Challenges and Remedies for Context-Aware Neural Machine Translation

Lorenzo Lupo

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

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

Défis et remèdes pour la traduction automatique neuronale en contexte

Challenges and Remedies for Context-Aware Neural Machine Translation

Présentée par :

Lorenzo LUPO

Direction de thèse :

Marco DINARELLI CHARGE DE RECHERCHE, Université Grenoble Alpes	Directeur de thèse
Laurent BESACIER INGENIEUR HDR, NAVER	Co-directeur de thèse

Rapporteurs :

François YVON
DIRECTEUR DE RECHERCHE, CNRS
Marcello FEDERICO
INGENIEUR DOCTEUR, Amazon Web Services AI Labs

Thèse soutenue publiquement le **28 mars 2023**, devant le jury composé de :

François YVON DIRECTEUR DE RECHERCHE, CNRS	Rapporteur
Marcello FEDERICO INGENIEUR DOCTEUR, Amazon Web Services AI Labs	Rapporteur
Rachel BAWDEN CHARGE DE RECHERCHE, INRIA de Paris	Examinatrice
Éric GAUSSIER PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Examinateur

Invités :

Marco DINARELLI
CHARGE DE RECHERCHE, CNRS
Laurent BESACIER
INGENIEUR HDR, Naver Labs Europe

Abstract

Current neural machine translation systems have reached close-to-human quality in translating stand-alone sentences. When it comes to translating documents, instead, machine translation has a significant margin of improvement ahead. In fact, some ambiguous elements of the discourse have multiple valid translations at the sentence level but only one at the document level, where they lose their ambiguity in the presence of extra-sentential context. Retrieving and exploiting such context to produce consistent document-level translations represents a challenging task. Many researchers have taken up this challenge in recent years and proposed approaches to context-aware neural machine translation. A common taxonomy divides them into two families: multi-encoding and single-encoding approaches, also known as concatenation approaches. The former family includes all the approaches that employ the standard encoder-decoder architecture to produce latent representations of the current sentence and that introduce additional learnable modules to encode and integrate its context, i.e., the previous or following sentences. Concatenation approaches, instead, rely entirely on the encoder-decoder architecture, but they concatenate the context to the current sentence before feeding it into the system.

In this work, we analyze both families of approaches to context-aware neural machine translation, identify some of their weaknesses, and address them with novel solutions. For multi-encoding systems, we identify two learning challenges faced by the modules that handle context: the sparsity of the training signal and the sparsity of disambiguating contextual elements. We introduce a novel pre-training setting in which sparsity is alleviated and demonstrate its effectiveness in fostering the learning process. For concatenation approaches, we address the challenge of dealing with long sequences by proposing a training objective that encourages the model to focus on the most relevant parts of each sequence. We couple this training objective with a novel technique to strengthen sentence boundaries and analyze their impact on the learned attention mechanism. Finally, we present a comparative study of various methods for discerning segments in the concatenation sequence, including novel variants of segment embeddings.

Résumé

Les systèmes actuels de traduction automatique neuronale ont atteint une qualité proche de celle d'un traducteur humain pour la traduction de phrases isolées. En revanche, lorsqu'il s'agit de traduire des documents, la traduction automatique dispose d'une marge d'amélioration importante. En fait, certains éléments ambigus du discours ont plusieurs traductions valides au niveau de la phrase mais une seule au niveau du document, car ils perdent leur ambiguïté en présence du contexte extra-sententiel. L'identification et l'exploitation du contexte utile pour produire des traductions cohérentes au niveau du document représentent une tâche difficile. De nombreux chercheurs ont relevé ce défi ces dernières années et ont proposé des approches de traduction automatique neuronale sensible au contexte. On peut les classer en deux familles : les approches à encodage multiple et les approches à encodage unique, également appelées approches de concaténation. La première famille comprend toutes les approches qui utilisent l'architecture standard d'encodeur-décodeur pour produire des représentations latentes de la phrase courante et qui introduisent des modules supplémentaires pour encoder et intégrer son contexte, c'est-à-dire les phrases précédentes ou suivantes. Les approches par concaténation, au contraire, reposent entièrement sur l'architecture standard d'encodeur-décodeur, mais elles concatènent le contexte à la phrase actuelle avant de l'introduire dans le système.

Dans ce travail, nous analysons les deux familles d'approches de traduction automatique neuronale sensible au contexte, nous identifions certaines de leurs faiblesses et nous y remédions par des solutions originales. Pour les systèmes à encodage multiple, nous identifions deux défis d'apprentissage auxquels sont confrontés les modules qui gèrent le contexte : la rareté du signal d'apprentissage et la rareté des éléments contextuels de désambiguïsation. Nous introduisons un nouveau cadre de pré-entraînement dans lequel la rareté est atténuée et nous démontrons son efficacité expérimentalement. Pour les approches de concaténation, nous relevons le défi de traiter de longues séquences en proposant un objectif d'entraînement qui encourage le modèle à se concentrer sur les parties les plus pertinentes de chaque séquence. Nous couplons cet objectif d'entraînement avec une nouvelle technique pour renforcer la séparation des phrases dans séquence traitée. Nous analysons l'impact de ces solutions sur le mécanisme d'attention appris. Enfin, nous présentons une étude comparative de diverses méthodes pour discerner les segments dans la séquence de concaténation, y compris des nouvelles variantes de plongement de segments.

Contents

Acronyms

MT	Machine Translation.
RMT	Rule-based Machine Translation.
SMT	Statistical Machine Translation.
NMT	Neural Machine Translation.
CANMT	Context-Aware Neural Machine Translation.
NLP	Natural Language Processing.
RNN	Recurrent Neural Networks.
WMT	Conference on Machine Translation.
IWSLT	International Conference on Spoken Language Translation.
BPE	Byte Pair Encoding.
d&r	*divide and rule.*
PSE	Position-Segment Embedding.

Chapter 1

Introduction

Machine Translation (MT) is a field of computational linguistics that studies the automatic translation of a source text into a target language. In the past decades, globalization of information and economies has fueled the need for translations at a large-scale, fostering advancements in MT. Three paradigms have guided the investigation of novel MT systems along the way: Rule-based Machine Translation (RMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT). While RMT is based on linguistic rules formulated by experts, SMT and NMT consist of machine learning methods that learn the translation task directly from a large amount of bitext data. In particular, NMT systems are based on neural networks that learn in a supervised fashion. Each training example consists of a source sentence and a reference translation, usually performed by a human translator. The network tries to translate the source, and its output is compared with the reference translation. A measure of the distance between the output and the reference quantifies the so-called *training loss*, the "error" of the system. Then, the learnable parameters of the network are adjusted based on the loss, employing standard optimization techniques like stochastic gradient descent so as to minimize the difference between the output and the reference translation in future iterations. NMT has seen substantial improvements in recent years, primarily fostered by the advent of the attention mechanism (Bahdanau et al., 2015a) and the Transformer model (Vaswani et al., 2017). While current NMT systems have reached close-to-human quality in the translation of de-contextualized sentences (Wu et al., 2016), they still have a wide margin of improvement ahead when it comes to translating documents (Läubli et al., 2018) such as articles, books, chats, transcripts of live events, or video subtitles.

1.1 Problem statement

Translating a document requires considering the linguistic relationships between its sentences. In other words, to contextualize them. Given a sentence in a document, we can refer to the rest of the document as the extra-sentential context of the sentence or, shortly, its context. The need to contextualize the translation of a given source sentence $\boldsymbol{x}^i$ can emerge even in concise documents. For example, if we take the English sentence:

$\boldsymbol{x}^i$ (EN): How are you today ?

There exist multiple French translations that are valid if the context is not provided, such as $\boldsymbol{y}^{i,1}$ and $\boldsymbol{y}^{i,2}$:

$\boldsymbol{x}^i$ (EN): How are you today ?
$\boldsymbol{y}^{i,1}$ (FR): *Comment vas-tu aujourd'hui?*
$\boldsymbol{y}^{i,2}$ (FR): *Comment allez-vous aujourd'hui?*

However, when we put $\boldsymbol{x}^i$ into context by providing the previous sentence $\boldsymbol{x}^{i-1}$, it becomes clear that $\boldsymbol{y}^{i,2}$ is the only valid option:

$\boldsymbol{x}^{i-1}$ (EN): Good morning Mr. President.
$\boldsymbol{x}^i$ (EN): How are you today ?
$\boldsymbol{y}^{i,1}$ (FR): *Comment vas-tu aujourd'hui?*
$\boldsymbol{y}^{i,2}$ (FR): *Comment allez-vous aujourd'hui?*

The title of the interlocutor ("Mr. President") requires the use of the formal "you" (T-V distinction) and, accordingly, the correct conjugation of the related verb.

The NMT paradigm has evolved around the task of context-agnostic translation. As such, until recently, many state-of-the-art NMT systems could not contextualize the current sentence beyond sentence boundaries (Bojar et al., 2017a,b). Unfortunately, context-agnostic NMT systems systematically produce inter-sentential inconsistencies, hindering the overall quality of the translation (Läubli et al., 2018; Toral et al., 2018; Voita et al., 2019b). As an example, until the end of December 2022, Google Translate, one of the most popular automatic translation systems globally, was still translating $\boldsymbol{x}^i$ with $\boldsymbol{y}^{i,1}$, even when $\boldsymbol{x}^{i-1}$ was provided as context. To address this limitation, researchers have been studying Context-Aware Neural Machine Translation (CANMT) since early 2017 (Jean and Cho, 2019; Tiedemann and Scherrer, 2017; Wang et al., 2017). Today this line of research is still vibrant, and no consensus has yet been reached on the best context-aware approaches and methods of evaluation.

1.2 Contributions

Our research aims to provide insights into some of the challenges faced by current approaches to CANMT and to propose new solutions.

Our contribution is threefold:

- In Chapter 3, we provide insights on the training challenges faced by multi-encoding translation models, a broad family of CANMT architectures, presented in Section 2.4.2. To overcome these challenges, we propose and analyze a pre-training method that enables consistent improvements in context-aware translation. The implementation of the experiments discussed in this chapter is publicly available at https://github.com/lorelupo/divide-and-rule.
- In Chapter 4, we tackle the learning complexity that is faced by concatenation approaches, another dominant family of CANMT approaches, discussed in Section 2.4.4. As a solution, we propose a training paradigm that focuses on the translation of the current sentence while considering the translation of context as a secondary task. We also propose introducing explicit information to differentiate the various sentences concatenated in the input sequence, with the goal of helping the model to optimize the new training objective. Both solutions are evaluated and analyzed on different domains and language pairs, showing consistent improvements in discourse phenomena disambiguation. The implementation of the experiments discussed in this chapter is publicly available at https://github.com/lorelupo/focused-concat.
- In Chapter 5, we deepen our study of the methods for differentiating the various sentences concatenated in concatenation approaches. We compare the method proposed in the previous chapter with three variants of segment embeddings and propose new ways to integrate them into the Transformer architecture. Our solutions are evaluated and analyzed on different domains and language pairs, showing that, despite being a very intuitive solution, they do not benefit the baselines except in a specific setting. The implementation of the experiments discussed in this chapter is publicly available at https://github.com/lorelupo/focused-concat.

In the Conclusions (§6), we present some research directions that emerge naturally from our work and that we deem valuable for the advancement of CANMT.

1.2.1 Publications

Most of the contributions presented in Chapter 3 are published in:

> Lupo et al. (2022a) - Lupo, L., Dinarelli, M. and Besacier, L. (2022). Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557-4672, Dublin, Ireland.

Most of the contributions presented in Chapter 4 are published in:

> Lupo et al. (2022b) - Lupo, L., Dinarelli, M. and Besacier, L. (2022). Focused Concatenation for Context-Aware Neural Machine Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, December 7–8, 2022. Association for Computational Linguistics.

The contributions presented in Chapter 5 will be submitted to the *Fourth Workshop on Insights from Negative Results in NLP*, co-located with EACL 2023.

Chapter 2

Background

In this chapter, we introduce the reader to the tasks of NMT, including details on the data used for training translation systems, the evaluation methods, and the state-of-the-art neural architectures for this task. Subsequently, we provide an overview of CANMT, starting with its motivation and formal definition, then moving to the datasets, the evaluation methods, and the CANMT approaches proposed in the literature. We conclude with a discussion on the strengths and weaknesses of such approaches.

2.1 Neural machine translation

2.1.1 Definition

A NMT system is a neural network with parameters θ that is trained end-to-end to model the conditional probability $P_\theta(\boldsymbol{y}|\boldsymbol{x})$ of a translation $\boldsymbol{y} = \{y_1, y_2, ..., y_{|\boldsymbol{y}|}\}$, given a sentence in the source language $\boldsymbol{x} = \{x_1, x_2, ..., x_{|\boldsymbol{x}|}\}$. This probability can be expressed as follows:

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

where $\boldsymbol{y}_{<t} = \{y_1, y_2, ..., y_{t-1}\}$ are the previously generated words. Given a parallel training corpus $\mathcal{C} = \{(\boldsymbol{x}^1, \boldsymbol{y}^1), (\boldsymbol{x}^2, \boldsymbol{y}^2), ..., (\boldsymbol{x}^{|\mathcal{C}|}, \boldsymbol{y}^{|\mathcal{C}|})\}$, where $\boldsymbol{x}^j$ is the jth source sentence of the corpus and $\boldsymbol{y}^j$ its reference translation, the standard training objective is to find parameter values $\hat{\theta}$ that maximize the log-likelihood of the training data:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\mathcal{C}; \theta). \tag{2.2}$$

By assuming that every sentence pair in the corpus is independent from the others, the log-likelihood can be written as follows:

$$\mathcal{L}(\mathcal{C};\theta) = \frac{1}{|\mathcal{C}|}\sum_{j=1}^{|\mathcal{C}|}\log P_\theta(\boldsymbol{y}^j|\boldsymbol{x}^j) \tag{2.3}$$

$$= \frac{1}{|\mathcal{C}|}\sum_{j=1}^{|\mathcal{C}|}\sum_{t=1}^{|\boldsymbol{y}^j|}\log P_\theta(y_t^j|\boldsymbol{y}_{<t}^j,\boldsymbol{x}^j). \tag{2.4}$$

At inference time, it is standard practice to use *beam search* (Graves, 2012; Luong et al., 2015) to find a translation hypothesis $\tilde{\mathbf{y}}$ that approximately maximizes Equation 2.1 for a given source $\boldsymbol{x}$:

$$\hat{\boldsymbol{y}} = \underset{\tilde{\mathbf{y}}\in beam(\boldsymbol{x},\theta)}{\operatorname{argmax}} P_\theta(\tilde{\mathbf{y}}|\boldsymbol{x}), \tag{2.5}$$

2.1.2 Data

Many training corpora $\mathcal{C}$ with aligned sentence pairs are freely available online for over 200 languages and dialects (Tiedemann, 2016). On top of this, the MT community has produced standardized datasets to compare different systems under similar experimental conditions. In this section, we describe the datasets adopted in our experiments.

WMT - Every year, the Conference on Machine Translation (WMT)[1] releases standardized datasets for many language pairs and specific domains, comprising multiple corpora that have been cleaned and pre-processed. The main goal is to incentivize the benchmarking of new MT models on specific translation domains. In our case, we employ the datasets that have been released for the news translation task. The testing and development sets are created from a sample of online newspapers, while data for training include different sources:

(i) Common Crawl,[2] consisting of sentences crawled from web pages of various domains.

(ii) Europarl (Koehn, 2005), extracted from the proceedings of the European Parliament. The sentences are organized in documents, each corresponding to a topic discussed in a parliamentary sitting.

(iii) News Commentary (Tiedemann, 2012), a corpus with political and economic commentaries crawled from the website Project Syndicate[3]. Each commentary represents a document.

[1] E.g., WMT22: https://www.statmt.org/wmt22/
[2] https://commoncrawl.org
[3] https://www.project-syndicate.org/

(iv) UN Parallel Corpus (Ziemski et al., 2016), composed of official records and other parliamentary documents of the United Nations.

IWSLT - The International Conference on Spoken Language Translation (IWSLT)[4] organizes evaluation campaigns focused on spoken language translation. Among others, they provide corpora consisting of parallel subtitles of TED and TEDx Talks (Cettolo et al., 2012), including document boundaries. The organizers provide standardized training, development, and testing splits.

OpenSubtitles2018 - OpenSubtitles2018 (Lison et al., 2018) is a corpus of movie and TV subtitles in a variety of languages, for which training, development, and testing splits are not available. Every subtitle file can be considered as a document, although other document boundaries can be identified. For instance, by considering each movie scene as a standalone document (c.f. Voita et al. (2019b)).

Details on data splits, language pairs, data versions, and statistics will be described in the following chapter's sections devoted to experiments.

2.1.3 Evaluation

The translation objective of a NMT system at inference time can be rewritten by means of the Bayes' rule as follows:

$$\underset{\tilde{\mathbf{y}}}{\operatorname{argmax}}\, P(\tilde{\mathbf{y}}|\boldsymbol{x}) = \underset{\tilde{\mathbf{y}}}{\operatorname{argmax}}\, \frac{P(\tilde{\mathbf{y}})P(\boldsymbol{x}|\tilde{\mathbf{y}})}{P(\boldsymbol{x})} \tag{2.6}$$

We can distinguish two concurrent components in this objective function:

(i) **Fluency**: $\operatorname{argmax}_{\boldsymbol{y}} P(\tilde{\mathbf{y}})$ selects for translation hypotheses that are syntactically appropriated and meaningful in the target language;

(ii) **Adequacy**: $\operatorname{argmax}_{\tilde{\mathbf{y}}} P(\boldsymbol{x}|\tilde{\mathbf{y}})$ selects for translation hypotheses that preserve the meaning of the source.

The most accurate translation evaluation along these two axes is that performed by bi-lingual readers, possibly professional translators. Humans can embrace the possibly vast diversity of alternative translations that can stem from the same source, accounting for meaning and style nuances. However, human evaluation is resource-intensive and is only sometimes worth it at the research and development stage. Here, automatic evaluation

[4] https://iwslt.org/

comes to the rescue. The most widespread automatic metrics are reference-based: they compute the similarity between the system output and one or more reference translations produced by humans. Some popular examples are BLEU (Papineni et al., 2002), which is the standard metric for MT, METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). More recently, another family of metrics relying on contextualized embeddings (Devlin et al., 2019) trained on large non-parallel corpora has been shown to be better correlated with human judgments. Among the most popular metrics belonging to this family are BERTscore (Zhang et al., 2020b), BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020).

2.1.3.1 BLEU

BLEU is based on the degree of overlap between the n-grams of a candidate translation $\tilde{\mathbf{y}}$ and those of the human reference $\boldsymbol{y}$. To compute an n-gram precision, we divide the number of correct n-grams by the total number of n-grams in the hypothesis. Evaluating a precision score per sequence leads to noisy scores. Therefore n-gram precision is macro-averaged by dividing the sum of correct n-grams by their total number in the overall corpus:

$$p_n = \frac{\sum_{j=1}^{|\mathcal{C}|} \sum_{g \in \texttt{n-grams}(\tilde{\mathbf{y}}^j)} \#(g, \boldsymbol{y}^j)}{\sum_{h=1}^{|\mathcal{C}|} \sum_{g' \in \texttt{n-grams}(\tilde{\mathbf{y}}^h)} \#(g', \tilde{\mathbf{y}}^h)}, \tag{2.7}$$

where $\texttt{n-grams}(\tilde{\mathbf{y}}^j)$ is the set of unique n-grams that are present in the jth translation hypothesis $\tilde{\mathbf{y}}^j$, and $\#(g, \boldsymbol{y})$ is the number of times an n-gram g appears in a sentence $\boldsymbol{y}$. BLEU is then calculated as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\frac{1}{n_{max}} \sum_{n=1}^{n_{max}} \log p_n\right), \tag{2.8}$$

where BP is a brevity penalty against short translations, which would otherwise be unduly rewarded by n-gram precisions. Typically, the maximum n-gram size is $n_{max} = 4$.[5] Given the word count r of the reference corpus and c of the machine-translated text, the brevity penalty is calculated as follows:

$$\text{BP} = \begin{cases} 1 & \text{if } c \geqslant r \\ e^{(1-r)/c} & \text{if } c < r \end{cases} \tag{2.9}$$

Hence, the final BLEU score lies between 0 and 1. The upper boundary is reached in case of a perfect match between hypothesis and reference. Despite being universally cited

[5] We use this value across all of our experiments.

across the literature, BLEU is often inadequate for MT evaluation. For instance, BLEU is unreliable by design for scoring a single sentence, or just a few of them, instead of a large corpus where the n-gram matches are averaged over large numbers. Averaging over many n-grams makes BLEU unfit also in the evaluation of discourse phenomena consisting in few words spanning over multiple sentences, as we will discuss in Section 2.3. Moreover, BLEU is usually calculated against a single reference, although it can support multiple references. Using a single reference leads to significant fluctuations in the score depending on the wording of the translation hypothesis. For example, if the hypothesis and the reference use different phrasings to express the same message, the BLEU score would be low despite the translation being equivalent.

Despite these problems, BLEU is an efficient metric to capture substantial differences between translation systems. Throughout this thesis, we shall use BLEU to illustrate the average translation quality of MT approaches under similar conditions, but mainly as a starting point for a more fine-grained evaluation. In Chapters 4 and 5, we couple BLEU with a more costly but better-performing metric for average translation quality called COMET (Rei et al., 2020).

2.1.3.2 COMET

COMET (Rei et al., 2020) is a learnable metric for machine translation evaluation based on large pre-trained multilingual language models like mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). The language model encodes the source, hypothesis, and reference sentences separately. The resulting word embeddings are passed through a pooling layer to create a sentence embedding for each input sentence. Given the concatenation of these sentence embeddings, a feed-forward network is trained to estimate the quality score that a human has assigned to the translation hypothesis. Being a learnable metric, COMET can be constantly improved and fine-tuned on specific translation quality metrics. In the recent shared tasks on metrics for machine translation evaluation, COMET has been ranked high for correlation with human judgment (Freitag et al., 2021). This success is probably attributable to its ability to model subtle translation quality properties that lexical overlapping methods like BLEU can not capture, as well as its ability to leverage the source text to inform its predictions.

2.1.4 Architectures

A NMT network is usually structured as an encoder-decoder architecture (Bahdanau et al., 2015b; Vaswani et al., 2017). The encoder reads the source sentence $\boldsymbol{x}$ and maps

it into a continuous representation $S = [\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_{|\boldsymbol{x}|}] \in \mathbb{R}^{d \times |\boldsymbol{x}|}$, where the ith token is represented by a vector s_i of size d. The decoder then exploits this latent representation to generate the corresponding translation one word at a time. Until recently, encoders and decoders were generally composed of stacked Recurrent Neural Networks (RNN), using either Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Units (Cho et al., 2014). The introduction of the *attention* mechanism (Bahdanau et al., 2015b) brought major improvements by enabling the joint learning of source-target word alignments and translation, and later by learning better word representations by means of *self-attention* (Lin et al., 2017). In 2017, Vaswani et al. (2017) proposed an encoder-decoder architecture entirely based on self-attention, which is state-of-the-art in NMT (Bojar et al., 2017b), as well as in many other Natural Language Processing (NLP) tasks (Devlin et al., 2019; Brown et al., 2020; Yang et al., 2019). Meanwhile, other research works proposed NMT architectures based on convolutional neural networks (Kalchbrenner et al., 2016; Gehring et al., 2017; Wu et al., 2019), but they were not as successful.

2.1.4.1 Attention

Attention is a function that takes in input a query vector $\boldsymbol{q} \in \mathbb{R}^d$, and two matrices: the keys and values $K, V \in \mathbb{R}^{d \times n}$. Then it computes a weighted sum of the value vectors, where a similarity function of the query with a key vector computes the weight assigned to the corresponding value:

$$Attention(\boldsymbol{q}, K, V) = V \cdot Similarity(\boldsymbol{q}, K) \in \mathbb{R}^d. \tag{2.10}$$

A common similarity function is the scaled dot-product (Vaswani et al., 2017):

$$Similarity(\boldsymbol{q}, K) = softmax\left(\frac{K^T \boldsymbol{q}}{\sqrt{d}}\right) \in \mathbb{R}^n, \tag{2.11}$$

where the softmax operator (Bridle, 1990), applied column-wise, normalizes the column vectors into probability distributions. Attention is commonly used to contextualize the query vector (the representation of a token belonging to the target sentence) with the source sentence, represented by the key-value pairs.

Self-attention - Attention can also be used to compute a representation of the tokens that encodes information from the whole sequence. For doing so, we can employ each token as a query and all the tokens belonging to the sequence (included itself) as key-value pairs. In practice, given the continuous representation of a sentence of $|\boldsymbol{x}|$ tokens, $S \in \mathbb{R}^{d \times |\boldsymbol{x}|}$, we can transform each token representation into query-key-value triplets by means of three

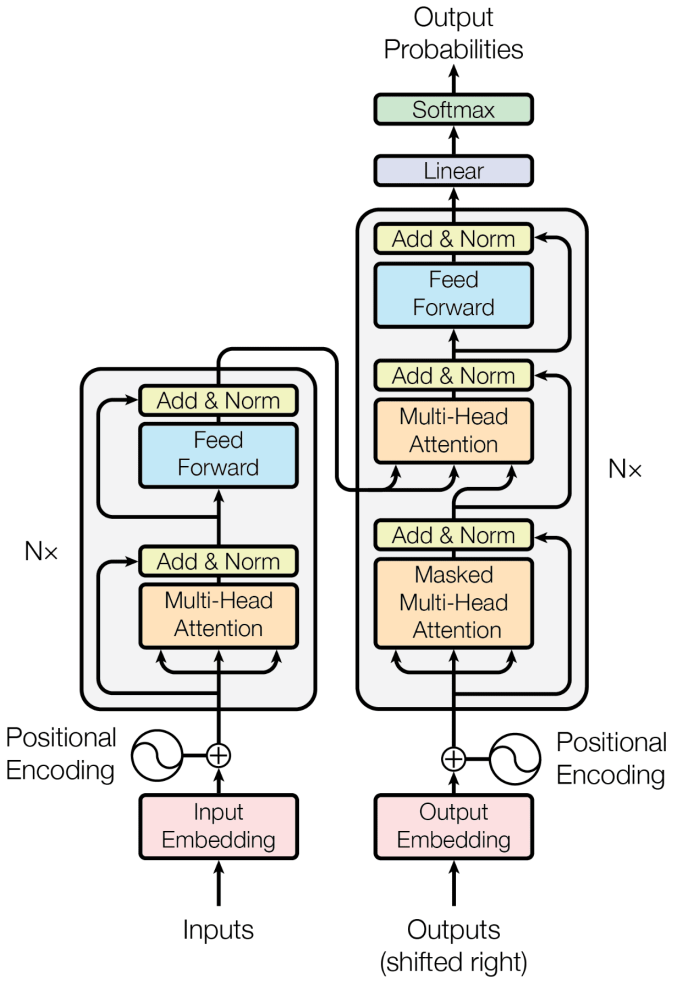

Figure 2.1 – The Transformer - model architecture (Vaswani et al., 2017)

learnable matrices $W_Q, W_K, W_V \in \mathbb{R}^{d \times |\boldsymbol{x}|}$ and perform scaled dot-product self-attention as follows:

$$Attention(W_Q S, W_K S, W_V S) = W_V S \cdot softmax\left(\frac{(W_K S)^T (W_Q S)}{\sqrt{d}}\right) \in \mathbb{R}^{d \times |\boldsymbol{x}|} \quad (2.12)$$

Self-attention allows the learning of token representations that are contextualised with the other tokens within the sentence, which is essential to capture their meaning.

2.1.4.2 The Transformer

The Transformer encoder-decoder architecture (Vaswani et al., 2017) is a stack of $2 \times N$ building blocks, depicted in Figure 2.1. On top of the decoder, there is a learned linear transformation and a softmax function to convert the decoder output to predicted token

probabilities. Each component of the encoder and decoder blocks is described below.

Token and position embeddings - Both the encoder and the decoder take in input a sum of learned token embeddings and position embeddings. The token embedding layer consists in a lookup table that stores a learnable vectorial representation for each token in the vocabulary. The input to this layer is a list of indices identifying the tokens of the input sequence, and the output is their word embeddings. The embedding layer can be shared between the encoder and the decoder. Position embeddings enable the model to capture the input sentence's sequentiality in the absence of recurrence. They are non-learnable parameter vectors of size d_{model}, one for each token position $t = \{1, 2, ..., |S|\}$, where S is the continuous representation of either the source $\boldsymbol{x}$ or translation hypothesis $\tilde{\mathbf{y}}$. The resulting position embedding matrix $P \in \mathbb{R}^{d_{model} \times |S|}$ is defined as:

$$PE_{(d,2t)} = sin(t/10000^{(2d/d_{model})}) \tag{2.13}$$

$$PE_{(d,2t+1)} = cos(t/10000^{(2d/d_{model})}) \tag{2.14}$$

Thus, each encoding dimension $d = \{1, 2, ..., d_{model}\}$ corresponds to two intertwined sinusoidal waves: a sine and a cosine. Some open-source implementations adopt a similar definition,[6] which, instead of intertwining the sinusoidal waves, assigns the first half of encoding dimensions $d = \{1, 2, ..., \frac{d_{model}}{2}\}$ a sine wave, and the second half a cosine wave:

$$PE_{(d,t)} = \begin{cases} sin(t/10000^{(2d/d_{model})}) & \text{if } 1 < d \leqslant \frac{d_{model}}{2} \\ cos(t/10000^{(2d/d_{model})}) & \text{if } \frac{d_{model}}{2} < d \leqslant d_{model} \end{cases} \tag{2.15}$$

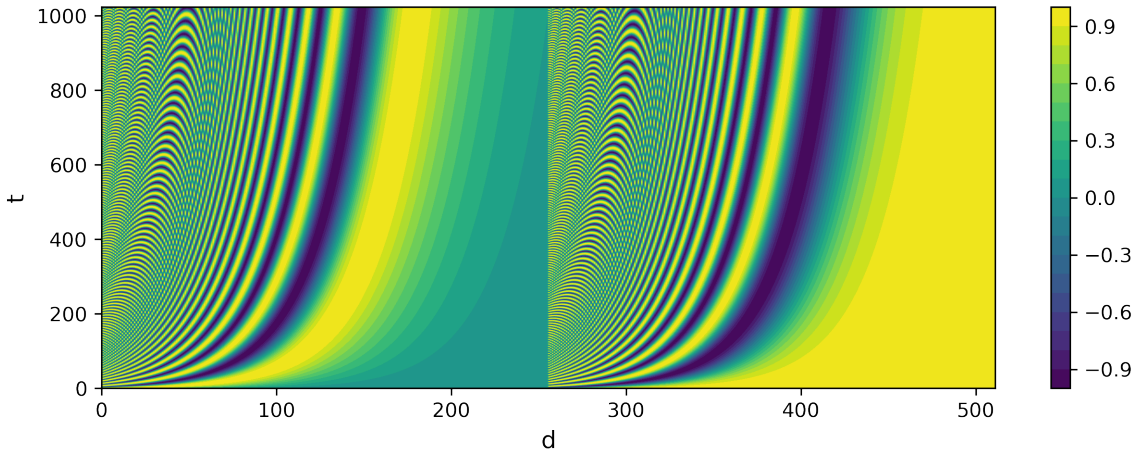

Figure 2.2 – Sinusoidal position embedding of 1024 positions with 512 encoding dimensions.

