are both the input sequence representation. During decoding, there are two attentions: one to the encoder representations, and another to the output sequence representation from previous time-steps. The $\mathsf{MultiHead}$ term refers to the partition of the attention vector into H sections, each called *head* and denoted with index h. This partition allows to each head to pay attention to different elements of the sequence (e.g. words of the sentence) reflecting independent and complementary information. This function is defined as follows:

$$\begin{aligned} \mathsf{MultiHead}(q_t, K, V) &= \mathsf{LayerNorm}(\hat{q}_t + q_t) \\ \text{where} \qquad \hat{q}_t &= W_a \cdot [A_1, ..., A_H] \\ A_h &= \text{Attention}(\ W_q^h \cdot q_t\ ,\ W_k^h \cdot K\ ,\ W_v^h \cdot V\ ) \end{aligned} \tag{6.2}$$

here W_a, W_q^h, W_k^h, W_v^h are parameter matrices, and the square brackets denote concatenation. This equation has a residual connection (He et al., 2016), and is normalized with a *layer normalization* function (Lei Ba et al., 2016), notated as $\mathsf{LayerNorm}$. The residual connection consists of adding the input vector to the layer's output to facilitate the flow of information, so the layer could be skipped if needed. The layer normalization is a technique to normalize the distribution of a vector. Both techniques allow smoother gradients, faster training, and better generalization accuracy. We refer the attention vector output as z_t:

$$z_t = \mathsf{MultiHead}(q_t, K, V) \tag{6.3}$$

The feed-forward layer function FFN receives a vector as argument, in this case z_t:

$$\begin{aligned} \mathsf{FFN}(z_t) &= \mathsf{LayerNorm}(\hat{z}_t\ +\ z_t) \\ \text{where} \quad \hat{z}_t &= W_f \cdot max(0, W_z \cdot z_t) \end{aligned} \tag{6.4}$$

here W_f, W_z are parameters of the network. The final hidden state h_t at time-step t is:

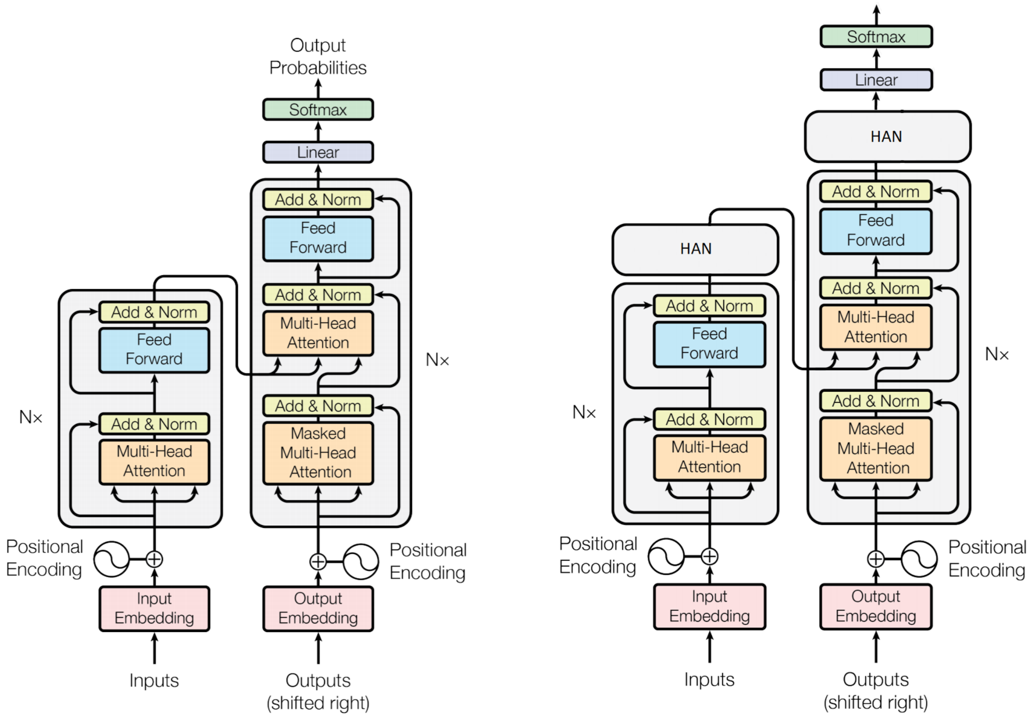
$$h_t = \text{FFN}(z_t) \tag{6.5}$$

The complete architecture is shown in Figure 6.1a. The left component is the encoder and the right is the decoder. The previous described layers, MultiHead and FFN, are repeated N times.

6.4 The Proposed Approach

The goal of NMT is to maximize the likelihood of a set of sentences in a target language represented as sequences of words $Y = [y_1, ..., y_t]$ given a set of input sentences in a source language $X = [x_1, ..., x_m]$ as:

$$\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} log(P_\Theta(Y^n | X^n)) \tag{6.6}$$



(a) Transformer Network

(b) Transformer Network with HAN

Figure 6.1 – Comparison between baseline and proposed architectures. *Image taken from* (Vaswani et al., 2017).

so, the translation of a document D is made by translating each of its sentences independently. In this study, we introduce dependencies on the previous sentences from the source and target sides:

$$\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} log(P_{\Theta}(Y^n|X^n, D_{x^n}, D_{y^n})) \tag{6.7}$$

where $D_{x^n} = (X^{n-k}, ..., X^{n-1})$ and $D_{y^n} = (Y^{n-k}, ..., Y^{n-1})$ denote the previous k sentences from source and target sides respectively. The contexts D_{x^n} and D_{y^n} are modeled with HANs.

6.4.1 Hierarchical Attention Network

The proposed HAN has two levels of abstraction. The word-level abstraction summarizes information from each previous sentence j into a vector s^j as:

$$q_w = f_w(h_t) \tag{6.8}$$

$$s^j = \text{MultiHead}(q_w, H^j, H^j) \tag{6.9}$$

where h_t denotes a hidden state of the NMT network at time-step t. $H^j = [h_1^j, ..., h_m^j]$ is the hidden state matrix representation of the j-th sentence of the context. The function f_w is a linear transformation to obtain the *query* q_w. We used the MultiHead attention function defined in Equation (6.2) to capture different types of relations among words. It matches the *query* against each of the hidden representations h_i^j.

The sentence-level abstraction summarizes the contextual information required at time t in d_t as:

$$q_s = f_s(h_t) \tag{6.10}$$

$$d_t = \mathsf{FFN}(\,\mathsf{MultiHead}(q_s, S, S)\,) \tag{6.11}$$

where $S = [s_1, ..., s_l]$ is the hidden state matrix representation of the previous sentences. f_s is a linear transformation, q_s is the query for the attention function, FFN is a position-wise feed-forward layer defined in Equation (6.4).

6.4.2 Context Gating

We use a gate (Tu et al., 2018, 2017) to regulate the information at sentence-level h_t and the contextual information at document-level d_t. The intuition is that different words require different amount of context for translation, e.g. ambiguous words vs. non-ambiguous ones.

$$\lambda_t = \sigma(W_h h_t + W_d d_t) \tag{6.12}$$

$$\widetilde{h}_t = \lambda_t h_t + (1 - \lambda_t) d_t \tag{6.13}$$

where W_h, W_p are parameter matrices, and $\widetilde{h}_t$ is the final hidden representation for a word x_t or y_t.

6.4.3 Integrated Model

The context can be used during encoding or decoding a word, and it can be taken from previously encoded source sentences (Bawden et al., 2018), previously decoded target sentences, or from previous alignment vectors (Tu et al., 2018) (i.e. context vectors (Bahdanau et al., 2015)). The different configurations will define the input *query* and *values* of the attention function. In this work we experiment with five of them: one at encoding time, three at decoding time, and one combining both. At encoding time the *query* is a function of the hidden state h_{x_t} of the current word to be encoded x_t, and the *values* are the encoded states of previous sentences $h_{x_i}^j$ (HAN encoder). At decoding time, the *query* is a function of the hidden state h_{y_t} of the current word to be decoded y_t, and the *values* can be (a) the encoded states of previous sentences $h_{x_i}^j$ (HAN decoder *source*), (b) the decoded states of previous sentences $h_{y_i}^j$ (HAN decoder), and (c) the alignment vectors c_i^j (HAN decoder *alignment*). Finally, we combine complementary target-source sides of the context by joining HAN encoder and HAN decoder. Figure 6.2 shows the integration of the HAN encoder with the NMT model; a similar architecture is applied to

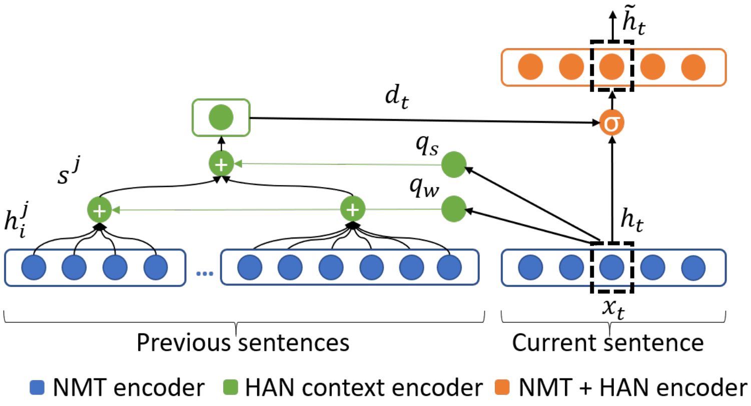


Figure 6.2 – Integration of HAN during encoding at time step t, $\tilde{h}_t$ is the context-aware hidden state of the word x_t. Similar architecture is used during decoding.

the decoder. The output $\tilde{h}_t$ is used by the NMT model as a replacement of h_t during the final classification layer. The complete architecture is show in Figure 6.1b.

6.5 Experimental Setting

In this section, we describe the datasets and evaluation metrics for our experiments. We also detail the model configurations, hyper-parameters, and training setup.

6.5.1 Datasets and Evaluation Metrics

We carry out experiments with Chinese-to-English (Zh-En), and Spanish-to-English (Es-En) sets on three different domains: talks, subtitles, and news where contextual information is relevant (Tu et al., 2018; Bawden et al., 2018).

TED Talks is part of the IWSLT 2014 and 2015 (Cettolo et al., 2012, 2015) evaluation campaigns[1]. We use *dev2010* for development; and *tst2010-2012* (Es-En), *tst2010-2013* (Zh-En) for testing. The Zh-En subtitles corpus is a compilation of TV subtitles designed for research on context (Wang et al., 2018b). [2]. In contrast to the other sets, it has three references to compare. The Es-En corpus is a subset of OpenSubtitles2018 (Lison and Tiedemann, 2016)[3]. We randomly select two episodes for development and testing each. Finally, we use the Es-En News-Commentaries11[4] corpus which has document-level delimitation. We evaluate on WMT sets (Bojar et al., 2013): *newstest2008* for development, and *newstest2009-2013* for testing. A similar corpus for Zh-En is too small to be comparable. Table 6.3 shows the corpus statistics.

[1] https://wit3.fbk.eu
[2] https://github.com/longyuewangdcu/tvsub
[3] http://www.opensubtitles.org
[4] http://opus.nlpl.eu/News-Commentary11.php

For evaluation, we use BLEU score (Papineni et al., 2002) on *tokenized* text, and we measure significance with the paired bootstrap re-sampling method proposed by Koehn (2004) (implementations by Koehn et al. (2007)) [5].

6.5.2 Model Configuration and Training

As baselines, we use an NMT transformer and a context-aware NMT transformer with cache memory, which we implemented for comparison following the best model described by Tu et al. (2018), with a memory size of 25 words. We used the OpenNMT (Klein et al., 2017) implementation of the transformer network. The configuration is the same as the model called "base model" in the original paper (Vaswani et al., 2017). The encoder and decoder are composed of 6 hidden layers each. All hidden states have a dimension of 512, dropout of 0.1, and 8 heads for the multi-head attention. The target and source vocabulary size is 30K. The optimization and regularization methods were the same as proposed by Vaswani et al. (2017). Inspired by Tu et al. (2018) we trained the models in two stages. First, we optimize the parameters for the NMT without the HAN, then we proceed to optimize the parameters of the whole network. We use $k = 3$ previous sentences, which gave the best performance on the development set.

6.6 Results

We performed a quantitative and qualitative analysis of results, including contextual discourse aspects such as coherence, cohesion, and pronoun translation.

6.6.1 Translation Performance

Table 6.1 shows the BLEU scores for different models. The baseline NMT transformer already has better performance than previously published results on these datasets, and we replicate previous improvements from the cache method over this stronger baseline. All of our proposed HAN models perform at least as well as the cache method. The best scores are obtained by the combined encoder and decoder HAN model, which is significantly better than the cache method on all datasets without compromising training speed (2.3K vs 2.6K tok/sec). A critical portion of the improvement comes from the HAN encoder, which can be attributed to the fact that the source-side always contains the correct information, while the target-side may contain erroneous predictions at testing time. However, combining HAN decoder with HAN encoder further improves translation performance, showing that they contribute complementary information. The three ways of incorporating information into the decoder all perform similarly.

[5] `multi-bleu.perl` and `bootstrap-hypothesis-difference-significance.pl`

	TED Talks				Subtitles				News	
	Zh–En		Es–En		Zh–En [6]		Es–En		Es–En	
Models	BLEU	Δ	BLEU	Δ	BLEU	Δ	BLEU	Δ	BLEU	Δ
NMT transformer	16.87		35.44		28.60		35.20		21.36	
+ cache (Tu et al., 2018)	17.32	(+0.45)***	36.46	(+1.02)***	28.86	(+0.26)	35.49	(+0.29)	22.36	(+1.00)***
+ HAN encoder	17.61	(+0.74)$^{***}_{\dagger\dagger}$	36.91	(+1.47)$^{***}_{\dagger\dagger}$	29.35	(+0.75)$^{*}_{\dagger}$	35.96	(+0.76)$^{*}_{\dagger}$	22.36	(+1.00)***
+ HAN decoder	17.39	(+0.52)***	37.01	(+1.57)$^{***}_{\dagger\dagger\dagger}$	29.21	(+0.61)*	35.50	(+0.30)	22.62	(+1.26)$^{***}_{\dagger\dagger\dagger}$
+ HAN decoder *source*	17.56	(+0.69)$^{***}_{\dagger\dagger}$	36.94	(+1.50)$^{***}_{\dagger\dagger}$	28.92	(+0.32)	35.71	(+0.51)*	22.68	(+1.32)$^{***}_{\dagger\dagger\dagger}$
+ HAN decoder *alignment*	17.48	(+0.61)$^{***}_{\dagger}$	37.03	(+1.60)$^{***}_{\dagger\dagger\dagger}$	28.87	(+0.27)	35.63	(+0.43)	22.59	(+1.23)$^{***}_{\dagger\dagger\dagger}$
+ HAN encoder + HAN decoder	**17.79**	(+0.92)$^{***}_{\dagger\dagger\dagger}$	**37.24**	(+1.80)$^{***}_{\dagger\dagger\dagger}$	**29.67**	(+1.07)$^{**}_{\dagger}$	**36.23**	(+1.03)$^{**}_{\dagger\dagger}$	**22.76**	(+1.40)$^{***}_{\dagger\dagger\dagger}$

Table 6.1 – BLEU score for the different configurations of the HAN model, and two baselines. The highest score per dataset is marked in bold. Δ denotes the difference in BLEU score with respect of the NMT transformer. The HAN uses $k = 3$ previous sentences. The significance values with respect to the NMT and the cache method are denoted by $*$, and † respectively. The repetitions correspond to the p-values: $^{*}_{\dagger} < .05, ^{**}_{\dagger\dagger} < .01, ^{***}_{\dagger\dagger\dagger} < .001$.