[6] The *fairseq* implementation (Ott et al., 2019) we employ in our experiments follows this definition.

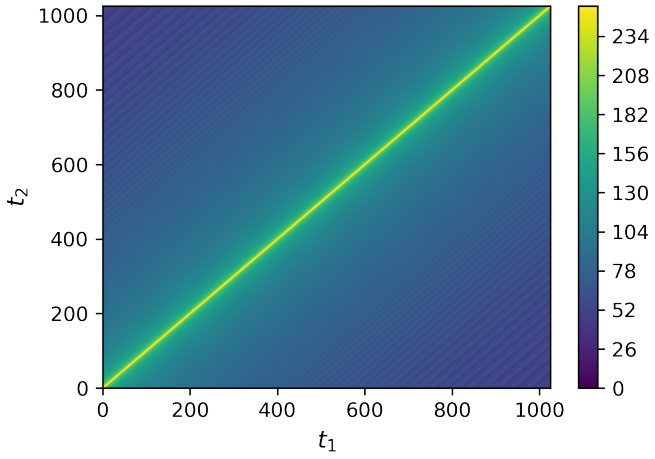

Figure 2.3 – The dot-product between sinusoidal position embedding is symmetrical and decays quickly with the distance between two positions $(t_1 - t_2)$.

Figure 2.2 represents the encoding of $|S| = 1024$ positions over $d_{model} = 512$ encoding dimensions, following this definition. Sinusoidal position embeddings have two interesting properties:

- $PE_{(d,t+k)}$ can be represented as a linear function of $PE_{(d,t)}$ for any fixed offset k. Therefore, the Transformer can easily learn to model relative distances between tokens (Vaswani et al., 2017; Denk, 2019).
- The dot-product between position embeddings is symmetrical and decays quickly with relative distance, as shown in Figure 2.3. Since the dot-product is the default similarity function used in the Transformer's attention mechanism (2.11), this property can be easily reflected in the attention output, resulting in a discounting of the weights attributed to distant values.

Multi-head attention - Instead of performing a single (self-)attention function with d_{model}-dimensional keys, values, and queries, multi-heading consists in projecting the query-key-value triplets h times, with different linear projections into $\mathbb{R}^{d_o}$. Subsequently, scaled dot-product attention is applied to each triplet. The output of each (self-)attention is called *head*:

$$head_i = Attention(W_Q^i S, W_K^i S', W_V^i S') \in \mathbb{R}^{d_o \times |S|} \tag{2.16}$$

where $W_Q^i \in \mathbb{R}^{d_o \times |S|}$ and $W_K^i, W_V^i \in \mathbb{R}^{d_o \times |S'|}$ are the projection matrices. The resulting vectors are then concatenated and linearly projected from $\mathbb{R}^{h \cdot d_o \times |S|}$ to $\mathbb{R}^{d_{model} \times |S|}$ with another learnable matrix W^o:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_h) W^o \in \mathbb{R}^{d_{model} \times |S|} \tag{2.17}$$

This process allows to jointly employ different attention patterns over the processed sequences, hence extracting different linguistic features. The Transformer uses multi-head attention in three different ways:

(i) Encoder blocks employ self-attention, applying the linear projections W_i^Q, W_i^K, and W_i^V to the same matrix $S = S'$, which is formed by the token representation output from the previous block.

(ii) Decoder blocks employ masked self-attention, a variant of self-attention in which all the elements of the target sequence that have not been decoded yet are masked out. The masking is implemented by setting to $-\infty$ the values in input of the softmax that need to be masked. The rationale for such masking is to preserve the auto-regressive property of the decoder.

(iii) Decoder blocks also employ encoder-decoder cross-attention, after the masked self-attention layer. Here, the queries come from the previous decoder layer, while the keys and values come from the output of the encoder. Thus, every token representation in the decoder can attend to the tokens representations in the encoder. In this case, $S \neq S'$.

Feed-forward - After the multi-head attention layers, each block in the Transformer applies two affine transformations with a ReLU activation (Nair and Hinton, 2010) in between:

$$FF(S) = W_2(\text{ReLU}(W_1 S + b_1)) + b_2 \in \mathbb{R}^{d_{model} \times |S|}. \tag{2.18}$$

The learnable matrix $W_1 \in \mathbb{R}^{d_{FF} \times d_{model}}$ maps to a larger space, usually $d_{FF} = 4 \times d_{model}$, and then $W_2 \in \mathbb{R}^{d_{model} \times d_{FF}}$ projects back to $\mathbb{R}^{d_{model}}$.

Add and normalize - Around each multi-head attention and feed-forward layer, there is a skip-connection (He et al., 2016), followed by a layer normalization (Ba et al., 2016):

$$AddNorm(S) = LayerNorm(S + Layer(S)) \tag{2.19}$$

2.1.4.3 Properties

The success of the Transformer is mainly ascribable to the use of self-attention for sequential learning and modeling. In fact, a self-attention layer compares favorably to its main alternatives, recurrent and convolutional layers, with respect to the three properties described below.

Layer Type	**Max Path Length**	**Sequential Ops**	**Complexity**
Self-attention	$O(1)$	$O(1)$	$O(\|S\|^2 \cdot d_{model})$
Recurrent	$O(\|S\|)$	$O(\|S\|)$	$O(\|S\| \cdot d_{model}^2)$
Convolution	$O(log_k(\|S\|))$	$O(1)$	$O(k \cdot \|S\| \cdot d_{model}^2)$

Table 2.1 – Maximum path length, sequential operations, and computational complexity of the three main neural architectures for sequence modeling. $|S|$ is the length of the sequence to be modeled, d_{model} is the representation dimensionality, and k the kernel size of convolutions.

Learnability of long-range dependencies - In many sequence modeling tasks, it is crucial to model temporal contingencies that span long intervals in the sequences. The ability to learn such dependencies depends on the length of the paths that forward and backward signals have to traverse in the network. The shorter these paths, the easier it is to learn long-range dependencies (Hochreiter et al., 2003). While self-attention relates all positions to one another in a single forward pass, a recurrent layer needs a number of forward passes equal to the sequence length $|S|$ to relate the sequence's last position to the first one. Moreover, information flows in one single direction for a recurrent layer. Therefore, relating the first and last positions to each other requires a bi-directional layer, which increases complexity. In the case of convolutions, we need a stack of $O(|S|/k)$ layers to relate all the positions to one another, where k is the kernel size. With a careful choice of dilations, the longest path can be shortened to $O(log_k(|S|))$ (Kalchbrenner et al., 2016).

Parallelizability - The fewer sequential operations are necessary for the forward pass, the faster it will be. Faster computation means faster training and inference. Self-attention and convolution operations can be fully parallelized to process the whole sentence, while each recurrent layer needs $O(|S|)$ sequential operations.

Computational complexity - Whenever the sequence length T is smaller than the representation dimensionality d_{model}, self-attention layers are less complex than recurrent and convolutional layers in terms of computation. Therefore, the Transformer is usually the winning architecture to process textual sentences, but it becomes unpractical to process long sequences such as full textual documents.

Table 2.1 represents these three properties with three variables: the maximum path length as a proxy of the difficulty to learn long-range dependencies; the number of sequential operations required to encode the sequence as a proxy of non-parallelizability; per-layer computational complexity.

Lexical ambiguity	
EN-1:	She is looking for a match.
EN-1.1:	She is looking for a partner.
EN-1.2:	She is looking for a wooden stick to set the fire.
Structural ambiguity	
EN-2:	Put the bottle on the table in the kitchen.
EN-2.1:	Put the bottle that is on the table in the kitchen.
EN-2.2:	Put the bottle on the table that is in the kitchen.
Scope ambiguity	
EN-3:	Every man loves a woman.
EN-3.1:	For every man, there is a loved woman.
EN-3.2:	There is one particular woman who is loved by every man.

Figure 2.4 – Linguistic ambiguity in its three main forms. Each ambiguous English sentence is followed by two legit disambiguations in the same language.

2.2 Context-aware neural machine translation

2.2.1 Motivation: the ambiguity in translation

Natural language can be ambiguous and open to multiple interpretations. As formulated by Chierchia and McConnell-Ginet (2000), "ambiguity arises when a single word or string of words is associated in the language system with more than one meaning". Ambiguity can be lexical, related to sentence structure, or to the multiplicity of possible scopes, as illustrated in Figure 2.4:

When it comes to translating a source language, the problem of ambiguity gets more complex because three sources of ambiguity interweave: the source language, the target language, and the cross-lingual interface (Bawden, 2018).

Ambiguity from the source language. As we have just seen, a sentence can be ambiguous for many reasons. In some cases, such as in poetic or comic language, the ambiguity is intentional, and it is desirable to preserve it during translation. However, preserving the same ambiguity in the target language is difficult and sometimes impossible because there is no analogous wording. For instance, the first example in Figure 2.4 cannot be translated into French with the same level of ambiguity. It must therefore be disambiguated in order to ensure the fluidity and adequacy of the translation:

Ambiguity from the target language. Specularly, an unambiguous expression in the

EN-1: She is looking for a match.
FR-1.1: Elle cherche un partenaire.
FR-2.1: Elle cherche une allumette.

source language may become ambiguous in the target language. Again, ambiguities can be lexical, structural, or of scope. For example, when a source word is translated with a polysemous word in the target language, a lexical ambiguity appears. E.g. :

EN-1: She fought for the best price.
FR-1: Elle s'est battue pour le meilleur prix.
Gloss: She fought for the best price/prize.

Ambiguity from the cross-lingual transfer. This type of ambiguity is specific to translation and concerns the mismatches in the conceptual spaces of the source and target languages. The ambiguity becomes problematic when transferring meaning from one language to another because it can hinder the translation's adequacy and fluency. For instance, the conceptual spaces of formality and familiarity are often problematic because they vary from one language to another. For example, English does not support the T-V (*Tu-Vos*) distinction, while French does. Therefore, the translation of the *you* pronoun into the French *tu* (familiar) or *vous* (formal) requires more contextual elements for the disambiguation. Inversely, the formality conveyed by the appropriate use of a French pronoun can not be entirely translated into English, where formality is conveyed by other morphological or discourse features, without there being a bijective mapping between formal French pronouns and formal English expressions.

In translation, it is generally desirable to correctly solve the ambiguities arising from the source sentence and the cross-lingual transfer of meaning. This is essential to guarantee the translation's fluency and adequacy. Sometimes, the ambiguities arising from the target language also need to be disambiguated, if the context available in the target text is insufficient for disambiguation by the reader. Usually, the longer the available text, the easier the task of disambiguation, thanks to the enlarged linguistic context, i.e., any linguistic information present in the text. Picking the right translation for a single word is hard without intra-sentential context, and this is why NMT systems produce word-by-word translations conditioned on the whole source sentence. However, these systems are trained in a context-agnostic fashion, relying on a strong independence assumption: each sentence of a document is independent from the others. In other words, every sentence in a document contains the information needed to solve the ambiguities that arise when translating it. Although this assumption holds for most sentences, it is wrong for others. For instance, in the examples above, the intra-sentential context is insufficient to select an

adequate translation.

Recently, Läubli et al. (2018) and Toral et al. (2018) showed that the ability to exploit extra-sentential context is a crucial challenge for NMT to reach human parity. Voita et al. (2019b) studied the output of a Transformer translating 2000 pairs of consecutive sentences from English to Russian. They found that at least 7% of the sentence pairs were not correctly translated because the system was agnostic to extra-sentential context.
In conclusion, we need a definition of NMT that is more aligned with how humans translate, i.e., by exploiting all the linguistic information available to guarantee the most adequate and fluent document translation.

2.2.2 Definition

A document-level NMT system models the conditional probability of the target document $Y = \{\boldsymbol{y^1}, \boldsymbol{y^2}, ..., \boldsymbol{y}^{|Y|}\}$ given the source document $X = \{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^{|X|}\}$:

$$P_\theta(Y|X) = \prod_{j=1}^{|X|} \prod_{t=1}^{|\boldsymbol{y}|} P_\theta(y_t^j | \boldsymbol{y}_{<t}^j, Y_{<j}, X), \tag{2.20}$$

where $\{\boldsymbol{y}_{<t}^j, Y_{<j}, X\}$ is the collection of all the sentences available in the source document, and the document translated so far. While for the target document, we usually have access to the past context only ($<j$), the future context ($>j$) is always available on the source side, except for the last sentence of the document. In practice, most of the existing CANMT systems make use of a smaller context than the available one, limiting themselves to a few sentences in the past or the future (see Section 2.4). Therefore, we prefer to employ the term **context-aware NMT** instead of document-level NMT. Using the neighbouring context is a legit approximation, as it contains most of the valuable contextual information (c.f. Müller et al. (2018); Lopes et al. (2020); see also Section 3.3.2), i.e., the one that helps disambiguate among similar translation alternatives. As for standard NMT, the source and target tokens are mapped to continuous representations before decoding, and the neural architecture is trained by maximizing the log-likelihood of the parallel corpus,

$$\mathcal{L}(\mathcal{C}, \theta) = \frac{1}{D} \sum_{d=1}^{D} \log P_\theta(Y^d | X^d) \tag{2.21}$$

$$= \frac{1}{D} \sum_{d=1}^{D} \sum_{j=1}^{|X^d|} \sum_{t=1}^{|\boldsymbol{y}^j|} \log P_\theta(y_t^j | \boldsymbol{y}_{<t}^j, Y_{<j}^d, X^d), \tag{2.22}$$

over a document-level parallel corpus $\mathcal{C} = \{(X^1, Y^1), (X^2, Y^2), ..., (X^D, Y^D)\}$ consisting of D documents. The main difference with the NMT training objective (Equation 2.3) is that the translation of each sentence is considered to be dependent on the context.

The generation of the translation is achieved by optimizing the context-aware translation probability of the document through beam search:

$$\hat{Y} = \underset{\tilde{Y} \in beam(X,\theta)}{\text{argmax}} \; P_\theta(\tilde{Y}|X), \tag{2.23}$$

2.2.3 Data

In Section 2.1.2 we have seen that some popular NMT corpora are organized in documents, like Europarl, News Commentary, IWSLT and OpenSubtitles2018. These corpora can also be employed for CANMT. In general, however, document-level corpora are rarer than sentence-level ones. Although the natural source of written texts is documents, many sentence-level datasets are assembled by scraping sentences from web pages with algorithms that do not preserve the information about document boundaries. This is the case for Common Crawl for example. Since deep learning models benefit from large training data (Hestness et al., 2017; Kaplan et al., 2020), researchers usually resort to a number of techniques to increase the total amount of training signal for learning CANMT. The two most common techniques adopted are: pre-training the model on large sentence-level corpora to perform context-agnostic NMT (also see the *two-step training* strategy presented in Section 2.4.2), and back-translation of monolingual data (Junczys-Dowmunt, 2019; Sugiyama and Yoshinaga, 2019; Ma et al., 2021b). In fact, large document-level monolingual corpora are freely available on the web, such as BookCorpus (Zhu et al., 2015) and PG-19 (Rae et al., 2020).

Details on data splits, language pairs, data versions, and statistics are described in the following chapter's sections devoted to experiments.

2.3 Evaluation

As for NMT (Section 2.1.3), the context-aware translation objective can be rewritten to highlight two concurrent desiderata, fluency and adequacy:

$$\underset{\tilde{Y}}{\text{argmax}} \, P_\theta(\tilde{Y}|X) = \underset{\tilde{Y}}{\text{argmax}} \, \frac{P_\theta(\tilde{Y})P_\theta(X|\tilde{Y})}{P_\theta(X)}. \tag{2.24}$$

The translation objective being extended to the whole document, also fluency $P_\theta(\tilde{Y})$ and adequacy $P_\theta(X|\tilde{Y})$ now refer to the document-level translation. This document-level objective is more aligned to the one of a human translator, who strives to contextualize each translated sentence with all the available information. As discussed in Section 2.2.1,

the extra-sentential linguistic context positively impacts fluency and adequacy by solving a range of ambiguities that affect translation. Such ambiguities pertain to a few words only, usually a tiny fraction of the whole document, but are nonetheless crucial to translation quality (Li et al., 2014). Consequently, most average translation quality metrics like BLEU are ill-equipped to measure improvements in solving them. Nevertheless, if such translation ambiguities are not handled correctly, the document-level translation quality is heavily damaged. Therefore, researchers in the field of CANMT have adopted evaluation methods that specifically target the translation of ambiguous inter-sentential linguistic phenomena (Popescu-Belis, 2019; Maruf et al., 2021). In this section, we discuss the discourse phenomena that may be ambiguous in translation, and then outline the existing evaluation methods for CANMT targeting such phenomena.

2.3.1 Translation ambiguities and discourse

In the linguistic literature, discourse is a general term for examples of language use, which usually refers to large units of language such as paragraphs, conversations, and interviews (Richards and Schmidt, 2013). Discourse can also be defined as a text that exhibits two properties across sentences: *cohesion* and *coherence*. Following this definition, every textual document amounts to a collage of discourse pieces, characterized by inter-sentential linguistic phenomena that contribute to making it coherent and cohesive. As it will be clear soon, cohesion and coherence during translation equates to resolving the translation ambiguities that arise at the sentence level thanks to context. In the absence of an appropriate use of context, cohesion and coherence in the target document are hampered by the ambiguities arising from the source language or the cross-lingual transfer, discussed in Section 2.2.1. Therefore, besides average translation quality scores, CANMT systems can be evaluated by analyzing how the inter-sentential linguistic phenomena[7] contributing to the cohesion and coherence of the source document are preserved in the translated document.

2.3.1.1 Cohesion

Cohesion is a surface property of text that refers to the relationships between its elements (Richards and Schmidt, 2013). These relationships can be grammatical or lexical.

Lexical cohesion is guaranteed by the usage of semantically related words. Two discourse

[7]The reader shall note that the linguistic phenomena discussed here are not purely inter-sentential. Our focus is on discourse phenomena, which span across multiple sentences by definition. However, the same kind of phenomena, can also be present within a single sentence.

phenomena contributing to lexical cohesion are *lexical repetition* and *collocation*. The former consists in the repeated use of synonyms or hyponyms. The latter consists in using series of words that co-occur more often than would be expected by chance (e.g., "Come to an end"). These phenomena might generate lexical discrepancies during translation, that only the context can solve. For instance, if the source repeats a word with multiple legit translations, the same translation must be used in the target sentence at every repetition.

$\boldsymbol{x}^{i-1}$ // $\boldsymbol{x}^{i}$ (EN): Would you like some soup? // Some soup?
$\boldsymbol{y}^{i-1}$ // $\boldsymbol{y}^{i,1}$ (FR): Tu veux de la soupe? // De la *soupe?*
$\boldsymbol{y}^{i-1}$ // $\boldsymbol{y}^{i,2}$ (FR): Tu veux de la soupe? // Du *potage?*

Notably, in this case, the MT system can not rely on source-side context to disambiguate the English "soupe" into either "soupe" or "potage" (which is a French hyponym of "soupe"), but it needs target-side context.

Grammatical cohesion, instead, is guaranteed by discourse phenomena like *deixis* and *ellipsis*.

Deixis - The use of an expression that directly relates to a time, place, person(s), or part of the same text and whose denotation depends upon context. In other words, deixis consists in the use of words and phrases such as "me" or "here" or "later" or "good question!". These phrases cannot be fully understood without additional contextual information: in this case, the identity of the speaker ("me"), the speaker's location ("here"), the time of the conversation ("later"), and a previous segment of the discourse ("good question!"). During translation, source-side deixis can generate ambiguities leading to translation errors. A typical case of problematic deixis is the one of *coreferential pronouns* (anaphoric or cataphoric) (Dylgjeri and Kazazi, 2013), that can raise discrepancies related to the gender or the T-V distinction. A translation ambiguity arises if the source language pronoun is neutral concerning gender or formality, but the target language requires to make the distinction. For instance:

$\boldsymbol{x}^{i-1}$ (EN): His cat eats so much.
$\boldsymbol{x}^{i}$ (EN): It always has a voracious appetite.
$\boldsymbol{y}^{i,1}$ (FR): *Il est toujours d'un appétit vorace.*
$\boldsymbol{y}^{i,2}$ (FR): *Elle est toujours d'un appétit vorace.*

Here, the neuter pronoun ("It") refers to an entity that was mentioned in the previous linguistic context ("His cat") and it raises a gender ambiguity for a translator that does not have access to it. In fact, French requires "It" to be translated into either "Il" or "Elle", according to the gender of the coreferent.

Ellipsis - The omission of words or phrases from sentences where they are unnecessary because they have already been referred to or mentioned. For example, when the subject of the verb in two co-ordinated clauses is the same, it may be omitted to avoid repetition. In translation, elliptical constructions in the source language raise ambiguity in two cases. Firstly, if the elided material affects the syntax of the sentence, which can lead to an incorrect inflection of some translated words. Secondly, if the target language does not allow the same types of ellipsis. For instance, French does not allow ellipsis with repetition of an auxiliary like in English. Hence the auxiliary has to be translated using an alternative expression that conveys the same meaning, which is different from the literal translation that a translator would pick in case context was unknown:

$\boldsymbol{x}^{i-1}$ (EN): I have met many people there.
$\boldsymbol{x}^{i}$ (EN): Have you?
$\boldsymbol{y}^{i,1}$ (FR): *Et toi?*
$\boldsymbol{y}^{i,2}$ (FR): *As tu?*

2.3.1.2 Coherence

Coherence is the ability of a text to convey meaning through the organization of its sentences, each conveying a part of the overall meaning. A coherent text can express a simple idea, a complex one, or even a whole narrative, without the necessity of either grammatical or lexical cohesion between its sentences (Richards and Schmidt, 2013). For instance, no cohesion devices are linking A's question and B's answer in this example:

A: What is the analysis of the software?
B: Unfortunately, I still have some bugs.

However, the exchange is coherent since the reader understands that the analysis depends on some software developed by B, which still suffers from some errors. This example highlights a critical facet of coherency: *lexical coherency*. Lexical coherency concerns how well a particular lexical choice fits semantically within the current discourse. Here, coherency is guaranteed by the use of the word "bugs", which is semantically related to "software" and makes us understand the situation. However, the word "bugs" can mean both "insects" or "errors in the code" and could raise a lexical discrepancy during translation should context not be adequately leveraged. Besides the appropriate use of the lexicon, coherency is guaranteed by how information is organized within the text. A meaningful organization requires appropriate sentence order and *discourse connectives*. While an NMT system usually preserves the number and order of sentences of a text,

Source:		
	context:	Oh, I hate flies. Look, there's another one!
	current:	Don't worry, I'll kill it for you.
Target:		
1	context:	Ô je déteste les mouches. Regarde, il y en a une autre !
	correct:	T'inquiète, je *la* tuerai pour toi.
	incorrect:	T'inquiète, je *le* tuerai pour toi.
2	context:	Ô je déteste les moucherons. Regarde, il y en a un autre !
	correct:	T'inquiète, je *le* tuerai pour toi.
	incorrect:	T'inquiète, je *la* tuerai pour toi.

Figure 2.5 – Example block from the contrastive test set on coreference by Bawden et al. (2018). "Mouches" and "moucherons" are alternative translations for "flies". The former alternative is feminine (requires "la"), while the latter is masculine (requires "le").

it might struggle with some discourse connectives whose meaning and function depend heavily upon context. For instance, the English word "since" can represent both causal or temporal discourse relation, according to the units of text that it is connecting, and therefore its translation needs contextual disambiguation.

2.3.2 Test Suites

Improvements in CANMT can be measured with targeted test sets. In the literature, we can find three kinds of test suites for this task: manual test suites, specialized test sets, and contrastive test suites.

Manual test suites - Standardized procedures or templates for the manual evaluation of a number of test cases based on a given machine translation task. For instance, WMT19 provided not only ratings for each system output but also detailed human analysis performed with manual test suites, such as the test suite on the coherence of English→Czech translations by Rysová et al. (2019). Manual test suites are the most accurate way to judge the quality of CANMT since human translators' knowledge, and *savoire faire* represent its target. However, manual evaluation is costly considering the volumes required by the ongoing research and development of novel CANMT systems, each trained on a different language pair.

Specialized test sets - The discourse phenomena that engender translation ambiguities

are sparse. Therefore, the quality of their translation has a minor impact on the average translation quality. Specialized test sets consist of sentences that are more densely populated with specific discourse phenomena than the average document. Consequently, average translation quality metrics like BLEU can be employed on these sets to evaluate context-aware translation. Usually, sentences are grouped in minimal documents where the target phenomenon is present at least once. For example, Voita et al. (2018) built a specialized English→Russian test set by extracting from OpenSubtitles2016 (Lison and Tiedemann, 2016) all the sentences containing at least an anaphoric pronoun whose nominal antecedent belongs to the previous sentence. As a result, this specialized test set consists of several minimal documents of two sentences, whose first sentence contains a nominal antecedent and the second a coreferential pronoun. Similarly, Cai and Xiong (2020) built a specialized test set focused on the English→Chinese translation of pronouns, discourse connectives, and ellipsis, while Wong et al. (2020) focused on the translation of cataphoric pronouns.

Contrastive test sets - Collections of *contrastive examples*, like those depicted in Figure 2.5. Each example consists of a source sentence paired with a reference translation and some corrupted versions of the reference. The reference is corrupted by substituting specific words representing the targeted discourse phenomenon (e.g., the third-person pronouns). The substitution is made with some pre-established alternatives (e.g., the third-person pronouns with a different gender than the reference). Models are assessed on their ability to rank the correct translation before the incorrect ones. The successful ranking depends on the ability of the system to exploit the relevant linguistic context, which is provided for both the source and the target side. In other words, the ranking accuracy reflects the ability of the CANMT system to disambiguate a specific kind of discourse-related translation ambiguity by leveraging context. However, a system that correctly ranks a contrastive example is not guaranteed to generate the uncorrupted reference. The system may well generate a different translation from all those provided in the contrastive example. In conclusion, contrastive sets provide a direct measure of a system's context-modeling ability but shall be used in conjunction with average translation quality metrics like BLEU or COMET. Unfortunately, contrastive test sets are expensive to build and are available for a limited set of discourse phenomena and language pairs.

2.3.2.1 Contrastive test sets in our experiments

Described below are the test suites that we employed for the experiments presented in the following chapters.

En→De ContraPro (Müller et al., 2018). A large-scale test set from OpenSubti-

Source

context-5:	I'm positive.
context-4:	Well, maybe our hacker removed the device.
context-3:	Wanted it to remain undetected.
context-2:	Wait, wait, guys, it just showed up.
context-1:	And... now it's gone.
current:	That means it's moving.

Target

context-5:	Sûr et certain.
context-4:	Peut-être que notre hacker a déplacé l'appareil.
context-3:	Voulant qu'il reste indétectable.
context-2:	Le voilà.
context-1:	Maintenant il n'est plus là.
correct:	Ça veut dire qu'*il* bouge.
incorrect:	Ça veut dire qu'*elle* bouge.

Figure 2.6 – Contrastive example from the En→Fr ContraPro by Lopes et al. (2020). "Device" is masculine, as well as its French translation "appareil". Therefore, the correct translation of the pronoun "it", appearing in the current sentence, is the masculine pronoun "il". The disambiguating context in underlined. Interestingly, this example showcases how the target-side context can be more informative than the source side in contrastive test sets.

tles2018 (Lison et al., 2018) that measures translation accuracy of the English anaphoric pronoun *it* into the corresponding German translations *er*, *sie* or *es*. Examples are balanced across the three pronoun classes (4,000 examples each). Each example requires identifying the nominal antecedent for a successful ranking. The nominal antecedent can be found either in the current sentence or in its context, which consists of up to five sentences in the past. Contrastive examples are created by changing the pronoun of the reference translation with wrong pronouns, while the nominal antecedent is kept unchanged. Therefore, this test suite can evaluate all kinds of CANMT systems (modeling either source-side context, target-side, or both).

En→Fr ContraPro (Lopes et al., 2020). A large-scale test set from OpenSubtitles2018, analogous to the previous one but focused on the translation of two English pronouns: *it* and *they*. It consists of 3,500 examples for each target pronoun type: *il* or *elle* for *it*, *ils* or *elles* for *they*. Figure 2.6 portrays an example from this test set.

Set	Total	d=0	d=1	d=2	d=3	d>3
En→De ContraPro	12000	2400	7075	1510	573	442
En→Fr ContraPro	14000	5986	4566	1629	880	939
Voita's Deixis	3000	0	1000	1000	1000	0
Voita's Lexical Cohesion	2000	0	855	630	515	0
Voita's Ellipsis inf	500	0	n.a.	n.a.	n.a.	0
Voita's Ellipsis vp	500	0	n.a.	n.a.	n.a.	0

Table 2.2 – Number of test istances of each contrastive set, and their distribution according to the distance (in number of sentences) of the disambiguating context. This detail is unknown for the subsets on ellipsis of Voita's contrastive set.

En→Ru Voita's test set (Voita et al., 2019b). It is a collection of sentence-pairs from OpenSubtitles2018 organized in 4 subsets that test for different discourse phenomenona needing contextual disambiguation: a test set for deixis with 3000 examples, one for lexical cohesion with 2000 examples, and two test sets for ellipsis. The first contains 500 contrastive examples on verb phrase ellipsis, the second 500 examples on "inflection" ellipsis, which refers to the kind of ellipsis affecting the morphological form of some words of the sentence with the elided content. Each contrastive example comes along with the three previous sentences, where the translation system can find the context necessary to the disambiguation of the discourse phenomenon. The contrastive examples are constructed by modifying the antecedent in the target context provided, as done for the examples shown in Figure 2.5, except for the verb phrase ellipsis subset. Therefore, CANMT systems modeling source-side context only can not be evaluated on deixis, lexical cohesion, and "inflection" ellipsis.

Table 2.2 summarizes some details about each of the above contrastive sets.

2.3.3 Automatic Metrics

Along the years, researchers have proposed some automatic metrics for machine translation targeting discourse phenomena. Back in the days of SMT, Hardmeier and Federico (2010) proposed an automatic evaluation metric for pronomial anaphora inspired by BLEU, called *AutoPRF*. Following their work, Miculicich Werlen and Popescu-Belis (2017) conceived *APT*, a metric to evaluate the accuracy of pronoun translation by aligning a candidate and a reference translation. The metric counts the number of identical and different pronouns, accounting for legitimate variations and omitted pronouns, and then combines all counts into one score. APT is language-specific and pronoun-specific, since it is based on specific

alignment software and heuristics. Unfortunately, (Guillou and Hardmeier, 2018) showed that these automatic metrics suffer from noisy alignments between the reference and the candidate translations, to the detriment of their correlation with human judgments. Later, Jwalapuram et al. (2019) open-sourced an automatic metric for context-aware pronoun translation from many source languages to English. The reference translations are coupled with a *noisy* version in which the reference pronoun has been replaced with a potentially incorrect one. The authors trained a neural model on this dataset that learns to rank English pronoun pairs, discerning the good from bad pronoun translations independently from the source language and the reference. Other metrics have been proposed for the evaluation of lexical cohesion (Wong and Kit, 2012) and discourse connectives (Hajlaoui and Popescu-Belis, 2013). Wong and Kit (2012) proposed to evaluate the document-level abundance of lexical cohesion devices with two simple metrics. A stemming algorithm (Porter, 1980) is used to identify word stems for each content word. Words with the same stem are identified and counted as *repetitions*. Then, synonyms and superordinates are clustered into semantic groups with WordNet (Fellbaum, 1998). Words belonging to the same semantic group or close semantic groups are counted as *synonyms*. The two metrics on lexical coherence and cohesion are then defined as *repetitions/contentwords* and *synonyms/contentwords*, respectively.