	Noun Translation					Pronoun Translation				
	TED Talks		Subtitles		News	TED Talks		Subtitles		News
Model	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En
NMT Transformer	40.16	65.97	46.65	61.79	47.94	63.44	68.00	69.71	65.83	47.22
+ cache	40.87	66.75	46.00	61.87	49.91	63.53	68.66	69.97	66.27	49.34
+ HAN encoder	41.93	67.75	46.78	61.52	50.06	64.05	69.17	**71.04**	**68.56**	49.57
+ HAN decoder	41.61	67.35	46.78	61.99	50.03	64.02	69.36	70.50	67.03	49.33
+ HAN encoder + HAN decoder	**42.99**	**67.81**	**47.43**	**62.30**	**50.40**	**64.35**	**69.60**	70.60	67.47	**49.59**
	Lexical cohesion					Coherence				
NMT Transformer	54.26	51.98	51.87	51.77	30.06	0.298	0.299	0.283	0.262	0.279
+ HAN encoder	54.87	52.35	51.89	52.33	30.34	0.304	0.299	0.285	0.262	0.280
+ HAN decoder	54.95	**52.43**	**52.33**	52.43	30.41	0.302	0.301	0.287	0.265	0.282
+ HAN enc. + HAN dec.	**55.40**	52.36	51.94	**52.75**	**30.58**	**0.305**	**0.302**	**0.287**	**0.265**	**0.282**
Human reference	56.08	57.02	54.81	58.19	35.12	0.310	0.314	0.296	0.270	0.298

Table 6.2 – Evaluation on discourse phenomena. Noun and pronoun translation: Accuracy with respect to a human reference. Lexical cohesion: Ratio of repeated and lexically similar words over the number of content words. Coherence: Average cosine similarity of consecutive sentences (i.e. average of LSA word-vectors)

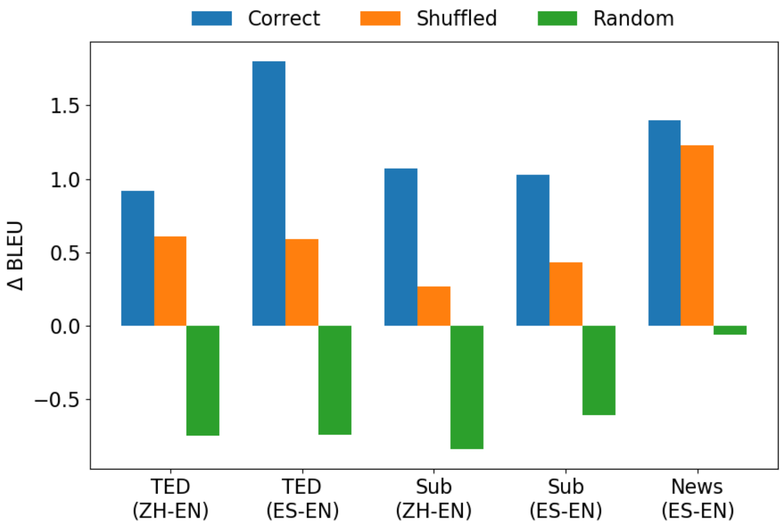


Figure 6.3 – Difference (Δ) BLUE score of the HAN model with respect to the transformer baseline, for *correct*, *shuffled* and *random* context

6.6.2 Effect of Context

We analyze two different questions related to the effect of context on translation. First, how much context is useful in terms of the number of sentences? Second, what happens when we change the context? To answer the first question, table 6.4 shows the performance of our best HAN model with a varying number k of previous sentences in the test-set. We can see that the best performance for TED talks and news is achieved with 3, while for subtitles, it is similar between 3 and 10. We can infer that HAN model can exploit a relatively short context of about 3 sentences; further context has no effect. This result could be caused by the limited memory capacity of the vector $\hat{h}_t$, which has to compress all contextual information. To answer the second questions, Figure6.3 shows the difference of BLEU scores of HAN model with respects to the baseline in three experimental cases: (a) *correct:* using the three previous sentences as context, (b) *shuffled:* using the three random sentences of the same document, (c) *random:* using the three random sentences of other documents,. We can see that the best results for test-sets are obtained in case *(a)* when having the correct context. In case *(b)*, the model can be seen as domain adaptation, so the results are better than baseline but worse than having context in correct order. In case *(c)*, the HAN model has worse performance than the baseline, with lesser impact in news. Thus, we can deduce that HAN model effectively exploits context in this scenario.

6.6.3 Accuracy of Pronoun/Noun Translations

We evaluate coreference and anaphora using the reference-based metric: accuracy of pronoun translation (Miculicich Werlen and Popescu-Belis, 2017b), which can be extended for nouns. The list of evaluated pronouns is predefined in the metric, while the list of nouns was extracted using

	TED Talks		Subtitles		News
	Zh–En	Es–En	Zh–En	Es–En	Es–En
Training	0.2M	0.2M	2.2M	4.0M	0.2M
Development	0.8K	0.8K	1.1K	1.0K	1.9K
Test	5.5K	4.7K	1.2K	1.0K	13.5K
Mean $\|s\|$ (src-tgt)	18-20	19-20	5-7	8-9	

Table 6.3 – Dataset statistics in # sentence pairs. $|s|$ denotes sentence lengths

	TED Talks		Subtitles		News
k	Zh–En	Es–En	Zh–En	Es–En	Es–En
1	17.70	37.20	29.35	36.20	22.46
3	**17.79**	**37.24**	29.67	**36.23**	**22.76**
5	17.49	37.11	**29.69**	36.22	22.54
7	17.00	37.22	29.64	36.21	22.64
10	10.50	13.71	28.89	36.18	

Table 6.4 – Performance for variable context sizes k with the HAN encoder + HAN decoder.

NLTK POS tagging (Bird, 2006). The upper part of Table 6.2 shows the results. For nouns, the joint HAN achieves the best accuracy with a significant improvement compared to other models, showing that target and source contextual information are complementary. Similarly, the joint model has the best result for TED talks and news for pronouns. However, HAN encoder alone is better in the case of subtitles. Here HAN decoder produces mistakes by repeating past translated personal pronouns. Subtitles is a challenging corpus for personal pronoun disambiguation because it usually involves a dialogue between multiple speakers.

6.6.4 Cohesion and Coherence Evaluation

We use the metric proposed by Wong and Kit (2012) to evaluate lexical cohesion. It is defined as the ratio between the number of repeated and lexically similar content words over the total number of content words in a target document. The lexical similarity is obtained using WordNet. Table 6.2 (bottom-left) displays the average ratio per tested document. In some cases, the HAN decoder achieves the best score because it produces a larger quantity of repetitions than other models. However, as previously demonstrated in 6.6.3, repetitions do not always make the translation better. Although HAN boosts lexical cohesion, the scores are still far from the human reference, so there is room for improvement in this aspect.

For coherence, we use a metric based on Latent Semantic Analysis (LSA) (Foltz et al., 1998). LSA is used to obtain sentence representations, then cosine similarity is calculated from one sentence to the next, and the results are averaged to get a document score. We employed the pre-trained LSA model *Wiki-6* from (Stefanescu et al., 2014). Table 6.2 (bottom-right) shows the average coherence score of documents. The joint HAN model consistently obtains the best

S:	y esto es un escape de su estado atormentado .
R:	and that is an escape from his tormented state .
B:	and this is an escape from its state .
C:	and this is an escape from their state .
H:	and this is an escape from **his** state .

Table 6.5 – Example of pronoun disambiguation (TED Talks Es-En).

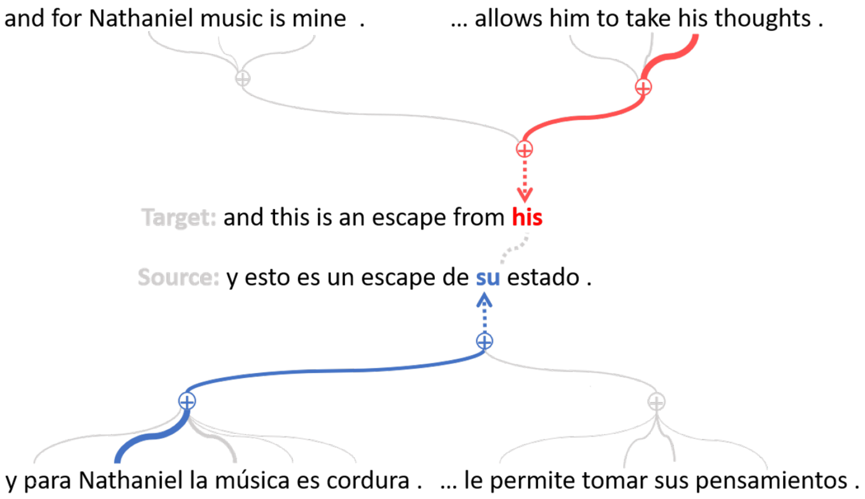

and for Nathaniel music is mine .
... allows him to take his thoughts .
Target: and this is an escape from his
Source: y esto es un escape de su estado .
y para Nathaniel la música es cordura .
... le permite tomar sus pensamientos .

Figure 6.4 – Example of pronoun disambiguation in context using HAN (TED Talks Es-En)

coherence score, but close to other HAN models. Most of the improvement comes from the HAN decoder. Although the absolute difference between human reference and baseline is small, this metric gives us a notion of the relative improvement, which places the HAN performance in between the human reference and baseline.

6.6.5 Qualitative Analysis

Table 6.5 displays an example where HAN helped to generate the correct pronoun translation. It displays the source S, reference R, translations of the baseline transformer B, the cache memory C, and the HAN model H. To visualize how the HAN model use the context, Figure 6.4 shows the source sentence and the translation to the target language together with two previous sentences at each side. The figure plots the hierarchical structure obtained when calculating the context vector for the pronoun '*his*' and its counterpart '*su*' in the source language. We see that HAN correctly translates the ambiguous Spanish pronoun '*su*' into the English '*his*'. The HAN decoder highlighted a previous mention of '*his*' (see red line), and the HAN encoder highlighted the antecedent '*Nathaniel*' (see blue line). HAN can capture interpretable inter-sentence connections. More samples with different attention heads are shown in the Appendix B.

[6] *NIST BLEU*: NMT transformer 35.99, cache 36.52, and HAN 37.15.

6.7 Conclusion

In this chapter, we proposed a multi-head hierarchical attention network (HAN) model for document-level machine translation. We integrated context from the source and target sides by directly connecting representations from previous sentence translations into the current sentence translation. The model significantly outperforms two competitive baselines, and the ablation study shows that the target and source context is complementary. It also improves lexical cohesion and coherence, and the translation of nouns and pronouns. Our experiments show that varying the input context to incorrect one harms the translations. Thus our model is sensitive to context. The qualitative analysis shows that the HAN model can identify meaningful previous sentences and words for the correct prediction and that the multi-head attention captures different types of connections. After the publication of this work, some studies improved over this model, for example, the hierarchical global context (Tan et al., 2019), and HAN with selective attention (Maruf et al., 2019). As mentioned in Section 6.2, research on document-level machine translation has recently become more important, and the human evaluation process for translation includes context further than one sentence.

Coreference Resolution Part II

7 Partially-supervised Mention Detection

This chapter is based on the following paper:

> Miculicich, L. and Henderson, J. (2020). Partially-supervised mention detection. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 91–98, Barcelona, Spain (online). Association for Computational Linguistics

7.1 Introduction

Mention detection is the task of identifying text spans referring to an entity: named, nominal or pronominal (Florian et al., 2004). It is a fundamental component for several downstream tasks, such as coreference resolution (Soon et al., 2001), and relation extraction (Mintz et al., 2009); and it can help to maintain coherence in large text generation (Clark et al., 2018), and contextualized machine translation (Miculicich et al., 2018).

Previous studies tackled mention detection jointly with named entity recognition (Xu et al., 2017; Katiyar and Cardie, 2018; Ju et al., 2018; Wang et al., 2018a). There, only certain types of entities are considered (e.g., person, location), and the goal is to recognize mention spans and their types. In this study, we are interested in discovering generic entity mentions, which can potentially be referred to in the text, without the use of syntactic parsing information. Our long term objective is to have a model that keeps track of entities in a document for word disambiguating language modeling and machine translation.

Data from coreference resolution is suitable for our task, but the annotation is partial in that it contains only mentions that belong to a coreference chain, not single entity-mentions. Nevertheless, the missing mentions have approximately the same distribution as the annotated ones, so we can still learn this distribution from the data. Figure 7.1 shows an example from Ontonotes V.5 dataset (Pradhan et al., 2012) where “the taxi driver” is annotated in sample 1 but not in 2.

Thus, we approach mention detection as a partially supervised problem and investigate two simple techniques to compensate for the fact that some negative examples are true mentions: weighted loss functions and soft-target classification. By doing this, the model is encouraged to predict more false-positive samples, so it can detect potential mentions which were not annotated. We implement two neural mention detection methods: a sequence tagging approach, and an exhaustive search approach. The first method is novel, whereas the other is similar to previous work Lee et al. (2017a). We evaluate both techniques for coreference resolution by implementing a multi-task learning system. We show that the proposed techniques help the model increase recall significantly with a minimal decrease in precision. In consequence, the F1 score of the mention detection and coreference resolution improves for both methods, and the exhaustive search approach yields a significant improvement over the baseline coreference resolver.

Our contributions are:

(i) We investigate two techniques to deal with partially annotated data.

(ii) We propose a sequence tagging method for mention detection that can model nested mentions.

(iii) We improve an exhaustive search method for mention detection.

(iv) We approach mention detection and coreference resolution as multi-task learning and improve both tasks’ recall.

```
1. [They] informed the [taxi driver] and asked
   [him] to take the vehicle outside ...

2. On [our] way back , the taxi driver gave
   [us] an explanation ...
```

Figure 7.1 – Samples from CoNLL 2012. Annotated mentions are within brackets, non-annotated ones are underlined.

The rest of the chapter is organized as follows. Section 7.2 contains related work to this study. Sections 7.3 and 7.4 describe the two mention detection approaches we use in our experiments. Section 7.5 presents the proposed methods to deal with partially annotated mentions. We use coreference resolution as a proxy task for testing our methods which is described in Section 7.6. Section 7.7 contains the experimental setting and the analysis of results. Finally, the final conclusion is drawn Section 7.8.

7.2 Related Work

Lee et al. (2017a) proposed the first end-to-end coreference resolution that does not require heavy feature engineering for word representations. Their mention detection is done by considering all spans in a document as the candidate mentions, and the learning signal is coming indirectly from the coreference annotation. Zhang et al. (2018) used a similar approach but introducing a direct learning signal for the mention detection, which is done by adding a loss for mention detection with a scaling factor as hyper-parameter. This allows a faster convergence at training time.

Name entity recognition has been largely studied in the community. However, many of these models ignored the nested entity names. Recently, Katiyar and Cardie (2018) presents a nested named entity recognition model using a recurrent neural network that includes extra connections to handle nested mention detection. Ju et al. (2018) uses stack layers to model the nested mentions, and (Wang et al., 2018a) use an stack recurrent network.

7.3 Sequence tagging model

Several studies have tackled mention detection and named entity recognition as a tagging problem. Some of them use one-to-one sequence tagging techniques (Lample et al., 2016; Xu et al., 2017), while others use more elaborate techniques to include nested mentions (Katiyar and Cardie, 2018; Wang et al., 2018a). Here, we propose a simpler yet effective tagging approach that can manage nested mentions.

We use a sequence-to-sequence model, which allows us to tag each word with multiple labels. The words are first encoded and contextualized using a recurrent neural network, and then a sequential decoder predicts the output tag sequence. During decoding, the model keeps a pointer into the encoder, indicating the word's position, which is being tagged at each time step. The tagging is done using the following set of symbols: $\{[,],+,-\}$. The brackets "[" and "]" indicate that the tagged word is the starting or ending of a mention respectively, the symbol "+" indicates that one or more mention brackets are open, and "-" indicates that none mention bracket is open. The pointer into the encoder moves to the next word only after predicting "+" or "-"; otherwise, it remains in the same position. Figure 7.2 shows a tagging example indicating the alignments of words with tags.

Given a corpus of sentences $X = (x_1, ..., x_M)$, the goal is to find the parameters Θ which maximize the log likelihood of the corresponding tag sequences $Y = (y_1, ..., y_T)$:

$$P_\Theta(Y|X) = \prod_{t=1}^{T} P_\Theta(y_t|X, y_1, ..., y_{t-1}) \tag{7.1}$$

The next tag probability is estimated with a softmax over the output vector of a neural network:

$$P_\Theta(y_t|X, y_1, ..., y_{t-1}) = softmax(o_t) \tag{7.2}$$

$$o_t = relu(W_o \cdot [d_t, h_i] + b_o) \tag{7.3}$$

where W_o, b_o are parameters of the network, d_t is the vector representation of the tagged sequence at time-step t, modeled with a long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997), and h_i is the vector representation of the pointer's word at time t contextualized with a bidirectional LSTM (Graves and Schmidhuber, 2005).

$$(h_1, ..., h_M) = BiLSTM(X) \tag{7.4}$$

$$d_t = LSTM(y_1, ..., y_{t-1}) \tag{7.5}$$

where the decoder is initialized with the last states of the bidirectional encoder, $d_0 = h_M$.

The i-th word pointed to at time t is given by:

$$i \leftarrow \begin{cases} 0, & \text{if}\, t = 0 \\ i+1, & \text{if}\, t > 0 \,\text{and}\, y_{t-1} \in \{+,-\} \\ i, & \text{otherwise} \end{cases} \tag{7.6}$$

At decoding time, we use a beam search approach to obtain the sequence. The complexity of the model is linear with respect to the number of words. It can be parallelized at training time, given that it uses ground-truth data for the conditioned variables. However, it cannot be parallelized

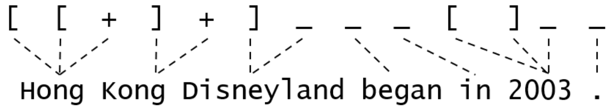


Figure 7.2 – Tagged sentence example

during decoding because of its autoregressive nature.

7.4 Span scoring model

Our span scoring model of mention detection is similar to the work of Lee et al. (2017a) for solving coreference resolution, and to Ju et al. (2018) for nested named mention detection, as both are exhaustive search methods. The objective is to score all possible spans m_{ij} in a document, where i and j are the starting and ending word positions of the span in the document. For this purpose, we minimize the binary cross-entropy with the labels y:

$$H(y, P_\Theta(m)) \quad = \quad -\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}(y_{m_{ij}} \, * \, log(P_\Theta(m_{ij})) \, + \, (1-y_{m_{ij}}) \, * \, log(1-P_\Theta(m_{ij}))) \tag{7.7}$$

where Θ are the parameters of the model, $y_{m_{ij}} \in [0,1]$ is one when there is a mention from position i to j. If $y_{m_{ij}}$ is zero when there is no mention annotated, this is the same as maximizing the log-likelihood. Nevertheless, we will consider models where this is not the case.

The probability of detection is estimated as:

$$P_\Theta(m_{ij}) = \sigma(V \cdot relu(W_m \cdot m_{ij} + b_m)) \tag{7.8}$$

$$m_{ij} = relu(W_h \cdot [h_i, h_j, \tilde{x}_{ij}] + b_h) \tag{7.9}$$

where V, W_m, W_h are weight parameters of the model, b_m, b_h are biases, and m_{ij} is a representation of the span from position i to j. It is calculated with the contextualized representations of the starting and ending words h_i, h_j, and the average of the word embeddings $\tilde{x}_{ij}$:

$$(h_1, ..., h_M) = BiLSTM(X) \tag{7.10}$$

$$\tilde{x}_{ij} = \frac{1}{j-i}\sum_{k=i}^{j} x_k \tag{7.11}$$

The complexity of this model is quadratic with respect to the number of words. However, it can be parallelized at training and decoding time. Lee et al. (2017a) uses an attention function over the embeddings instead of an average. That approach is less memory efficient and requires the maximum length of spans as a hyperparameter. Also, they include embeddings of the span lengths which are learned during training. As shown in the experimental part, these components do not improve the performance of our model.

7.5 Partially annotated data

The partial annotation of coreference data for mention detection means that not labeled spans may be true mentions of entities. Thus, the approach of treating spans without mention annotations as true negative examples would be incorrect. On the other hand, the ideal solution of sampling all possible mention annotations, which are consistent with the given partial annotation, would be intractable. We want to modify the model's loss function in such a way that, if the system predicts a false-positive, the loss is reduced. This encourages the model to favor recall over precision by predicting more mention-like spans, even when they are not labeled. We assume that it is possible to learn the true mention distribution using the annotated mention samples by extrapolating the non-annotated mentions, and we propose two ways to encourage the model to do so.

Weighted loss function: We use a weighted loss function with weight $w \in]0,1[$ for negative examples only. The *sequence tagging* model makes word-wise decisions; thus, we consider words tagged as "out of mention", y_t="-", as negative examples, while the rest are positives. Although this simplification has the potential to increase inconsistencies, e.g., having non-ending or overlapping mentions, we observe that the LSMT-based model can capture the simple grammar of the tag labels with very few mistakes. For *span scoring*, the distinction between negative and positive examples is clear, given that the decisions are made for each span.

Soft-target classification: Soft-targets allow us to have a distribution over all classes instead of having a single class annotation. Thus, we applied soft-targets to negative examples to reflect the probability that they could actually be positive ones. For *sequence tagging*, we set the target of negative examples, y_t="-", to $(\rho, \rho, \rho, 1-3\rho)$ corresponding to the classes $([,], +, -)$. For *span scoring*, we change the target of negative examples to $y_{neg}=\rho$. In both cases, ρ is the probability of the example being positive.

7.6 Coreference Resolution

We use multi-task learning to train the mention detection together with coreference resolution. The weights to sum the loss functions of each task are estimated during training, as in (Cipolla et al., 2018). The sentence encoder is shared, and the output of mention detection serves as input to coreference resolution. We use the coreference resolver proposed by Lee et al. (2017a). It uses a pair-wise scoring function s between a mention m_k and each of its candidate antecedents m_a, defined as:

$$s(m_k, m_a) = s_c(m_k, m_a) + s_m(m_k) + s_m(m_a) \tag{7.12}$$

where s_c is a function that assesses whether two mentions refer to the same entity. We modified the mention detection score s_m.

For the *sequence tagging* approach, the function s_m serves as a bias value and it is calculated as:

$$s_m = v.P(y_{t_i} = \text{“[”}).P(y_{t_j} = \text{“]”}) \tag{7.13}$$

where y_{t_i} and y_{t_j} are the labels of the first and last words of the span, and v is a scalar parameter learned during training. At test time, only mentions in the one-best output of the mention detection model are candidate mentions for the coreference resolver. During training, the set of candidate mentions includes both the spans detected by the mention detection model and the ground truth mentions. The mention decoder is run for one pass with ground-truth labels in the conditional part of the probability function (Eq. 7.2), to get the mention detection loss, and run for a second pass with predicted labels to provide input for the coreferece task and compute the coreference loss.

For the *span scoring* approach, s_m is a function of the probability defined in Eq. 7.8, scaled by a parameter v learned during training.

$$s_m = v.P(m_{i,j}) \tag{7.14}$$

Instead of the end-to-end objective of Lee et al. (2017a), we use a multi-task objective, which adds the loss function of mention detection. We do not prune mentions with a maximum length, nor impose any maximum number of mentions per document. We use the probability of the mention detector with a threshold of τ for pruning.

7.7 Experiments and Results

We evaluate our model on the English OntoNotes set from the CoNLL 2012 shared-task (Pradhan et al., 2012), which has 2802 documents for training, 343 for development, and 348 for testing. The setup is the same as Lee et al. (2017a) for comparison purposes, with the hyper-parameters ρ, w, τ optimized on the development set. We use the average F1 score as defined in the shared-task (Pradhan et al., 2012) for evaluation of mention detection and coreference resolution.

7.7.1 Mention detection

First, we evaluate our stand-alone mention detectors. For this evaluation, all unannotated mentions are treated as negative examples. Table 7.1 show the results on the test set with models selected

Model	Rec.	Prec.	F1
Sequence tagging	73.7	77.5	75.6
Span scoring	72.7	79.2	75.8
+ span size emb.	71.6	80.1	75.6
- avg. emb. + att. emb.	72.1	78.9	75.4

Table 7.1 – Mention detection evaluation

	Mention			Coref.
Model	Rec.	Prec.	F1	Avg. F1
Lee et al. (2017a)	–	–	–	67.2
Sequence tagging	73.1	84.9	78.6	59.9
+ wt. loss w=0.01	77.3	83.2	80.1	64.1
+ soft-target ρ=0.1	74.3	84.0	78.8	61.2
Span scoring	75.3	88.3	81.3	67.0
+ wt. loss w=0.3	76.3	88.1	81.8	67.1
+ soft-target ρ=0.1	78.4	87.9	82.9	67.6

Table 7.2 – Coreference resolution evaluation (CoNLL 2012)

using the best F1 score with τ=0.5, on the development set. We can see that *sequence tagging* performs almost as well as *span scoring* in F1 score, even though the latter is an exhaustive search method. We also evaluate the *span scoring* model with different components from Lee et al. (2017a). By adding the span size vector, the precision increases but the recall decreases. Replacing the average embedding $\tilde{x}$ with attention over the embeddings requires a limited span size for memory efficiency, resulting in decreased performance.

7.7.2 Coreference Resolution

Table 7.2 shows the results obtained for our multi-task systems for coreference resolution and mention detection with and without the loss modification. The *sequence tagging* method obtains lower performance compared to *span scoring*. This result can be attributed to its one-best method to select mentions, in contrast to *span scoring*, where uncertainty is fully integrated with the coreference system. The *span scoring* method performs similarly to the coreference resolution baseline, showing that the naive introduction of a loss for mention detection does not improve performance (although we find it does decrease convergence time). However, adding the modified mention loss does improve coreference performance. For *sequence tagging*, the weighted loss results in higher performance, while for the *span scoring*, soft-targets work best. In both cases, the recall increases with a small decrease in precision, which improves the F1 score of mention detection and improves coreference resolution.

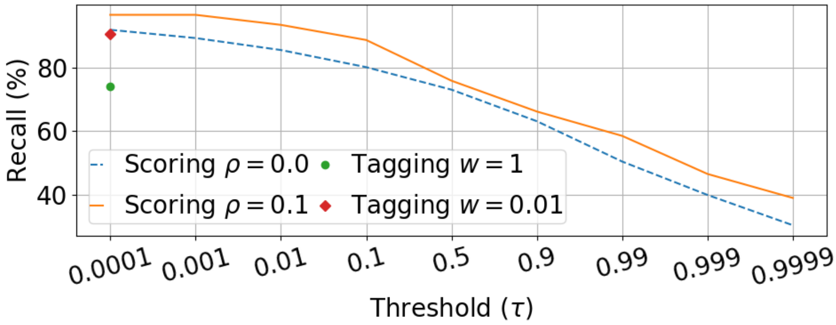


Figure 7.3 – Recall of the *mention scoring* function with respect to the detection threshold τ. Values for the *sequence tagging* are referential

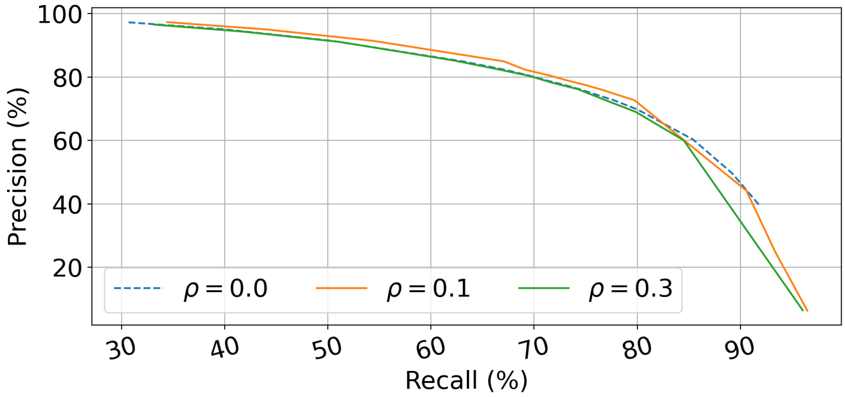


Figure 7.4 – Precision vs. Recall graph for the *mention scoring* function with soft-targets.

7.7.3 Recall performance

Figure 7.3 shows a comparison of the mention detection methods in terms of recall. The unmodified *sequence tagging* model achieves 73.7% recall, and by introducing a weighted loss at w=0.01, it reaches 90.5%. The lines show the variation of recall for the *span scoring* method with respect to the detection threshold of τ. The dotted line represents the unmodified model, while the continuous line represents the model with soft-targets at ρ=0.1, which shows higher recall for every τ. Figure 7.4 show the precision-recall plot of the *mention scoring* with soft-targets method for different values of ρ. ρ=0 corresponds to the baseline. The model with ρ=0.1 presents better results in comparison with others.

7.8 Conclusion

We investigate two simple techniques to deal with partially annotated data for mention detection and propose two methods to approach it: a Weighted loss function and a soft-target classification. We evaluate them on coreference resolution and mention detection with a multi-task learning approach. We show that the techniques effectively increase the recall of mentions and coreference links with a small decrease in precision, thus, improving the F1 score. In the future, we plan to use these methods to maintain coherence over long distances when reading, translating, and generating large text, by keeping track of abstract representations of entities.

8 Graph based Coreference Resolution

8.1 Introduction

Coreference resolution is the task of grouping together the expressions that refer to the same entity in a text. This task consists of two parts: mention detection and coreference linking. In the previous chapter, we explored the task of mention detection alone. Here, we propose a novel modeling approach for coreference resolution, and we include methods from the previous chapter.

Current state-of-the-art solutions for coreference resolution are based on neural networks (Lee et al., 2017a, 2018). The problem is formulated in an end-to-end manner where the models jointly learn to detect mentions and link coreferent mentions. The objective is to predict the antecedent of each mention-span in a document, so the model performs pair-wise decisions of all mentions. After having the model predictions, related mentions are grouped into clusters. Under this scenario, each decision (i.e., whether two mentions are related to the same entity or not) is independent of other decisions. Lee et al. (2018) proposed an iterative method to update the representation of a mention with information of its probable antecedents. However, decisions are still made locally.

Following this thesis's core idea, we propose a modeling approach that learns coreference at the document-level and takes global decisions. We propose to model mentions and coreference links in a graph structure where the nodes are tokens in the text, and the edges represent the relationship between them. Figure 8.1 shows a short example of this approach, where the graph edges are drawn with directed arrows. Our model receives a document as input and predicts the graph of mentions and coreference links. We follow a similar approach to the graph-to-graph transformer proposed in (Mohammadshahi and Henderson, 2020) for parsing, but instead of encoding sentences, we encode complete documents. Our model predicts the graph in a non-autoregressive manner, then iteratively refines it based on previous predictions. This recursive process introduces global dependencies between decisions. We define different structures for input and output graphs, which is new compared to (Mohammadshahi and Henderson, 2020). The input encodes all possible relations between nodes or tokens to make the learning process easier while the output encodes the minimal set of connections to facilitate the prediction. We initialize the transformer with the pre-trained language model BERT (Devlin et al., 2019). During the first iteration, the model predicts edges that identify mention-spans only, unlike (Mohammadshahi and Henderson, 2020) who predict the whole graph at each iteration. This is because mention detection has a different level of representation whose outputs are input to the coreference resolution (i.e. discourse) level. The model predicts the complete graph from the second to last iteration. The iterative process finishes when there are no more changes in the graph or when a maximum number of iterations is reached.

Ideally, the whole document should be encoded at once, but there is a limitation in the maximum length in practice. In order to deal with this issue, we applied two strategies: overlapping windows and reduced document. In the first one, we split documents in overlapping windows of the maximum allowed size N. The segments overlap for a value $N/2$. At decoding time, segments are input in order, and we construct the final graph by joining all graphs from different segments.

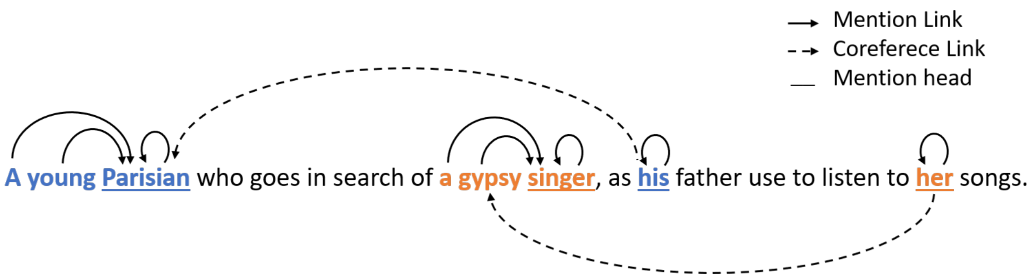

Mention Link
Coreferece Link
Mention head
A young Parisian who goes in search of a gypsy singer, as his father use to listen to her songs.