Automatic metrics are inexpensive compared to human evaluation, and they can also be used for tuning CANMT systems besides evaluation. Moreover, despite being language-specific, automatic metrics can be extended to other language pairs, given the availability of the needed software like language parsers, alignment software, stemmers, lexical databases. However, such software is intrinsically prone to errors, especially for less common language pairs, and the heuristics guiding these metrics are often too simplistic. These two elements often entail poor correlation with human judgments (Wong and Kit, 2012; Guillou and Hardmeier, 2018).

2.3.4 Statistical significance testing

When comparing the performance of different trained systems, following one of the above evaluation techniques, we must ensure that performance differences are not coincidental before stating that one system is better. System evaluation is always performed on a reduced sample of translation pairs, which is meant to represent a larger, and potentially boundless population of documents to be translated. Therefore, should we measure a difference between two systems' performance, it might be simply due to the specific sample that we have selected. A slightly different sample might have produced the opposite result, even if both samples represent the same population with the same degree

of representativeness. Statistical significance tests are used to decide whether we can reject the *null hypothesis*, stating that the performance difference between two trained systems is null on the population represented by the sample, under the assumption that the sample is indeed representative of it. In order to do so, significance tests compute the probability (*p-value*) of obtaining a performance difference greater or equal to the observed difference, on another sample from the same population, if the null hypothesis were true. If the *p-value* is lower than a certain threshold, usually 0.05, then we can reject the null hypothesis and affirm that the difference in performance measured between the two systems is statistically significant. Evidently, the *p-value* is inversely proportional to both sample size and performance difference. The closer the size of the sample to the size of the population, the more representative it is presumably. The wider the difference between two systems on the sample, the more likely we will measure a similar difference on similar samples.

In practice, several tests exist for statistical significance, some of which are based on certain hypotheses about the distribution of the performance differences observed in the sample. When comparing to CANMT systems, we do not have any prior knowledge about such distribution, and therefore we must resort to so-called *non-parametric tests* (Dror et al., 2018). In particular, in our experiments, we will use two different non-parametric tests:

- **McNemar's test** McNemar (1947) for comparing accuracy results on the contrastive test sets. This test is specifically designed for paired nominal observations, which is exactly the situation encountered in contrastive test sets: each system obtains a binary outcome (correct/incorrect ranking) for each contrastive example where it has to rank the correct translation higher than the incorrect ones.
- **Approximate randomization** (Riezler and Maxwell, 2005) for all the other cases, e.g., for comparing BLEU scores, with the sole exception of COMET scores.[8] Approximate randomization is based on resampling and it can be applied to non-binary, non-paired scores without requiring compliance to any hypothesis about their distribution (contrarily to, for instance, the Wilcoxon test (Wilcoxon, 1946)).

It is worth noticing that, a statistically significant improvement of a trained system over another, on a particular test set, is not enough to affirm that the first system is superior to the second (Carver, 2012). The reason is twofold. First, a single test set might not be representative of the population that we target with our approach. Second, a single trained

[8]Whose official library offers a built-in tool for the calculation of statistical significance with the paired T-Test and bootstrap resampling (Koehn, 2004).

system can not be fully representative of an approach such as a neural architecture or a training strategy because randomness affects it at various levels, from the initialization of the trainable parameters to the shuffling of the data. For this reason, we always adopt more than one test set, possibly sampled from different linguistic domains and language pairs. In cases where we measure minor differences between concurring systems, before making conclusions about the superiority of one system over the other, we run multiple experiments with different random seeds and then average the results. This procedure should alleviate the noise affecting a single experiment.

2.4 Approaches

In this section, we outline the approaches and architectures to context-aware NMT that have been proposed in the literature. According to a popular taxonomy (Kim et al., 2019), we group them in two families: concatenation approaches (Section 2.4.1) and multi-encoding approaches (Section 2.4.2), following . In Section 2.4.3 we will overview some approaches that do not fit in either of these families.

2.4.1 Concatenation approaches

Strategy	**Training**	**Inference**
Sliding2to1	$\boldsymbol{y}^j$	$\boldsymbol{y}^j$
Sliding2to2	$\boldsymbol{y}^{j-1}$<S>$\boldsymbol{y}^j$	$\boldsymbol{y}^j$
Jumping2to2	$\boldsymbol{y}^{j-1}$<S>$\boldsymbol{y}^j$	$\boldsymbol{y}^{j-1}$<S>$\boldsymbol{y}^j$

Figure 2.7 – System output of the three main concatenation strategies with $\boldsymbol{x}^{j-1}$<S>$\boldsymbol{x}^j$ as input. <S> is a special token that marks the boundaries between sentences. For simplicity, we omit the end-of-sequence tag <E>. At inference time, the output reported for SlidingKtoK is what is kept after discarding the translation of the context.

The most straightforward approach to CANMT consists in concatenating the context to the current sentence before feeding it to the standard encoder-decoder architecture (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Junczys-Dowmunt, 2019). Concatenation approaches have the advantage of using the same architecture of standard context-agnostic NMT, so that all the parameters can be easily initialized with those of a pre-trained context-agnostic NMT system. Besides the end-of-sequence token <E>, a special token <S> is introduced to mark the boundaries between the concatenated sentences. Both past and

future sentences can be concatenated to the current sentence $\boldsymbol{x}^j$, until document boundaries are reached. In practice, concatenating the full document X might be impossible with the Transformer model since self-attention's complexity scales quadratically with sequence length (see Section 2.1.4.3). Usually, only a fixed number of sentences $K : K < |X|$ is concatenated. Decoding can then follow three strategies: SlidingKto1, SlidingKtoK, or JumpingKtoK.

SlidingKto1 (Tiedemann and Scherrer, 2017; Agrawal et al., 2018) - The model decodes a single sentence: the current one. In this case, the model does not have access to the target-side context.

SlidingKtoK (Tiedemann and Scherrer, 2017; Agrawal et al., 2018) - The model generates the translation of a window of K source sentences: the current (jth) sentence and the $K-1$ sentences concatenated as context on the source side. By decoding the full sequence, the model also has access to the target-side past context. Despite the increased complexity, exploiting target-side context proved useful in most of the studies on concatenation approaches (Agrawal et al., 2018; Scherrer et al., 2019a; Lopes et al., 2020; Ma et al., 2021b). The training loss is calculated over the whole output. However, the translation of the context is discarded at inference time. Then, the window is slid by one position forward to repeat the process for the $(j+1)$th sentence and its context. The downside of this approach is the necessity to translate every sentence in the document as many times as the size K of the sliding window (except for sentences close to document boundaries).

JumpingKtoK (Junczys-Dowmunt, 2019) - Similarly to the SlidingKtoK strategy, the whole concatenated sequence is translated by the model. However, the translated context is not discarded at inference time: the entire translated sequence of K sentences is kept. Then, the model jumps K positions forward, to the next window of K sentences, and repeats the process. Therefore, every sentence is translated a single time at inference, differently from SlidingKtoK. Instead, the training phase can be identical. The input sequences can be formed by either sliding a window of K sentences over the training data or by jumping K positions forward for each example. The former training strategy is arguably more convenient because it results in more training instances. The downside of JumpingKtoK is not distinguishing between current and context sentences. Each sentence in the window is "current" because it is kept at inference time. Hence, the available context is different for each sentence in the translated window. The first sentence in the window has only access to $K-1$ source sentences in the future, while the last sentence can see the past $K-1$ sentences, both source and target, but no future. This strategy is suboptimal whenever the size of the window $K < |X|$ because the model loses access to important context for some sentences.

Figure 2.7 outlines these three concatenation approaches, in the case of one previous sentence as context ($K = 2$).

Concatenation approaches are trained by optimizing the same objective function as standard NMT, defined in Equation 2.3:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \log P_\theta(\boldsymbol{y}_K^j | \boldsymbol{x}_K^j) \tag{2.25}$$

$$= \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \sum_{t=1}^{|\boldsymbol{y}_K^j|} \log P_\theta(y_{K,t}^j | \boldsymbol{y}_{K,<t}^j, \boldsymbol{x}_K^j) \tag{2.26}$$

The only difference consists in the source and target sequences under consideration, which are concatenations of K source and target sentences, respectively, so that the likelihood of the current source and target sentences is conditioned on their context:

$$\boldsymbol{x}_K^j = \boldsymbol{x}^{j-K+1}{}_{\texttt{<S>}} \boldsymbol{x}^{j-K+2}{}_{\texttt{<S>}} \ldots {}_{\texttt{<S>}} \boldsymbol{x}^{j-1}{}_{\texttt{<S>}} \boldsymbol{x}^{j}{}_{\texttt{<E>}}, \tag{2.27}$$

$$\boldsymbol{y}_K^j = \boldsymbol{y}^{j-K+1}{}_{\texttt{<S>}} \boldsymbol{y}^{j-K+2}{}_{\texttt{<S>}} \ldots {}_{\texttt{<S>}} \boldsymbol{y}^{j-1}{}_{\texttt{<S>}} \boldsymbol{y}^{j}{}_{\texttt{<E>}}. \tag{2.28}$$

Both past and future context can be concatenated to the current pair $\boldsymbol{x}^j, \boldsymbol{y}^j$, although here we consider only the past context for simplicity. In the case of SlidingKto1, the likelihood of the target sequence is not conditioned on contextual translations, i.e., $\boldsymbol{y}_K^j = \boldsymbol{y}^j$.

2.4.2 Multi-encoding approaches

Multi-encoding models couple a self-standing sentence-level NMT system, with parameters θ_S, with additional parameters for modeling the context either on the source side, target side, or both. We refer to these parameters as the *contextual parameters* θ_C. The complete context-aware architecture has parameters $\Theta = [\theta_S; \theta_C]$, and it can model context from the past, future, or both. Most of the literature focuses on a few previous sentences, where the relevant context is concentrated.
Most of the multi-encoding models can be described as instances of two architectural families (Kim et al., 2019), which only differ in how the encoded representations of the context and the current sentence are integrated. These two families are depicted in Figure 2.8.

Outside integration - In this approach, the encoded representations are merged outside the decoder (Maruf et al., 2018; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Maruf et al., 2019; Zheng et al., 2020). Merging can happen in different ways, such as by simply concatenating the encodings, by summing them with a gate, or with an

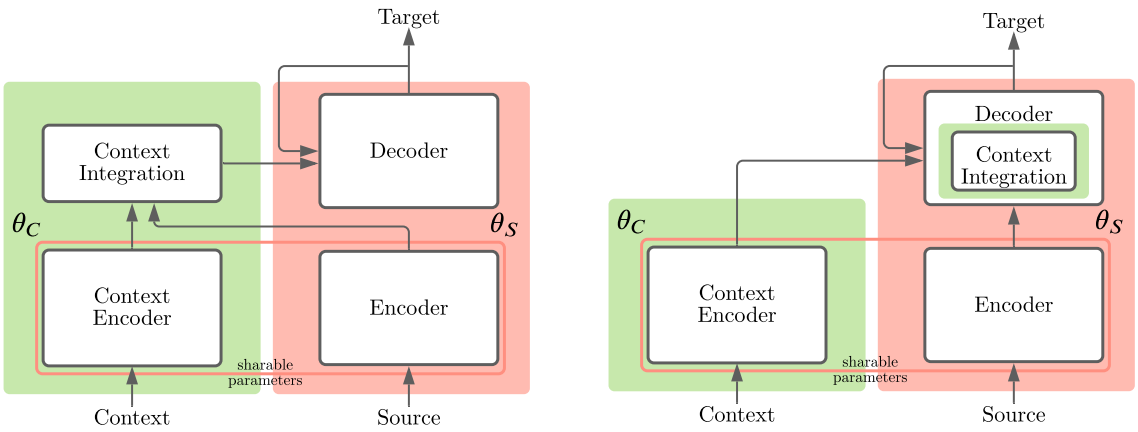

Figure 2.8 – General outline of multi-encoding models with outside (left) and inside (right) context integration. Modules with sentence-level parameters θ_S have a red background, while the green background is for contextual parameters θ_C.

attention mechanism whose queries from the current sentence attend to context token representations. A popular multi-encoding architecture with outside integration is one proposed by Miculicich et al. (2018). They adopt a hierarchical attention network to encode the context and integrate it with the information processed by the standard NMT encoder, as depicted in Figure 2.9.

Inside integration - The decoder attends to the context representations directly, using its internal representation of the decoded history as queries (Tu et al., 2018; Kuang et al., 2018; Bawden et al., 2018; Voita et al., 2019b; Tan et al., 2019). The attention layers in charge of integrating the context are usually embedded in the decoder and intertwined with its other layers.

Including past target-side context can be harmful because of the error propagation problem (Zhang et al., 2018, 2020a), but most of the literature shows it to be important to make the most out of context. Past works have successfully included target-side context information in two ways:

(i) Translating past sentences along with the current one and then discarding them, as in SlidingKtoK (Bawden et al., 2018).

(ii) By making the decoder attend the target-side hidden representations or embeddings of previously decoded sentences (Miculicich et al., 2018; Voita et al., 2019b; Maruf et al., 2019; Zheng et al., 2020). In this case, some extra parameters can be added

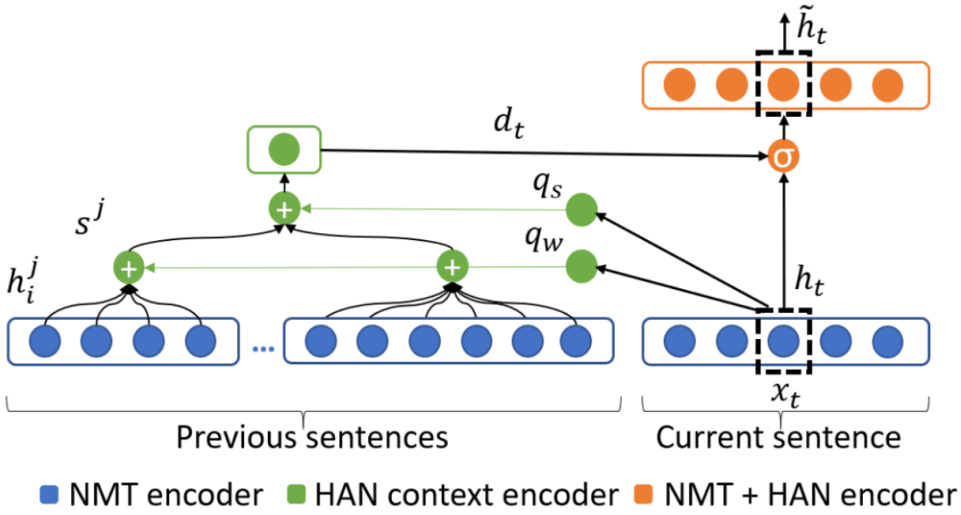

Figure 2.9 – Illustration of the hierarchical attention network (HAN) adopted by Miculicich et al. (2018) to encode context and integrate it with the context-agnostic token representations. The + symbol denotes attention, while σ is a gate function. The context-agnostic representation h_t of the token x_t is transformed into two queries: q_w and q_s. The former is employed to attend the tokens h_i^j of the jth context sentence. As a result, a contextualized representation s_j of h_t is generated for each context sentence. q_s attends to these representations to form a more global contextual representation d_t. This is then merged with the context-agnostic representation h_t by means of a gate, resulting in the context-aware representation $\tilde{h}_t$.

to encode and integrate target-side context. The disadvantage of this approach is that sentences can not be batched at inference time, but they have to be translated one at a time, following the order of the document.

Many works have found it helpful to share parameters between the standard encoder and the context encoder (Voita et al., 2018; Li et al., 2020), which is equivalent to having a single encoder for both current and context sentences. Thus, the number of contextual parameters to learn, $|\theta_C|$, is drastically reduced. Moreover, sharing parameters allows caching token representations from the current sentence to employ them in later steps as context representations. Two-pass approaches represent an extreme variant of representation caching with multi-encoding models (Voita et al., 2019a; Zheng et al., 2020).

Two-pass approaches - A context-agnostic model (or encoder) makes a first sentence-level translation (or encoding) of the entire document. All the context-agnostic token representations are cached. Then, a second pass is performed with the context-integration modules, which exploit the context-agnostic drafts as contextual information for the current sentence. This technique has the advantage of making future target-side context available, while usually, it is not.

Two-step training - Multi-encoding models are commonly trained following a two-step strategy (Tu et al., 2018; Zhang et al., 2018; Miculicich et al., 2018; Li et al., 2020), in order to exploit sentence-level parallel corpora before document-level training. The first step involves training θ_S independently on a sentence-level parallel corpus $\mathcal{C}_S$. Secondarily, contextual parameters θ_C are trained on a document-level parallel corpus $\mathcal{C}_D$ while fine-tuning or freezing θ_S. Note that $\mathcal{C}_S$ can also include sentences from $\mathcal{C}_D$.

Encoding sentence position - Multi-encoding CANMT can benefit from knowing the distance of context sentences from the current one. Knowing the order of contextual information is essential to understand the context fully and to select between alternative contextual clues. The standard positional encoding proposed by Vaswani et al. (2017) is insufficient because sentences are encoded separately, re-initializing token positions. Different strategies have been proposed in the literature to include sentence position information in multi-encoding architectures:

(i) Adding a *sentence distance embedding* to each token, that tells the model how far away tokens are from the current sentence (Voita et al., 2019b).

(ii) Adding a *sentence position embedding*, similar to classical positional encoding but for the position of the segment within the document (Zheng et al., 2020).

(iii) Assign token positional embeddings progressively to the current sentence, then to

the previous one, and so on, so that far away sentences have high values of positional embeddings (Li et al., 2019).

2.4.3 Other approaches

A few CANMT approaches do not fall into either of the two families presented above. For instance, some approaches focus on leveraging document-level monolingual data to learn to contextualize NMT. Martínez Garcia et al. (2019) and Yu et al. (2020) train a context-aware language model on a target-language corpus and then generate translations by fusioning the softmax output of the NMT decoder and the language model. Voita et al. (2019a) devised an automatic post-editing system called DocRepair, trained to turn the context-agnostic translation of a document into a contextualized, consistent translation. Training data for DocRepair are generated in two steps, each involving a context-agnostic NMT system. First, the target-language monolingual documents are translated into the source language. Then, they are translated back from the source to the target language. Thus, the training set consists of document pairs comprising the original document and a context-agnostic translation of it. Being an automatic post-editing system, DocRepair can work on top of whatever MT system.

Some concatenation or multi-encoding approaches try to integrate discourse-related information as additional input features. Examples of extra features are lexical chains of semantically similar words to promote word sense disambiguation (Rios Gonzales et al., 2017), or coreference chains to promote coreference resolution (Stojanovski and Fraser, 2018).

Finally, other research works looked at the problem of CANMT from a learning perspective, trying to include context in the standard learning objective (Saunders et al., 2020; Jean and Cho, 2019). For example, Jean and Cho (2019) designed a regularisation term that is applied at the token, sentence, and corpus levels and that is based on a pair-wise ranking loss that pushes the model to assign a higher log-probability to a translation paired with the correct context than to a translation without context. More recently, (Hwang et al., 2021) proposed a similar approach based on a contrastive loss.

2.4.4 Challenges

Both concatenation and multi-encoding approaches have strengths and weaknesses, that are almost complementary. Concatenation approaches have the advantage of employing the standard encoder-decoder architecture without any additional learnable parameter. Hence,

learning intra-sentential contextualization can be easily transfered to extra-sentential token contextualization. In fact, translating a concatenation of sentences is equivalent to translating a long sentence from an architectural standpoint. Nonetheless, concatenating sentences results in processing longer sequences, which brings three main downsides.

Error accumulation - If a standard NMT system generates some wrong tokens, these will negatively impact the generation of the rest of the sequence because generation follows an auto-regressive strategy (Ranzato et al., 2016). The longer the sequence to generate, the higher the risk of accumulating errors. In the case of SlidingKtoK and JumpingKtoK, the risk of error accumulation increases with K (Zhang et al., 2020a).

Computational complexity - Self-attention's complexity increases quadratically with sequence length, as discussed in Section 2.1.4.3, slowing down training and forcing the selection of smaller batch sizes, which may be sub-optimal for Transformers (Popel and Bojar, 2018). Ma et al. (2020) proposed to lighten the computational burden of processing long sequences by adopting a SlidingKto1 approach where the K-1 source context sentences are treated in the first self-attention block of the encoder only. Subsequently, the latent representations of the context sentences are discarded, and the remaining layers deal with the representation of the current sequence. Another possible solution is the adoption of self-attention approximations with sub-quadratic complexity (Kitaev et al., 2020; Rae et al., 2020; Beltagy et al., 2020; Wang et al., 2020; Zaheer et al., 2020). In a preliminary analysis, we have trained and tested the Luna architecture (Ma et al., 2021a), which is equivalent to the Transformer apart from the adoption of *Luna attention*, as a drop-in replacement for the regular self-attention. *Luna attention*'s complexity is linear with respect to sequence length, although it requires sequential computation for each time-step. Although Luna is not the first self-attentive alternative with linear complexity (Tay et al., 2020), it compares favorably in terms of performance on long-sequence tasks to other linearly-complex architectures such as the Linear Transformer (Katharopoulos et al., 2020) and the Performer (Choromanski et al., 2021). Unfortunately, our preliminary results were not encouraging since we measured degraded performance with *Luna attention* both on short and long concatenations, in line with Petrick et al. (2022)'s work on CANMT with sub-quadratic attention.

Learning challenge - When sequences are long, it is challenging for the attention mechanism to match tokens within the sequence correctly, and the risk of paying attention to irrelevant elements increases. Paying attention to the "wrong tokens" can harm intra-sentential and extra-sentential contextualization, associating queries with the wrong latent features. Liu et al. (2020b) showed that learning to translate long sequences comprised of many sentences fails without the employment of large-scale pre-training or data-augmentation (Junczys-Dowmunt, 2019). Bao et al. (2021) provided some ev-

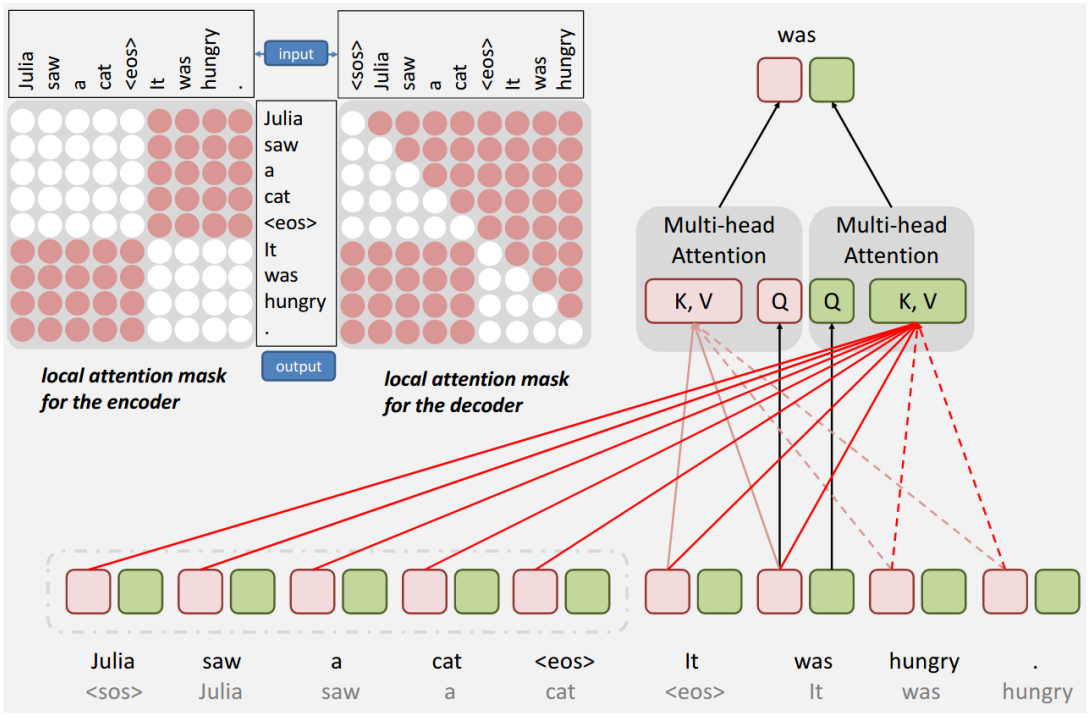

Figure 2.10 – Illustration of the Long-Short Term Masking Self-Attention by Zhang et al. (2020a). Green nodes: global self-attention,which is the same as the standard self-attention. Pinknodes: local self-attention, which does not have access to the information from the document context. The red dash lines are removed in the decoder attention.

idence about this learning challenge. They analyzed the distribution of the attention weights in concatenation models stuck in local minima with a high validation loss. Their attention weights distribution is flat (with high entropy), both in the encoder and the decoder. In other words, attention struggles to learn the *locality properties* of language, i.e., the principles of proximity between linguistic elements characterizing many linguistic structures (Rizzi, 2013; Hardmeier, 2012). Instead, successful models converging to low validation loss present a distribution of the attention weights that is less flat and more peaked on a few tokens within the sequence. As a solution, Zhang et al. (2020a) and Bao et al. (2021) proposed two modifications of the standard Transformer architecture that encourage a certain degree of local focus of the attention module, at the cost of adding some learnable parameters. Zhang et al. (2020a) introduced a self-attention mask to enable *local* self-attention on top of the standard self-attention. They refer to standard self-attention as *global* self-attention because each token can attend to all the tokens belonging to the concatenation. Instead, the proposed masking prevents queries to attend the tokens outside their own sentence, as pictured in Figure 2.10. Thus, two hidden representations are produced for every token: the local representation that is context-agnostic and the global representation that is context-aware but noisier, "distracted" by the aboundant contextual information. The two representations are then concatenated, and the model learns how to trade off between them with a linear projection. Bao et al. (2021) proposed a very similar solution, which extends the idea of local self-attention to cross-attention. Instead of using the combination of local and global attention in every layer, they use *local* attention in every layer and *global* attention only in the top layers of the decoder. They explain this choice by mentioning that the standard NMT Transformer architecture models long distance syntactic relations in its top layers, while the lower layers mostly catch local syntactic relations (Jawahar et al., 2019). In Chapters 4 and 5 we will propose and evaluate some light-weight approaches to tackle this learning challenge.

Multi-encoding architectures are more flexible considering context length. First, computational complexity does not grow quadratically with context length, which can translate into increased efficiency during inference. Second, they all separate the local encoding of the current sentence from the global, contextualized encoding, thus preventing self-attention from getting too "distracted" by context. Finally, they usually generate the current sentence alone, avoiding the problem of error accumulation. Another advantage of multi-encoding models is that they can encode context with a different network than the one used for the current sentence, potentially in a more efficient way. Intuitively, encoding the context does not require the same sophistication as encoding the current sentence. We do not need to translate context but only extract a few features that are helpful for the current sentence. Therefore, the contextual encoder could be a shallower, more efficient version

of the current-sentence encoder (Zhang et al., 2018). Alternatively, it could be based on a different architecture, such as one that approximates attention with sub-quadratic complexity, while the current sentence could still be processed with the original attention mechanism (see Section 6.2.2).

However, multi-encoding architectures also present some non-trivial challenges related to the fact of having to learn additional parameters. These parameters are in charge of contextualization, which requires to learn three different tasks:

(i) encoding the context meaningfully;

(ii) identifying the relevant context;

(iii) merging the relevant context with the local information.

The first task can be bypassed by sharing the parameters between the standard and context encoder. Instead, the selection and integration of context have to be learned from scratch on document-level data. However, learning to identify the relevant tokens among the many available in the context is challenging and presents a significant obstacle: the available training signal is as sparse as the discourse phenomena involved. We will formalize and discuss this learning challenge thoroughly in Chapter 3.

In conclusion, it is not yet clear which approach is best, and the scientific community is still researching both concatenation, and multi-encoding approaches (Ma et al., 2021b; Zhang et al., 2021; Yin et al., 2021; Zhang et al., 2022; Sun et al., 2022; Guo et al., 2022; Tan et al., 2022). We hope to contribute to this collective effort with this thesis.

Chapter 3

Divide and rule pre-training for multi-encoding approaches

Most of the contributions presented in this chapter are published in:

> Lupo, L., Dinarelli, M. and Besacier, L. (2022). Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557-4672, Dublin, Ireland.

Contents

3.1 Introduction

As discussed in Section 2.4.4, multi-encoding models are more flexible than concatenation approaches and potentially more efficient, but they have been criticized for acting as mere regularization methods (Kim et al., 2019; Li et al., 2020). In some cases, they have even been shown to perform worse than sentence-level systems on contrastive test sets for the disambiguation of discourse phenomena (Lopes et al., 2020). In this chapter, we address this criticism by showing that training multi-encoding models is challenging for two reasons:

- the sparsity of *contextual training signal*, i.e., the signal that pushes systems to translate in a context-aware fashion, which comes from the words that need context to be correctly translated;
- the sparsity of relevant context words, the ones needed to disambiguate translation.

A trivial way to improve context-aware learning is by increasing the amount of document-level training data. Large document-level parallel corpora are not always available, but some works have proposed data augmentation techniques to remedy this lack (Sugiyama and Yoshinaga, 2019; Stojanovski et al., 2020; Huo et al., 2020). However, as we will show in our experimental section, this solution is not efficient and often sub-optimal. Therefore, we introduce a novel pre-training strategy, *divide and rule* (*d&r*), that is based on a simple and yet powerful technique to augment the contextual training signal and to ease learning efficiently: splitting parallel sentences into segments (see Figure 3.1). Simply put, feeding a context-aware model with a sequence of incomplete, shorter, consecutive segments forces it to look for context (i.e., surrounding segments) more frequently and makes it easier to retrieve relevant context because segments are shorter. This results in faster and improved learning. We pre-train multi-encoding models on split datasets and evaluate them in two ways: BLEU score and a contrastive evaluation of translation of discourse phenomena.

Our main contributions are the following:

- we show that context-aware multi-encoding models need to be trained carefully because the contextual training signal is sparse, as well as the context elements useful for contextualization;
- we propose the *d&r* pre-training strategy, which facilitates the training of contextual parameters by splitting sentences into segments, with four splitting variants;
- we support this strategy with an analysis of the impact of splitting on the distribution of discourse phenomena;

$\mathbf{x}^{i,1}$	He said that it was a project of peace
$\mathbf{x}^{i,2}$	and unity and that it brought people together .
$\mathbf{y}^{i,1}$	*Il disait que c' était un projet de paix*
$\mathbf{y}^{i,2}$	*et d' unité et qu' il réunissait les gens .*
$\mathbf{x}^{j,1}$	I think single-cell organisms **are**
$\mathbf{x}^{j,2}$	possible within two years .
$\mathbf{y}^{j,1}$	*Je pense que les organismes unicellulaires*
$\mathbf{y}^{j,2}$	***sont** possibles dans 2 ans .*

Figure 3.1 – Example of sentence pairs from En→Fr IWSLT17, after being tokenized and split in the middle. After the splitting, some syntactic relations span across two segments (underlined). Also, some source-side words are not parallel with their reference (**in bold**).