Figure 8.1 – Example of a graph structure for coreference.

In the second approach, we use two networks. The first is the previously described overlapping model, and we use it for predicting the first graph. The second one reduces the document by concatenating tokens of mention-spans separated by a particular token. This network refines the initial graph for the following iterations.

The experiments show improvements over various baselines. Our analysis indicates that the models reach the best solution in a maximum of three iterations. Given that we predict the graph at once for each iteration, our model's complexity is lower than the baseline.

Our contributions are the following:

(i) We propose a novel modeling approach to coreference resolution using a graph structure.

(ii) We propose two non-autoregressive graph models that can predict the complete entity coreference structure of a document.

(iii) We show improvements over various baselines models.

The rest of the chapter is organized as follows. Section 8.2 presents a summary of coreference resolution approaches and related research work to this chapter. Section 8.3 briefly describes the fundamentals of the state-of-the-art approach. In Section 8.4, we define entity mentions and their coreference links as a graph and fomulate the task as a sequence-to-graph problem. In Section 8.5, we present our iterative refinement solution to increase the accuracy of the model and, in Section 8.6, we present two proposed architecture models. Sections 8.7 and 8.8 contain the experimental setup and results respectively, Finally, Section 8.9 draws the conclusions of this chapter.

8.2 Related Work

The first approaches were rule-based systems (Lappin and Leass, 1994; Manning et al., 2014), but eventually, they were outperformed by machine learning approaches (Aone and William, 1995; McCarthy, 1995; Mitkov, 2002) due to annotated corpora's creation. We can classify the coreference approaches in three main types: mention-pair, entity-mention, and ranking model. Mention-pair models set coreference as a binary classification problem. The input is a pair of mentions, and the output is the prediction of whether they are related or not. The initial state is the mention detection (Chapter 7), where the input is raw text, and the output is the locations of each entity mention in the text. Mention detection is done as an independent task in a pipeline model (Soon et al., 2001) or as part of an end-to-end model (Lee et al., 2017a). The next stage is the classification. At first, the best classifiers were decision trees (Soon et al., 2001; McCarthy, 1995; Aone and William, 1995), but later, neural networks became the state-of-art (Lee et al., 2017a, 2018). The final stage is reconciling the pair-wise decisions to create entity chains, usually by utilizing greedy algorithms or clustering approaches. Entity-mention models focus on maintaining single underlying entity representation for each cluster, contrasting the independent pair-wise decisions of mention-pair approaches (Clark and Manning, 2015, 2016). Ranking models aim at ranking the possibles antecedent of each mention instead of making binary decisions (Ng, 2005; Wiseman et al., 2016). An alternative modeling approach is to perform clustering instead of classification (Fernandes et al., 2012).

The state-of-the-art models for CR are based on the work of Lee et al. (2017a, 2018). They introduced the first end-to-end model that jointly optimizes mention detection and coreference resolution tasks. These neural network-based models also simplify the mention input representation to be word embedding vectors, instead of the traditional pipeline of different linguistic feature extraction tools such as part-of-speech (POS) tagging and dependency parsing. The following models proposed improvements over this work. Fei et al. (2019) use reinforcement learning to directly optimize the model on the evaluation metrics (Section 2.4.2 of the background chapter). Joshi et al. (2019) uses BERT embeddings (Devlin et al., 2019) as input. Joshi et al. (2020) introduced a new SpanBERT embedding model, which is shown to outperform BERT for CR task.

8.3 Neural Coreference Resolution

Neural coreference resolution, as formulated in (Lee et al., 2017a, 2018), is a mention-pair approach (described in the previous section). It uses an exhaustive method defining mentions as any text span of any size in a document. There, a document D represents a sequence of tokens of size N. The objective is to assign an antecedent y_i to each text span m_i in D. The set of possible antecedents of the span m_i is denoted as $\mathrm{Y}(i)$. This set contains all text spans with index less than i, plus a null antecedent ϵ, $\mathrm{Y}(i) = \{\epsilon, m_1, ..., m_{i-1}\}$. The null antecedent is assigned when: (a) the span is not an entity mention, (b) the span is the first mention of an entity in the document. The final mention clusters are constructed greedily by grouping connected spans based on the model predictions during decoding time.

The model is trained to learn a conditional probability distribution over documents $p(y_1, ..., y_n|D)$, assuming independence among each decision of antecedent assignment y_i, as follows:

$$p(y_1, ..., y_M|D) = \prod_{i=1}^{M} p(y_i|D) \tag{8.1}$$

In (Lee et al., 2018), the probability distribution $p(y_i|D)$ is inferred over T iterations of the model over the same input document. At each iteration t, the span representations are updated with a convex combination of two components: (a) the representation of the span at time $t-1$, (b) the weighted average of all possible antecedents at time $t-1$ where the weights are given by the probability distribution of the model at time $t-1$. They called this model high-order coreference resolution since each mention representation considers information from its probable antecedents.

The training optimization is done using cross-entropy. Given that a mention-span m_i can have more than one true antecedent, the loss considers the sum of probabilities of all true antecedents in the annotated data:

$$log \prod_{i=1}^{M} \sum_{y_i \in \mathrm{Y}(i) \cap \mathrm{C}(i)} p(y_i|D) \tag{8.2}$$

where $\mathrm{C}(i)$ indicates the cluster of mention-spans that includes m_i in the annotated data. If the span does not belong to any cluster or all its antecedents have been pruned, then the span is assigned to the null cluster $\mathrm{C}(i) = \{\epsilon\}$.

This model's complexity is of the order $\mathcal{O}(N^4)$, where N is the document length. The complexity is computed by considering all possible text spans M of the document, so $\mathcal{O}(M) = \mathcal{O}(N^2)$. Then, it considers all possible combinations of span-antecedents $\mathcal{O}(M^2)$. The model prunes spans and candidate antecedents to predetermine maximum numbers in order to maintain computation efficiency.

8.4 Graph Modeling

We propose to model the set of coreference links of a document in a graph structure where the nodes are words and the edges are links of different types. Given a document $D = [x_1, ..., x_N]$ of size N, the coreference graph is defined as the matrix $G \subset \mathbb{N}^{N \times N}$ of links between tokens. Here, the relation type between two tokens, x_i and x_j, is encoded with natural numbers and is denoted as $g_{i,j} \in \{0, 1, 2\}$. We define three relation types: (0) no link, (1) mention link, and (2) coreference link.

Mention link This type of link serves to identify mentions. It relates tokens belonging to the same mention-span m. We define mention links in two different manners depending on whether the graph is an input or output of the model (refer to Section 8.5 for details). When the graph is an input G^{in}, there is a directed link from each mention's token to the mention head[1], including the head to itself. Let $\mathsf{start}(m)$, $\mathsf{end}(m)$, and $\mathsf{head}(m)$ be the functions to obtain the mention's starting, ending, and head index. Then:

$$g^{in}_{i,j} = 1 \iff i \in [\mathsf{start}(m), ..., \mathsf{end}(m)] \land j = \mathsf{head}(m)$$

[1] In linguistics, the head of a phrase is the word that determines the syntactic category. Every mention has a unique head even in the case of nested mentions.

When the graph is the model's output G^{out}, there is only one directed link from the last token of the mention-span to the first token. Then:

$$g^{out}_{i,j} = 1 \iff i = \mathsf{end}(m) \wedge j = \mathsf{start}(m)$$

The difference between input and output encoding is functional. Both encoding methods define a mention-span uniquely, even when having nested mentions; every mention has a unique start-end combination and a unique head. The model utilizes the output for prediction, so it is simpler to predict one single link, whereas, in the input, the model uses links to all tokens to represent every mention token properly.

Mention heads We simplified the head identification process by considering the first token of a mention span as the head. Although this method is naive, experiments show that this approximation works well enough in practice. However, as some spans can potentially have the same head in case of nested mentions, we fix this issue by defining heuristics. First, we order the mentions by size. Then, we iterate over this list to assign heads; if the first token of the mention is already a head of another mention, we assign the next token as the head. In this manner, every mention has a unique head to be identifiable.

Coreference link This type of link defines the relationship between a mention m and each of its antecedents a. Here, we define coreference links in two different manners depending on whether the graph is an input or output of the model. When the graph is input, there is a directed link from a mention head token to each one of its antecedents' head token. Then:

$$g^{in}_{i,j} = 2 \iff i = \mathsf{head}(m) \wedge j = \mathsf{head}(a) \wedge a = \mathsf{antecedent}(m)$$

When the graph is a model's output, the mention m should be connected to at least one of its antecedents a. Here, we only consider antecedents that appear before the mention. If the mention has no antecedent, or correspond to the first mention of an entity in the text, then it is connected to a null antecedent ϵ.

$$\exists j \in \{\epsilon, \mathsf{head}(a) \mid a = \mathsf{antecedent}(m) \wedge \mathsf{start}(a) < \mathsf{start}(m)\} :$$

$$g^{in}_{i,j} = 2 \wedge i = \mathsf{head}(m)$$

We consider all possible connections between mentions in an entity cluster at the input, so the model has complete information. On the other hand, we consider that predicting at least one connection of the mention to its cluster is sufficient during output.

Figure 8.1 shows an example taken from CoNLL 2012 dataset (Section 2.5.2). There, mention spans are shown in bold, and colors represent entity clusters. The mention links are indicated with continuous line arrows and the coreference links with dotted arrows.

The objective is to learn the conditional probability distribution $p(G|D)$. This distribution is approximated by assuming independence among each relation $g_{i,j}$ as:

$$p(G|D) = \prod_{i=1}^{N} \prod_{j=1}^{i} p(g_{i,j}|D) \tag{8.3}$$

The probability $p(g_{i,j}|D)$ is split in two parts: one for mention links p_m and the other for coreference links p_c. The mention link probability is defined as:

$$p_m(g_{i,j} = 1|D) = \sigma(W_m \cdot [h_i, h_j]) \tag{8.4}$$

where W_m is a parameter matrix, and h_i and h_j are the hidden state representations of the tokens x_i and x_j respectively. This probability indicates whether there is a mention starting at position i and ending at position j of the document D. The optimization is done using binary-cross-entropy loss, defined as follows:

$$loss_m = \sum_{i=1}^{N} \sum_{j=1}^{i} \hat{g}_{i,j} \cdot p_m(g_{i,j}) + (1 - \hat{g}_{i,j}) \cdot (1 - p_m(g_{i,j})) \tag{8.5}$$

$$\tag{8.6}$$

where $\hat{g}_{i,j}$ is the annotated label; it is 1 when there is a mention starting on j and finishing at i and 0 otherwise.

The coreference link probability is defined as:

$$p_c(g_{i,j} = 2|D) = \frac{exp(W_c \cdot [h_i, h_j])}{\sum_{j' \in \mathrm{A}(i)} exp(W_c \cdot [h_i, h_{j'}])} \tag{8.7}$$

where W_m is a parameter matrix, and h_i and h_j are the hidden state representations of the tokens x_i and x_j respectively. Similar to the baseline, we denote $\mathrm{A}(i)$ as the set of all candidate antecedents of x_i. This set contains all mention heads with an index less than i, plus a null head ϵ, $\mathrm{A}(i) = \{\epsilon, x_k | k < i$ and $x_k \in \mathrm{H(D)}\}$, and H(D) is the set of all mention heads in the document. The optimization is done with cross-entropy loss. Given that a mention-span m_i can have more than one true antecedent, the loss considers the sum of probabilities of all true antecedents in the annotated data (as in Equation(8.2)):

$$loss_c = log \prod_{i \in \mathrm{H}(D)} \sum_{j \in \mathrm{Y}(i) \cap \hat{\mathrm{C}}(i)} p_c(g_{i,j}|D) \tag{8.8}$$

where $\hat{\mathrm{C}}(i)$ indicates the annotated cluster of mention-spans that includes m_i in the annotated data. If the mention does not belong to any cluster, then the span is assigned to the null cluster $\hat{\mathrm{C}}(i) = \{\epsilon\}$. The final loss is the sum of $loss_m$ and $loss_c$.

We use BERT (Devlin et al., 2019) to encode the document and obtain the token's hidden state representations $\{h_1,..,h_N\}$. BERT is a pre-trained masked language model that uses a transformer architecture (Section 6.3). This model is trained on a large quantity of text data to predict masked tokens on the text based on the unmasked ones. BERT representations were used to improve the baseline coreference resolution described in Section 8.3 with significant results (Joshi et al., 2019) .

8.5 Iterative Refinement

The strong independence assumption made in Equation (8.3) does not reflect real scenario and could lead to poor performance. Therefore, we use an iterative refinement approach to include dependencies similar to (Mohammadshahi and Henderson, 2020). We called our model *graph-to-graph transformer* (G2GT). Under this approach, the model makes T iterations over the same document D. At each iteration t, the predicted coreference graph G_t is conditioned on the previously predicted one G_{t-1}. The model's conditional probability distribution is now defined as follows:

$$p(G^t|D,G^{t-1}) = \prod_{i=1}^{N}\prod_{j=1}^{i} p(g_{i,j}|D,G^{t-1}) \tag{8.9}$$

It means that the graph should be input to the transformer model. Following (Mohammadshahi and Henderson, 2020), the graph is encoded in the self-attention function of the transformer, this architecture is described in Section 6.3. We modify Equation (6.1) for our purpose as follows:

$$\begin{aligned} \mathsf{Attention}(Q,K,V,L_k,L_v) &= \text{softmax}(\frac{Q\cdot(K+L_k)^\intercal}{\sqrt{d}})\cdot(V+L_v) \qquad (8.10)\\ \text{where} \qquad L_v &= E(G_{t-1})\cdot W_v \\ L_k &= E(G_{t-1})\cdot W_k \end{aligned}$$

where E is an embedding matrix to encode the types of links in the graph. Thus, the relationship between a pair of tokens is encoded as an embedding vector. W_k, W_v are weight parameters that serve to specialize $E(G_{t-1})$ to be either *key* or *value* vector. The complexity of our model is of the order of $\mathcal{O}(N^2 \times T)$, where N is the document length, and T is the number of refinement iterations of the model.

To illustrate the iterative refinement of a graph, Figure 8.2 shows an example of two iterations of the model. The mention links are indicated with continuous line arrows and the coreference links with dotted arrows. The initial graph matrix G_0^{in} is full of zeros, so no connections are drawn. The first predicted graph G_1^{out} only has mention-links because there were no mention heads to be connected initially. This graph is transformed to serve as input G_1^{in} for the next iteration. Finally, during the second iteration, the model predicts the coreference graph G_2^{out}. The model can continue iterating for a maximum of T times.

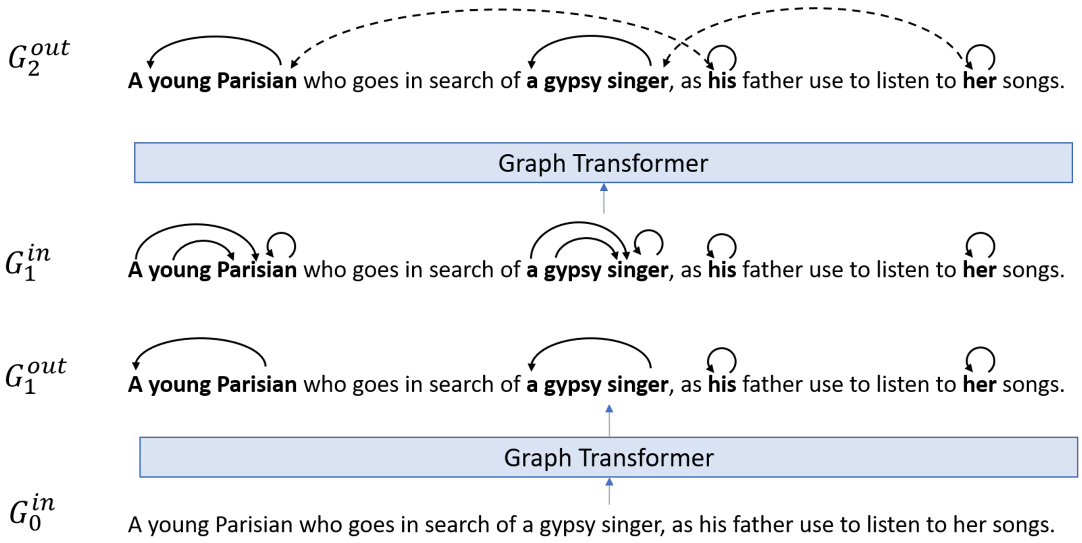

G_2^{out}
A young Parisian who goes in search of a gypsy singer, as his father use to listen to her songs.
Graph Transformer
G_1^{in}
A young Parisian who goes in search of a gypsy singer, as his father use to listen to her songs.
G_1^{out}
A young Parisian who goes in search of a gypsy singer, as his father use to listen to her songs.
Graph Transformer
G_0^{in}
A young Parisian who goes in search of a gypsy singer, as his father use to listen to her songs.

Figure 8.2 – Example of iterations with graph-to-graph transformer.

8.6 Architectures

There exists in practice a maximum length for encoding a document due to the limited computational memory. In this section, we describe two strategies to manage this issue: overlapping window and reduced document.

8.6.1 Overlapping Windows

Here, we split the documents into overlapping segments of the maximum size K, with an overlap of $K/2$ tokens. The segments are encoded individually on our graph-to-graph transformer model. During training, each segment is treated as an independent sample. However, during decoding, the segments are decoded in order. The subgraph corresponding to the overlapping part is input to the next segment. The following segments predict new links starting only on the non-overlapping part. The union of the segmented graphs forms the final graph.

8.6.2 Reduced Document

This model has two parts; one to detect mentions and the other to perform coreference resolution. The mention detection is similar to the previously described model. The coreference resolution part receives a shorter version of the document as input. The complete model is described in the following:

Mention Detection This transformer is non-iterative so it corresponds to the definition in Equation (8.3). To encode the document, we apply overlapping windows, as in the previous section. For prediction, we used the *soft-target* method proposed in Section 7.5 of the

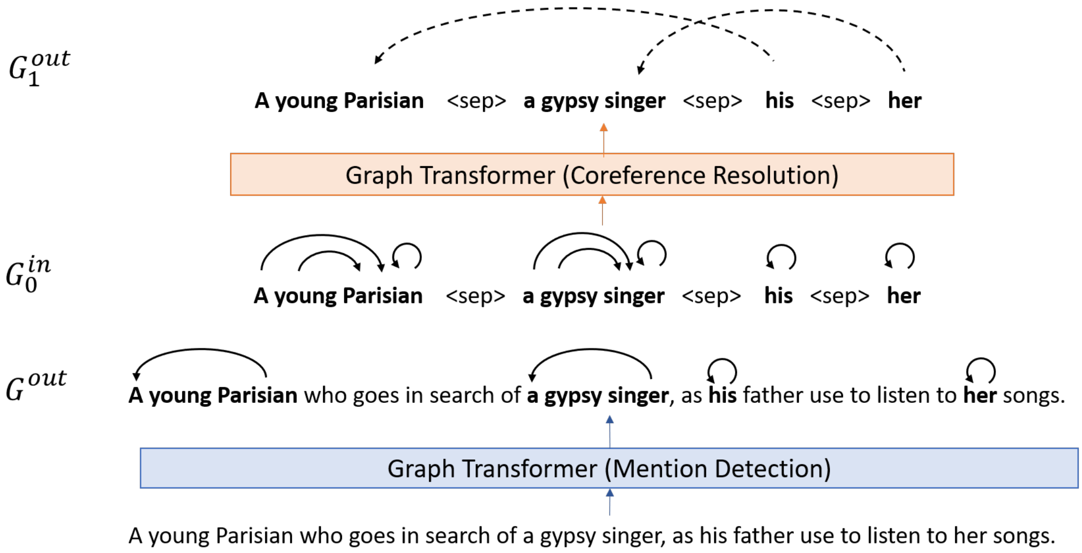

G_1^{out}
A young Parisian <sep> a gypsy singer <sep> his <sep> her
Graph Transformer (Coreference Resolution)
G_0^{in}
A young Parisian <sep> a gypsy singer <sep> his <sep> her
G^{out}
A young Parisian who goes in search of a gypsy singer, as his father use to listen to her songs.
Graph Transformer (Mention Detection)
A young Parisian who goes in search of a gypsy singer, as his father use to listen to her songs.

Figure 8.3 – Example of iterations with graph-to-graph transformer in pipeline.

previous chapter. This method enables the model to increase the recall of detection. Given that the candidate mentions will be fixed for the coreference resolution part, we need to detect most of them here.

Coreference Resolution This part is a graph-to-graph transformer with iterative refinement. The input is a shorter version of the document obtained by concatenating the tokens from mention-spans with a separation token in between and removing all other tokens. To maintain coherence in the document, we modify the token input representation as to the sum of three vectors: (a) a token embedding, (b) an embedding of the token's position in the original document, so we retain information of distance between mentions, and (c) the token's contextualized representation obtained from the mention detection part where the original document is encoded. This second part predicts only coreferent links, but the input graph contains both mentions and coreference links.

Figure 8.3 shows an example of this architecture with one iteration over a document. The mention links are indicated with continuous line arrows and the coreference links with dotted arrows. The first model predicts the graph of mention-spans G^{out}. This graph is transformed into the input format for the next model G_0^{in}. Then, the second model predicts the graph of coreference G_1^{out}. Note that this coreference resolution model can continue iterating for T times.

8.7 Experimental Setting

This section describes the dataset, the model configuration, hyper-parameters, and training setup used in our experiments.

	Train	**Dev.**	**Test**	**Total**
Number of documents	2,802	343	348	3,493
Number of words	1.3 M	160 K	170 K	1.6 M
Average word length per document	464	466	488	458
Number of entity changes/clusters	35 K	4.5 K	4.5 K	44 K
Number of coreference links	120 K	14 K	15 K	150 K
Number of mentions	155 K	19 K	19 K	194 K

Table 8.1 – Dataset statistics and splits.

8.7.1 Dataset

We use the CoNLL 2012 corpus described in Section 2.5.2. It contains data from diverse domains e.g., newswire, magazine, conversations. We experiment only with the English part. Table 8.1 shows the statistics of the dataset; the average length per document does not exceed the 500 words, however this length increases during preprocessing. We pre-processes the text to extract sub-word units (Section 2.2.1) with BERT tokenizer (Wu et al., 2016). Sub-words are smaller token units used to manage out of vocabulary words. That means that the actual document length is even larger in terms of sub-word units. We map the positional annotation of mentions from words to sub-words and retain this mapping for back transformation during evaluation. We do not use any annotation for word markers (e.g., POS, syntax), given in the dataset, only raw words. There are no annotations for stand-alone mentions i.e., entities mentioned only once in the document are not considered.

8.7.2 Model configuration

We adopted the `bert_base_uncased` pre-trained model (Devlin et al., 2019) and the implementation of Wolf et al. (2019) [2]. All hyper-parameters follow this implementation unless specified otherwise.

Training The graph-to-graph transformer considers an independent loss for each different refinement iteration. There is no back-propagation between refinement iterations because the model makes discrete decisions when predicting the graph for the next refinement step. There are two stopping criteria for the refinement: (a) when a maximum number of iterations T is reached, or (b) when there are no more changes in the graph, $G_t = G_{t-1}$. This criteion is for both training and testing. Our models are trained with a maximum segment length of $K = 512$ and a batch size of 1 document. We use BertAdam (Kingma and Ba, 2014; Wolf et al., 2019) optimizer with a base learning rate of $2e-3$ and no warm-up. As our graphs are directed, we use only the lower triangle of G for predictions. The components of the reduced models are trained independently. The coreference resolution follows the currently described training schema. The mention detection model has no

[2] https://huggingface.co/transformers/

		Development				Test			
Model	**Iter.**	**MUC**	**B^3**	**CEAF$_{\phi_4}$**	**Avg. F1**	**MUC**	**B^3**	**CEAF$_{\phi_4}$**	**Avg. F1**
GT BERT-base	$T = 1$	75.7	68.4	65.2	69.8	76.1	67.0	66.3	69.8
	$T = 2$	76.9	69.3	66.0	70.7	77.9	69.9	67.0	71.6
	$T = 3$	77.2	69.7	66.3	71.0	78.1	70.3	67.0	71.8
G2GT BERT-base *reduced*	$T = 1$	79.2	76.1	68.5	71.6	80.7	70.1	69.3	73.4
	$T = 3$	80.0	69.6	70.2	73.3	83.2	71.9	71.0	75.4

Table 8.2 – F1 scores by varying the number of refinement iterations T.

	MUC			B^3			CEAF$_{\phi_4}$			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**Avg. F1**
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Clark and Manning (2016)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Lee et al. (2017a)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Fei et al. (2019)	85.4	77.9	81.4	77.9	66.4	71.7	70.6	66.3	68.4	73.8
Baseline (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
+ BERT-base (Joshi et al., 2019)	80.4	82.3	81.4	69.6	73.8	71.7	69.0	68.5	68.8	73.9
+ BERT-large (Joshi et al., 2019)	84.7	82.4	83.5	76.5	74.0	75.3	74.1	69.8	71.9	76.9
G2GT BERT-base	78.4	77.9	78.1	69.6	71.0	70.3	66.8	67.3	67.0	71.8
G2GT BERT-base *overlap*	81.2	82.8	82.0	69.8	73.6	71.6	69.6	69.3	69.4	74.4
G2GT BERT-base *reduced*	83.4	83.1	83.2	70.1	73.7	71.9	72.1	70.1	71.0	75.4

Table 8.3 – Coreference resolution evaluation (CoNLL 2012)

iterative refinement step and follows the training schema of our *span scoring soft-target* approach described in Section 7.5 of the previous chapter, with $\rho = 0.1$.

Evaluation At evaluation time, we map back all sub-words units to words and reconstruct the document in CoNLL 2012 format (Section 2.5.2). We use the precision, recall, and F1 score calculated in three different manners: MUC that counts the number of links between mentions, B^3 that counts the number of mentions, and CEAF that counts the entity clusters. These metrics are standard for coreference resolution and are defined in Section 2.4.2.

8.8 Results Analysis

This section describes the results of various baselines and our models. First, we analyze the optimum number of refinement iterations, and then we show results using the best models.

Table 8.2 shows the performance of our graph-to-graph transformer (G2GT) models when varying the maximum number of refinement iterations T from 1 to 3. The results are in terms of the F1 score of the three coreference metrics and the average. Both implementations shown in the table perform the best when using $T = 3$. There is a significant decrease in performance when the graphs are not refined $T = 1$. We chose the best models on the development set, but we also show the test set results.

Table 8.3 shows the evaluation results on the test set in terms of precision (P), recall (R), and F1 score for each metric. The last column displays the average F1 of the three metrics. The first section of the table exhibits scores of different coreference resolution systems from the literature. The second section shows the result of the 'Baseline' (Lee et al., 2018) system described in Section 8.3. This model uses ELMo[3] (Peters et al., 2018) instead of BERT to obtain word representations. Baseline plus 'BERT-base' and 'Base BERT-large' (Joshi et al., 2019) correspond to the baseline model replacing the ELMo word representations with the indicated BERT architectures. We copy all these values from the original papers. The last section of the table presents scores of our graph-to-graph transformer (G2GT) models with iterative refinement. 'G2GT Bert-base' is our model with no special treatment for document length; the documents are truncated at the maximum segment length of K for both training and testing. 'G2GT Bert-base *overlap*' and 'G2GT Bert-base *reduce*' are the models described in Section 8.6.

G2GT Bert-base performs poorly in comparison to the baseline. Both G2GT Bert-base *overlap* and *reduce* perform better than the comparable baseline Base BERT-base. The baseline with BERT-large is not directly comparable with our results because the improvement is orthogonal to the CR methods. We expect that our model would improve when using BERT-large[4] in similar proportion as the baseline. The reduce document strategy performs better than overlapping window. We believe that encoding all the document mentions together helps to coordinate the model decisions, and these results can vary in favor of G2GT Bert-base with better computational resources and larger segment length.

8.9 Conclusion

We proposed a graph-to-graph transformer model with iterative refinement for coreference resolution. For this purpose, we define a graph structure to encode coreference links contained in a document. That enables our model to predict the complete coreference graph at once. The graph is then refined in a recursive manner, iterating over the model conditioned on the document and the graph prediction from previous step. We experimented with two methods to manage long documents and maintain the computation efficient. The first method encodes the document in overlapping segments. The second one reduces the document size to a more manageable length. The evaluation shows that both methods outperform a comparable baseline and that the second method has better performance than the first one. This experiment shows that document-level information is useful to improve coreference and should not be dismissed. The models can be further improved by replacing the underlying language model BERT-base (Devlin et al., 2019) with a more powerful pre-trained model such as BERT-large (Devlin et al., 2019), XLNet (Yang et al., 2019a), SpanBERT (Joshi et al., 2020), Longformer (Beltagy et al., 2020), or BART (Lewis et al., 2020). Also, the decoding approach of graphs obtained from the overlapping segments can be improved.

[3] Embeddings from Language Models (ELMo) (Peters et al., 2018) is a pre-trained language model based on LSTMs.

[4] We do not report BERT-large for now due to hardware limitations

9 Conclusions and Future Work

9.1 General Conclusions

In this thesis, we addressed the problem of discourse phenomena in machine translation. Particularly, we focused on the coherent translation of entity mentions. We divided the thesis into two parts. The first part describes our main work on machine translation task. The second one presents our complementary work on coreference resolution. Our main conclusions are the following:

- We proposed a simple metric to evaluate the accuracy of pronoun translation (APT) and show that it has a high correlation with the human judgment of correctness. APT is a reference-based metric thus it has strong constraints on the correctness of the pronouns, but when averaged over a large number of translations the scores reflect similar results as human evaluation. APT has limitations when evaluating challenging instances, thus for finer evaluation test suites can be used instead, when available. APT can be used without the restriction of language and can be adapted to evaluated nouns or other specific word types. This metric is utilizes to evaluate our proposed approaches.

- We tested the hypothesis that discourse information can effectively improve machine translation. We showed that this is the case for coreferences. We work under the assumption that the clusters of mentions referring to the same entity should be equal in the source and target languages. We proposed two methods that improve entity translation based on this assumption: re-ranking and post-editing. Our post-editing approach brought a significant improvement for pronoun translation as evaluated by APT, and improved the general translation quality according to a manual evaluation.

- We integrated discourse phenomena in the translation models. We proposed a novel decoder based on self-attention and residual connections. The self-attention allows selective access to contextual information and the residual connection facilities the flow of information. Additionally, we adapted similar architectures, designed for language modeling, to translation. Experiments in various datasets showed improvement of our method in comparison to different baselines. Finally, a qualitative analysis showed that the proposed model learns structures resembling syntactic ones.

- We established the problem as document-level machine translation. For this purpose, we proposed to use a multi-head hierarchical attention network (HAN) model. We integrated context from source and target sides by directly connecting representations from previous sentence translations into the current sentence translation. The model significantly outperforms two competitive baselines, and the ablation study shows that the target and source context is complementary. It also improves lexical cohesion and coherence, and the translation of nouns and pronouns. Our experiments show that varying the input context to incorrect one has a negative effect on the translations, thus our model is sensitive to context. With this work, we showed that discourse information improves significantly translation verifying our thesis hypothesis that discourse and context improves translation.

- Finally, we were interested in coreference resolution as it is a particularly challenging discourse phenomenon. First, we worked on mention detection. We evaluated two models, one of which was novel, and proposed a method to learn from partially-annotated data. We showed that our method boost detection recall without decreasing precision. Second, we proposed a graph-based coreference resolution model. This model works with iterative refinement and showed to outperform the baseline model. The model can encode discourse relations (in this case coreferences) explicitly as a substructure of the language model. This has the potential to (a) extract discourse information which is usually embedded in language, and (b) manipulate the discourse information for certain goals.