- we demonstrate that this strategy is both effective and efficient, as it allows multi-encoding models to learn better and faster than by simply increasing the training data.

3.2 The double challenge of sparsity

Some works criticized multi-encoding methods (Kim et al., 2019; Li et al., 2020), arguing that they do not improve sentence-level baselines in terms of BLEU when the baseline is well regularized. When there are improvements, it is argued that the context-encoder works as a noise generator, making training more robust. The improvements are not to be attributed to better context modeling. Along this path, Lopes et al. (2020) showed that multi-encoding architectures struggle to model contextual information and even deteriorate the performance of a sentence-level baseline on contrastive test sets. Many proponents of multi-encoding models only show BLEU improvements without providing any discourse-targeted evaluation. This does not allow an assessment of their context-modeling capability. We posit that training the contextual parameters of multi-encoding models is non-trivial because of two challenges: (i) the sparsity of the training signal, which comes from the words that need context to be correctly translated (most of the words of a sentence can be translated without context); (ii) the sparsity of context words that are useful for contextualization (most of the context is useless). As such, missing the right experimental setting can bring unsuccessful training and unconvincing results.

Algorithm 1: Split parallel corpus

input: Parallel corpus $\mathcal{C}$, minimum source length l_{min}, function wheresplit()
for $i = 1, \ldots, |\mathcal{C}|$ **do**
  if $len(\mathbf{x}^i) \geqslant l_{min}$ **then**
    $m_x, m_y = \text{wheresplit}(\mathbf{x}^i, \mathbf{y}^i, ...)$
    $\mathbf{x}^{i,1} = \mathbf{x}^i_{<m_x}$ and $\mathbf{x}^{i,2} = \mathbf{x}^i_{\geqslant m_x}$
    $\mathbf{y}^{i,1} = \mathbf{y}^i_{<m_y}$ and $\mathbf{y}^{i,2} = \mathbf{y}^i_{\geqslant m_y}$
  end if
end for
return Split corpus $\mathcal{C}_D$

3.2.1 More data?

A trivial way to offset sparsity is to increase the volume of training data. In fact, existing works that report strong results with discourse-targeted evaluation train their contextual parameters with millions of document-level sentence pairs (Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019b; Zheng et al., 2020; Wong et al., 2020; Kang et al., 2020). In contrast, many works in the literature train models with the TED talks' subtitles released by the IWSLT shared tasks (Cettolo et al., 2012), which only consist of a couple of hundred thousand parallel sentences. In the experimental section (3.5), we will show that IWSLT's subtitles are not sufficient to train multi-encoding models effectively. It follows that one cannot make fair comparisons between alternative architectures in such experimental settings. On the other hand, we will provide empirical confirmation of the intuition that increasing the volume of training data helps in learning contextual parameters. However, this solution is inefficient and only partial to the double sparsity problem. Moreover, it is not always possible: large document-level training sets may not be available in many languages. In the following section, we propose a pre-training solution that makes efficient use of the available data for learning contextual parameters effectively.

3.3 Proposed approach

One way to simulate document-level data is to split sentences into two or more segments (Luong et al., 2016). In this way, intra-sentential syntactic relations are broken, and a word previously disambiguated by looking at its neighbors in the sentence now requires contextual information to be correctly translated. Moreover, splitting sentences increases the concentration of relevant context words, as shown in Section 3.3.2. Within the

framework of MT, if we split the source sentence, its corresponding reference has to be split too. The proposed approach, *divide and rule* (*d&r*), consists in pre-training the model on a dataset $\mathcal{C}_D$ that results from splitting all the sentences of a parallel corpus $\mathcal{C}$ that have at least l_{min} tokens, as described by Algorithm 1. Each source-side sentence $\mathbf{x}^i$, with index $i = 1, ..., |\mathcal{C}|$, is split into $\mathbf{x}^{i,1}$ and $\mathbf{x}^{i,2}$. Its corresponding reference $\mathbf{y}^i$ is split into $\mathbf{y}^{i,1}$ and $\mathbf{y}^{i,2}$. The resulting corpus is a document-level parallel corpus $\mathcal{C}_D$, such that, if the original corpus C was itself document-level, then $\mathcal{C}_D$ keeps the same document boundaries than C. Figure 3.1 illustrates two examples of parallel sentences split in the middle. In both instances, a CANMT system needs to look at $\mathbf{x}^{i,1}$ for translating $\mathbf{x}^{i,2}$ correctly, i.e., to look at the past context. In the first one, the English neuter pronoun “it" could be translated into “il" or “elle" according to the gender of its antecedent (there is no singular neuter 3rd-person in French). The antecedent “a project", which is in the previous segment, allows for disambiguating it into “il". In the second example, the adjective “possible” can be correctly translated into its plural version “possibles” by looking back at the noun it refers to: “organisms”.

3.3.1 Splitting methods

In Algorithm 1, the wheresplit function returns the token indices m_x and m_y of $\mathbf{x}^i$ and $\mathbf{y}^i$, where the sentence is split. In this work, we propose and experiment with four variants of this function.

Middle-split. The simplest strategy is to split the source and the target in the middle. In this case, wheresplit = middlesplit$(\mathbf{x}^i, \mathbf{y}^i)$ returns $m_x = \lfloor len(\mathbf{x}^i)/2 \rfloor$ and $m_y = \lfloor len(\mathbf{y}^i)/2 \rfloor$. Following this method, it can happen that $\mathbf{x}^{i,j}$ and $\mathbf{y}^{i,j}$, with $j = 1, 2$, are not parallel, as illustrated in the second example of Figure 3.1. The verb “are” belongs to $\mathbf{x}^{i,1}$, but its translation “sont” does not belong to its corresponding reference segment $\mathbf{y}^{i,1}$. This problem arises whenever the splitting separates a set of words from their reference, which end up in the other segment. Evidently, this method requires that the two languages do not have strong syntactic divergence, to avoid too large mismatches between $\mathbf{x}^{i,j}$ and $\mathbf{y}^{i,j}$, with $j = 1, 2$.

Aligned-split. As a solution to the misalignment problem between source and target segments, we can calculate word alignments A^i and use them to inform our splitting strategy by setting wheresplit = alignedsplit$(\mathbf{x}^i, \mathbf{y}^i, A^i)$, where alignedsplit splits each sentence close to the middle while avoiding to separate aligned words in different segments. The word alignments of the ith sentence are a set of tuples:

$$A^i = \{(j, k) | x^i_j \text{ and } y^i_k \text{ are aligned}\},$$

where $j = 1, ..., |\mathbf{x}^i|$ and $k = 1, ..., |\mathbf{y}^i|$ are the indices of the words belonging to $\mathbf{x}^i$ and $\mathbf{y}^i$, respectively. The alignedsplit method sets $m_x = \lfloor len(\mathbf{x}^i)/2 \rfloor$ and $m_y = \max\{k \colon (j,k) \in A^i, j \leqslant m_S\}$. Then, it checks whether this choice is not breaking apart two aligned words. Formally, it checks that:

$$x^i_j \in \mathbf{x}^{i,1} \wedge y^i_k \in \mathbf{y}^{i,1} \text{ or } x^i_j \in \mathbf{x}^{i,2} \wedge y^i_k \in \mathbf{y}^{i,2}. \tag{3.1}$$

If this condition is not encountered, it tries to split the sentence pairs closely, where condition (3.1) is met. If the condition cannot be met (e.g., because one of the two segments would be too short (<3 tokens)), alignedsplit falls back on middlesplit.

Synt-split. Splitting aims at breaking intra-sentential discourse phenomena in order to force the model to exploit the context more frequently. Therefore, we propose a splitting method that maximizes this objective. This method consists in retrieving syntactic and semantic relations L (i.e., syntactic dependencies and some discourse phenomena) in the training set, and leveraging this information to split sentences as close to the middle as possible while breaking at least a relation, if present. Since not all discourse phenomena raise translation ambiguities when broken, one can choose which phenomena should be prioritized; in this work, we chose pronominal coreferences. The function wheresplit = syntsplit$(\mathbf{x}^i, \mathbf{y}^i, L^i)$ takes as input the coreference relation L^i detected by CoreNLP (Manning et al., 2014) on the source sentence i. If L^i is not empty, a relevant intra-sentential coreferential relation is present (in our experiments, we look at pronominal coreferences). In this case, the algorithm checks whether splitting in the middle ($m_S = \lfloor len(\mathbf{x}^i)/2 \rfloor$) allows breaking L^i, i.e., to separate the two syntactically-related tokens in different segments. If *middle-split* does not achieve this goal, m_x is set to the closest index from the middle that breaks the relation, except for the case in which breaking the relation would mean generating a too-short segment (<3 tokens). In this case, the algorithm falls back to *middle-split*.

Multi-split. The methods above can be extended to splitting sentences into more than two segments. The more we split sentences, the more likely it is that context is needed for each segment, thus increasing the training signal for contextual parameters.

For more details, refer to Section 3.5.3.

3.3.2 Impact on discourse phenomena

To give an explicit picture of how and why splitting sentences helps to learn contextual parameters, we processed the source training data of En→Fr IWSLT17 with CoreNLP (Manning et al., 2014), and we computed some statistics on coreference chains and dependency

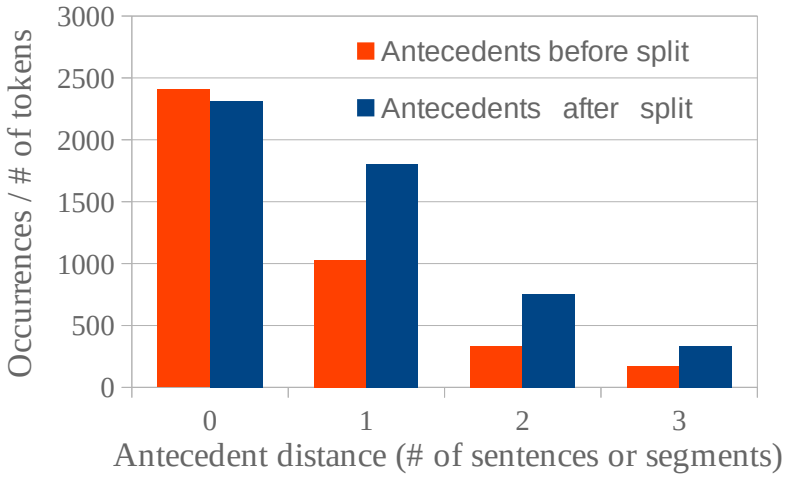

Figure 3.2 – En→Fr IWSLT17: number of antecedents of anaphoric pronouns at a given distance in terms of sentences or segments, normalized by the number of tokens the model needs to attend to resolve the coreference.

parse trees, before and after applying the *middle-split* method. Statistics show how splitting the sentences of a document helps in two ways:

More cases. Splitting generates new cases that require context for disambiguation, making training signal more abundant. When syntactic dependencies are split into two segments, the model needs to access the context for reconstructing the syntactic structure of the source sentence and correctly translate it, as shown in Figure 3.1. To have an idea of the magnitude of this effect, we calculated the percentage of the sentences where the splitting method breaks at least one syntactic dependency between the main verb of the sentence (the root) and : (i) the subject or object (18.1% of the sentences); (ii) any complement (9.5%); (iii) any modifier (9.3%). Considering all the dependencies with the root, except punctuation, we find that in 84.8% of the sentences at least a syntactic dependency is broken. Given such a high proportion, the *middle-split* variant is a good approximation of a syntactically supported splitting approach. These cases add up to the many other cases of broken relations, such as coreferences, which make the overall contextual training signal more abundant.

Denser cases. The splitting also shortens the average length of text sequences, which eases the job of CANMT systems because they have to attend to fewer words while looking for context. In Figure 3.2, we show how many antecedents of an anaphoric pronoun are present in the data at a given distance d, expressed in number of sentences from the current one for original data, and in number of segments for split data. $d = 0$ means that both the pronoun and its antecedent are in the same sentence (or segment); $d = 1$ means that the antecedent is in the previous sentence (or segment), and so on. We show

statistics up to $d = 3$, the maximum context distance we experiment with. The absolute number of antecedents is normalized by the average length of a sentence or segment. The resulting bar plot shows that splitting sentences into segments makes pronominal antecedents denser in the set of context tokens the model is attending, which fosters the learning of contextual parameters. The same effect applies to the other discourse phenomena that require contextual disambiguation.

Coreferences - original data			
d	#tokens	Occurrences	
		All	Pronouns
0	21.01	67,864 (3230)	50,556 (2406)
1	42.02	68,703 (1635)	43,220 (1029)
2	63.03	35,780 (568)	21,234 (337)
3	84.04	25,533 (304)	14,284 (170)
Coreferences - split data			
d	#tokens	Occurrences	
		All	Pronouns
0	10.51	32,190 (3063)	24,328 (2315)
1	21.02	54,424 (2589)	37,966 (1806)
2	31.53	37,837 (1200)	23,732 (753)
3	42.04	22,529 (536)	14,035 (334)
Dependency trees			
Split dependency			Occurrences
subj or obj			41,065
complement			21,726
modifier			21,144
any			147,066

Table 3.1 – Occurrences of coreferential antecedents at a given distance d (in number of sentences) from a mention in the current sentence, in the En→Fr IWSLT17 training data. In brackets, the same figure is normalized by the average number of tokens the model has to attend to resolve the coreference (#tokens). At the bottom, the number of sentences for which at least one syntactic dependency is split into two segments when using the split data.

Coreferences - original data			
d	#tokens	Occurrences	
		All	Pronouns
0	8.32	36,628 (4402)	27,179 (3267)
1	16.64	60,204 (3618)	41,652 (2503)
2	24.96	26,397 (1058)	16,142 (647)
3	33.28	11,571 (348)	6,654 (200)
Coreferences - split data			
d	#tokens	Occurrences	
		All	Pronouns
0	4.16	13,322 (3202)	9,134 (2196)
1	8.32	46,227 (5556)	34,104 (4099)
2	12.48	33,566 (2690)	22,676 (1817)
3	16.64	18,961 (1139)	12,248 (736)

Table 3.2 – Occurrences of coreferential antecedents at a given distance d (in number of sentences) from a mention in the current sentence, in a sample of 1/10th of the En→Ru OpenSubstitles2018 data curated by Voita et al. (2019b). In brackets, the same figure is normalized by the average number of tokens the model has to attend to resolve the coreference (#tokens). At the bottom, the number of sentences for which at least one syntactic dependency is split into two segments when using the split data.

In Table 3.1, we provide details on the syntactic features and the impact of splitting (with *middle-split*) for En→Fr IWSLT17, while Table 3.2 shows the equivalent figures for a subset of the En→Ru OpenSubtitles2018 prepared by Voita et al. (2019b). The subset was built by randomly selecting 1/10th of the available documents. Figure 3.3 portrays a visual comparison of the two datasets. This complementary information confirms that the *middle-split* method effectively strengthens the contextual training signal and facilitates its exploitation by CANMT systems in different text domains.

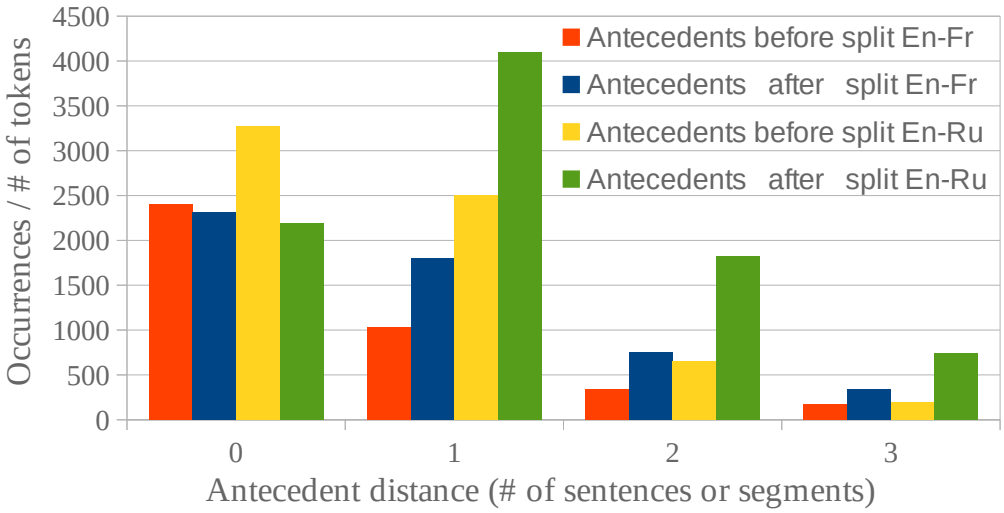

Figure 3.3 – En→Fr IWSLT17 and 1/10th of En→Ru OpenSubtitles2018: comparison of the number of antecedents of anaphoric pronouns at a given distance in terms of sentences or segments, normalized by the number of tokens that the model needs to attend for resolving the coreference. Since sentences are much shorter in the En-Ru corpus than En-Fr (8.32 vs. 21.02 tokens on average), the density of discourse phenomena within the sentence is much higher.

3.4 Experimental setup

3.4.1 Data

We conduct experiments for three language pairs, English→Russian, English→German, and English→French, on different domains. Following Kim et al. (2019), we pre-train sentence-level baselines on large sentence-level parallel data to make them as robust as possible. In particular, we employ data released by Voita et al. (2019b) for En→Ru (6.0M sentences from OpenSubtitles2018), data from the WMT17[1] news translation shared task for En→De (∼5.2M sentences), and data from WMT14[2] for En→Fr (∼35.8M sentences). We train the contextual parameters of context-aware models in two settings while freezing the rest of their parameters:

High resource. For En→Ru, it consists of the documents extracted from OpenSubtitles2018 and pre-processed by Voita et al. (2019b). For the other two language pairs, we build the training set by assembling (i) News-Commentary-v12 for En→De and News-Commentary-v9 for En→Fr; (ii) Europarl-v7 for En→De/Fr; (iii) TED talks subtitles

[1]http://www.statmt.org/wmt17/translation-task.html
[2]http://www.statmt.org/wmt14/translation-task.html

	En→Ru		En→De		En→Fr	
	Low Res	Hig Res	Low Res	Hig Res	Low Res	Hig Res
Sentence-level train	OpenSubs2018	OpenSubs2018	WMT17	WMT17	WMT14	WMT14
Context-aware train	1/10th of OpenSubs2018	OpenSubs2018	IWSLT17	News-v12 Europarl-v7 IWSLT17	IWSLT17	News-v9 Europarl-v7 IWSLT17
Fine-tuning	-	-	-	IWSLT17	-	IWSLT17
Test (BLEU)	OpenSubs2018	OpenSubs2018	IWSLT17	IWSLT17	IWSLT17	IWSLT17
Contrastive test	EllipsisVP	EllipsisVP	ContraPro	ContraPro	ContraPro	ContraPro

Table 3.3 – Summary of the datasets used at each stage of training and evaluation of the models.

released by IWSLT17 (Cettolo et al., 2012) for En→De/Fr.

Low resource. For En→Ru, it consists of a sample of 1/10th of the data from a random shuffle of the high-resource setting. For En→De/Fr, we use IWSLT17's TED talks alone.

We recap in Table 3.3 the datasets adopted at each stage of training and evaluation. The sentence-level training concerns the baselines, whose parameters are also used to initialize the encoder and decoder of the context-aware models (Θ_S). Concerning En→Ru, Voita et al. (2019b) released two datasets extracted from OpenSubtitles2018: a document-level dataset of 1.5M sentences with context and document boundaries (used for document-level training) and a sentence-level dataset of 6M sentences (used for sentence-level training), which includes the sentences of the document-level dataset. Since these data have already been pre-processed, we only apply Byte Pair Encoding (BPE) (Sennrich et al., 2016) with 32k merge operations jointly for English and Russian. For the other two language pairs, instead, we tokenize data with the Moses toolkit (Koehn et al., 2007), clean them by removing long sentences, and encode them with byte pair encoding, using 32k merge operations jointly for source and target languages.
While IWSLT provides document boundaries for TED subtitles, the WMT releases of New-Commentary and Europarl do not provide them. Therefore, a small fraction of sentences in the High Resource setting will be paired with incoherent context. However, we found the models to be robust against occasional incoherent context (see also Voita et al. (2018) and Müller et al. (2018)). In order to teach the models to translate headlines (the first line in a document), we need to have headlines in the training set. As such, we

Corpus	Tgt	Docs	Sents	Doc Length			Sent Length			Sent Length (BPE)		
				mean	std	max	mean	std	max	mean	std	max
Low	De	1.7k	0.2M	117.0	58.4	386	20.8	14.3	153	23.3	16.3	195
Low	Fr	1.9k	0.2M	118.6	56.7	390	21.0	14.3	153	23.5	16.3	202
Low	Ru	150k	0.6M	4.0	0.0	4	8.3	4.7	64	8.6	4.9	69
High	De	12.2k	2.3M	188.4	36.2	386	27.3	16.1	249	29.1	17.4	408
High	Fr	12.4k	2.3M	187.3	37.0	390	27.6	16.3	250	29.2	17.3	503
High	Ru	1.5M	6.0M	4.0	0.0	4	8.3	4.7	64	8.6	4.9	69
Both	De	62	5.4k	87.6	53.5	296	19.0	12.5	114	21.1	14.0	132
Both	Fr	66	5.8k	88.2	51.5	297	19.3	12.5	90	21.6	14.1	106
Both	Ru	10k	40k	4.0	0.0	4	8.2	4.8	50	8.5	5.0	58
Both	De	12	1.1k	90.0	29.2	151	19.3	12.7	102	21.6	14.3	116
Both	Fr	12	1.2k	100.8	28.5	156	19.8	13.2	89	22.2	14.9	105
Both	Ru	10k	40k	4.0	0.0	4	8.2	4.8	42	8.5	5.0	50

Table 3.4 – Statistics for the training (1st block), validation (2nd block) and test set (3rd block) after pre-processing, and after BPE tokenization. All figures refer to the English text (source side).

set artificial document boundaries in News-Commentary and Europarl every 200 sentences. Details on the datasets after pre-processing are reported in Table 3.4. In the case of En→De/Fr, baselines and context-aware models trained on high resources are also fine-tuned on the low-resource setting (IWSLT17) so that both high and low-resource settings can be bench-marked on the IWSLT17's test set 2015. Test sets 2011-2014 are used as development sets. For En→Ru, we use the validation and test sets provided by Voita et al. (2019b).

3.4.1.1 Evaluation

Besides evaluating average translation quality with BLEU (Papineni et al., 2002),[3] we employ three contrastive test sets for the evaluation of translation of discourse phenomena (described in details in Section 2.3.2.1):

En→Ru EllipsisVP: the subset on verb phrase ellipsis of the broader Voita's contrastive set (Voita et al., 2019b). The subset contains 500 examples of verb phrase ellipsis from OpenSubtitles2018. Each example contains multiple contrastive hypotheses to evaluate

[3]Moses' *multi-bleu-detok* (Koehn et al., 2007) for De/Fr, *multi-bleu* on lowercased Ru as Voita et al. (2019b).

the translation of the ellipsis. Source sentences contain an auxiliary verb (e.g., "do") and an omitted main verb, which can be imputed thanks to one of the three context sentences. The complete test suite released by Voita et al. (2019b) contains other subsets for evaluating other discourse phenomena. Still, we restrain our evaluation to verb phrase ellipsis because the other examples are conceived for systems using target-side context too.

En→De ContraPro (Müller et al., 2018) - A large-scale test set from OpenSubtitles2018 (Lison et al., 2018), that measures translation accuracy of the English anaphoric pronoun *it* into the corresponding German translations *er*, *sie* or *es*. Examples are balanced across the three pronoun classes (4,000 examples each). Each example requires identification of the pronominal antecedent, either in the source or target side, that can be found in the current sentence or any of the previous ones.

En→Fr ContraPro (Lopes et al., 2020) - A large-scale test set from OpenSubtitles2018, completely analogous to the previous one but focused on the translation of two English pronouns: *it* and *they*. It consists of 3,500 examples for each target pronoun type: *il* or *elle* for *it*, *ils* or *elles* for *they*.

We verify the statistical significance of the differences between models' accuracies with the paired McNemar test (McNemar (1947); see Section 2.3.4 for more details on statistical hypothesis testing).

3.4.2 Models

We experiment with a context-agnostic baseline and two multi-encoding models:

(i) *base*: A sentence-level baseline, following the *Transformer-base* by Vaswani et al. (2017).

(ii) *K2*: A context-aware multi-encoding architecture with *outside integration* (see Section 2.4.2) that encodes a single past source sentence as context.

(iii) *K4*: A context-aware multi-encoding architecture with *outside integration*, that encodes three past source sentences as context.[4]

It should be noted that we experiment with multi-encoding systems that exploit the source-side context only. Some works in the literature have found that target-side context

[4]Although the splitting does not increase the number of inter-segment phenomena for $d > 1$, it strengthens the signal by making it denser (see Section 3.3.2). Thus, *K4* and any wider-context model can profit from the proposed approach.

boosts the performance of multi-encoding systems (Bawden et al., 2018; Miculicich et al., 2018; Voita et al., 2019b; Maruf et al., 2019; Zheng et al., 2020). However, including target-side context requires adding contextual parameters that integrate it into the decoder. Besides adding complexity to the architecture, these parameters are as much affected by the *double challenge of sparsity* as the contextual parameters that integrate source-side context. Therefore, while having access to target-side context can facilitate the disambiguation of inter-sentential discourse phenomena, it can also represent a further learning complexity. In fact, critical works in the literature have targeted both source-side-only multi-encoding systems (Kim et al., 2019; Li et al., 2020) and systems including target context too (Lopes et al., 2020), showing that they also suffer from learning difficulties. Thus, the minimum viable experimental setting for the proposed approach is to apply it to source-side-only multi-encoding systems. This setting is free from the confounding factors that would derive from including target context and it allows to draw conclusions that can reasonably be transferred to target-side multi-encoding systems too.

For both *K2* and *K4*, sentence-level parameters θ_S follow the *Transformer-base* configuration (hidden size of 512, feed-forward size of 2048, 6 layers, 8 attention heads, total of 60.7M parameters), while contextual parameters θ_C follow a hierarchical architecture with source-side encoder proposed by Miculicich et al. (2018) (hidden size of 512, feed-forward size of 2048, 8 attention heads, total of 4.7M parameters). Context-aware models are trained following the *two-step strategy* described in Section 2.4.2. Sentence-level parameters θ_S of both *K2* and *K4* are initialized with *base* and frozen. This has the advantage of saving time and computation since only a small fraction of parameters (θ_C) is trained (4.7M over a total of 65.2M).

More details about the models' implementation and training are discussed in the Appendix A.1.

3.5 Results and analysis

3.5.1 Training contextual parameters is hard

In this section, we provide evidence about the difficulty of training contextual parameters on document-level data. In the second block of Table 3.5, below the results of the sentence-level baseline *base*, we report the performance of context-aware models trained on original document-level data, comparing low and high-resource settings. When trained on little resources, models display good BLEU on the test set, generally without any significant degradation with respect to *base*, or even with some improvements. However, such marginal

		En→Ru		En→De		En→Fr		
Model	Setting	BLEU	EllipsisVP	BLEU	ContraPro	BLEU	ContraPro	Hours
Concat2to1	Low Res	31.12	31.00	33.41	47.38	41.27	80.42	2.17
Concat2to1	High Res	29.92	62.6	33.05	59.49	40.99	85.57	35.32
Zhang2018	Low Res	n.a.	n.a.	31.03	42.60	40.95	59.00	n.a.
base	-	31.37	25.40	32.97	46.37	41.44	79.46	-
K2	Low Res	30.89	32.20	33.14	47.05	41.87	79.24	2.40
K4	Low Res	31.00	29.20	32.86	46.48	41.32	80.53	2.80
K2	High Res	31.15	44.00	33.16	57.75	41.49	84.32	19.10
K4	High Res	31.23	39.20	33.10	51.14	41.73	82.94	21.50
K2-d&r	Low Res	31.09	47.00*	33.44	60.21*	41.78	84.06	6.5
K4-d&r	Low Res	32.12	46.60*	33.36	56.22*	41.68	85.50*	6.8
K2-d&r	High Res	31.09	59.40*	32.82	**61.09***	41.81	84.17	32.8
K4-d&r	High Res	31.27	**60.40***	33.07	59.56*	41.91	**85.66***	33.1

Table 3.5 – BLEU score on testsets and accuracy (%) on contrastive sets. The last column reports the total context-aware training time spent on En→Fr, including the time for *d&r* pre-training. The symbol * denotes statistically significant ($p<0.01$) improvements w.r.t non-*d&r* counterparts (second block) and *base*.

fluctuations in BLEU are difficult to interpret, as they do not necessarily correspond to better or worse translations (Freitag et al., 2020). Accuracy on the contrastive test sets also increases only marginally over the baseline, if at all, for En→De/Fr. *K2* even shows a slight degradation of performance over the sentence-level baseline for En→Fr. These results highlight the struggle of contextual parameters to learn to exploit context for better translations, other than acting as mere regularizers, as it was suggested by Kim et al. (2019) and Li et al. (2020).

Instead, En→Ru models trained on low resources improve over *base*, in line with our expectations. In fact, the sentences belonging to the En→Ru Low Res setting are 2.5x shorter than those belonging to the En→Fr/De Low Res setting. This mitigates the *double challenge of sparsity* since useful contextual elements can be retrieved more easily in shorter sentences.

When passing from a low-resource setting to a high-resource setting, we measure substantial improvements in context-modeling capabilities across all language pairs. These results confirm the intuition discussed in Section 3.2: increasing the volume of data compensates for sparsity. Instead, BLEU improves by a few decimal points only in the high-resource setting, showing its inefficacy in measuring improvements in context-aware translation.

For the sake of benchmarking, we report in the first block the results obtained by two other source-side context-aware models trained on low resources following the same procedure.[5] *Concat2to1*[6] is a single-encoder approach that concatenates the previous context sentence to the current one and outputs the translation of the current sentence. *Zhang2018* is a multi-encoding model that looks at 2 previous sentences as context, proposed by Zhang et al. (2018).[7] *Concat2to1*'s performance on test suites are comparable to *K2/K4* on low resources, or slightly better since concatenation models are less affected by the problem of sparsity. In fact, they do not have to learn parameters that are specialized in contextualization. This advantage is better highlighted in the high-resource setting, where *Concat2to1* is stronger on the test suites (although BLEU lacks behind). *Zhang2018* performs very poorly, confirming the difficulty of multi-encoding models to learn contextualization on low resources without any help against the problem of sparsity.

3.5.2 Main results

In this section, we show that the proposed pre-training strategy is an effective solution to the double challenge of sparsity and an efficient one compared to simply increasing the training data. The third block of Table 3.5 reports the performance of models that have undergone *d&r* pre-training on the same document-level data as the models in the previous block, but where sentences were split into two segments following the *middle-split* method with $l_{min} = 7$. After *d&r* pre-training, models have been finetuned on the original, non-split data. The pre-training proves to be very effective, as all models belonging to the third block show substantial improvements in accuracy on the test suites, with the sole exception of *K2-d&r* on En→Fr High Res. The average gain is of **+10.79** accuracy points on Low Res, **+8.49** on High Res, showing that *d&r* brings substantial improvements even when data are abundant. Interestingly, gains are not uniformly distributed across language pairs and domains: **+17.20** on average for En→Ru, **+8.67** for En→De, **+3.09** for En→Fr. While context-aware translation measurements improve significantly, we keep measuring minor fluctuations in BLEU.

It is clear that a proper comparison between single and multi-encoding models cannot be made without proper training of the multi-encodings' contextual parameters, which targets the problem of sparsity. Here, *d&r* pre-training allows *K2/4* to achieve results on test suites comparable to *Concat2to1* (*K4* is consistently better), along with better BLEU scores (except for *K2* on german).[8] A comparison between -*d&r* models trained on Low

[5] We do not compare with target-side approaches as we experimented with source-side only.
[6] The implementation is our own.
[7] Results reported are by Lopes et al. (2020)
[8] A detailed comparison between single and multi-encoding models is beyond this chapter's scope.

	ContraPro	$d = 0$	$d = 1$	$d = 2$	$d = 3$	$d > 3$
base	46.37	**83.3**	32.4	44.8	48.9	71.9
K2	47.05	82.5	33.9	45.3	48.0	69.9
K4	46.48	82.4	32.8	45.0	48.9	71.7
K2-d&r	**60.21**	81.1	**56.5**	44.9	48.7	**73.3**
K4-d&r	56.22	81.7	46.8	**55.2**	**56.2**	72.4
Sample Size	12000	2400	7075	1510	573	442
Relative Size	100.0%	20.0%	59.0%	12.6%	4.8%	3.7%

Table 3.6 – Accuracy (%) by pronoun antecedent distance of Low Res models on ContraPro (En→De). The first column represents the weighted average, calculated based on the sample size of each group.

Res against models trained on High Res without *d&r* shows another quality of the *d&r* pre-training strategy: efficiency. In fact, the same context-aware models achieve superior performances with 1/10th of the document-level data and 1/3rd of the total training time (see time reported in the last column of Table 3.5, which includes pre-training time when present).

To strengthen the empirical evidence about the improved context-modeling capacity of the systems pre-trained with *d&r*, we present in the following sections an analysis of pronoun translation by antecedent distance and an ablation study in which we test models on ContraPro with inconsistent context.

3.5.2.1 Accuracy by antecedent distance

We investigate more in detail the performance of the proposed approach on pronoun translation disambiguation. We report in Table 3.6 the accuracy on En→De ContraPro, detailed by varying antecedent distance. We notice that all the improvements achieved by *-d&r* models are related to those pronouns whose antecedent is in the context ($d \geqslant 1$), which is in line with the expectations of context-aware models exploiting context for disambiguation. *K2-d&r* is very strong in translating pronouns with antecedent distance $d = 1$, surpassing *base* and *K2* baselines by 22+ points of accuracy. Similarly, *K4-d&r* surpasses baselines by a large margin on $0 \leqslant d \leqslant 3$, beating all the other models on $d = 2, 3$, as expected. We notice however that *K4-d&r* lacks behind *K2-d&r* on $d = 1$. On one side, this could be explained by the fact that *K2-d&r* is more specialized at modeling a single past sentence. On the other side, we also notice that the hierarchical

context-encoding architecture by Miculicich et al. (2018) of our multi-encoding systems does not encode context distance with any kind of embedding[9]. Hence, *K4-d&r* might perform worse on $d = 1$ than *K2-d&r* because it gives the same importance to further away context ($d = 2, 3$). Since pronouns with antecedent distance $d = 1$ are the most frequent in the test set, *K2-d&r* has the highest average result (reported in "Total"). To test this hypothesis, we perform an ablation study by adding different kinds of segment embeddings to *K4-d&r*, presented in the Appendix A.2. Unfortunately, we do not see major changes in performance when enabling the model to discriminate segment distances. Therefore, we posit that *K4* underperforms *K2* on $d = 1$ mostly because it is more affected by the challenge of sparsity. In fact, it has to spot relevant context among $3x$ more tokens, on average. This might be the reason why *K4* starts catching up *K2* when the training conditions are the most favorable to context-aware learning: with *d&r* pre-training plus high resources (see Table 3.5).

3.5.2.2 Ablation: shuffling context

Here we describe an ablation study that again suggests that the proposed approach pushes the models to exploit context more frequently. Table 3.7 shows the performance of models trained on Low Res when the evaluation is undertaken by randomly shuffling the context of every sentence with other sentences from the same dataset (c.f. Scherrer et al. (2019b)). In brackets, the delta w.r.t. the results with consistent context, presented in Table 3.5. A random context is inconsistent with the current sentence and thus misleading for a context-aware system. Indeed, *-d&r* models display a significant drop in accuracy when evaluated with inconsistent context, which confirms that they rely on context information to improve pronoun translations. However, their performance doesn't drop below baselines, which suggests that *d&r* doesn't make multi-encoding models over-reliant on context.

Instead, BLEU is not affected by shuffled context, showing once again that average translation quality metrics are ill-equipped to detect changes in context-aware translation.

3.5.3 Impact of the splitting method

Following Section 3.3.1, we study the impact of using a different splitting method other than *middle-split*. All the variants are applied to the En→De/Fr low-resource setting (IWSLT), with $l_{min} = 7$, and the *d&r* pre-trained models are evaluated on ContraPro. The *aligned-split* method is based on alignments learned with *fast_align* (Dyer et al., 2013). For the *synt-split* method, we retrieve intra-sentential pronominal coreferences with

[9]See Section 2.4.2 for a brief presentation of the techniques adopted in the literature.

	En→De		En→Fr	
Model	BLEU	ContraPro	BLEU	ContraPro
base	32.97 (+0.00)	46.37 (0.00)	41.44 (-0.00)	79.46 (0.00)
K2	33.06 (+0.06)	46.7 (-0.35)	41.75 (-0.12)	79.05 (-0.19)
K4	32.73 (-0.13)	46.21 (-0.27)	41.47 (+0.15)	79.24 (-1.29)
K2-d&r	33.1 (-0.34)	47.6 (-12.61)	41.64 (-0.14)	78.94 (-5.12)
K4-d&r	33.05 (-0.31)	47.96 (-8.26)	41.55 (-0.13)	79.05 (-6.45)

Table 3.7 – BLEU and accuracy results on ContraPro when the context provided to the model is inconsistent. In brackets, their changes w.r.t. the results achieved with consistent context presented in Table 3.5. All models are trained in the low-resource setting.

	En→De			
	Middle-split	Aligned-split	Synt-split	Multi-split
K2-d&r	60.21	+0.69*	-2.67*	-
K4-d&r	56.22	-1.38*	+1.33*	+1.13*
	En→Fr			
K2-d&r	84.06	+0.27	+0.15	-
K4-d&r	85.50	+0.20	+0.33**	-0.09

Table 3.8 – Comparison between the *middle-split* method and the other splitting methods (relative difference) on ContraPro. *: $p < 0.01$, **: $p < 0.05$.

CoreNLP (Manning et al., 2014), and we try to split them whenever possible. If there are multiple occurrences in the same sentence, we split as close to the middle as possible while attempting to break the maximum number of coreferences.[10] Finally, for the *multi-split* method, we split sentence-pairs in a half for $len(\mathbf{x}^i) \geqslant 7$, and also in three segments of identical size for $len(\mathbf{x}^i) \geqslant 15$. The performance differences between models pre-trained with *middle-split* and the other variants are reported in Table 3.8.

As we can see, splitting variants allow small improvements in 7 cases out of 10, although variations are marginal: the simple *middle-split* method seems to be close to optimal already. Multiple elements can explain this observation. Firstly, *middle-split* produces segment pairs that are already well aligned: most of the source and target segments are aligned except for one or two words. Having only a few misplaced words might act as

[10]More sophisticated *synt-split* methods could be devised, targeting other discourse phenomena, or several simultaneously, with different degrees of priority.

a regularization factor. Secondly, *middle-split* breaks a syntactic relation for the vast majority of sentences already, as explained in Section 3.3.1, which means that improvements achieved with syntactically driven splitting can only be marginal. Thirdly, splitting in more than one segment can be beneficial in some cases, because it allows to break more syntactic relations and increase the density of training signal, but it also increases the risk of misalignment between source and target and might make the task too hard. Finally, tools like *fast_align* and CoreNLP are characterized by a certain language-dependent error rate, which affects the performance of the splitting methods. In conclusion, *d&r* pre-training with *middle-split* seems to be the most convenient alternative for most use cases because of its efficacy, its simplicity, and its language independence. Nonetheless, one variant (or a combination of them!) could be more convenient for some specific applications that strive for optimal performance.

3.5.4 On the scope of middle-split

Even though *middle-split* relies on the syntactic similarity between the source and the target languages, we argue that this condition is met by a large number of language pairs, in the order of millions. In fact, there are around 4,000 written languages in the world (Eberhard et al., 2021), and most of them can be grouped into a few types with similar word orders, as shown by the ample literature on word order typologies (Tomlin, 2014; Dryer and Haspelmath, 2013).

The most significant structural feature in a language is the *constituent order*, concerning the relative order of subject (S), object (O), and verb (V) in a clause. There are seven possible language types concerning the constituent order (Dryer, 2013b): SOV, SVO, VSO, VOS, OVS, OSV, and NDO (non-dominant order). Tomlin (2014) estimates that more than 40% of the world languages belong to the SOV type (languages adopting the SOV order), another 40% belong to the SVO type, while almost 10% of languages adopt VSO order. The other types are rarer. In this chapter, we have shown that the middle-split method is beneficial both in the case of language pairs of the same type that deploy the same constituent order, like En-Fr/Ru, which all adopt SVO order, as well as for languages that belong to different types, as for En-De, where English is SVO and German is NDO, deploying both SOV and SVO according to the use cases (Dryer, 2013b).

Similar observations also apply when we look at other word order categories. For instance, when looking at the order of modifiers or adverbials, languages can be clustered into a few types. Here again, the top two types represent the vast majority of languages (Dryer, 2013a,c). Therefore, we believe that our method could be beneficial for millions of language pairs, including many low-resource languages belonging not only to the same word order

types but also to slightly different ones (as in the case of SOV and SVO).

3.6 Conclusions

3.6.1 Takeaways

Multi-encoding models are a broad family of context-aware NMT models. In this work we have discussed the difficulty of training contextual parameters due to the sparsity of the tokens in need of context, and their relevant context. We have proposed a pre-training approach called *divide and rule*, based on splitting the training sentences, with four variants. After having analyzed the implications of splitting on discourse phenomena, we have shown empirically that *d&r* pre-training allows to learn contextual parameters better and faster than by simply adding training data. We have shown that the simplest and language-independent splitting variant, *middle-split*, is a strong baseline that can be easily applied for pre-training any multi-encoding NMT model in settings with weak (like En-Fr/Ru) or moderate (like En-De) word order divergence. Arguably, to millions of language pairs.

3.6.2 Limitations and future works

The main limitation of our experimental section is the lack of experiments with multi-encoding systems that integrate target-side context on top of source-side context. We have explained in Section 3.4.2 the reasons for this choice. Having measured substantial improvements with the application of the proposed approach to multi-encoding systems that handles different lengths of source-side context, on a varied set of training conditions (languages, domains, volumes of data, structures of documents), we can reasonably expect improvements for multi-encoding systems handling target-side context too. However, an empirical quantification of the gains achievable in that setting would be valuable. Likewise, it would be valuable to conduct experiments with multi-encoding architectures other than (Miculicich et al., 2018)'s.

Our experiments are limited to language pairs with weak or moderate word order divergence because of the unavailability of discourse-targeted test suites for language pairs with a substantial word order divergence. As we have repeatedly shown in the paper, targeted evaluation is essential for evaluating context-aware NMT systems, which cannot be undertaken solely with average translation quality metrics (e.g., BLEU). One possible solution for employing *d&r* pre-training on language pairs with strong syntactic divergence

could be a preliminary application of word ordering tools before splitting. This solution can be easily implemented for low-resource languages too, and it has been proven very effective for improving machine translation in these settings (Zhou et al., 2019).

Recently, (Fernandes et al., 2021) have proposed a training strategy for context-aware models that pushes them to exploit context more consistently. The idea is to randomly mask tokens in the current sentence in order to encourage the model to use extra-sentential information to compensate for them. It would be valuable to compare *d&r* with this approach and to analyze the possibility to combine the two approaches. This would require a re-implementation of (Fernandes et al., 2021)'s approach, since the results reported in their published paper refer to a target-context-only multi-encoding system and to concatenation systems. Thus, they are not directly comparable with our work.

Finally, future works could also explore the impact of *d&r* pre-training on other tasks, such as next-sentence prediction or document-level coreference resolution, which also require models to be context-aware.

In the next chapter, we will study the other family of approaches to CANMT: concatenation approaches. Similarly to the work conducted on multi-encoding systems, we will propose solutions to tackle some key limitations of concatenation and shed light on some of its aspects through the analysis of the proposed approaches.

Chapter 4

Focused concatenation

Most of the contributions presented in this chapter are published in:

> Lupo, L., Dinarelli, M. and Besacier, L. (2022). Focused Concatenation for Context-Aware Neural Machine Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, December 7–8, 2022. Association for Computational Linguistics.

Contents

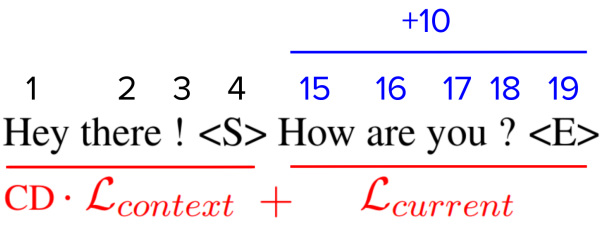

Figure 4.1 – Example of the proposed approach applied over a window of 2 sentences, with context discount CD and segment-shifted positions by a factor of 10.

4.1 Introduction

As discussed in Section 2.4.1, concatenation approaches have the advantage of treating the task of CANMT in the same way as context-agnostic NMT, which eases learning because the learnable parameters responsible for extra-sentential contextualization are the same that undertake intra-sentential contextualization. Indeed, as we have seen in the previous chapter, learning the parameters responsible for extra-sentential contextualization in multi-encoding approaches (θ_C) has been shown to be challenging because the training signal is sparse and the task of retrieving useful context elements is difficult. Despite its simplicity, the concatenation approach has been shown to achieve competitive or superior performance than more sophisticated, multi-encoding systems (Lopes et al., 2020; Ma et al., 2021b).

Nonetheless, encoding current and context sentences together comes at a cost, as discussed in Section 2.4.4. Transformer-based NMT systems struggle to learn locality properties (Rizzi, 2013) of both the language itself and the source-target alignment when the input sequence grows in length, as in the case of concatenation. Unsurprisingly, the presence of context makes learning harder for concatenation models by distracting attention. Moreover, we know from Chapter 3 that NMT systems only require context for a sparse set of inter-sentential discourse phenomena. Therefore, it is desirable to make concatenation models more focused on local linguistic phenomena to improve performance. Recent works (Zhang et al., 2020a; Bao et al., 2021) have demonstrated that a viable solution is the introduction of strong inductive biases on locality in the NMT architecture, such as the partial masking of context (see Section 2.4.4 for more details). Based on these premises, we propose an improved concatenation approach to CANMT that is more focused on the translation of the current sentence by means of two simple, parameter-free solutions:

- context discounting: a simple modification of the NMT loss that improves context-aware translation of a sentence by making the model less distracted by its concatenated context;

- Segment-shifted positions: a simple, parameter-free modification of position embeddings that facilitates the achievement of the context-discounted objective by supporting the learning of locality properties in the document translation task.

We support our solutions with extensive experiments, analysis, and benchmarking.

4.2 Proposed approach

As discussed in Section 2.4.1, a typical strategy to train a concatenation approach and generate translations is using sliding windows (SlidingKtoK). For the sake of clarity, we outline this approach here once again. The model decodes the translation $\boldsymbol{y}_K^j$ of a source window $\boldsymbol{x}_K^j$, formed by K consecutive sentences belonging to the same document: the current (jth) sentence and $K-1$ sentences concatenated as source-side context. A special token `<S>` is introduced to mark sentence boundaries in the concatenation:

$$\boldsymbol{x}_K^j = \boldsymbol{x}^{j-K+1}{}_{\texttt{<S>}}\boldsymbol{x}^{j-K+2}{}_{\texttt{<S>}}\ldots{}_{\texttt{<S>}}\boldsymbol{x}^{j-1}{}_{\texttt{<S>}}\boldsymbol{x}^{j}{}_{\texttt{<E>}},$$
$$\boldsymbol{y}_K^j = \boldsymbol{y}^{j-K+1}{}_{\texttt{<S>}}\boldsymbol{y}^{j-K+2}{}_{\texttt{<S>}}\ldots{}_{\texttt{<S>}}\boldsymbol{y}^{j-1}{}_{\texttt{<S>}}\boldsymbol{y}^{j}{}_{\texttt{<E>}}.$$

Both past and future contexts can be concatenated to the current pair $\boldsymbol{x}^j, \boldsymbol{y}^j$, although in the above equations we consider the past context only $\boldsymbol{x}^{j-K<i<j}, \boldsymbol{y}^{j-K<i<j}$, for simplicity. At training time, the loss is calculated over the whole output $\boldsymbol{y}_K^j$. At inference time, the translation of the entire sequence $\boldsymbol{x}_K^j$ is generated, but only the translation $\boldsymbol{y}^j$ of the current sentence is eventually kept. In contrast, the translation of the context is discarded. Then, the window is slid by one position forward to repeat the process for the $\boldsymbol{x}^{j+1}, \boldsymbol{y}^{j+1}$ sentence pair and its context. Concatenation approaches are trained by optimizing the same objective function as standard NMT over a window of sentences, i.e., Equation 2.25:

$$\mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) = \sum_{t=1}^{|\boldsymbol{y}_K^j|} \log P(y_{K,t}^j | \boldsymbol{y}_{K,<t}^j, \boldsymbol{x}_K^j), \tag{4.1}$$

so that the likelihood of the current target sentence is conditioned on the source and target context.

4.2.1 Context discounting

The above objective function does not factor in the fact that we only care about the translation of the current sentence, $\boldsymbol{y}^j$, because the translation of the context will be

discarded during inference. Moreover, in practical terms, we only need the translation of the context for disambiguating relatively sparse inter-sentential discourse phenomena that are ambiguous at sentence level (Voita et al. (2019b), Chapter 3). Hence, we propose to encourage the model to focus on the translation of the current sentence $\boldsymbol{x}^j$ by applying a discount $0 \leqslant \text{CD} < 1$ to the loss generated by context tokens:

$$\begin{aligned} \mathcal{L}_{\text{CD}}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) &= \text{CD}{\cdot}\mathcal{L}_{context} + \mathcal{L}_{current} \\ &= \text{CD}{\cdot}\mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}_{K-1}^{j-1}) + \mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}^j). \end{aligned} \tag{4.2}$$

This is equivalent to considering a SlidingKtoK concatenation approach as the result of a multi-task sequence-to-sequence setting (Luong et al., 2016), where a SlidingKto1 model performs the *reference task* of translating the current sentence given a concatenation of its source with K-1 context sentences, while the translation of the context sentences is added as a secondary, complementary task. The reference task is assigned a more significant weight than the secondary task in the multi-task composite loss. As we will see in Section 4.3.5.2, this simple modification of the loss allows the model to learn a self-attentive mechanism that is less distracted by noisy context information, thus achieving net improvements in the translation of inter-sentential discourse phenomena occurring in the current sentence (Section 4.3.3), and helping concatenation systems to generalize to broader context after training (Section 4.3.5.4).

4.2.2 Segment-shifted positions

Context discounting pushes the model to discriminate between the current sentence and the context. Such discrimination can be undertaken by cross-referencing the information provided by two elements: sentence separation tokens `<S>` and sinusoidal position encodings, as defined in (Vaswani et al., 2017). To facilitate this task, we propose to provide the model with extra information about sentence boundaries and their relative distance. (Devlin et al., 2019) achieve this goal by adding segment embeddings to every token representation in input to the model, on top of token and position embeddings, such that every segment embedding represents the sentence position in the window of sentences. We propose an alternative solution that does not require any extra learnable parameter or memory allocation: segment-shifted positions. As shown in Figure 4.1, we apply a constant shift after every separation token `<S>` so that the resulting token position is equal to its original position plus a total shift depending on the chosen constant *shift* and the index $k = 1, 2, ..., K$ of the sentence the token belongs to: $t' = t + k * shift$. As a result, the position distance between tokens belonging to different sentences increases. For example, the distance between the first token of the current sentence and the last

Src	Tgt	Docs	Sents	Doc Length			Sent Length			Sent Length (BPE)		
				mean	std	max	mean	std	max	mean	std	max
En	Ru	1.5M	6.0M	4.0	0.0	4	8.3	4.7	64	8.6	4.9	69
En	De	1.7k	0.2M	117.0	58.4	386	20.8	14.3	153	23.3	16.3	195
En	Ru	10k	40k	4.0	0.0	4	8.2	4.8	50	8.5	5.0	58
En	De	62	5.4k	87.6	53.5	296	19.0	12.5	114	21.1	14.0	132
En	Ru	10k	40k	4.0	0.0	4	8.2	4.8	42	8.5	5.0	50
En	De	12	1.1k	90.0	29.2	151	19.3	12.7	102	21.6	14.3	116

Table 4.1 – Statistics for the training set (1st block), development set (2nd block) and test set (3rd block) after pre-processing, and after BPE tokenization. All figures refer to the English text.

token of the preceding context sentence increases from 1 to 1 + *shift*. By increasing the distance between sinusoidal position embeddings[1] of tokens belonging to different sentences, their dot product, which is at the core of the attention mechanism, becomes smaller (Figure 2.3), possibly resulting in smaller attention weights. In other words, the resulting distribution of attention weights could become more localized, as demonstrated by the empirical analysis reported in Section 4.3.6.1. In Section 4.3.3, we present the impact of segment-shifted positions on performance. In Chapter 5, we will also study their impact on non-context-discounted concatenation models, and compare them with a bunch of segment embedding variants.

4.3 Experiments

4.3.1 Setup

We experiment with three models:

- *base*: a context-agnostic baseline following *Transformer-base* (Vaswani et al., 2017) (see Section 2.1.4.2);
- *s4to1*: short for Sliding4to1, a context-aware baseline, consisting of a concatenation approach with the same architecture as *base*, but that processes sliding windows of

[1]Positions can be shifted by segment also in the case of learned position embeddings, both absolute and relative. We leave such experiments to future works.

4 concatenated sentences as the source, and it translates the 4th sentence into the target language (see Section 2.4.1);

- *s4to4*: short for Sliding4to4, a context-aware concatenation approach with the same architecture as *base*, but that processes sliding windows of 4 concatenated sentences as the source, and it decodes the whole window into the target language (see Section 2.4.1). We will study the impact of context discounting and segment-shifted positions on this architecture.

Models are trained and evaluated on two language pairs covering two different domains. For En→Ru, we adopt the document-level corpus released by Voita et al. (2019b) and based on OpenSubtitles2018, comprising training, development, and test set. For En→De, we train models on the TED talks subtitles released by IWSLT17 (Cettolo et al., 2012) and test them on the IWSLT17's test set 2015, while test-sets 2011-2014 are used for development, following prior works in the literature. Detailed figures about the data sets are reported in Table 4.1.
Interestingly, the document structure differs greatly between the two language pairs. The En→Ru data set is comprised of very short documents, each consisting of four sentences. As a consequence, the sequences processed by the sliding window model are concatenations of 1, 2, 3, or 4 consecutive sentences, in equal volume.[2] Instead, the En→De dataset is comprised of long documents of dozens of sentences, and therefore the majority of training and test sequences processed by Sliding4to4 consists of 4-sentence-long windows.

Besides evaluating average translation quality with BLEU[3] (Papineni et al., 2002) and COMET[4] (Rei et al., 2020), we employ two contrastive test suites to evaluate the translation of inter-sentential discourse phenomena. For En→Ru, we adopt Voita et al. (2019b)'s test suite for evaluating deixis, lexical cohesion, verb-phrase ellipsis, and inflection ellipsis (details in Section 2.3.2.1). This test suite contains also a development set with examples of deixis and lexical cohesion, which we adopted for a preliminary analysis of context discounting. For En→De, we evaluate models on ambiguous pronoun translation with ContraPro (Müller et al., 2018), a large contrastive set of ambiguous pronouns whose antecedents belong to context (see Section 2.3.2.1). To validate the improvements achieved by our approaches on the test sets, we perform statistical hypothesis tests following the

[2]The 1st sentence of the document is translated without past context (because it doesn't exist), then the 2nd sentence is translated using the 1st as context, then the 3rd using the 1st and 2nd as context, and finally the 4th sentence is translated, with the remaining 3 sentences concatenated as past context.

[3]Moses' *multi-bleu-detok* (Koehn et al., 2007) for De, *multi-bleu* for lowercased Ru as Voita et al. (2019b).

[4]Default model: wmt20-comet-da.

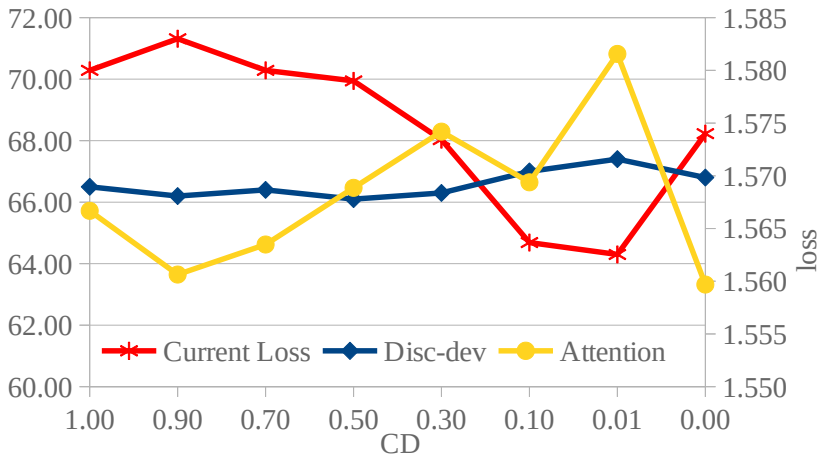

Figure 4.2 – Evaluation of En→Ru s4to4 trained with various levels of context discounting, ranging from 1 to 0. We plot the best *current loss* obtained by each model on the development set (red) and its average accuracy on the development portion of Voita's contrastive set on discourse phenomena (blue). In yellow is the average portion of attention that is focused on the current sentence (see Section 4.3.5.2).

methodology presented in Section 2.3.4. Appendix B.1 and B.1.1 report more details on the experimental setup and the statistical significance testing, respectively.

4.3.2 Preliminary analysis

As a preliminary analysis, we evaluate the impact of various values of context discounting on the performance of concatenation approaches with sliding windows, in order to choose one value for all the subsequent experiments. We train s4to4 models with context discounts ranging from 1 (no context discounting) to 0 (context loss is completely ignored): $\texttt{CD} = 1.0, 0.9, 0.7, 0.5, 0.3, 0.1, 0.01, 0$. We evaluate these models on the En→Ru development set by means of their average loss calculated over the current target sentence (*current loss*, i.e. $\mathcal{L}_{current}$) and the average accuracy on the disambiguation of discourse phenomena. The results are plotted in Figure 4.2. We find out that the stronger the context discounting, the better the performance, with an improving trend from $\texttt{CD} = 1$ to $\texttt{CD} = 0.01$. Performance drops on the extreme case of $\texttt{CD} = 0$, likely because too much training signal is lost in this situation. Therefore, we set $\texttt{CD} = 0.01$ for all our following experiments. In Section 4.3.5.3, we expand this analysis with *ex-poste* results, confirming that a strong context discounting translates into improved performance.

			En→Ru				
System	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	BLEU	COMET
base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
s4to1	50.00	45.87	57.60	71.40	51.66	32.64	0.322
s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
s4to4 + CD	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
s4to4 + shift + CD	85.76	**48.33***	**81.40**	**80.4***	**73.56***	32.45	0.334*
			En→De				
	d=1	d=2	d=3	d>3	ContraPro	BLEU	COMET
base	32.89	43.97	47.99	70.58	37.27	29.63	0.546
s4to1	36.90	46.55	49.38	69.68	40.67	29.28	0.526
s4to4	68.89	74.96	79.58	**87.78**	71.35	29.48	0.536
s4to4 + CD	**72.86***	75.96	80.10	84.38	74.31*	29.32	0.522
s4to4 + shift + CD	72.56*	**77.15***	**80.27**	86.65	**74.39***	29.20	0.528

Table 4.2 – Accuracy on the contrastive sets for the evaluation of discourse phenomena; BLEU and COMET scores on the corresponding test sets. The accuracy on discourse phenomena is detailed on the left with the accuracy on each subset. For Voita, each subset corresponds to a specific discourse phenomenon. For ContraPro, each subset contains examples of anaphoric pronouns with antecedents at a specific distance $d \in [1, 2, ...]$ (in number of sentences). The symbol * denotes statistically significant ($p < 0.05$) improvements w.r.t. baselines (base, s4to1, s4to4).

4.3.3 Main results

Table 4.2 illustrates the main evaluation results measured in terms of accuracy on contrastive test sets (Voita's and ContraPro), BLEU, and COMET, for the En→Ru and En→De language pairs. We first observe that s4to4 is a strong context-aware baseline as it improves accuracy on contrastive sets by a large margin compared to the context-agnostic *base* and the context-aware s4to1. This is in line with the findings of previous works (Voita et al., 2019b; Zhang et al., 2020a; Lopes et al., 2020).

As measured by BLEU, the average translation quality is virtually the same for all models. Indeed, our primary focus is on the contrastive evaluation of discourse translation since average translation quality metrics like BLEU have been repeatedly shown to be scarcely sensitive to improvements in CANMT (Hardmeier, 2012). Learned average translation quality metrics like COMET might be more sensitive to inter-sentential discourse phenomena when applied at the document level, as we do. However, COMET differences are also

System	base	s4to4	s4to4+CD	s4to4+shift+CD
Deixis	50.00 – 50.00	84.37 – 86.13	86.42 – 87.69	85.98 – 88.06
Lex co.	45.87 – 45.87	46.04 – 46.32	46.30 – 47.12	46.15 – 47.80
Ell. inf	51.59 – 53.01	76.71 – 79.56	79.00 – 81.94	79.38 – 81.29
Ell. vp	26.41 – 28.72	71.77 – 73.90	76.32 – 78.15	75.55 – 79.45
Voita	46.61 – 46.89	71.12 – 72.04	72.98 – 73.64	73.10 – 73.68
BLEU	31.96 – 32.07	32.29 – 32.50	32.38 – 32.52	32.34 – 32.50

Table 4.3 – 95% confidence intervals for the mean accuracy on the contrastive set (Voita, %) and the mean BLEU score on the test set. The intervals are based on 6 training and evaluation runs for each model, with a different random seed at each run.

negligible. On En→Ru, all models perform on par according to statistical significance tests, with the sole exception of s4to4 + shift + CD. On En→De, our approaches perform slightly worse than baselines in terms of COMET, but again by a small margin.
Instead, we remark relevant performance improvements when evaluating the accuracy of inter-sentential discourse phenomena. Adding a 0.01 context discounting (+ CD) improves the accuracy on all of the four discourse phenomena under evaluation in En→Ru, and for all distances of pronoun's antecedents in En→De, with the sole exception of $d > 3$, proving to be an effective solution. Adding segment-shifted positions further improves performance for three discourse phenomena out of four and for pronouns with antecedents at distances $d = 1, 2$, showing that sliding windows systems often benefit from enhanced sentence position information to achieve the discounted CANMT objective. For both language pairs, we adopt a segment-shifting equal to the average sentence length, calculated over the entire training corpus, i.e., +8 positions for En→Ru and +21 positions for En→De. Experiments with other shifting values are reported in Section 4.3.6.2.