9.2 Directions for Future Research

Although the work presented here demonstrates improvement in machine translation, there is an immense space for further amelioration. We identify some research directions in the following:

- In our document-level translation model, the maximum effective context used is around three sentences. We believe it is because the context is summarized in a single vector with limited capacity. Recently, with more powerful hardware, NMT models can process larger sequences, thus a document can be fully encoded (Junczys-Dowmunt, 2019). Processing a document as a single sequence is simpler and does not limit the available context. Thus, neural networks can learn internal discourse connections easily. Moreover, these models can profit from research on efficient attention-based networks that require less memory and have better algorithm complexity.

- Although in some cases a document could be encoded in a single sequence sample, in practice, there is always a limit for the sequence length. The previous strategy is not sufficient for all scenarios e.g. when translating a book. Thus, the problem of extracting, summarizing, and retaining meaningful contextual information remains of interest for future research. This problem overlaps with several other tasks such as dialog generation, summarization, machine reading, etc.

- The evaluation of document-level machine translations remains a challenging problem. Even though automatic metrics are helpful, they are not sensitive enough to evaluate difficult translation cases that require text comprehension. There is a line of research focusing on test suites (Bawden et al., 2018) and human evaluation (Bojar et al., 2019).

- We have investigated context at the document-level. However, context is a more generic concept that involves all circumstances around the produced language. For instance previous word knowledge, background, and environment. This is also an interesting topic for any natural language task and specifically for translation.

A Self-attentive Networks Architecture

This appendix describes in detail the implementation of the *self-attentive residual decoder* for NMT, which builds on the attention-based NMT implementation of `dl4mt-tutorial`[1].

The input of the model is a source sentence denoted as 1-of-k coded vector, where each element of the sequence corresponds to a word:

$$x = (x_1, x_2, ..., x_m), x_i \in \mathbb{R}^V$$

and the output is a target sentence denoted as well as 1-of-k coded vector:

$$y = (y_1, y_2, ..., y_n), y_i \in \mathbb{R}^V$$

where V is the size of the vocabulary of target and source side, m and n are the lengths of the source and target sentences respectively. We omit the bias vectors for simplicity.

A.1 Encoder

Each word of the source sentence is embedded in a e-dimensional vector space using the embedding matrix $\bar{E} \in \mathbb{R}^{e \times V}$. The hidden states are $2d$-dimensional vectors modeled by a bi-directional GRU. The forward states $\overrightarrow{h} = (\overrightarrow{h}_1, ..., \overrightarrow{h}_m)$ are computed as:

$$\overrightarrow{h}_i = \overrightarrow{z}_i \odot \overrightarrow{h}_{i-1} + (1 - \overrightarrow{z}_i) \odot \overrightarrow{h}'_i$$

where

$$\begin{aligned}
\overrightarrow{h}'_i &= tanh(\overrightarrow{W}\bar{E}x_i + \overrightarrow{U}[\overrightarrow{r}_i \odot \overrightarrow{h}_{i-1}]) \\
\overrightarrow{z}_i &= \sigma(\overrightarrow{W}_z\bar{E}x_i + \overrightarrow{U}_z\overrightarrow{h}_{i-1}) \\
\overrightarrow{r}_i &= \sigma(\overrightarrow{W}_r\bar{E}x_i + \overrightarrow{U}_r\overrightarrow{h}_{i-1})
\end{aligned}$$

Here, $\overrightarrow{W}, \overrightarrow{W}_z, \overrightarrow{W}_r \in \mathbb{R}^{d \times e}$ and $\overrightarrow{U}, \overrightarrow{U}_z, \overrightarrow{U}_r \in \mathbb{R}^{d \times d}$ are weight matrices. The backward states $\overleftarrow{h} = (\overleftarrow{h}_1, ..., \overleftarrow{h}_m)$ are computed in similar manner. The embedding matrix $\bar{E}$ is shared for both passes, and the final hidden states are formed by the concatenation of them:

$$h_i = \begin{bmatrix} \overrightarrow{h}_i \\ \overleftarrow{h}_i \end{bmatrix}$$

[1] https://github.com/nyu-dl/dl4mt-tutorial

A.2 Attention Mechanism

The *context vector* at time t is calculated by:

$$c_t = \sum_{i=1}^{m} \alpha_i^t h_i$$

where

$$\alpha_i^t = \frac{exp(e_i^t)}{\sum_j exp(e_j^t)}$$
$$e_i^t = v_a^\intercal tanh(W_d s_{t-1} + W_e h_i)$$

Here, $v_a \in \mathbb{R}^d$, $W_d \in \mathbb{R}^{d\times d}$ and $W_e \in \mathbb{R}^{d\times 2d}$ are weight matrices.

A.3 Decoder

The input of the decoder are the previous word y_{t-1} and the *context vector* c_t, the objective is to predict y_t. The hidden states of the decoder $s = (s_1, ..., s_n)$ are initialized with the mean of the *context vectors*:

$$s_0 = tanh(W_{init} \frac{1}{m} \sum_{i=1}^{m} c_i)$$

where $W_{init} \in \mathbb{R}^{d\times 2d}$ is a weight matrix, m is the size of the source sentence. The following hidden states are calculated with a GRU conditioned over the *context vector* at tine t as follows:

$$s_t = z_t \odot s_t' + (1 - z_t) \odot s_t''$$

where

$$s_t'' = tanh(Ey_{t-1} + U[r_t \odot s_{t-1}] + Cc_t)$$
$$z_i = \sigma(W_z E y_{t-1} + U_z s_{t-1} + C_z c_t)$$
$$r_i = \sigma(W_r E y_{t-1} + U_r s_{t-1} + C_r c_t)$$

Here, $E \in \mathbb{R}^{e\times V}$ is the embedding matrix for the target language. $W, W_z, W_r \in \mathbb{R}^{d\times e}$, $U, U_z, U_r \in \mathbb{R}^{d\times d}$, and $C, C_z, C_r \in \mathbb{R}^{d\times 2d}$ are weight matrices. The intermediate vector s_t' is calculated from a simple GRU:

$$s_t' = GRU(y_{t-1}, s_{t-1})$$

In the attention-based NMT model, the probability of a target word y_t is given by:

$$p(y_t|s_t, y_{t-1}, c_t) = softmax(W_o tanh(W_{st} s_t + W_{yt} y_{t-1} + W_{ct} c_t))$$

Here, $W_o \in \mathbb{R}^{V\times e}$, $W_{st} \in \mathbb{R}^{e\times d}$, $W_{yt} \in \mathbb{R}^{e\times e}$, $W_{ct} \in \mathbb{R}^{e\times 2d}$ are weight matrices.

A.3.1 Self-Attentive Residual Connections

In our model, the probability of a target word y_t is given by:

$$p(y_t|s_t, d_t, c_t) = softmax(W_o tanh(W_{st} s_t + W_{dt} d_t + W_{ct} c_t))$$

Here, $W_o \in \mathbb{R}^{V\times e}$, $W_{st} \in \mathbb{R}^{e\times d}$, $W_{dt}, W_{yt} \in \mathbb{R}^{e\times e}$, $W_{ct} \in \mathbb{R}^{e\times 2d}$ are weight matrices. The summary vector d_t can be calculated in different manners based on previous words y_1 to y_{t-1}. First, a simple average:

$$d_t^{avg} = \frac{1}{t-1}\sum_{i=1}^{t-1} y_i$$

The second, by using an attention mechanism:

$$d_t^{cavg} = \sum_{i=1}^{t-1} \alpha_i^t y_i$$
$$\alpha_i^t = \frac{exp(e_i^t)}{\sum_{j=1}^{t-1} exp(e_j^t)}$$
$$e_i^t = v^\intercal tanh(W_y y_i)$$

where $v \in \mathbb{R}^e$, $W_y \in \mathbb{R}^{e\times e}$ are weight matrices.

A.3.2 Memory RNN

This model modifies the recurrent layer of the decoder as follows:

$$s_t = z_t \odot s_t' + (1 - z_t) \odot s_t''$$

where

$$s_t'' = tanh(Ey_{t-1} + U[r_t \odot \tilde{s}_t] + Cc_t)$$
$$z_i = \sigma(W_z E y_{t-1} + U_z \tilde{s}_t + C_z c_t)$$
$$r_i = \sigma(W_r E y_{t-1} + U_r \tilde{s}_t + C_r c_t)$$

Here, $E \in \mathbb{R}^{e\times V}$ is the embedding matrix for the target language. $W, W_z, W_r \in \mathbb{R}^{d\times e}$, $U, U_z, U_r \in \mathbb{R}^{d\times d}$, and $C, C_z, C_r \in \mathbb{R}^{d\times 2d}$ are weight matrices. The intermediate vector s_t' is calculated from a simple GRU:

$$s_t' = GRU(y_{t-1}, \tilde{s}_t)$$

The recurrent vector $\tilde{s}_t$ is calculated as following:

$$\begin{aligned} \tilde{s}_t &= \sum_{i=1}^{t-1} \alpha_i^t s_i \\ \text{where} \qquad \alpha_i^t &= \frac{exp(e_i^t)}{\sum_{j=1}^{t-1} exp(e_j^t)} \\ e_i^t &= v^\intercal tanh(W_m s_i + W_s s_t) \end{aligned}$$

where $v \in \mathbb{R}^d$, $W_m \in \mathbb{R}^{d \times d}$, and $W_s \in \mathbb{R}^{d \times d}$ are weight matrices.

A.3.3 Self-Attentive RNN

The formulation of this decoder is as following:

$$p(y_t|y_1,...,y_{t-1},c_t) \approx softmax(W_o tanh(W_{st}s_t + W_{yt}y_{t-1} + W_{ct}c_t + W_{mt}\tilde{s}_t))$$

Here, $W_o \in \mathbb{R}^{V \times e}$, $W_{st} \in \mathbb{R}^{e \times d}$, $W_{yt} \in \mathbb{R}^{e \times e}$, $W_{ct} \in \mathbb{R}^{e \times 2d}$, and $W_{mt} \in \mathbb{R}^{e \times d}$ are weight matrices.

$$\begin{aligned} \tilde{s}_t &= \sum_{i=1}^{t-1} \alpha_i^t s_i \\ \alpha_i^t &= \frac{exp(e_i^t)}{\sum_{j=1}^{t-1} exp(e_j^t)} \\ e_i^t &= v^\intercal tanh(W_m s_i + W_s s_t) \end{aligned}$$

where $v \in \mathbb{R}^d$, $W_m \in \mathbb{R}^{d \times d}$, and $W_s \in \mathbb{R}^{d \times d}$ are weight matrices.

B Hierarchical Attention Network Examples

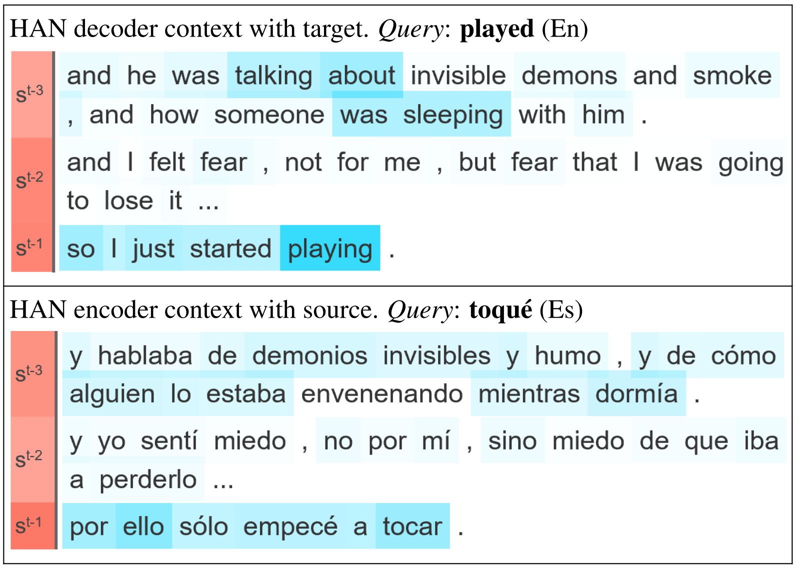
These examples were taken from the Spanish-English TED talks corpus. We show the behavior of the attention function of HAN. First, we show the attention to context for HAN encoder and HAN decoder respectively. Second, we show the multi-head attention only for HAN decoder (English) for better understanding.

B.1 Encoder and Decoder Attention

Currently Translated Sentence

Src.:	y **toqué** el primer movimiento del concierto para violín de Beethoven .
Ref.:	and I **played** the first movement of the Beethoven Violin Concerto .
Base:	and I ***touched*** the first move from the concert to Beethoven .
Cache:	and I ***touched*** the first move of Beethoven 's violin .
HAN:	and I **played** the first move of Beethoven 's violin .

Context from Previous Sentences



HAN decoder context with target. *Query*: **played** (En)

s^{t-3} and he was talking about invisible demons and smoke , and how someone was sleeping with him .

s^{t-2} and I felt fear , not for me , but fear that I was going to lose it ...

s^{t-1} so I just started playing .

HAN encoder context with source. *Query*: **toqué** (Es)

s^{t-3} y hablaba de demonios invisibles y humo , y de cómo alguien lo estaba envenenando mientras dormía .

s^{t-2} y yo sentí miedo , no por mí , sino miedo de que iba a perderlo ...

s^{t-1} por ello sólo empecé a tocar .

Table B.1 – In this example, the HAN model disambiguates correctly the word "*toqué*", which can be translated as "*touched*" or "*played*". We can see that the HAN decoder uses the semantically close word "*playing*" from the previous sentence. In similar manner, the HAN encoder focused on "*tocar*" which is coherent with "*toqué*".

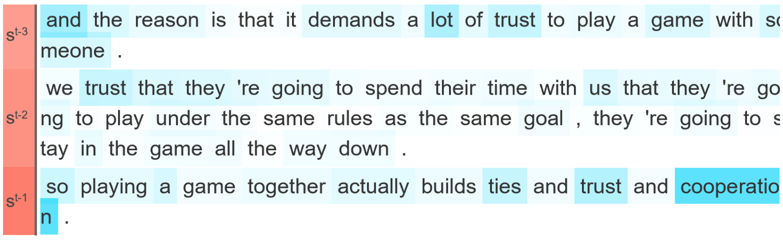
B.2 Multi-Head Attention

Currently Translated Sentence

Src:	y como resultado **construimos** relaciones sociales más fuertes .
Ref:	and we actually **build** stronger social relationships as a result .
Base:	and as a result , we ***construct*** stronger social relationships .
HAN:	and as a result , we **build** stronger social relationships .

Context from Previous Sentences. *Query*: **build**

Head 1: Attention to related words "*construimos*", "*trust*"...

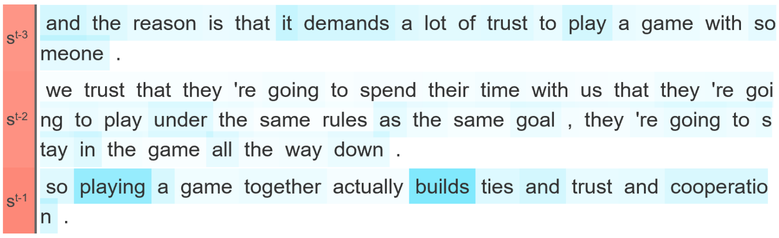


s^{t-3} and the reason is that it demands a lot of trust to play a game with s
meone .

s^{t-2} we trust that they 're going to spend their time with us that they 're go
ng to play under the same rules as the same goal , they 're going to s
tay in the game all the way down .

s^{t-1} so playing a game together actually builds ties and trust and cooperatio
n .

Head 4: Attention to a similar translation "*builds*" in s^{t-1}



s^{t-3} and the reason is that it demands a lot of trust to play a game with so
meone .

s^{t-2} we trust that they 're going to spend their time with us that they 're goi
ng to play under the same rules as the same goal , they 're going to s
tay in the game all the way down .