In order to strengthen the evidence on the significance of the improvements achieved by our solutions, we trained each En→Ru model (except s4to1) six times, each time with a different random seed[5] for the initialization of the learnable parameters and the shuffling of the training data (c.f. section 2.3.4). We present in Table 4.3 the 95% confidence level interval for the mean performance achieved by each model on the test sets. These intervals confirm that context discounting consistently improves the performance of the s4to4 model on the contrastive set. Instead, for segment-shifted positions, the conclusion is not so clear-cut. The s4to4+shift+CD system has a confidence interval that is slightly higher than s4to4+CD's, but with ample overlapping with it.

[5]The selected seeds are $\{0, 12, 54, 345, 876, 6789\}$.

System	Voita	BLEU
s2to2	59.10	32.73
s2to2 + CD	60.28*	32.69
s2to2 + shift + CD	**60.54***	32.41
s3to3	65.58	32.34
s3to3 + CD	**67.02***	32.42
s3to3 + shift + CD	66.98*	32.45

Table 4.4 – Accuracy on the En→Ru contrastive set for the evaluation of discourse phenomena (Voita, %), and BLEU score on the test set. The symbol * denotes statistically significant ($p < 0.05$) improvements w.r.t. s2to2/s3to3. Our approach can be applied effectively to different concatenation windows.

System	Deixis	Lex co.	Ell. inf	Ell. vp	Voita
Chen et al. (2021)	62.30	47.90	64.90	36.00	55.61
Sun et al. (2022)	64.70	46.30	65.90	53.00	58.13
Zheng et al. (2020)	61.30	58.10	72.20	80.00	63.30
Kang et al. (2020)	79.20	62.00	71.80	80.80	73.46
Zhang et al. (2020a)	**91.00**	46.90	78.20	**82.20**	**75.61**
s4to4 + shift + CD	85.76	**48.33**	**81.40**	80.40	73.56

Table 4.5 – Benchmarking on En→Ru: accuracy on the contrastive sets for the evaluation of discourse phenomena (Voita, %).

As a further experiment, we apply our solutions to En→Ru concatenation models trained with concatenated windows shorter than four sentences.[6] The results presented in Table 4.4 show that context discounting is effective for s2to2 and s3to3 too, while adding segment-shifted positions only helps s2to2+CD. As in the case of s4to4, BLEU only displays minor fluctuations.

System	d=1	d=2	d=3	d>3	ContraPro
Maruf et al. (2019)	34.70	46.40	51.10	70.10	39.15
Voita et al. (2018)	39.00	48.00	54.00	66.00	42.55
Stojanovski and Fraser (2019)	53.00	46.00	50.00	71.00	52.55
Lupo et al. (2022a)	56.50	44.90	48.70	73.30	54.98
Müller et al. (2018)	58.00	55.00	55.00	75.00	58.13
s4to4 + shift + `CD`	**72.56**	**77.15**	**80.27**	**86.65**	**74.39**

Table 4.6 – Benchmarking on En→De: accuracy on the contrastive sets for the evaluation of pronominal anaphora (ContraPro, %).

4.3.4 A comparison with the literature

For a wider contextualization of our results, we compare in Tables 4.5 and 4.6 our best systems with other CANMT systems from the literature. For the En→Ru language pair (Table 4.5), we compare with all the systems from the literature that were trained and evaluated under the same experimental conditions as ours, to the best of our knowledge. In particular, we report the results by Chen et al. (2021), Sun et al. (2022)'s *MR Doc2Doc*, Zheng et al. (2020), Kang et al. (2020)'s *CADec + DCS-pf* and Zhang et al. (2020a). All of them are sophisticated CANMT systems that add extra trainable parameters to the Transformer architecture. Despite being the simplest and the only parameter-free approach, our method outperforms all the others on lexical cohesion and noun phrase inflection based on elided context. At the same time, it is only second to Zhang et al. (2020a) on deixis and verb-phrase ellipsis. BLEU scores were not available for comparison on the same test set, except for Zhang et al. (2020a), which scored 31.84 BLEU points against the 32.45 BLEU points of our method.

For the En→De language pair (Table 4.6), we compare to the works in the literature that adopts Müller et al. (2018)'s test set and provide details about their accuracy on pronouns with antecedents at $d > 1$. In particular: Maruf et al. (2019)'s best offline system, Stojanovski and Fraser (2019)'s *pron-25→pron-0**, Lupo et al. (2022a)'s *K1-d&r*, Müller et al. (2018)'s *s-hier-to-2.tied* and their evaluation of Voita et al. (2018)'s architecture.[7]

[6]We cannot evaluate with more sentences because 4 is the maximum size of documents in the contrastive set specialized on discourse phenomena Voita et al. (2019b).

[7]Whenever the cited works present and evaluate multiple systems, we compare to the best performing one. To the best of our knowledge, we are including all the relevant works available in the literature. BLEU scores are not compared because, besides using different training data, the cited works don't adopt the same test set either, with the sole exception of (Lupo et al., 2022a).

All of these works but Maruf et al. (2019) adopt the much larger WMT17[8] dataset for training. Despite this advantage, our system outperforms each of them on all the discourse phenomena under evaluation by a large margin.

Notably, from this comparison, it might seem that our approach is proposed in opposition to those used as benchmarks. Instead, it can be complementary to many of these approaches, such as (Zhang et al., 2020a)'s, hopefully in a synergistic way. We encourage future research to investigate this possibility.

4.3.5 Analysis of context discounting

In this section, we analyze the proposed context discounting method to widen our understanding of it and draw lessons on the concatenation approach more generally. To this aim, we undertake a threefold analysis of the context-discounted training objective, investigating its impact on training, on the learned attention function, and on the robustness of the learned CANMT system with respect to conditions unseen during training.

4.3.5.1 Loss distribution

In this section, we analyze the impact of context discounting on the ability of the model to predict the translation of the current sentence. In Figure 4.3 we plotted the evolution along training epochs of the loss calculated on the current target sentence (*current loss*) for the En→Ru language pair. The plot shows that context discounting enables better predictions of the current correction and that substantial discounting works best. This evidence empirically supports our idea of context discounting as a solution to improve model performance on the current sentence.

Figure 4.4, instead, represents the ratio between the *current loss* and the average loss-per-sentence calculated on the context sentences belonging to the same sliding window. Interestingly, predictions are improved on the current sentence (Figure 4.3) partially as a result of a trade-off with context quality (Figure 4.4). In fact, the current/context loss ratio of context-discounted models increases along training even when the *current loss* is decreasing, indicating that, at the beginning of training, context discounting pushes the model to only care about predictions of the tokens belonging to the current sentence. Still, later it allows for good predictions of the context too. Such behavior is in line with the intuition that a good translation of the current sentence, even if strongly prioritized, also requires a good translation of the context. Otherwise, it is not possible to systematically

[8]http://www.statmt.org/wmt17/translation-task.html

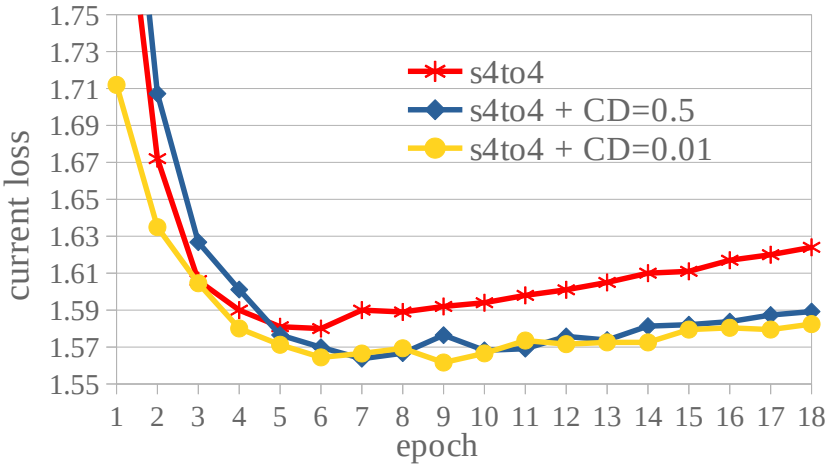

Figure 4.3 – Context discounting enables better predictions of the current sentence (lower validation loss). Language pair: En→Ru.

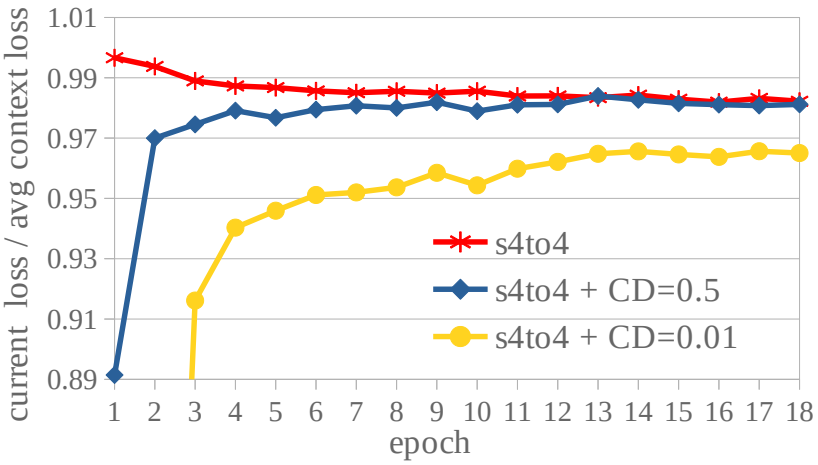

Figure 4.4 – Context discounting encourages the model to improve predictions of the current sentence relative to the context (lower current/context validation loss ratio). Language pair: En→Ru.

CD	Loss	Voita^{dev}	% Attn	Entropy
1.00	1.580	66.50	65.72	2.218
0.90	1.583	66.20	63.65	2.314
0.70	1.580	66.40	64.62	2.289
0.50	1.579	66.10	66.47	2.248
0.30	1.573	66.30	68.29	2.199
0.10	1.564	67.00	66.65	2.199
0.01	**1.563**	**67.40**	**70.82**	**2.109**
0.00	1.574	66.80	63.32	2.280

Table 4.7 – Numerical values corresponding to Figure 4.2, plus the corresponding average entropy of self-attention weights. A strong context discount of 0.01 results in the best *current loss* and average accuracy on the contrastive development and test set on discourse phenomena. *Current queries* attend more consistently to current tokens in self-attention (higher % Attn). The distribution of self-attention weights is generally more focused (lower entropy).

solve the translation ambiguities referring to context.

Equivalent insights can be drawn from the same analyzes on the En→De language pair, visualized in Figures 4.5 and 4.6. Another consequence of context discounting becomes evident in this case: the stronger the context discounting, the longer training takes. Unsurprisingly, weakening the training signal coming from a portion of the processed sequence (i.e., the context) slows down the learning process.

4.3.5.2 Attention distribution

In this section, we show some empirical evidence in favor of our intuition that context discounting improves performance by helping the self-attentive mechanism to be more focused on the current sentence (less distracted by context). We analyzed the distribution of self-attention weights generated when the queries are tokens belonging to the current sentence (*current queries*) and how it is impacted by context discounting. Figure 4.2 clearly shows that context discounting impacts the distribution of self-attention weights by skewing it towards the current sentence: a higher percentage of the total attention from *current queries* is directed towards tokens belonging to the (same) current sentence. As expected, the higher the context discounting, the higher the portion of attention weights not dispersed toward context. This is also reflected by the average entropy of the distribution of self-attention weights, which is minimal in the case of $CD = 0.01$, as reported in

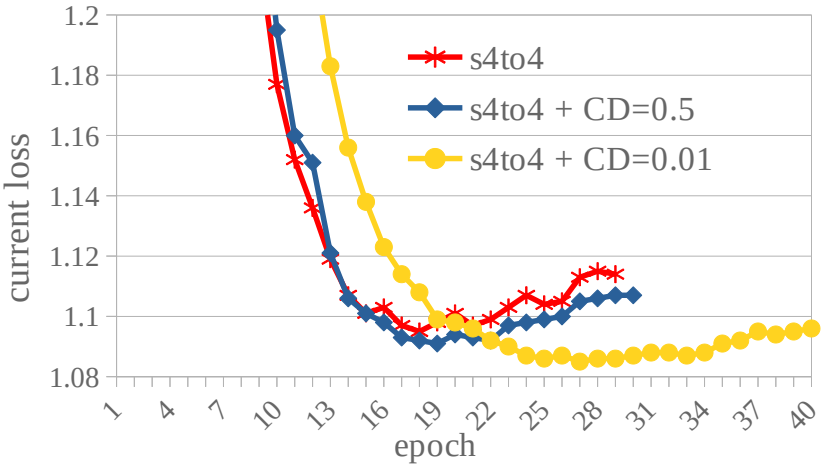

Figure 4.5 – Context discounting enables better predictions of the current sentence (lower validation loss). Language pair: En→De.

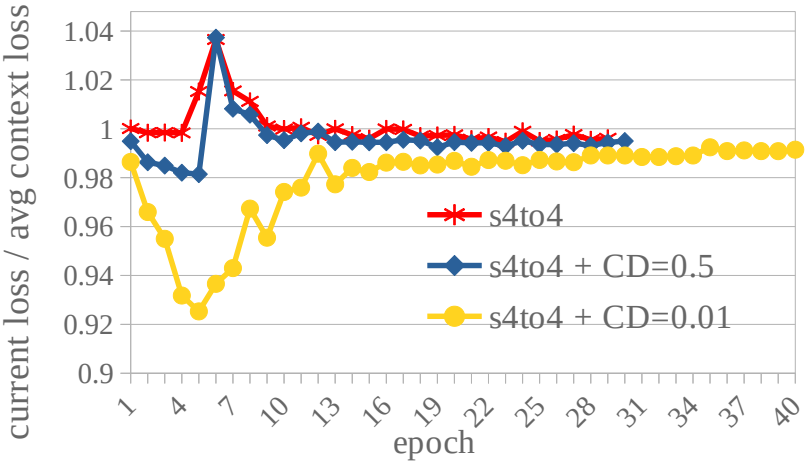

Figure 4.6 – Context discounting encourages the model to improve predictions of the current sentence relative to the context (lower current/context validation loss ratio). Language pair: En→De.

		En→Ru		En→De	
CD	Loss	Voita^{test}	Voita^{dev}	Loss	ContraPro
1.000	1.580	69.99	66.50	1.097	70.43
0.900	1.583	70.26	66.20	1.096	69.44
0.700	1.580	70.96	66.40	1.093	70.52
0.500	1.579	70.89	66.10	1.092	70.38
0.300	1.573	71.59	66.30	1.089	72.49
0.100	1.564	71.86	67.00	**1.086**	69.58
0.010	1.563	**73.19**	67.40	1.090	**74.31**
0.009	1.563	67.30	67.30	**1.086**	71.93
0.007	**1.562**	67.90	**67.90**	1.091	72.72
0.005	**1.562**	67.00	67.00	1.110	71.25
0.003	1.563	67.20	67.20	1.105	71.13
0.001	1.563	67.50	67.50	1.104	64.53
0.000	1.574	70.34	66.80	1.191	61.14

Table 4.8 – Ex-poste analysis of context discounting with different intensities. A strong context discounting results in the best performances of s4to4, both in terms of *current loss* on the development set and accuracy on contrastive sets for the evaluation of discourse phenomena.

Table 4.7. However, the limit case of $CD = 0$ is not aligned with this trend. We suspect that the self-attention distribution is flatter in this case because the model encounters learning difficulties due to the training signal from the context being completely ignored (c.f. Bao et al. (2021) on non-fully-converged models having a flatter attention distribution).

4.3.5.3 Discounting value

In Figure 4.2 and Table 4.7, we have documented the impact of different values of context discounting on the performance of a model over a development set in terms of best *current loss* and accuracy on discourse phenomena. As we have seen, the performance improves between $CD = 1$ and $CD = 0.01$, then drops at $CD = 0$. Here, we provide a more fine-grained analysis of discounting values in the $[0.01, 0]$ range and report accuracy on the test set. Moreover, we expand the analysis to the En→De language pair. Results are presented in Table 4.8. For both language pairs, all the best performance measurements correspond to context discounting values in the range $[0.1, 0.005]$, confirming the conclusion

			En→Ru			
System	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	BLEU
s4to1	50.00	45.87	57.60	71.40	51.66	**32.64**
s4to4 + *CD=0*	**86.48**	**46.27**	**70.00**	**78.60**	**71.98**	28.55
			En→De			
System	d=1	d=2	d=3	d>3	ContraPro	BLEU
s4to1	36.90	46.55	49.38	69.68	40.67	**29.28**
s4to4 + *CD=0*	**57.35**	**67.81**	**71.72**	**85.29**	**61.14**	11.85

Table 4.9 – Complete context discounting is better than no target context at all.

of the preliminary analysis (Section 4.3.2): a substantial discounting of context works best.

The extreme case of $CD = 0$ is worth additional investigation, especially compared to an s4to1 system trained in the same conditions. In Table 4.9 we compare the performance of a fully context-discounted s4to4 system (s4to4+*CD=0*) and s4to1. Despite both systems being trained without including the loss from the context in the training signal, the former strongly outperforms the latter in terms of accuracy on the contrastive sets. Indeed, there is a significant difference between the s4to4+*CD=0* approach and s4to1: the former is allowed to decode the target context, even if the context is discarded before calculating the loss, while the latter is penalized from any decoding of the context. The whole sequence decoded by s4to1 is compared with the reference translation of the current sentence only. Therefore, contrary to s4to4+*CD=0*, the trained s4to1 system is not used to decode any context, which penalizes it in terms of accuracy on the contrastive test sets because they contain target context. However, when it comes to generating a translation (instead of simply scoring sentences from a contrastive set), s4to4+*CD=0* largely lags behind. This is because s4to4+*CD=0* is not trained to produce a meaningful translation of the context nor to avoid the translation of the context. Therefore it translates context poorly, and the errors propagate to the current sentence too. We can conclude that while substantial context discounting works well, full context discounting ($CD = 0$) results in systems that can model inter-sentential phenomena but can not generate consistent translations.

4.3.5.4 Robustness

In this section, we provide evidence showing that context discounting increases the robustness of concatenation approaches to context sizes unseen during training. This outcome is in line with the analysis of the attention distribution presented above (4.3.5.2) since more robustness to context variations is expected from a model that relies less

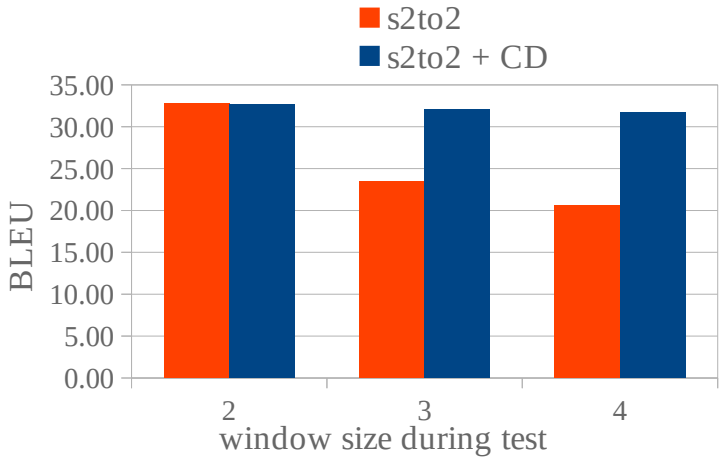

Figure 4.7 – Our approach improves the robustness of En→Ru s2to2 to window sizes unseen during training.

on context. Figure 4.7 shows that the s2to2 model is not robust to the translation of concatenation windows longer than those seen during training, i.e. longer than 2 sentences. Indeed, s2to2 loses 9.23 BLEU points when translating the same test set with windows of 3 sentences, and 12.14 BLEU points when translating with windows of 4. Instead, the context discounted model (blue bars) is very robust to unseen context lengths, being capable of translating them with minor degradation in average translation quality (−0.68 and −1.06 BLEU points for windows of 3 and 4, respectively). Similarly, s3to3 loses 1.74 BLEU points when tested with windows of size 4, while s3to3+`CD` is perfectly robust (see Table B.5 in the Appendix).

4.3.6 Analysis of segment-shifted positions

In this section, we analyze the impact of adding segment-shifted positions on top of context discounting, along with some alternatives to this approach.

4.3.6.1 Attention distribution

Segment-shifted positions are meant to help context-discounted models to learn the locality properties of the processed languages (Hardmeier, 2012). In other words, we expect segment-shifted positions to increase the localization of the distribution of self-attention weights, i.e., to make it less uniform. We can evaluate this effect by computing the average entropy of the distribution of self-attention weights generated by all queries (both from current and context sentences), equivalently to what we have done for Table 4.7. Table 4.10

System	Attn entropy
s4to4	2.218
s4to4 + `CD`	2.109
s4to4 + shift + `CD`	**2.062**

Table 4.10 – The average entropy of self-attention weights decreases with context discounting and segment-shifted positions. All of the three values are different from one another with statistical significance ($p<0.01$).

System	Shift	Voita	BLEU
s4to4 + shift + `CD`	100	73.46	32.41
s4to4 + shift + `CD`	avg-sequence	**73.86**	32.37
s4to4 + shift + `CD`	avg-corpus	73.56	32.45

Table 4.11 – Accuracy on the En→Ru contrastive set for the evaluation of discourse phenomena (Voita, %), and BLEU score on the test set. Differences across models are not statistically significant ($p>0.05$).

shows the results: context discounting slightly reduces the average entropy, which is further lowered with the adoption of segment-shifted positions. Segment-shifted positions make self-attention more focused locally, as intended, which also explains why the job of context discounting is eased by this solution.

4.3.6.2 Segment-shifting variants

In the experiments reported above, we have adopted a shifting value equal to the average sentence length calculated over the entire training corpus (*avg-corpus*). I.e., +8 positions for En→Ru, +21 positions for En→De. In this section, we evaluate two alternatives:

- *100*: applying a big shift of 100 units, one order of magnitude bigger than the average sentence length in the corpus;
- *avg-sequence*: applying a shift equal to the average sentence length of the window, calculated dynamically for each window of 4 concatenated sentences.

The results of this study are reported in Table 4.11. We do not observe relevant differences in average translation quality (BLEU) or accuracy in the translation of discourse phenomena.

Therefore, we confirm that selecting a shift equal to the average sentence length calculated on the corpus (avg-corpus) approach is a good alternative.

4.3.7 Synergies with *divide and rule*

Before concluding this chapter, we investigate possible synergies between the focused concatenation approach and the *divide and rule* (*d&r*) pre-training technique discussed in Chapter 3. The *d&r* technique facilitates the learning process of multi-encoding approaches, which encode current and context sentences separately, and does not apply to plane concatenation approaches. In fact, feeding a standard concatenation approach with a concatenation of split sentences would be virtually the same as feeding it with complete sentences. Nonetheless, the focused concatenation approach acts more like multi-encoding systems and could benefit from *d&r* pre-training. Indeed, the context-discounted model is encouraged to process context and current sentences differently, as we have observed in Section 4.3.5.1.

Thus, we pre-train on the training data with sentences split in a half (*middle-split*) both the concatenation baseline s4to4 and its focused counterparts (s4to4+CD and s4to4+shift+CD). After *d&r* pre-training, we keep training the models on the original document-level data (see Appendix B.1 for details). We compare the performances of these systems with the same models trained on original data only (see also Table 4.2). The results, displayed in Table 4.12, are mixed. As expected, *d&r* is not helpful for the standard s4to4. Its performance is slightly degraded in terms of accuracy on the contrastive sets and BLEU for both language pairs. Instead, the context-discounted approach (s4to4+CD) improves its performance on Voita and ContraPro after undergoing *d&r* pre-training. Interestingly, for En→Ru, this improvement is entirely driven by the gain on deixis, while the performance on the other discourse phenomena is degraded. Deixis is the most frequent phenomenon in the contrastive set (see Section 2.3.2.1 or Appendix B.2.1), and therefore the most impactful on the overall accuracy on Voita. We do not have a clear explanation for this behavior, except noticing that in our experiments, we have often measured an inverse correlation between the performance on deixis and the performance on the other discourse phenomena present in the contrastive set. The BLEU score is slightly degraded on En→Ru, while it is slightly improved in En→De. Surprisingly, the beneficial effects of *d&r*on s4to4+CD do not transfer on s4to4+shift+CD. In conclusion, *d&r*improves the performance of the context-discounted model only on certain discourse phenomena (deixis and pronominal anaphora), for both language pairs. The pre-trained s4to4+CD achieves the best results on the contrastive sets presented so far while maintaining a comparable average translation quality. However, this synergy dissolves with the adoption of segment-shifted positions.

		En→Ru					
System	*d&r*	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	BLEU
s4to4	no	85.80	46.13	79.60	73.20	72.02	32.45
s4to4 + CD	no	**87.16**	46.40	81.00	78.20	73.42	32.37
s4to4 + shift + CD	no	85.76	**48.33**	**81.40**	**80.40**	**73.56**	32.45
s4to4	yes	86.20	46.07	72.20	67.00	70.84	32.07
s4to4 + CD	yes	**90.24**	45.93	80.00	76.00	**74.50**	31.95
s4to4 + shift + CD	yes	87.08	**46.33**	**81.60**	**77.60**	73.36	32.08
		En→De					
System	*d&r*	d=1	d=2	d=3	d>3	ContraPro	BLEU
s4to4	no	68.89	74.96	79.58	**87.78**	71.35	29.48
s4to4 + CD	no	**72.86**	75.96	80.10	84.38	74.31	29.32
s4to4 + shift + CD	no	72.56	**77.15**	**80.27**	86.65	**74.39**	29.20
s4to4	yes	67.49	73.90	78.53	87.10	70.06	29.08
s4to4 + CD	yes	**72.97**	**77.01**	**78.88**	**87.55**	**74.63**	29.78
s4to4 + shift + CD	yes	71.15	75.89	76.43	86.42	72.91	28.98

Table 4.12 – Comparison between the systems presented in Table 4.2 and the same models pre-trained with the approach presented in Chapter 3: *divide and rule* (*d&r*). Metrics: accuracy on the contrastive sets for evaluating discourse phenomena (Voita and ContraPro) and BLEU. The accuracy on contrastive sets is detailed on the left with the accuracy on each subset. For En→Ru, each subset corresponds to a specific discourse phenomenon. For En→De, each subset contains examples of anaphoric pronouns with antecedents at a specific distance $d \in [1, 2, ...]$ (in number of sentences).

4.4 Conclusions

4.4.1 Takeaways

We presented a simple, parameter-free modification of the NMT objective for context-aware translation with sliding windows of concatenated sentences: context discounting. We analyzed the impact of our approach in the trade-off between current sentence predictions and context sentence predictions, showing that context discounting helps the model to focus on the current sentence, as intended. As a result, the concatenation model significantly improves its ability to disambiguate inter-sentential discourse phenomena, and becomes more robust to new context sizes. As an additional inductive bias towards locality, we equip our model with segment-shifted positions, marking clearer boundaries between sentences. This solution brings further improvements on contrastive test sets, although only marginal. In the attempt to explain the proposed solutions' empirical functioning, we investigated their impact on the distribution of the self-attention weights, showing that they make it more focused and skewed toward the current sentence, as intended. Finally, we have analyzed the impact of *divide and rule* pre-training on focused concatenation approaches. We found it beneficial for the context-discounted model without segment-shifted positions, which achieves the best performance on contrastive sets obtained so far in our work.

4.4.2 Limitations and future works

Our experiments are limited to the use case of short concatenated windows (up to 4 sentences). This is enough for capturing most of the ambiguous inter-sentential discourse phenomena that usually span across a few sentences only (Müller et al., 2018; Voita et al., 2019b; Lupo et al., 2022a). However, recent works suggest that longer context windows might be helpful to increase the average translation quality (BLEU) of concatenation approaches (Junczys-Dowmunt, 2019; Bao et al., 2021; Sun et al., 2022), and long-range discourse phenomena could be handled. We hope to investigate the impact of context discounting on longer sequences in future works. We also encourage to test the effectiveness of our approach on a wider range of data scenarios: from very limited document-level data to very abundant, including back translation (Ma et al., 2021b) and monolingual pre-training techniques (Junczys-Dowmunt, 2019; Sun et al., 2022), to understand whether these methods are only alternative to context discounting or there exist synergies. Furthermore, experimenting with future context is also needed (c.f. Wong et al. (2020)).

Chapter 5

Encoding sentence position in concatenation approaches

The contributions presented in this chapter will be submitted to the *Fourth Workshop on Insights from Negative Results in NLP*, co-located with EACL 2023.

Contents

5.1 Introduction

In the previous chapter, we have introduced the segment-shifted position embeddings as a way to help context-discounted concatenation approaches to discern the sentences in the concatenation window. Explicitly telling the model which tokens belong to the different sentences of the processed sequence is not a new idea, but a pretty intuitive one that has already been tested successfully in other tasks and approaches (Devlin et al., 2019; Voita et al., 2018; Zheng et al., 2020). We believe that providing explicit information about the position of the sentences at the token level helps concatenation approaches to overcome the learning challenge presented in Section 2.4.4. In fact, if the tokens' latent representations contain information about the position of the sentences they belong to, they can be processed by the attention function accordingly. For instance, attention can recognize more readily the tokens belonging to the same sentence, which have higher chances to be related to one another, as well as the distance of their context. As it is fundamental for the Transformer model to have a notion of the sequentiality of the tokens, it is also valuable to know the order of the sentences in the window. The temporal structure of the document constitutes essential information for its understanding and the correct disambiguation of inter-sentential discourse phenomena. It could be argued that the information needed to determine which sentence a token belongs to, and its position within the window, is already available thanks to token position embeddings. However, the position embeddings do not constitute direct information about their sentence membership because such information can only be retrieved through a comparison with the position of separator tokens. We propose to inject sentence-membership information directly into token representations in order to bypass the need to learn and perform such a comparison.

This chapter presents a comparative study of various approaches to encoding sentence membership and position in concatenation approaches. Besides segment-shifted position embeddings, we will evaluate three different kinds of segment embeddings: one hot, sinusoidal (Vaswani et al., 2017) and learned (Devlin et al., 2019). Inspired by the literature on position embeddings (Chen et al., 2021; Liu and Zhang, 2020), we propose to make segment embeddings persistent over layers, adding them to the input of every layer in addition to the first. Moreover, we propose fusing position and segment embeddings into a single vector where token and segment positions are encoded in two orthogonal sets of dimensions, allowing a clearer distinction between them, along with memory savings.

To the best of our knowledge, this is the first comparative study on the employment of sentence position encodings for concatenation approaches to CANMT. Although a few studies in the literature have adopted fixed or learned segment embeddings for multi-encoding approaches (Voita et al., 2018; Zheng et al., 2020; Bao et al., 2021), a

comprehensive comparison is missing in this case too.

5.2 Proposed approach

Besides segment-shifted position embeddings, described in Section 4.2.2, we propose to equip SlidingKtoK with three different kinds of segment embeddings, described below. Segment embeddings encode the position k of each sentence within the window of K concatenated sentences into a vector of size d. We attribute sentence positions $k = 1, 2, ..., K$ starting from right to left. The first sentence of the window is given the position $k = K$, while the sentence following the last separation token is given $k = 1$. The underlying rationale is always to attribute the position $k = 1$ to the current sentence, no matter how many sentences are combined. Indeed, the SlidingKtoK approach is trained with concatenation windows formed by *up to* K sentences, according to the available context. This methodology is equivalent to attributing sentence distance embeddings to each token, where the distance is expressed in terms of the number of sentences from the current sentence plus one.

One-hot encoding - This is a standard method to encode a categorical variable into a d-dimensional vector. The one-hot encoding of position k is a vector where all the elements equal zero except the kth element, which equals one.

Sinusoidal encoding - This method encodes every sentence position with d sinusoids, one for each embedding dimension. The same approach is employed to encode token positions in the Transformer model (Vaswani et al., 2017). Intuitively, the sinusoidal representation is a continuous equivalent of the binary representation. For a detailed description of sinusoidal embeddings and their properties, refer to Section 2.1.4.2.

Learned embedding - The d-dimensional embedding is randomly initialized and then learned with the rest of the model, like in Devlin et al. (2019).

The simplest strategy to integrate segment embeddings (SE) with position embeddings (PE) and token embeddings (TE) is by adding them up, as in Devlin et al. (2019). For a given token x_t^k at position t, belonging to the kth source sentence of the concatenation window $X = \{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^K\}$, its continuous representation input to the model results from the sum of the token, position, and segment embeddings:

$$s_t^k = TE_v + PE_t + SE_k$$

where $v \in \mathcal{V}$ is a token belonging to the model's vocabulary $\mathcal{V}$. This operation requires that all three embeddings have the same dimensionality $d = d_{model}$. In the next section,

we present an alternative strategy for integrating segment and position embeddings in a unique vector before adding it to the token embedding. Both strategies will be evaluated in the experimental section (5.3).

We propose to encode the sentence position with absolute position embeddings instead of relative ones (Shaw et al., 2018) because the literature suggests that both perform similarly in applications like ours (Liu et al., 2020a; Rosendahl et al., 2019; Likhomanenko et al., 2021; Chen et al., 2021). Moreover, absolute position embeddings can be fixed, as in the case of one-hot and sinusoidal encodings, while relative ones are always learned. While early works leaned in favor of relative encodings (Shaw et al., 2018), later research suggested that relative encodings might be advantageous only in the case of long sequences or unseen sequence lengths (Rosendahl et al., 2019; Likhomanenko et al., 2021). However, Chen et al. (2021) found that the argued superiority of relative position embeddings might simply be due to their being added to each attention head. When applying the same procedure with absolute position embeddings, they find the best performance across a range of natural language understanding tasks. Liu et al. (2020a) also found an increased performance of absolute sinusoidal position embeddings when adding them to the input of each block in the Transformer architecture.

5.2.1 Persistent encodings

For a SlidingKtoK model, the maximum amount of concatenated sentences K is usually tiny and fixed between training and inference. Therefore, in view of the above findings, we believe that absolute positions are adapted for marking sentence positions within a concatenation window. In addition, we propose to test the efficacy of making sentence position encodings persistent across Transformer's blocks, as (Liu et al., 2020a) did for position embeddings. In other words, we propose adding segment-shifted position embeddings or segment embeddings to each block's input instead of just the first. This option will be tested empirically in Section 5.3 and benchmarked against persistent token position embeddings in Section 5.3.3.

5.2.2 Position-segment embeddings

The usual sum of token and position embeddings in the Transformer is based on the premise that the model can still distinguish both signals after being added up. This is likely accomplished by learning token embeddings in a way that guarantees they can be distinguished from non-learnable position representations. However, adding a third non-learnable representation (segment embedding) to the same vector could make distinguishing

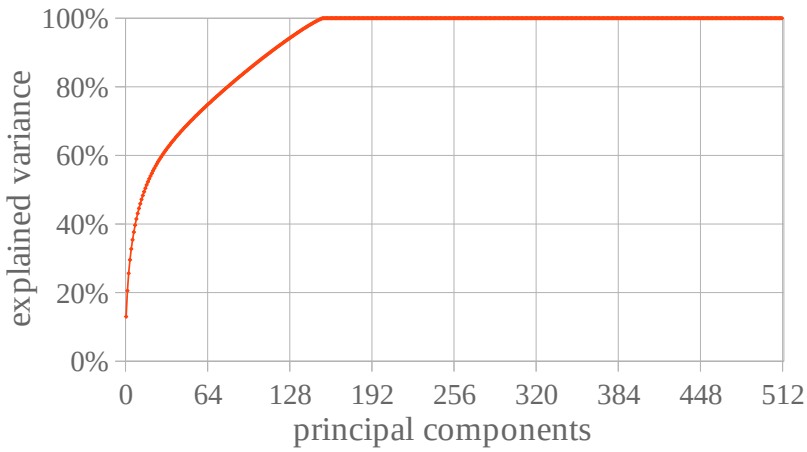

Figure 5.1 – Cumulative ratio of the variance explained by the principal components of the 1024×512 sinusoidal position embedding matrix.

between the token, position, and segment embeddings hard, if not impossible, for the model. For instance, if both segment and position embeddings are sinusoidal, distinguishing between $PE_t + SE_k$ and $PE_k + SE_t$ is impossible. Instead, if position and segment embeddings were concatenated to one another, they would be perfectly distinguishable because they would belong to orthogonal spaces. Unfortunately, concatenating two embeddings with d_{model} dimensions would then oblige to project the concatenated vector back to a d_{model}-dimensional space. To avoid this expensive operation, we propose to reduce the dimensionality of position and segment embeddings from $d_{PE} = d_{SE} = d_{model}$ to values that sum up to the model dimension, i.e., $d_{PE} + d_{SE} = d_{model}$. Thus, position and segment embeddings can be concatenated into a unique vector that we call Position-Segment Embedding (PSE), of size $d_{PSE} = d_{model}$:

$$PSE_{t,k} = [PE_t, SE_k].$$

Reducing the embedding dimensionality for both token and sentence positions can be made without loss of information. As mentioned earlier, sinusoidal embeddings are the continuous counterpart of the binary representation for natural numbers. Being continuous, they can represent a much larger set of positions with the same number of dimensions. In practice, for the Transformer-base architecture (Vaswani et al., 2017), with $d_{model} = 512$, all the positions $t = 1, 2, ..., t_{max}$ can be modeled with unique embeddings in the usual applications, where $t_{max} = 1024$. Even in the case of much longer sequences, the bottleneck is attention's complexity rather than sinusoidal embeddings' representativeness. This can be easily shown with a Principal Component Analysis (Jolliffe and Cadima, 2016) of the sinusoidal position embedding matrix $PE \in \mathcal{R}^{1024 \times 512}$ representing 1024 positions with 512 dimensions. In Figure 5.1, we plot the cumulative ratio of the variance explained by

each component. Less than half of the principal components can explain the entirety of the variance represented in the sinusoidal embeddings.[1] In other words, 1024 positions can be represented with the same resolution using less than half the dimensions.

In the experimental section, we will empirically evaluate the impact of representing token and sentence positions with PSE, where the former are encoded with sinusoids and the latter with either one-hot, sinusoidal or learned encodings.

5.3 Experiments

5.3.1 Setup

The experimental setup in this chapter follows closely the previous chapter's. We briefly outline it here again. We experiment with two models:

- *base*: a context agnostic baseline following *Transformer-base* (Vaswani et al., 2017) (see section 2.1.4.2);
- *s4to4*: short for Sliding4to4, a context-sensitive concatenation approach with the same architecture as *base*, processing sliding windows of 4 concatenated sentences as the source, and decoding the whole window into the target language (see Section 2.4.1).

We equip the s4to4 model with the various sentence position encodings proposed above in order to evaluate their impact and compare them. When experimenting with PSE, we allocate four dimensions to segment embeddings ($d_{SE} = 4$), which is enough to encode the position of each of the four sentences in the concatenation window, with both one-hot and sinusoidal encodings. Since $d_{model} = 512$, this leaves $d_{PE} = d_{model} - d_{SE} = 508$ dimensions available to encode token positions.

The models are trained and evaluated on two language pairs covering two different domains: movie subtitles for En→Ru Voita et al. (2019b), IWSLT17's TED talk subtitles for En→De. In addition to evaluating the average translation quality with BLEU, we employ two contrastive sets to evaluate the translation of inter-sentential discourse phenomena. For En→Ru, we adopt Voita et al. (2019b)'s set for the evaluation of deixis, lexical cohesion, verb-phrase ellipsis, and inflectional ellipsis (details in Section 2.3.2.1). For En→De, we evaluate the models on the translation of ambiguous pronouns with ContraPro (Müller

[1] With $d_{PE} = 128$, 94% of the variance is still explained; with $d_{PE} = 155$, the entirety of it.

et al., 2018), a large contrastive set of ambiguous pronouns whose antecedents belong to the context (see Section 2.3.2.1). More details are reported in Section 4.3.1 and Appendix B.1.

5.3.2 Results

We first study the impact of sentence position encodings on the s4to4 model trained on En→De data. In Table 5.1, we compare models equipped with a combination of the following elements:

- encoding method (Enc.): either segment-shifted positions, one-hot, sinusoidal or learned encodings;
- persistence (Pers.): sentence position encodings are added to the input of each model's block, or just to the first (if the option is not checked);
- position-segment embeddings (PSE): position and segment embeddings are concatenated into a unique vector, or added together (if the option is not checked).

As usual, we primarily focus on the accuracy on contrastive test sets, as BLEU displays minor fluctuations throughout the whole table. However, the performance on the contrastive sets is not encouraging either: non of the encoding variants proposed outperform s4to4 consistently. The one-hot encoding helps, but only by a thin margin. Making encoding persistent or fusing them into PSE does not help either. The only exception is s4to4+lrn+pers+pse (last line), which gains more than two accuracy points over baseline. However, this result is solely driven by the net improvement on deixis disambiguation (almost +5 points), while the performance is degraded on the other three discourse phenomena. In conclusion, sentence position encodings do not seem to benefit the vanilla s4to4 approach.

5.3.2.1 With context discounting

Encouraged by the outcomes of segment-shifted positions in the previous chapter, we hypothesize that context-discounted concatenation approaches could leverage sentence position encodings more effectively. Indeed, the context-discounted objective function incentivizes distinguishing among different sentences. Table 5.2 displays the results of the s4to4+`CD` model equipped with the various combinations of encodings tested before, except the *non-persistent* PSE.[2] In this case too, vanilla encoding methods do not significantly

[2]Since preliminary experiments where not encouraging, we do not provide results for the non-persistent PSE combination in order to economize experiments.

System	Enc.	Pers.	PSE	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	BLEU
base				50.00	45.87	51.80	27.00	46.64	31.98
s4to4				85.80	46.13	79.60	73.20	72.02	32.45
s4to4	shift			85.24	46.07	77.20	71.20	71.28	32.27
s4to4	shift	✓		**85.96**	**46.33**	75.20	**74.00**	71.80	31.93
s4to4	1hot			**86.08**	47.07	78.00	**75.60**	**72.52**	32.61
s4to4	1hot	✓		83.76	**47.53**	78.00	75.00	71.44	32.42
s4to4	1hot		✓	84.56	46.13	78.20	73.00	71.24	32.33
s4to4	1hot	✓	✓	84.56	46.47	76.00	73.40	71.16	32.41
s4to4	sin			**86.36**	45.80	76.40	73.60	71.92	32.39
s4to4	sin	✓		84.96	46.13	74.80	74.00	71.20	32.38
s4to4	sin		✓	84.64	**46.40**	76.60	73.60	71.26	32.56
s4to4	sin	✓	✓	85.24	46.33	76.40	**75.20**	71.68	32.38
s4to4	lrn			85.48	46.27	76.20	**75.60**	71.80	32.56
s4to4	lrn	✓		84.84	45.93	77.60	74.40	71.40	32.50
s4to4	lrn		✓	83.60	**46.67**	74.80	70.80	70.36	32.37
s4to4	lrn	✓	✓	**90.52**	46.00	74.80	66.60	**73.20**	32.38

Table 5.1 – s4to4 trained on En→Ru OpenSubtitles with different sentence position encodings (Enc.) and two options: persistency of the encodings (Pers.) and concatenation of position and segment embeddings (PSE). Accuracy on Voita's En→Ru contrastive set and BLEU on the test set. The accuracy on the contrastive set is detailed on the left, with the accuracy on each subset corresponding to a specific discourse phenomenon. The values in bold obtain the best performance within their block of rows and outperform the baselines (first block).

System	Enc.	Pers.	PSE	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	BLEU
base				50.00	45.87	51.80	27.00	46.64	31.98
s4to4				85.80	46.13	79.60	73.20	72.02	32.45
s4to4+CD				87.16	46.40	81.00	78.20	73.42	32.37
s4to4+CD	shift			85.76	48.33	81.40	80.40	73.56	32.45
s4to4+CD	shift	✓		**88.76**	**52.13**	**83.00**	76.20	**75.94**	31.98
s4to4+CD	1hot			86.40	46.73	82.00	76.40	73.06	32.35
s4to4+CD	1hot	✓		87.68	46.80	81.60	**78.60**	73.90	32.56
s4to4+CD	1hot	✓	✓	**88.88**	**47.67**	**82.20**	75.40	**74.50**	32.33
s4to4+CD	sin			87.96	46.80	78.00	76.60	73.48	32.53
s4to4+CD	sin	✓		86.80	**47.00**	80.80	78.20	73.40	32.52
s4to4+CD	sin	✓	✓	**89.28**	46.67	**83.20**	77.20	**74.68**	32.27
s4to4+CD	lrn			88.12	46.47	81.20	75.60	73.68	32.45
s4to4+CD	lrn	✓		86.84	**52.27**	**84.60**	**80.00**	**75.56**	32.43
s4to4+CD	lrn	✓	✓	**93.20**	47.40	72.20	64.40	74.48	32.35

Table 5.2 – Context-discounted s4to4 trained on En→Ru OpenSubtitles with different sentence position encodings (Enc.) and two options: persistency of the encodings (Pers.) and concatenation of position and segment embeddings (PSE). Accuracy on Voita's En→Ru contrastive set and BLEU on the test set. The accuracy on the contrastive set is detailed on the left, with the accuracy on each subset corresponding to a specific discourse phenomenon. The values in bold obtain the best performance within their block of rows and outperform the baselines (first block).

System	Enc.	Pers.	PSE	d=1	d=2	d=3	d>3	ContraPro	BLEU
base				32.89	43.97	47.99	70.58	37.27	29.63
s4to4				68.89	74.96	79.58	87.78	71.35	29.48
s4to4+CD				72.86	75.96	80.10	84.38	74.31	29.32
s4to4+CD	shift			72.56	**77.15**	**80.27**	**86.65**	**74.39**	29.20
s4to4+CD	shift	✓		69.15	74.23	77.13	86.42	71.22	27.50
s4to4+CD	sin			71.83	**76.82**	**80.97**	**87.55**	73.88	29.23
s4to4+CD	sin	✓		72.08	76.35	79.23	85.97	73.82	29.26
s4to4+CD	sin	✓	✓	71.22	76.42	78.88	86.87	73.22	28.73
s4to4+CD	lrn			70.21	75.29	77.66	85.06	72.14	28.35
s4to4+CD	lrn	✓		68.53	72.51	75.74	**86.65**	70.42	27.87
s4to4+CD	lrn	✓	✓	68.40	**79.07**	**80.27**	83.48	71.48	28.63

Table 5.3 – Context-discounted s4to4 trained on En→De IWSLT17 with different sentence position encodings (Enc.) and two options: persistency of the encodings (Pers.) and concatenation of position and segment embeddings (PSE). Accuracy on ContraPro and BLEU on the test set. The accuracy on pronoun translation is detailed on the left with the accuracy on each subset, corresponding to pronouns with a specific antecedent's distance (in number of sentences). The values in bold obtain the best performance within their block of rows and outperform the baselines (first block).

help the s4to4+CD model. However, making the encodings persistent boosts performance in all cases but for sinusoidal embeddings. Employing shift+pers improves performance by 2.52 accuracy points over s4to4+CD, while lrn+pers brings a +2.14 improvement. Instead, one-hot segment embeddings benefit only slightly (+0.48) from being persistent. We believe that the reason is the same as the reason why sinusoidal segment embeddings do not benefit from persistence. As discussed in Section 5.2.2, one-hot or sinusoidal segment embeddings might not be distinguishable from sinusoidal position embeddings once they are added together. Instead, when one-hot and sinusoidal segment embeddings are concatenated to position embeddings into a unique PSE and made persistent, they boost the performance of s4to4+CD by +1.08 and +1.26 accuracy points, respectively.

With the aim of evaluating the generalizability of these results to another language pair and domain, we train and test the context-discounted approach on the En→De IWSLT17 dataset. In this case, we avoid experimenting with one-hot encodings since it was found to be the less promising approach on the En→Ru setting. Table 5.3 summarizes the results on En→De. Unfortunately, the improvements achieved on En→Ru do not transfer to

System	Enc.	Pers.	PSE	d=1	d=2	d=3	d>3	ContraPro	BLEU
s4to4+CD				**81.79**	82.11	82.19	90.04	**82.24**	31.69
s4to4+CD	shift	✓		79.61	81.45	83.42	86.65	80.45	30.71
s4to4+CD	sin	✓	✓	79.85	**82.38**	**84.46**	86.87	80.85	31.40
s4to4+CD	lrn	✓		79.13	79.73	82.19	88.00	79.82	31.58

Table 5.4 – Context-discounted s4to4 trained on the En→De high-resource setting with different sentence position encodings (Enc.) and two options: persistency of the encodings (Pers.) and concatenation of position and segment embeddings (PSE). Accuracy on ContraPro and BLEU on the test set. The accuracy on pronoun translation is detailed on the left with the accuracy on each subset, corresponding to pronouns with a specific antecedent's distance (in number of sentences).

this setting. The context-discounted s4to4 slightly benefits from segment-shifted position embeddings, but the other approaches degrade its performance.

5.3.2.2 Increasing training data for the English to German pair

We hypothesize that the model does not undergo sufficient training in this setting to reap the benefits of segment embeddings. In En→De IWSLT17, the training data volume is smaller than in the En→Ru setting: 0.2 million sentences versus 6 millions (see Table 4.1). Therefore, we choose to experiment with the En→De high-resource training set employed in Chapter 3 and presented in Section 3.4.1. This setting expands the IWSLT17 training data (Cettolo et al., 2012) by adding the News-Commentary-v12 and Europarl-v7 sets released by WMT17[3]. The resulting training set comprises 2.3M sentences.[4] Training on this data is more expensive than training on the En→Ru setting, considering that the average sentence length is of 27.3 tokens versus 8.3 tokens, respectively. Therefore, we only train the most promising approaches.[5] Their performances are compared in Table 5.4. As expected, the s4to4+CD model drastically improves its performance compared to training on IWSLT17 alone: +7.93 accuracy points on ContraPro and +2.37 BLEU points on the test set (c.f. Table 5.3). However, even with larger training volumes, segment position encodings do not seem to help s4to4+CD on the En→De language pair.

[3]http://www.statmt.org/wmt17/translation-task.html

[4]More data statistics are presented in Table 3.4.

[5]We set shift= 27 for segment-shifted position embeddings, consistently with the average sentence length of the training data.

System	Enc.	Pers.	PSE	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	BLEU
s4to4				85.80	46.13	**79.60**	**73.20**	72.02	32.45
s4to4		✓		**86.08**	**47.27**	79.40	72.80	**72.44**	32.29
s4to4 + CD				87.16	46.40	81.00	**78.20**	73.42	32.37
s4to4 + CD		✓		88.24	46.87	81.40	77.80	74.10	32.12
s4to4 + CD	segshift	✓		**88.76**	**52.13**	**83.00**	76.20	**75.94**	31.98

Table 5.5 – Making positions persistent across Transformer's blocks improve discourse disambiguation performance both for vanilla and context-discounted s4to4. Segment-shifting positions further improve performance.

5.3.3 Persistent positions

Making sentence position encodings persistent across the layers have been found beneficial for context-discounted models on the En→Ru setting (Table 5.2). The best-performing model, s4to4+CD+shift+pers, shifts token positions by a constant factor every time we pass from one sentence to the next and makes the resulting position embeddings persistent throughout Transformer's blocks. In Table 5.5, we benchmark this model against models employing persistent token position embeddings but without segment-shifting. Both vanilla and context-discounted s4to4 perform better when positions are persistent across Transformer's blocks, as suggested by Liu et al. (2020a) and Chen et al. (2021). Segment-shifting further enhances performance, which confirms that the model benefits from a sharper distinction between sentences.

5.4 Conclusions

5.4.1 Takeaways

In this chapter, we proposed to encode into token representations their sentence membership and its position within a window of concatenated sentences, intending to improve concatenation approaches to CANMT. Besides adopting existing encoding methods, we proposed a novel way to integrate token and sentence position embeddings in a unique vector called position-segment embedding (PSE). Moreover, we proposed to make sentence position encodings persistent throughout the model's layers, adding them to the input of each Transformer's block.

We compared the various combinations of encoding methods, persistence, and integration

with position embeddings on the English-to-Russian language pair. Unfortunately, none of the proposed encoding variants improved the performance of the vanilla concatenation approach. Instead, improvements were observed when sentence position encoding methods were used in conjunction with the context-discounted training objective, which encourages the model to exploit the information about sentence position. In particular, we found the best results with persistent segment-shifted positions and persistent learned segment embeddings. Moreover, persistent sinusoidal and one-hot segment embeddings were found to benefit the context-discounted approach, but only when they are fused with position embeddings into PSE. Unfortunately, these results did not transfer to the English-to-German language pair, neither in the IWSLT setting nor the high-resource setting. Sentence position encodings were found unhelpful in this case.

In conclusion, the employment of persistent encodings was found valuable in a specific setting, as well as the adoption of PSE for non-learnable segment embeddings, but the results did not generalize. Since encoding sentence positions persistently is easy to implement and requires little or no additional parameters to learn, we still encourage developers to test this option if they strive to get the most out of their CANMT systems. To the best of our knowledge, this work constitutes the first comparative study on sentence position encodings for CANMT.

5.4.2 Limitations and future works

Further analysis would be needed to understand why the improvements achieved on En→Ru are not transferable to En→De. To begin with, transferability to a third language pair could be investigated to gain insights into whether En→De is an exception or the other way around. Then, one could investigate whether the model integrates sentence position encodings to identify sentence boundaries and positions better, as intended. To this end, one could evaluate the effectiveness of a shallow neural network in predicting the position of a token's sentence based on the token representation output by the encoder. If the addition of sentence position encodings in the input improves accuracy on this task, we can reasonably conclude that the model exploits sentence position information as intended. Through such an analysis, differences between pairs of languages could be revealed, providing explanatory elements about the differences in translation performances that have been observed. Finally, it would be valuable to conduct an equivalent study in the case of multi-encoding approaches, comparing the impact of employing different sentence position encoding solutions.

Chapter 6

Conclusions

In this doctoral thesis, we identified some key challenges faced by the current deep-learning architectures for context-aware translation and proposed some solutions to tackle them. In particular, we have focused on the learning aspect of both multi-encoding and concatenation approaches, proposing a pre-training solution in the former case, and a modified loss function in the latter. Then, we continued the work on concatenation approaches by proposing to extend the standard Transformer architecture with segment-shifted positions or segment embeddings, undertaking a comparative evaluation of different configurations. We conclude by summarizing the contributions made and identifying potential directions for further research.

6.1 Summary of contributions

Chapter 3: Divide and Rule. We have examined the challenges associated with training multi-encoding models, a broad category of context-aware neural machine translation models. The sparsity of tokens requiring context and the sparsity of their relevant context have been identified as primary challenges. To address them, we have proposed a pre-training approach referred to as *divide and rule*, which involves splitting the training sentences to increase each segment's linguistic dependency from the others. The increased inter-segmental dependency constitutes additional training signal for the learnable parameters in charge of modeling the context. Moreover, having to deal with shorter sequences, the context-handling parameters can retrieve relevant context more easily. An analysis of the impact of splitting on the distribution of discourse phenomena was conducted, showing empirically that *d&r* pre-training enables the learning of contextual parameters with greater efficiency than simply increasing training data. Besides splitting sentences in the

middle, we have proposed more sophisticated sentence-splitting methods aiming at maximizing the resulting training signal or fixing source-target misalignments. An empirical comparison of these splitting variants showed that *middle-split*, which is the simplest and language-independent method of splitting, is a strong baseline performing almost on par with the more sophisticated variants. Moreover, being *middle-split* language-independent, it can be easily applied to pre-training multi-encoding NMT system on many language pairs with weak or moderate word order divergence.

Chapter 4: Focused Concatenation. We have identified a limitation of the concatenation approach with sliding windows: it is not trained with an explicit focus on the translation quality of the current sentence, which is the only one kept after generation. To overcome this limitation, we have proposed a parameter-free modification of the NMT objective function, referred to as *context discounting*. We analyzed the quality of the model's predictions at the sentence level. We found that context discounting improves the model's predictions of the current target sentence both absolutely and relatively to the context predictions' quality. Moreover, the attention mechanism was found to be less "distracted" by context when trained with context discounting. As a result, the concatenation model exhibited an improved ability to disambiguate inter-sentential discourse phenomena and became robust to larger context sizes. In addition, we introduced a mechanism to increase the distance between the representations of tokens belonging to different sentences: segment-shifted position embeddings. They resulted in a more focused distribution of attention weights and marginal improvements on contrastive test sets. Finally, we investigated the interaction between *context discounting* and *divide and rule* applied to the concatenation approach. We found them to be synergistic when position embeddings are not segment-shifted.

Chapter 5: Sentence Position Encodings. We proposed to encode sentence membership and sentence position into token representations to improve context-aware translation on concatenation sequences. We conducted a comparative evaluation of four approaches: segment-shifted position embeddings, one-hot segment embeddings, sinusoidal segment embeddings, and learned segment embeddings. Moreover, we introduced and evaluated two novel methods to integrate sentence position embeddings into the model: *persistent embeddings* and *position-segment embeddings*. All the combinations of encodings and integration methods have been evaluated empirically on the English-to-Russian language pair. We found that the standard concatenation model does not benefit from these approaches, while the context-discounted concatenation approach does. Persistent segment-shifted position embeddings and persistent segment embeddings were found particularly effective in improving performance on the disambiguation of inter-sentential discourse phenomena. Subsequently, we have trained and evaluated the most promising approaches on English-

to-German data. Unfortunately, we did not observe any improvement over the vanilla or context-discounted concatenation approach in this case. Therefore we could not make general conclusions about the proposed approaches.

6.2 Perspectives

In the course of this research, several new questions have emerged that warrant further investigation. This section delves into potential avenues for future research and open questions.

6.2.1 Online translation with concatenation

Some applications require translating documents in real time. As new sentences become available in the source language, they have to be translated into the target language. For example, some commercial exchange platforms, such as Airbnb,[1] offer an automatic chat translation service, translating each user's dialogue turn into their interlocutor's language to ease communication. DeepL[2] constitutes another example. Their machine translation interface translates a text as soon as it is typed. If the user continues to type sentences in succession, the new sentences are translated while keeping the existing context unchanged. Avoiding re-translating the context saves computational resources. Moreover, re-translating past messages in a chat room like Airbnb's is unnecessary and confusing for the user.

The concatenation approach could be adapted to these online scenarios to produce context-aware translations more efficiently and coherently. As we have seen in Section 2.4.1, the SlidingKtoK approach translates both the current sentence and its context, then discards the translation of the context. In cases where the document is translated sequentially, the past context has already been translated and can be used to guide the generation of the current sentence. Instead of generating the past target context again, one could provide it to the decoder to condition the generation of the current sentence. In the case of a chat, for instance, the system can translate the current message by conditioning on the past messages that have already been translated, employing them as both source-side and target-side context. Restraining generation to the current sentence would make inference more efficient. Moreover, conditioning on what has already been translated should improve the coherency of the target text. Evaluating translation quality in this scenario could also

[1] https://www.airbnb.it/
[2] https://www.deepl.com/translator

provide insights for some non-online scenarios. For example, if the sequential translation of sentences is found to improve quality, batch translation efficiency and speed could be traded off in non-online scenarios.

6.2.2 Efficient attention for long context

As discussed in Section 2.4.4, a major drawback of concatenation approaches is related to the computational complexity of self-attention, which scales quadratically with sequence length. In recent years, several studies have focused on modifying self-attention to reduce its complexity to a subquadratic scale with respect to sequence length (Lin et al., 2022; Tay et al., 2022). Recent works have also investigated the employment of efficient attention alternatives for long-context-aware translation with concatenation (Petrick et al., 2022; Wu et al., 2022). Petrick et al. (2022) adopted the locality sensitive-hashing approach by Kitaev et al. (2020) for the Transformer's attention layers, while Wu et al. (2022) employed the random feature attention by Peng et al. (2021). Both works observed relevant speedups during inference compared to the Transformer with standard attention but at the expense of the accuracy on contrastive test sets for the disambiguation of discourse phenomena (on ContraPro (Müller et al., 2018), and Voita's (Voita et al., 2019b), respectively). Petrick et al. (2022) also found efficient attention to underperform standard attention in terms of BLEU on both short and long concatenation windows, while Wu et al. (2022) found performances to be comparable on short windows. In some preliminary experiments, we also found discouraging performances on the English-French language pair when substituting standard attention with the efficient Luna attention by Kitaev et al. (2020). Efficient self-attention approximations might be bound to underperform it on the translation task. However, efficient attention may effectively solve long-distance dependencies currently unattainable by standard attention.

To the best of our knowledge, efficient self-attention alternatives have not yet been studied in the context of multi-encoding approaches to CANMT. We think that a multi-encoding architecture that employs both standard and efficient self-attention can exploit the advantages of both. Following a multi-resolution approach, standard attention could be used to process the current sentence alone, which requires the highest precision, while the concatenation of the current and context sentences could be processed with an efficient attention alternative in order to produce a contextualized representation that will be merged with the sentence-level one. Alternatively, both standard and efficient attention could be adopted in a modified concatenation approach similar to the one proposed by Zhang et al. (2020a) (see Section 2.4.4). In this case, local self-attention could be performed as usual, while global attention would be based on an efficient alternative.