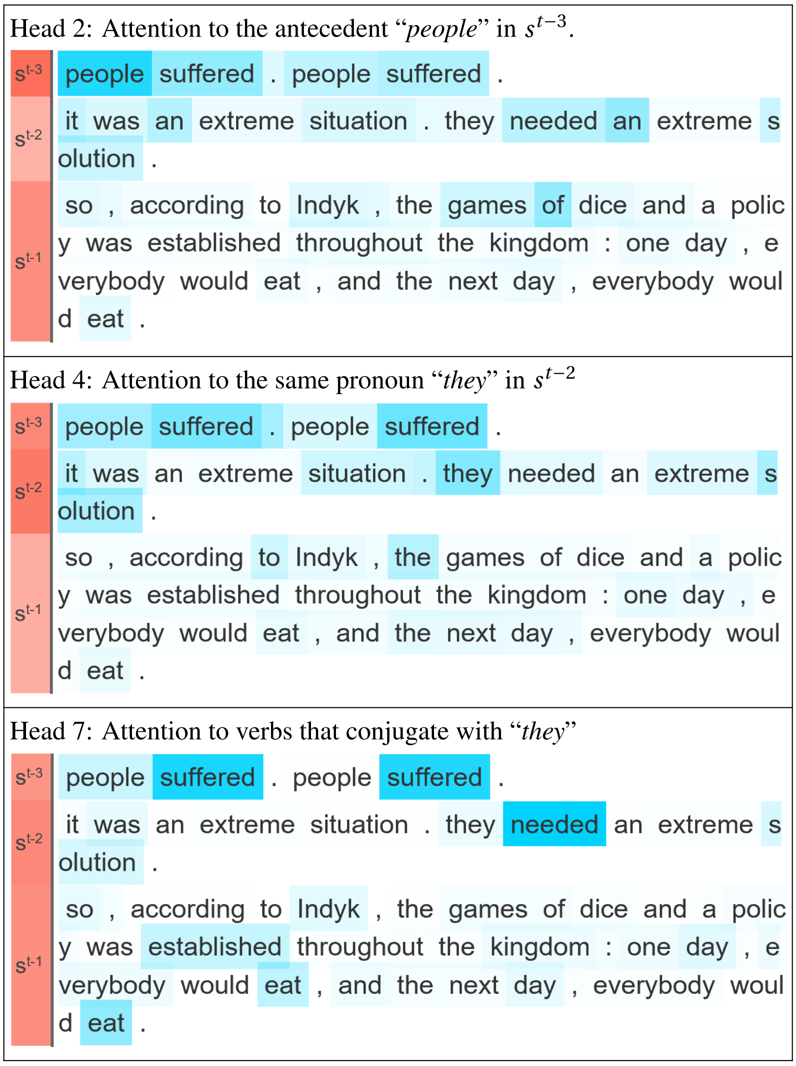
s^{t-1} so playing a game together actually builds ties and trust and cooperatio
n .

Table B.2 – This example displays the translation of the ambiguous Spanish word "*construimos*", which can be translated as "*construct*" or "*build*'. HAN translates this word correctly according to the context using for example related words "*trust*", "*ties*", and "*cooperation*" on previous sentences with *head 1*, and a previous translation "*builds*" in the previous sentence with *head 4*.

Currently Translated Sentence

Src.:	y $< ellos >$ estarían tan compenetrados en la partida de dados porque los juegos son tan atractivos ...
Ref:	and **they** would be so immersed in playing the dice games because games are so engaging ..
Base:	and ***you*** would be so $< unk >$ in the start of it because games are so attractive ...
HAN:	and **they** would be so $< unk >$ in the start of dice because games are so attractive ...

Context from Previous Sentences. *Query*: **they**



Head 2: Attention to the antecedent "*people*" in s^{t-3}.

s^{t-3} people suffered . people suffered .

s^{t-2} it was an extreme situation . they needed an extreme s olution .

s^{t-1} so , according to Indyk , the games of dice and a polic y was established throughout the kingdom : one day , e verybody would eat , and the next day , everybody woul d eat .

Head 4: Attention to the same pronoun "*they*" in s^{t-2}

s^{t-3} people suffered . people suffered .

s^{t-2} it was an extreme situation . they needed an extreme s olution .

s^{t-1} so , according to Indyk , the games of dice and a polic y was established throughout the kingdom : one day , e verybody would eat , and the next day , everybody woul d eat .

Head 7: Attention to verbs that conjugate with "*they*"

s^{t-3} people suffered . people suffered .

s^{t-2} it was an extreme situation . they needed an extreme s olution .

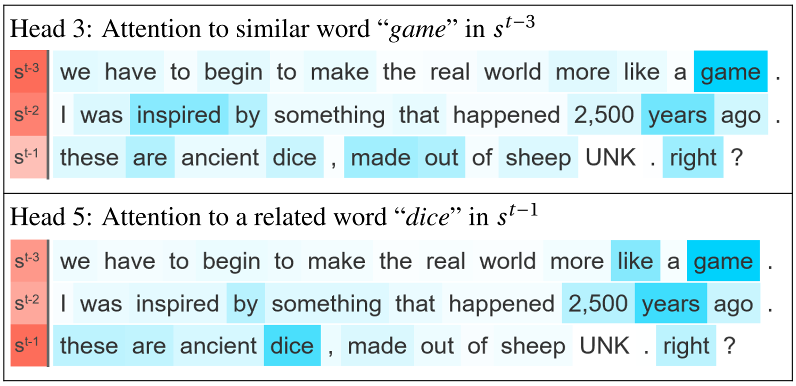
s^{t-1} so , according to Indyk , the games of dice and a polic y was established throughout the kingdom : one day , e verybody would eat , and the next day , everybody woul d eat .

Table B.3 – This example displays the translation of Spanish pronoun "*ellos*", which is a dropped-pronoun which is implicit in the verb conjugation of "*estarían*". As we can observe, HAN translates correctly the dropped-pronoun into the English "*they*". Each head focuses on a different aspect during translation, for example *head 2* seems to attend to the antecedent of the pronoun "*people*" in the third previous sentence, *head 4* attends to the same pronoun on the second previous sentence, and *head 7* attends to different verbs on all previous sentences.

Currently Translated Sentence

Src:	antes de los fantásticos controladores de **juegos** teníamos tabas de oveja .
Ref:	before we had awesome **game** controllers , we had sheep ’s knuckles .
Base:	before the fantastic ***TV*** controllers , we had $< unk >$.
HAN:	before the fantastic **game** controllers , we had $< unk >$.

Context from Previous Sentences. *Query*: **game**



Head 3: Attention to similar word “*game*” in s^{t-3}

Head 5: Attention to a related word “*dice*” in s^{t-1}

Table B.4 – This example shows the translation of the Spanish word “*juegos*”. The baseline translates it incorrectly, while HAN translates it correctly by spotting a similar translation “*game*” in the third previous sentence with *head 3*, and a related word “*dice*” on previous sentence with *head 5*.

.

Bibliography

Aggarwal, C. C. et al. (2018). *Neural networks and deep learning*. Springer.

Aone, C. and William, S. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, volume 1, pages 563–566, Granada, Spain.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, USA.

Banerjee, S. and Lavie, A. (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*, pages 65–72.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (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)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bengtson, E. and Roth, D. (2008). Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Honolulu, Hawaii. Association for Computational Linguistics.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Bird, S. (2006). Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72. Association for Computational Linguistics.

Black, E. W., Abney, S., Flickenger, D. P., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R. J. P., Jelinek, F., Klavans, J. L., Liberman, M. Y., Marcus, M. P., Roukos, S., Santorini, B., and Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove*, California, USA.

Blackwood, G., Ballesteros, M., and Ward, T. (2018). Multilingual neural machine translation with task-specific attention. *arXiv preprint arXiv:1806.03280*.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors (2019). *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Florence, Italy. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Born, L., Mesgar, M., and Strube, M. (2017). Using a graph-based coherence model in document-level machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 26–35, Copenhagen, Denmark. Association for Computational Linguistics.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, California, USA. Association for Computational Linguistics.

Bussmann, H. (2006). *Routledge dictionary of language and linguistics*. Routledge.

Callin, J., Hardmeier, C., and Tiedemann, J. (2015). Part-of-speech driven cross-lingual pronoun prediction with feed-forward neural networks. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 59–64, Lisbon, Portugal. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Cancedda, N., Dymetman, M., Foster, G., and Goutte, C. (2009). A statistical machine translation primer. *Learning machine translation*, pages 1–38.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., and Federico, M. (2015). The IWSLT 2015 evaluation campaign. In *In proceedins of the International Workshop on Spoken Language Translation*.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2014). Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*, volume 57.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 33–40.

Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas. Association for Computational Linguistics.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Cipolla, R., Gal, Y., and Kendall, A. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491. IEEE.

Clark, E., Ji, Y., and Smith, N. A. (2018). Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, New Orleans, Louisiana. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (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)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (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)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Dabre, R., Puzikov, Y., Cromieres, F., and Kurohashi, S. (2016). The Kyoto University cross-lingual pronoun translation system. In *Proceedings of the First Conference on Machine Translation*, pages 571–575, Berlin, Germany. Association for Computational Linguistics.

Daniluk, M., Rocktäschel, T., Welbl, J., and Riedel, S. (2016). Frustratingly short attention spans in neural language modeling. In *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., et al. (2012). Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, Baltimore, MD.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R. M., and Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *ACL (1)*, pages 1370–1380. Citeseer.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Fei, H., Li, X., Li, D., and Li, P. (2019). End-to-end deep reinforcement learning based coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Fernandes, E., dos Santos, C., and Milidiú, R. (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Jeju Island, Korea. Association for Computational Linguistics.

Florian, R., Hassan, H., Ittycheriah, A., Jing, H., Kambhatla, N., Luo, X., Nicolov, N., and Roukos, S. (2004). A statistical model for multilingual entity detection and tracking. In *HLT-NAACL 2004: Main Proceedings*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Fonseca, E., Yankovskaya, L., Martins, A. F. T., Fishel, M., and Federmann, C. (2019). Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Forcada, M. L. and Ñeco, R. P. (1997). Recursive hetero-associative memories for translation. In *International Work-Conference on Artificial Neural Networks*, pages 453–462. Springer.

Garcia, E. M., Creus, C., Espana-Bonet, C., and Màrquez, L. (2017). Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):85–96.

Garcia, E. M., España-Bonet, C., and Màrquez, L. (2015). Document-level machine translation with word vector models. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 59–66, Antalya, Turkey.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323.

Gong, Z., Zhang, M., and Zhou, G. (2011). Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Gong, Z., Zhang, Y., and Zhou, G. (2010). Statistical machine translation based on LDA. In *2010 4th International Universal Communication Symposium*, pages 286–290. IEEE.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Grevisse, M. and Goosse, A. (2007). *Le bon usage et son édition Internet*. Grevisse de la langue française. De Boeck Supérieur, Louvain-la-Neuve.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., and Socher, R. (2018). Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Guillou, L. (2012). Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France. Association for Computational Linguistics.

Guillou, L. (2016). *Incorporating Pronoun Function into Statistical Machine Translation*. PhD thesis, University of Edinburgh, UK.

Guillou, L. and Hardmeier, C. (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)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Guillou, L. and Hardmeier, C. (2018). Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016). Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 525–542, Berlin, Germany. Association for Computational Linguistics.

Hajlaoui, N. and Popescu-Belis, A. (2013). Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 236–247. Springer-Verlag, LNCS 7817, Samos, Greece.

Hardmeier, C. (2012). Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, 1(11).

Hardmeier, C. (2014). *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala University, Sweden.

Hardmeier, C. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289, Paris, France.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Hardmeier, C., Nivre, J., and Tiedemann, J. (2012). Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea. Association for Computational Linguistics.

Hardmeier, C., Stymne, S., Tiedemann, J., and Nivre, J. (2013). Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia, Bulgaria. Association for Computational Linguistics.

Hasler, E., Blunsom, P., Koehn, P., and Haddow, B. (2014). Dynamic topic adaptation for phrase-based mt. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 328–337.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv*, pages arXiv–1803.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Isabelle, P., Cherry, C., and Foster, G. (2017). A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017). Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Jelinek, F. (1980). Interpolated estimation of markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Joshi, M., Levy, O., Zettlemoyer, L., and Weld, D. (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)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Ju, M., Miwa, M., and Ananiadou, S. (2018). A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics.

Junczys-Dowmunt, M. (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)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Katiyar, A. and Cardie, C. (2018). Nested named entity recognition revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 861–871, New Orleans, Louisiana. Association for Computational Linguistics.

Kim, J., El-Khamy, M., and Lee, J. (2017). Residual lstm: Design of a deep recurrent architecture for distant speech recognition. *arXiv preprint arXiv:1701.03360*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P., Martin, J., Mihalcea, R., Monz, C., and Pedersen, T., editors (2005). *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54, Edmonton, Canada.

Kothur, S. S. R., Knowles, R., and Koehn, P. (2018). Document-level adaptation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 64–73, Melbourne, Australia. Association for Computational Linguistics.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Lapshinova-Koltunski, E. and Hardmeier, C. (2017). Discovery of discourse-related language contrasts through alignment discrepancies in english-german translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81, Copenhagen, Denmark. Association for Computational Linguistics.

Läubli, S., Sennrich, R., and Volk, M. (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*, Brussels, Belgium. Association for Computational Linguistics.

Le Nagard, R. and Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden. Association for Computational Linguistics.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA. Association for Computational Linguistics.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017a). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Lee, K., He, L., and Zettlemoyer, L. (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)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Lee, K., Levy, O., and Zettlemoyer, L. (2017b). Recurrent additive networks. *arXiv preprint arXiv:1705.07393*.

Lei Ba, J., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. In *NISP 2016 - Deep Learning Symposium paper*.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting* *of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Li, J. J., Carpuat, M., and Nenkova, A. (2014). Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288, Baltimore, Maryland. Association for Computational Linguistics.

Liepins, R., Germann, U., Barzdins, G., Birch, A., Renals, S., Weber, S., van der Kreeft, P., Bourlard, H., Prieto, J., Klejch, O., Bell, P., Lazaridis, A., Mendes, A., Riedel, S., Almeida, M. S. C., Balage, P., Cohen, S. B., Dwojak, T., Garner, P. N., Giefer, A., Junczys-Dowmunt, M., Imran, H., Nogueira, D., Ali, A., Miranda, S., Popescu-Belis, A., Miculicich Werlen, L., Papasarantopoulos, N., Obamuyide, A., Jones, C., Dalvi, F., Vlachos, A., Wang, Y., Tong, S., Sennrich, R., Pappas, N., Narayan, S., Damonte, M., Durrani, N., Khurana, S., Abdelali, A., Sajjad, H., Vogel, S., Sheppey, D., Hernon, C., and Mitchell, J. (2017). The SUMMA platform prototype. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–119, Valencia, Spain. Association for Computational Linguistics.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. In *Proceedings of the International Conference on Learning Representations*, Toulon, France.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association.

Liu, Y. and Lapata, M. (2018). Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Loáiciga, S. and Grisot, C. (2016). Predicting and using a pragmatic component of lexical aspect. *LiLT (Linguistic Issues in Language Technology)*, 13.

Loáiciga, S., Stymne, S., Nakov, P., Hardmeier, C., Tiedemann, J., Cettolo, M., and Versley, Y. (2017). Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation (DiscoMT)*, Copenhagen, Denmark.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, B.C., Canada.

Luong, N. Q., Miculicich Werlen, L., and Popescu-Belis, A. (2015a). Pronoun translation and prediction with or without coreference links. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 94–100, Lisbon, Portugal. Association for Computational Linguistics.

Luong, N. Q. and Popescu-Belis, A. (2016). Improving pronoun translation by modeling coreference uncertainty. In *Proceedings of the First Conference on Machine Translation*, pages 12–20, Berlin, Germany. Association for Computational Linguistics.

Luong, N. Q. and Popescu-Belis, A. (2017). Machine translation of Spanish personal and possessive pronouns using anaphora probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain.

Luong, N. Q., Popescu-Belis, A., Rios Gonzales, A., and Tuggener, D. (2017). Machine translation of spanish personal and possessive pronouns using anaphora probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 631–636, Valencia, Spain. Association for Computational Linguistics.

Luong, T., Pham, H., and Manning, D. C. (2015b). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.

Luotolahti, J., Kanerva, J., and Ginter, F. (2016). Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proceedings of the First Conference on Machine Translation*, pages 596–601, Berlin, Germany. Association for Computational Linguistics.

Ma, Q., Wei, J., Bojar, O., and Graham, Y. (2019). Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Mann, W. and Thompson, S. (1988). Rethorical structure theory: Toward a functional theory of text organization. *Text: Interdisciplinary Journal for the Study of Discourse*, 8:243–281.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Marcu, D. (1996). Building up rhetorical structure trees. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1069–1074.

Marcu, D. (1997). The rhetorical parsing of unrestricted natural language texts. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–103, Madrid, Spain. Association for Computational Linguistics.

Marcus, R. W. E. H. M., Palmer, M., Ramshaw, R. B. S. P. L., and Xue, N. (2011). Ontonotes: A large training corpus for enhanced processing. In *Joseph Olive, Caitlin Christianson, andJohn*

McCary, editors,Handbook of Natural LanguageProcessing and Machine Translation: DARPA GlobalAutonomous Language Exploitation.

Màrquez, L., Recasens, M., and Sapena, E. (2013). Coreference resolution: an empirical study based on SemEval-2010 Shared Task 1. *Language Resources and Evaluation*, 47(3):661–694.

Maruf, S. and Haffari, G. (2018). Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284. Association for Computational Linguistics.

Maruf, S., Martins, A. F. T., and Haffari, G. (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)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Mascarell, L. (2017a). *Crossing Sentence Boundaries in Machine Translation*. PhD thesis, University of Zurich.

Mascarell, L. (2017b). Lexical chains meet word embeddings in document-level statistical machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 99–109, Copenhagen, Denmark. Association for Computational Linguistics.

McCarthy, J. (1995). Using decision trees for coreference resolution. In *Proc. 14th International Joint Conf. on Artificial Intelligence (IJCAI), Quebec, Canada, Aug. 1995*.

Merity, S. (2016). Peeking into the neural network architecture used for google's neural machine translation.

Meyer, T. and Popescu-Belis, A. (2012). Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138. Association for Computational Linguistics.

Meyer, T. and Webber, B. (2013). Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Miculicich, L. and Henderson, J. (2020). Partially-supervised mention detection. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 91–98, Barcelona, Spain (online). Association for Computational Linguistics.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Miculicich Werlen, L., Pappas, N., Ram, D., and Popescu-Belis, A. (2018). Self-attentive residual decoder for 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)*, pages 1366–1379, New Orleans, Louisiana. Association for Computational Linguistics.

Miculicich Werlen, L. and Popescu-Belis, A. (2017a). Using coreference links to improve spanish-to-english machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Miculicich Werlen, L. and Popescu-Belis, A. (2017b). Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn discourse treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Mitkov, R. (2002). *Anaphora Resolution*. Longman, London, UK.

Mohammadshahi, A. and Henderson, J. (2020). Recursive non-autoregressive graph-to-graph transformer for dependency parsing with iterative refinement. *Transactions of the Association for Computational Linguistics*, 8.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (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*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Nadejde, M., Reddy, S., Sennrich, R., Dwojak, T., Junczys-Dowmunt, M., Koehn, P., and Birch, A. (2017). Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Copenhagen, Denmark. Association for Computational Linguistics.

Nakaiwa, H. and Ikehara, S. (1995). Intrasentential resolution of japanese zero pronouns in a machine translation system using semantic and pragmatic constraints. In *Proceeedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95)*, pages 96–105.

Ng, V. (2005). Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 157–164, Ann Arbor, Michigan. Association for Computational Linguistics.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden. Association for Computational Linguistics.

Nielsen, M. A. (2015). *Neural networks and deep learning*, volume 2018. Determination press San Francisco, CA.

Nirenburg, S. (1989). Knowledge-based machine translation. *Machine Translation*, 4(1):5–24.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pappas, N., Miculicich, L., and Henderson, J. (2018). Beyond weight tying: Learning joint input-output embeddings for neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 73–83, Brussels, Belgium. Association for Computational Linguistics.

Pappas, N. and Popescu-Belis, A. (2017). Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58:591–626.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Popescu-Belis, A. (1999). Evaluation numérique de la résolution de la référence: Critiques et propositions. *TAL: Traitement automatique des langues*, 40(2):117–146.

Popović, M. (2015). chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Pouget-Abadie, J., Bahdanau, D., van Merriënboer, B., Cho, K., and Bengio, Y. (2014). Overcoming the curse of sentence length for neural machine translation using automatic segmentation. *arXiv preprint arXiv:1409.1257*.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Pu, X., Mascarell, L., and Popescu-Belis, A. (2017a). Consistent translation of repeated nouns using syntactic and semantic cues. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 948–957. Association for Computational Linguistics.

Pu, X., Pappas, N., and Popescu-Belis, A. (2017b). Sense-aware statistical machine translation using adaptive context-dependent clustering. In *Proceedings of the Second Conference on Machine Translation*, pages 1–10.

Rafalovitch, A. and Dale, R. (2009). United Nations General Assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299.

Recasens, M. and Hovy, E. (2011). BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics.

Recasens, M. and Martí, M. A. (2010). Ancora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 72–82.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Rios, A. (2015). *A Basic Language Technology Toolkit for Quechua*. PhD thesis, University of Zurich.

Rios Gonzales, A. and Tuggener, D. (2017). Co-reference resolution of elided subjects and possessive pronouns in spanish-english statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 657–662, Valencia, Spain. Association for Computational Linguistics.

Ruiz, N. and Federico, M. (2011). Topic adaptation for lecture translation through bilingual latent semantic models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 294–302.

Scarton, C. and Specia, L. (2015). A quantitative analysis of discourse phenomena in machine translation. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Schwenk, H., Dchelotte, D., and Gauvain, J.-L. (2006). Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 723–730.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL (Vol. 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sim Smith, K. (2017). On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Soltani, R. and Jiang, H. (2016). Higher order recurrent neural networks. *arXiv preprint arXiv:1605.00064*.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Stanchev, P., Wang, W., and Ney, H. (2019). EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

Stanojević, M. and Sima'an, K. (2015). BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401, Lisbon, Portugal. Association for Computational Linguistics.

Stefanescu, D., Banjade, R., and Rus, V. (2014). Latent semantic analysis models on wikipedia and tasa. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Sukhbaatar, S., szlam, a., Weston, J., and Fergus, R. (2015). End-to-end memory networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Tam, Y.-C., Lane, I., and Schultz, T. (2007). Bilingual lsa-based adaptation for statistical machine translation. *Machine translation*, 21(4):187–207.

Tan, X., Zhang, L., Xiong, D., and Zhou, G. (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)*, pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.

Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., et al. (2016). Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*.

Tiedemann, J. (2010). Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Tran, K., Bisazza, A., and Monz, C. (2016). Recurrent memory networks for language modeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 321–331, San Diego, California. Association for Computational Linguistics.

Tu, Z., Liu, Y., Lu, Z., Liu, X., and Li, H. (2017). Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

Tu, Z., Liu, Y., Shi, S., and Zhang, T. (2018). Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6.

Tuggener, D. (2016). *Incremental Coreference Resolution for German*. PhD thesis, University of Zurich.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-sense disambiguation for machine translation. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 771–778.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52, Columbia, MD, USA.

Voita, E., Sennrich, R., and Titov, I. (2019a). Context-aware monolingual repair for 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)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Voita, E., Sennrich, R., and Titov, I. (2019b). 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*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (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)*, pages 1264–1274. Association for Computational Linguistics.

Wang, B., Lu, W., Wang, Y., and Jin, H. (2018a). A neural transition-based model for nested mention recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1011–1017, Brussels, Belgium. Association for Computational Linguistics.

Wang, C. (2017). Rra: Recurrent residual attention for sequence learning. *arXiv preprint arXiv:1709.03714*.

Wang, L., Tu, Z., Shi, S., Zhang, T., Graham, Y., and Liu, Q. (2018b). Translating pro-drop languages with reconstruction models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1–9, New Orleans, Louisiana, USA. AAAI Press.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017). Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Wang, M., Lu, Z., Li, H., and Liu, Q. (2016a). Memory-enhanced decoder for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 278–286, Austin, Texas. Association for Computational Linguistics.

Wang, W., Peter, J.-T., Rosendahl, H., and Ney, H. (2016b). CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Wang, Y. and Tian, F. (2016). Recurrent residual learning for sequence classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 938–943, Austin, Texas. Association for Computational Linguistics.

Weaver, W. (1955). Translation. *Machine translation of languages*, 14(15-23):10.

Weston, J., Chopra, S., and Bordes, A. (2015). Memory networks. In *Proceedings of the International Conference on Learning Representations*, San Diego, USA.

Williams, P., Sennrich, R., Nadejde, M., Huck, M., Hasler, E., and Koehn, P. (2014). Edinburgh's Syntax-Based Systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 207–214, Baltimore, Maryland, USA. Association for Computational Linguistics.

Williams, P., Sennrich, R., Nadejde, M., Huck, M., and Koehn, P. (2015). Edinburgh's Syntax-Based Systems at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 199–209, Lisbon, Portugal. Association for Computational Linguistics.

Wiseman, S., Rush, A. M., Shieber, S., and Weston, J. (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), pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Wiseman, S., Rush, A. M., and Shieber, S. M. (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*, pages 994–1004, San Diego, California. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Wong, B. T. M. and Kit, C. (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*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Woosung Kim and Khudanpur, S. (2004). Cross-lingual latent semantic analysis for language modeling. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–257.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). Document-level consistency verification in machine translation. In *Machine Translation Summit*, volume 13, pages 131–138.

Xiong, D., Ben, G., Zhang, M., Lv, Y., and Liu, Q. (2013a). Modeling lexical cohesion for document-level machine translation. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Xiong, D., Ding, Y., Zhang, M., and Tan, C. L. (2013b). Lexical chain based cohesion models for document-level statistical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1563–1573, Seattle, Washington, USA. Association for Computational Linguistics.

Xu, M., Jiang, H., and Watcharawittayakul, S. (2017). A local detection approach for named entity recognition and mention detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1247, Vancouver, Canada. Association for Computational Linguistics.

Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., and Rutherford, A. (2015). The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019a). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Yang, Z., Hu, Z., Deng, Y., Dyer, C., and Smola, A. (2017). Neural machine translation with recurrent attention modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 383–387, Valencia, Spain. Association for Computational Linguistics.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (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*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Yang, Z., Zhang, J., Meng, F., Gu, S., Feng, Y., and Zhou, J. (2019b). Enhancing context modeling with a query-guided capsule network for document-level 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)*, pages 1527–1537, Hong Kong, China. Association for Computational Linguistics.

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, B., Xiong, D., Su, J., and Duan, H. (2017). A context-aware recurrent encoder for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2424–2432.

Zhang, B., Xiong, D., su, j., Duan, H., and Zhang, M. (2016a). Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

Zhang, R., Nogueira dos Santos, C., Yasunaga, M., Xiang, B., and Radev, D. (2018). Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107, Melbourne, Australia. Association for Computational Linguistics.

Zhang, Y., Chen, G., Yu, D., Yaco, K., Khudanpur, S., and Glass, J. (2016b). Highway long short-term memory RNNs for distant speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5755–5759.

Zhao, B. and Xing, E. P. (2008). Hm-bitam: Bilingual topic exploration, word alignment, and translation. In *Advances in Neural Information Processing Systems*, pages 1689–1696.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and*

Evaluation (LREC'16), pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Zilly, J. G., Srivastava, R. K., Koutník, J., and Schmidhuber, J. (2017). Recurrent highway networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4189–4198, International Convention Centre, Sydney, Australia. PMLR.

Lesly Miculicich

+41 76 44 77 561 – lmiculicich@idiap.ch

Research Interest

Natural Language Understanding, Neural Machine Translation, Machine Learning, Deep Leaning

Education

PhD Student in Electrical Engineering (GPA 5.8/6)
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland *2016-2020*
Thesis: Context-aware Neural Machine Translation
Supervisors: James Henderson, Hervé Bourlard

M.Sc. in Computer Science (GPA 5.6/6)
Universities of Bern, Neuchâtel and Fribourg, Switzerland *2012-2014*

B.Sc. in Computer Science (2nd best)
Pontificia Universidad Católica del Perú (PUCP) *2002-2008*

Research Experience

Google AI, Pragma Group
Software Engineering Intern *2019*
Work on user-initiated repairs for a task-oriented dialogue system.

Microsoft Research AI, Machine Translation Group
Research Intern *2019*
Proposed a model for table to text generation and translation.

Idiap Research Institute, NLU group
Research Assistant *2016-2020*
Extending contextual information for NMT.

University of Neuchâtel, IIUN
Graduate Researcher *2015-2016*
Author profiling, identification of sociolect aspects based on stylistic features.

Idiap Research Institute, NLP group
Research Intern *2014-2015*
Improving the translation of anaphoric pronouns in machine translation.

Professional Experience

Information Security Analyst
Scotiabank International, Latam Tech. Department, Peru *2009-2012*
Proposed and leaded the implementation of a dashboard to manage security indicators of all branches.

Business Intelligence Analyst
PUCP, Statistics and Institutional Intelligence Office, Peru *2008-2009*
Proposed and implemented an institutional balanced scorecard application.

Business Intelligence Intern
PUCP, Statistics and Institutional Intelligence Office, Peru *2007-2008*
Statistical data analysis of graduated students to predict prospective scholarship recipients.

Technical skills

Programming: Python, Java, C++ **OS:** Linux, Windows **Toolkits:** PyTorch, TensorFlow

Academic Service

Member of the Program Committee of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)
Conference reviewer at ACL, NAACL, EMNLP, and RANLP.

Teaching Experience

Graduate Teaching Assistant
University of Neuchâtel, Faculty of Science, Switzerland *2015-2016*
Lab instructor for "Introduction to Statistical Learning" and "Seminar of Natural Language Processing".

Undergraduate Teaching Assistant
PUCP, Faculty of Science and Engeneering, Peru *2005-2009*
Lab and practical exam instructor of different introductory courses in computer science.

Honors and Awards

2012-2014: Swiss Government Scholarship for foreign students for master studies.

2008: Finished 2nd best in bachelor studies in Computer Engineering PUCP.

2004-2007: Scholarship for academic performance by the Faculty of Science and Engineering PUCP.

2002-2004: Awards for academic excellence (finished 1st) in General Studies of Science PUCP.

Publications

Miculicich L., Marone M., Hassan H., *"Selecting, Planning, and Rewriting: A Modular Approach for Data-to-Document Generation and Translation"*. In EMNLP-IJCNLP, 2019.

Miculicich L., Henderson J., *"Partially-supervised Mention Detection"*. to be Submitted to ACL.

Ram D., Miculicich L., Bourlard H., *"Neural Network based End-to-End Query by Example Spoken Term Detection"*. Submitted to IEEE/ ACM TASLP, 2019.

Ram D., Miculicich L., Bourlard H., *"Multilingual Bottleneck Features for Query by Example Spoken Term Detection"*. In ASRU, 2019.

Miculicich L., Ram D., Pappas N., Henderson J., *"Document-Level Neural Machine Translation with Hierarchical Attention Networks"*. In EMNLP, 2018.

Pappas N., Miculicich L., Henderson J., *"Beyond Weight Tying: Learning Joint Input-Output Embeddings for Neural Machine Translation"*. In WMT, 2018.

Miculicich L., Pappas N., Ram D., Popescu-Belis A., *"Self-Attentive Residual Decoder for Neural Machine Translation"*. In NAACL, 2018.

Ram D., Miculicich L., Bourlard, H., *"CNN based Query by Example Spoken Term Detection"*. In Interspeech, 2018.

Miculicich L., Popescu-Belis A., *"Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT)"*. In DiscoMT at EMNLP, 2017.

Miculicich L., Popescu-Belis A., *"Using Coreference Links to Improve Spanish-to-English Machine Translation"*. In CORBON at EACL, 2017.

Miculicich L., *"Statistical Learning Methods for Profiling Analysis"*. In CLEF, 2015.

Luong N., Miculicich L., Popescu-Belis A., *"Pronoun Translation and Prediction with or without Coreference Links"*. In DiscoMT at EMNLP, 2015.

(see Google Scholar profile)

Open Source Software

`https://github.com/idiap/HAN_NMT`

`https://github.com/idiap/Attentive_Residual_Connections_NMT`

`https://github.com/idiap/APT`

`https://github.com/idiap/joint-embedding-nmt`

`https://github.com/idiap/CNN_QbE_STD`