6.2.3 Long-range arena

As research progresses in the field of deep learning and CANMT, it will be possible to accurately model more and more context, with the hope of achieving further performance improvements. However, it is still unclear how valuable long context (more than four sentences) is to improve the performance of existing CANMT models. In fact, although some attempts to develop long-context-aware NMT architectures have already been published (Zheng et al., 2020; Sun et al., 2020; Petrick et al., 2022; Wu et al., 2022), there is a lack of evaluation methods to measure the improvements achieved with the inclusion of long context. Their results showed degraded performance in the disambiguation of discursive phenomena spanning a few sentences, but the results on discursive phenomena beyond a few sentences remain unknown. Therefore, we deem it worthwhile to analyze the marginal performance achievable with the inclusion of long context, following the methodologies adopted for similar investigations on the value of few-sentences context (Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019b). We expect some improvement to be achievable, albeit marginal. We also expect that as machines approach human quality in document-level translation, such improvements will become necessary to close the gap. If our expectations are confirmed, developing new contrastive sets for long-range discourse phenomena would be very useful for the research community. Following the same line of research, it would be interesting to evaluate the effectiveness of recently published automatic metrics (Jiang et al., 2022; Vernikos et al., 2022) in measuring improvements in long-range discursive phenomena and possibly adapt them for this purpose.

Appendix A

Divide and Rule

A.1 Details on experimental setup

All models are implemented in *fairseq* (Ott et al., 2019). After having pre-trained the baseline on 4 Tesla V100 for 200k steps, we train all models on a single Quadro RTX 6000, with a fixed batch size of approximately 16k tokens,[1] as it has been shown that Transformers need a large batch size for achieving the best performance (Popel and Bojar, 2018). We stop training after five consecutive non-improving validation steps (regarding the loss on dev). In Table A.1 we indicate the performance on the development set of the models reported in Table 3.5. We train models with the optimizer configuration and learning rate schedule described in Vaswani et al. (2017). The maximum learning rate is 0.0007 for baselines on En→Ru/De, 0.001 for models on En→De/Fr low resource settings, and 0.0005 for all the others. In the En→De/Fr High Resource setting, contextual parameters are finetuned on IWSLT17 with an initial learning rate of 0.0002 that shrinks by a factor of 0.99 at every epoch. We use label smoothing with an epsilon value of 0.1 (Pereyra et al., 2017) for all settings. Since the sentence-level parameters are pre-trained on a large amount of parallel data (WMT), the models are pretty robust to generalization, and dropout can be set to 0.1, which gave the best results for the non-contextual baseline *K1*. At inference time, we use beam search with a beam of 4 for all models. We adopt a length penalty of 0.6 for all models ($P_{len} < 1$ favors shorter sentences), except *-d&r* En→Fr models, to which we assign $P_{len} = 1$. The learning rate (searched in $\{0.001, 0.0007, 0.0005, 0.0002\}$) and the length penalty (searched in $[0.4, 1.5]$) have been finetuned manually for each model to achieve the best accuracy on the validation set. The other hyper-parameters were set accordingly to the relevant literature (Vaswani et al., 2017; Popel and Bojar, 2018; Voita

[1] The optimizer update is delayed to simulate the 16k tokens.

Model	Setting	En→Ru	En→De	En→Fr
Concat2to1	Low Res	3.624	3.628	3.207
Concat2to1	High Res	3.659	3.734	3.228
base	-	3.626	3.629	3.230
K2	Low Res	3.599	3.617	3.216
K4	Low Res	3.605	3.618	3.215
K2	High Res	3.596	3.617	3.210
K4	High Res	3.597	3.617	3.211
K2-d&r	Low Res	3.595	3.617	3.213
K4-d&r	Low Res	3.595	3.616	3.212
K2-d&r	High Res	3.593	3.616	3.211
K4-d&r	High Res	3.592	3.615	3.211

Table A.1 – Corresponding loss on the development set for each model reported in Table 3.5.

et al., 2019b; Lopes et al., 2020).

A.2 Ablation: segment embeddings

In Table A.2, we show the results of context-aware models equipped with the addiction of sinusoidal segment embeddings that allow the models to tell the distance of each context sentence from the current one. In general, we measure minor differences with respect to the performance of the models without segment embeddings, reported in Table 3.5. We also experimented with learned segment embeddings, like the ones employed by Devlin et al. (2019), but the results were no better.

		En→De		En→Fr	
Model	Setting	BLEU	**ContraPro**	BLEU	**ContraPro**
K2	Low Res	33.14 (+0.00)	46.32 (-0.73)	n.a.	n.a.
K4	Low Res	33.16 (+0.30)	46.42 (-0.06)	n.a.	n.a.
K4	High Res	33.16 (+0.06)	47.5 (-3.64)	41.55 (-0.18)	83.50 (+0.56)
K2-d&r	Low Res	33.33 (-0.11)	58 (-2.21)	41.77 (-0.01)	83.89 (-0.17)
K4-d&r	Low Res	33.31 (-0.05)	58.5 (+2.28)	41.81 (+0.13)	85.41 (-0.09)

Table A.2 – BLEU and accuracy results on ContraPro of multi-encoding models equipped with sinusoidal segment embeddings. In brackets, we report the changes with respect to the performance obtained without segment embedding, as reported in Table 3.5.

Appendix B

Focused Concatenation

B.1 Details on experimental setup

All models are implemented in *fairseq* (Ott et al., 2019) and follow the *Transformer-base* architecture (Vaswani et al., 2017): hidden size of 512, feed forward size of 2048, 6 layers, 8 attention heads, for a total of 60.7M parameters. They are trained on 4 Tesla V100, with a fixed batch size of approximately 32k tokens for En→Ru and 16k for En→De, following the recommendation of using a large batch size for Transformers (Popel and Bojar, 2018). We stop training after 12 consecutive non-improving validation steps, following a slightly different stopping criterion for each language pair. For En→Ru we adopt the discounted loss $\mathcal{L}_{\text{CD}}$, while for En→De we adopt the full loss $\mathcal{L}$, equivalent to $\mathcal{L}_{\text{CD}}$ with $CD = 1$. Before testing, we average the weights of the 5 model checkpoints that are closest to the best-performing checkpoint, included. We train models with the optimizer configuration and learning rate (LR) schedule described in Vaswani et al. (2017). The maximum LR is optimized for each model over the search space {9E-5, 2E-4, 5E-4, 7E-4, 9E-4, 1E-3, 2E-3, 3E-3}. We select the LR converging to the best loss on the validation set. We use label smoothing with an epsilon value of 0.1 (Pereyra et al., 2017) for all settings. We adopt strong model regularization (dropout=0.3) following Kim et al. (2019) and Ma et al. (2021b). At inference time, we use beam search with a beam of 4 for all models. We adopt a length penalty of 0.6 for all models. The other hyperparameters were set according to the relevant literature (Vaswani et al., 2017; Popel and Bojar, 2018; Voita et al., 2019b; Ma et al., 2021b; Lopes et al., 2020).

B.1.1 Statistical hypothesis tests

We perform statistical hypothesis testing with McNemar's test (McNemar, 1947) for comparing accuracy results on the contrastive test sets. For comparing BLEU performances and mean entropy (Table 4.10), we use approximate randomization (Riezler and Maxwell, 2005) with 10000 and 1000 permutations, respectively. For COMET, the official library[1] has a built in tool for the calculation of statistical significance with Paired T-Test and bootstrap resampling (Koehn, 2004).

B.2 Details on experimental results

Below, we report more details about the results presented in our Tables and Figures.

B.2.1 Details of the evaluation on discourse phenomena

In Tables B.1 and B.2, we document the accuracy achieved on the different subsets of the contrastive sets by each model that we have presented in Chapter 4. For Voita's En→Ru contrastive set (Voita et al., 2019b), we report the accuracy on each of the 4 discourse phenomena included in it; for the En→De ContraPro (Müller et al., 2018), the accuracy on anaphoric pronouns with antecedents at different distances $d = 1, 2, ...$ (in number of sentences). Our approach mostly outperforms baselines and other variants on the majority of the evaluation subsets. We complement Voita/ContraPro with two other metrics on discourse phenomena: Voita/ContraPro$_{\text{avg}}$ and Voita/ContraPro$_{\text{alld}}$, calculated as follow:

$$\text{Voita/ContraPro} = \frac{7075 * (\text{d=1}) + 1510 * (\text{d=2}) + 573 * (\text{d=3}) + 442 * (\text{d>3})}{9600} \tag{B.1}$$

$$\text{Voita/ContraPro}_{alld} = \frac{2400*(\text{d=0})+7075*(\text{d=1})+1510*(\text{d=2})+573*(\text{d=3})+442*(\text{d>3})}{12000} \tag{B.2}$$

$$\text{Voita/ContraPro}_{\text{avg}} = \frac{(\text{d=1}) + (\text{d=2}) + (\text{d=3}) + (\text{d=4})}{4} \tag{B.3}$$

Voita/Contrapto represents the overall accuracy on the contrastive set equivalent to the average of each subset weighted by its sample size (indicated in the last row). For En→De we exclude d=0, as it is the subset of anaphoric pronouns whose antecedent belongs to the same sentence, and thus they don't require context. We include d=0 in Voita/ContraPro$_{\text{alld}}$. Instead, Voita/ContraPro$_{\text{avg}}$ represents the average accuracy as if each subset of the contrastive sets had the same size. While Voita/ContraPro is a proxy

[1] https://github.com/Unbabel/COMET

of the ability to correctly translate a distribution of inter-sentential discourse phenomena as represented in the contrastive set, Voita/ContraPro$_{\text{alld}}$ is a proxy for the average ability to translate each of the inter-sentential phenomena under evaluation. Interestingly, Voita/ContraPro$_{\text{alld}}$ captures more evidently than Voita/ContraPro the improvement achieved by adding segment-shifted positions to the context-discounted concatenation models.

B.2.2 Numerical values of presented figures

In this section, we report the numerical values corresponding to the figures displayed in Chapter 4, if they were not already reported.

	En→Ru						
System	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	$\text{Voita}_{\text{avg}}$	Max LR
base	50.00	45.87	51.80	27.00	46.64	43.67	1E-03
s4to1	50.00	45.87	57.60	71.40	51.66	56.22	9E-04
s4to4	85.80	46.13	79.60	73.20	72.02	71.18	1E-03
s4to4 + CD	**87.16**	46.40	81.00	78.20	73.42	73.19	9E-04
s4to4 + shift + CD	85.76	**48.33**	**81.40**	**80.40**	**73.56**	**73.97**	1E-03
s4to4 + 100 + CD	85.60	**48.73**	**80.80**	**79.60**	**73.46**	**73.68**	1E-03
s4to4 + avg-seq + CD	**87.24**	48.07	79.40	78.80	73.86	73.38	1E-03
s2to2	61.84	46.07	74.60	69.00	59.10	62.88	2E-03
s2to2 + CD	**62.88**	46.27	78.00	**71.60**	60.28	64.69	1E-03
s2to2 + shift + CD	62.60	**46.60**	**81.20**	71.40	**60.54**	**65.45**	2E-03
s3to3	73.52	45.87	78.00	72.60	65.58	67.50	9E-04
s3to3 + CD	73.88	**46.80**	**82.40**	**78.00**	**67.02**	**70.27**	9E-04
s3to3 + shift + CD	**75.24**	46.07	79.40	76.00	66.98	69.18	1E-03
Chen et al. (2021)	62.30	47.90	64.90	36.00	55.61	52.78	n.a.
Sun et al. (2022)	64.70	46.30	65.90	53.00	58.13	57.48	n.a.
Zheng et al. (2020)	61.30	58.10	72.20	80.00	63.30	67.90	n.a.
Kang et al. (2020)	79.20	**62.00**	71.80	80.80	73.46	73.45	n.a.
Zhang et al. (2020a)	**91.00**	46.90	**78.20**	**82.20**	**75.61**	**74.58**	n.a.
s4to4 + CD=1.000	84.80	46.00	75.80	75.00	71.28	70.40	9E-04
s4to4 + CD=0.900	85.64	46.20	77.40	71.80	71.60	70.26	9E-04
s4to4 + CD=0.700	85.52	46.13	79.20	73.00	71.82	70.96	9E-04
s4to4 + CD=0.500	86.04	46.33	77.40	73.80	72.04	70.89	9E-04
s4to4 + CD=0.300	86.68	46.07	77.00	76.60	72.52	71.59	9E-04
s4to4 + CD=0.100	86.56	46.47	79.20	75.20	72.66	71.86	9E-04
s4to4 + CD=0.010	87.16	46.40	81.00	78.20	73.42	73.19	9E-04
s4to4 + CD=0.009	87.80	**47.00**	79.80	76.60	73.64	72.80	9E-04
s4to4 + CD=0.007	**87.96**	46.93	81.40	**79.00**	**74.10**	**73.82**	9E-04
s4to4 + CD=0.005	87.64	46.87	**82.80**	76.20	73.78	73.38	9E-04
s4to4 + CD=0.003	87.52	46.47	80.60	78.20	73.58	73.20	9E-04
s4to4 + CD=0.001	87.28	**47.00**	81.40	77.80	73.66	73.37	9E-04
s4to4 + CD=0.000	86.48	46.27	70.00	78.60	71.98	70.34	9E-04
s4to4 + *d&r*	86.20	46.07	72.20	67.00	70.84	67.87	9E-04→2E-04
s4to4 + CD + *d&r*	**90.24**	45.93	80.00	76.00	**74.50**	73.04	2E-03→1E-04
s4to4 + shift + CD + *d&r*	87.08	**46.33**	**81.60**	**77.60**	73.36	**73.15**	9E-04→9E-05
Sample size	2500	1500	500	500	5000	5000	-

Table B.1 – Accuracy on the En→Ru contrastive set for the evaluation of discourse phenomena (Voita, %), and on its 4 subsets: deixis, lexical cohesion, inflection ellipsis, and verb phrase ellipsis. $\text{Voita}_{\text{avg}}$ denotes the average on the 4 discourse phenomena, while Voita represents the average weighted by the frequency of each phenomenon in the test set (see row "Sample size"). The last column represents the maximum learning rate that was selected for training each model. In the case of models pre-trained with divide and rule (*d&r*), we provide the Max LR for both pre-training and training.

	En→De								
System	d=0	d=1	d=2	d=3	d>3	CP	CP_{avg}	CP_{alld}	Max LR
base	68.75	32.89	43.97	47.99	70.58	37.27	48.86	43.57	3E-03
s4to1	73.83	36.90	46.55	49.38	69.68	40.67	50.63	47.30	7E-04
s4to4	75.20	68.89	74.96	79.58	**87.78**	71.35	77.80	72.12	1E-03
s4to4 + `CD`	**76.66**	**72.86**	75.96	80.10	84.38	74.31	78.33	**74.78**	9E-04
s4to4 + shift + `CD`	75.25	72.56	**77.15**	**80.27**	86.65	**74.39**	**79.16**	74.56	9E-04
(Maruf et al., 2019)	68.60	34.70	46.40	51.10	70.10	39.15	50.58	45.04	n.a.
(Müller et al., 2018)	75.00	39.00	48.00	54.00	66.00	42.55	51.75	49.04	n.a.
(Stojanovski and Fraser, 2019)	74.00	53.00	46.00	50.00	71.00	52.55	55.00	56.84	n.a.
(Lupo et al., 2022a)	**81.10**	56.50	44.90	48.70	73.30	54.98	55.85	**60.21**	n.a.
(Müller et al., 2018)	65.00	**58.00**	**55.00**	**55.00**	**75.00**	**58.13**	60.75	59.51	n.a.
s4to4 + `CD=1.000`	74.08	67.97	74.63	78.88	84.61	70.43	76.52	71.16	9E-04
s4to4 + `CD=0.900`	73.87	66.60	74.50	79.40	84.61	69.44	76.28	70.32	9E-04
s4to4 + `CD=0.700`	73.25	67.74	75.09	79.93	87.10	70.52	77.47	71.06	9E-04
s4to4 + `CD=0.500`	73.33	67.92	73.90	78.53	87.10	70.38	76.86	70.97	9E-04
s4to4 + `CD=0.300`	75.16	69.92	**76.75**	**82.19**	86.42	72.49	**78.82**	73.02	9E-04
s4to4 + `CD=0.100`	74.45	66.93	74.03	76.96	87.33	69.58	76.31	70.56	9E-04
s4to4 + `CD=0.010`	**76.66**	**72.86**	75.96	80.10	84.38	**74.31**	78.33	**74.78**	9E-04
s4to4 + `CD=0.009`	76.16	70.06	74.83	76.61	85.97	71.93	76.87	72.78	9E-04
s4to4 + `CD=0.007`	75.16	70.57	76.49	79.75	85.06	72.72	77.97	73.20	9E-04
s4to4 + `CD=0.005`	73.91	68.94	75.23	78.70	85.06	71.25	76.98	71.79	9E-04
s4to4 + `CD=0.003`	75.04	68.62	75.43	78.01	**87.78**	71.13	77.46	71.92	9E-04
s4to4 + `CD=0.001`	74.37	61.49	69.27	74.34	84.38	64.53	72.37	66.50	9E-04
s4to4 + `CD=0.000`	67.62	57.35	67.81	71.72	85.29	61.14	70.54	62.44	9E-04
s4to4 + *d&r*	67.49	73.90	78.53	87.10	70.06	75.24	77.40	73.69	1E-03→1E-04
s4to4 + `CD` + *d&r*	**72.97**	**77.01**	**78.88**	**87.55**	**74.63**	**77.82**	**79.52**	**76.85**	1E-03→1E-04
s4to4 + shift + `CD` + *d&r*	71.15	75.89	76.43	86.42	72.91	76.47	77.91	75.40	9E-04→2E-04
Sample size	2400	7075	1510	573	442	9600	9600	12000	-

Table B.2 – Accuracy on the En→De contrastive set for the evaluation of anaphoric pronouns (CP = ContraPro, %). The first 5 columns represent the accuracy for each subset of pronouns with antecedents at a specific distance $d \in [0, 1, 2, 3, > 3]$ (in number of sentences). It should be noted that here we are reporting results on $d = 0$ too, differently from what we did in the tables of Chapter 4. CP_{avg} denotes the average on the 4 subsets of pronouns with extra-sentential antecedents ($d > 0$) while CP represents the average weighted by the size of each of the 4 subsets (see row "Sample size"). CP_{alld} is equivalent to CP, but it includes the accuracy on $d = 0$. The last column represents the maximum learning rate that was selected for training each model. In the case of models pre-trained with divide and rule (*d&r*), we provide the Max LR for both pre-training and training.

	Current Loss			Loss Ratio		
epoch	CD=1	CD=0.5	CD=0.01	CD=1	CD=0.5	CD=0.01
1	2.570	2.765	2.701	0.997	0.891	0.705
2	1.859	1.923	1.895	0.994	0.970	0.607
3	1.672	1.707	1.712	0.989	0.975	0.916
4	1.606	1.627	1.635	0.987	0.979	0.940
5	1.590	1.601	1.605	0.987	0.977	0.946
6	1.581	1.577	1.580	0.986	0.979	0.951
7	1.580	1.570	1.571	0.985	0.981	0.952
8	1.590	1.564	1.564	0.986	0.980	0.954
9	1.589	1.567	1.566	0.985	0.982	0.958
10	1.592	1.576	1.569	0.986	0.979	0.954
11	1.594	1.568	1.562	0.984	0.981	0.960
12	1.598	1.569	1.567	0.984	0.981	0.962
13	1.601	1.576	1.574	0.983	0.984	0.965
14	1.605	1.574	1.572	0.984	0.983	0.966
15	1.610	1.581	1.573	0.983	0.982	0.965
16	1.611	1.582	1.573	0.982	0.981	0.964
17	1.617	1.584	1.579	0.983	0.981	0.966
18	1.620	1.587	1.580	0.982	0.981	0.965
19	1.624	1.589	1.579	0.984	0.982	0.966
20	n.a.	1.593	1.582	n.a.	0.981	0.966
21	n.a.	1.594	1.587	n.a.	0.981	0.967
22	n.a.	1.591	1.584	n.a.	0.983	0.967
23	n.a.	n.a.	1.589	n.a.	n.a.	0.969
24	n.a.	n.a.	1.587	n.a.	n.a.	0.971
25	n.a.	n.a.	1.587	n.a.	n.a.	0.971
26	n.a.	n.a.	1.591	n.a.	n.a.	0.970

Table B.3 – Numerical values corresponding to Figures 4.3 and 4.4.

	Current Loss			Loss Ratio		
epoch	CD=1	CD=0.5	CD=0.01	CD=1	CD=0.5	CD=0.01
1	2.606	2.604	2.601	1.002	0.997	0.988
2	2.306	2.307	2.314	1.000	0.988	0.968
3	2.105	2.100	2.133	1.000	0.987	0.957
4	1.969	1.973	1.985	1.000	0.984	0.934
5	1.755	1.814	1.858	1.017	0.983	0.927
6	1.598	1.707	1.773	1.039	1.039	0.939
7	1.410	1.466	1.631	1.018	1.010	0.945
8	1.301	1.336	1.505	1.013	1.008	0.969
9	1.221	1.250	1.406	1.003	0.999	0.957
10	1.177	1.195	1.310	1.002	0.997	0.976
11	1.152	1.160	1.253	1.003	1.000	0.978
12	1.136	1.151	1.220	0.999	1.001	0.992
13	1.119	1.121	1.183	1.002	0.997	0.979
14	1.107	1.106	1.156	0.999	0.997	0.986
15	1.101	1.101	1.138	0.998	0.996	0.984
16	1.103	1.098	1.123	1.002	0.996	0.988
17	1.097	1.093	1.114	1.002	0.997	0.988
18	1.095	1.092	1.108	0.999	0.997	0.987
19	1.098	1.091	1.099	0.999	0.994	0.987
20	1.101	1.094	1.098	1.000	0.997	0.989
21	1.097	1.093	1.096	0.998	0.996	0.986
22	1.099	1.092	1.092	0.999	0.996	0.989
23	1.103	1.097	1.090	0.997	0.995	0.989
24	1.107	1.098	1.087	1.001	0.997	0.987
25	1.104	1.099	1.086	0.997	0.996	0.989
26	1.105	1.100	1.087	0.998	0.996	0.989
27	1.113	1.105	1.085	1.000	0.996	0.988
28	1.115	1.106	1.086	0.997	0.995	0.991
29	1.114	1.107	1.086	0.998	0.996	0.991
30	n.a.	1.107	1.087	n.a.	0.997	0.991
31	n.a.	n.a.	1.088	n.a.	n.a.	0.990
32	n.a.	n.a.	1.088	n.a.	n.a.	0.990
33	n.a.	n.a.	1.087	n.a.	n.a.	0.991
34	n.a.	n.a.	1.088	n.a.	n.a.	0.991
35	n.a.	n.a.	1.091	n.a.	n.a.	0.994
36	n.a.	n.a.	1.092	n.a.	n.a.	0.993
37	n.a.	n.a.	1.095	n.a.	n.a.	0.993
38	n.a.	n.a.	1.094	n.a.	n.a.	0.993
39	n.a.	n.a.	1.095	n.a.	n.a.	0.993
40	n.a.	n.a.	1.096	n.a.	n.a.	0.993

Table B.4 – Numerical values corresponding to Figures 4.5 and 4.6.

System	Inference	$\text{Voita}_{\text{avg}}$	BLEU
s2to2	2.00	62.88	32.73
s2to2	3.00	62.78	23.50
s2to2	4.00	61.47	20.59
s2to2 + `CD`	3.00	64.43	32.05
s2to2 + `CD`	4.00	62.15	31.67
s3to3	3.00	67.50	32.34
s3to3	4.00	70.10	30.60
s3to3 + `CD`	3.00	67.02	32.42
s3to3 + `CD`	4.00	70.25	32.42

Table B.5 – Numerical values corresponding to Figure 4.7, with extended details. Our approach improves robustness of En→Ru s2to2 and s3to3 to window sizes unseen during training.

Appendix C

Encoding Sentence Position

C.1 Allocating more space to segments in PSE

For the English-to-Russian language pair, we have found that one-hot and sinusoidal segment embeddings need to be integrated into PSE for being leveraged by s4to4+CD (Section 5.3.2.1). Instead, learned embeddings worked best when added to position segment embeddings.

Here, we evaluate whether PSE with learned segment embeddings would perform better if more dimensions were allocated to segments. In particular, we try $d_{SE} = 128$, which leaves $d_{PE} = d_{model} - d_{SE} = 384$ dimensions to position embeddings, largely enough as shown in Section 5.2.2.

As shown in Table C.1, increasing the number of dimensions allocated to segment embeddings deteriorates the performance on Voita's contrastive set. The reason could be that learning a meaningful representation of sentence positions with more dimensions is harder.

System	Enc.	Pers.	PSE	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	BLEU
s4to4+CD	lrn	✓	4	93.20	47.40	72.20	64.40	74.48	32.35
s4to4+CD	lrn		128	83.88	46.33	65.20	50.20	67.38	32.43
s4to4+CD	lrn	✓	128	78.20	46.40	40.60	30.60	60.14	32.35

Table C.1 – s4to4 trained on En→Ru OpenSubtitles. Accuracy on Voita's En→Ru contrastive set and BLEU on the test set. The accuracy on the contrastive set is detailed on the left, with the accuracy on each subset corresponding to a specific discourse phenomenon. Result: allocating more dimensions to segments in PSE deteriorates performance.

C.1.1 Details of the evaluation on discourse phenomena

Table C.2 reports more details about the evaluation on the En→De settings, accordingly to the discussion presented in Appendix B.2.1.

System	Enc.	Pers.	PSE	d=0	d=1	d=2	d=3	d>3	CP	CP_{avg}	CP_{alld}
base				68.75	32.89	43.97	47.99	70.58	37.27	48.86	43.57
s4to4				75.20	68.89	74.96	79.58	87.78	71.35	77.80	72.12
s4to4+CD				76.66	72.86	75.96	80.10	84.38	74.31	78.33	74.78
s4to4+CD	shift			75.25	72.56	77.15	80.27	86.65	74.39	79.16	74.56
s4to4+CD	shift	✓		72.41	69.15	74.23	77.13	86.42	71.22	76.73	71.46
s4to4+CD	sin			76.75	71.83	76.82	80.97	87.55	73.88	79.29	74.46
s4to4+CD	sin	✓		76.50	72.08	76.35	79.23	85.97	73.82	78.41	74.35
s4to4+CD	sin	✓	✓	77.25	71.22	76.42	78.88	86.87	73.22	78.35	74.02
s4to4+CD	lrn			73.91	70.21	75.29	77.66	85.06	72.14	77.06	72.49
s4to4+CD	lrn	✓		73.66	68.53	72.51	75.74	86.65	70.42	75.86	71.07
s4to4+CD	lrn	✓	✓	73.54	68.40	79.07	80.27	83.48	71.48	77.81	71.89
High Resource Setting											
base				82.83	35.18	44.90	51.13	66.28	39.09	49.37	47.84
s4to4				82.41	80.66	81.72	84.29	88.00	81.38	83.67	81.59
s4to4+CD				83.70	81.79	82.11	82.19	90.04	82.24	84.03	82.54
s4to4+CD	shift	✓		81.70	79.61	81.45	83.42	86.65	80.45	82.78	80.70
s4to4+CD	sin	✓	✓	84.12	79.85	82.38	84.46	86.87	80.85	83.39	81.50
s4to4+CD	lrn	✓		83.12	79.13	79.73	82.19	88.00	79.82	82.26	80.48
Sample size				2400	7075	1510	573	442	9600	9600	12000

Table C.2 – Accuracy on the En→De contrastive set for the evaluation of anaphoric pronouns (CP = ContraPro, %). The first d=* columns represent the accuracy for each subset of pronouns with antecedents at a specific distance $d \in [0, 1, 2, 3, > 3]$ (in number of sentences). It should be noted that here we are reporting results on $d = 0$ too, differently from what we did in the tables of Chapter 5. CP_{avg} denotes the average on the 4 subsets of pronouns with extra-sentential antecedents ($d > 0$) while CP represents the average weighted by the size of each of the 4 subsets (see row "Sample size"). CP_{alld} is equivalent to CP, but it includes the accuracy on $d = 0$.

Appendix D

Résumé du mémoire

D.1 Introduction

La traduction automatique (MT) est un domaine de la linguistique computationnelle qui étudie la traduction par un logiciel d'un texte source dans une langue cible. Au cours des dernières décennies, la mondialisation de l'information et des économies a alimenté le besoin de traductions à grande échelle, ce qui a favorisé les progrès de la MT. Trois paradigmes ont guidé l'étude de nouveaux systèmes de traduction en cours de route : la traduction automatique basée sur des règles (RMT), basée sur la statistique (SMT) et sur les réseaux neuronaux (NMT). Alors que la RMT repose sur des règles linguistiques formulées par des experts, la SMT et la NMT consistent en des méthodes d'apprentissage automatique qui apprennent la tâche de traduction directement à partir d'une grande quantité de données textuels accompagnés par leur traduction. En particulier, les systèmes NMT sont basés sur des réseaux neuronaux qui apprennent de manière supervisée. Chaque exemple d'apprentissage consiste en une phrase source et une traduction de référence, généralement réalisée par un traducteur humain. Le réseau tente de traduire la phrase source, et son résultat est comparé à la traduction de référence. Une mesure de la distance entre la sortie et la référence quantifie ce que l'on appelle le *perte d'apprentissage* (*training loss* en anglais), l'"erreur" du système. Ensuite, les paramètres d'apprentissage du réseau sont ajustés en fonction de la perte, en utilisant des techniques d'optimisation standard comme la descente de gradient stochastique, afin de minimiser la différence entre la sortie et la traduction de référence dans les itérations futures. La NMT a connu des améliorations substantielles ces dernières années, principalement favorisées par l'avènement du mécanisme d'attention (Bahdanau et al., 2015a) et du modèle Transformer (Vaswani et al., 2017). Si les systèmes actuels de NMT ont atteint une qualité proche de celle de l'homme dans la

traduction de phrases décontextualisées (Wu et al., 2016), ils ont encore une grande marge d'amélioration lorsqu'il s'agit de traduire des documents (Läubli et al., 2018) tels que des articles, des livres, des chats, des transcriptions d'événements ou des sous-titres vidéo.

La traduction d'un document nécessite de prendre en compte les relations linguistiques entre ses phrases. En d'autres termes, de les contextualiser. Étant donné une phrase dans un document, nous pouvons nous référer au reste du document comme au contexte *au delà* de la phrase ou, en bref, à son contexte. La nécessité de contextualiser la traduction d'une phrase source donnée $\boldsymbol{x}^i$ peut apparaître même dans des documents concis. Par exemple, si nous prenons la phrase anglaise :

$\boldsymbol{x}^i$ (EN): How are you today ?
$\boldsymbol{y}^{i,1}$ (FR): *Comment vas-tu aujourd'hui?*
$\boldsymbol{y}^{i,2}$ (FR): *Comment allez-vous aujourd'hui?*

Il existe plusieurs traductions françaises qui sont valables si le contexte n'est pas fourni, comme $\boldsymbol{y}^{i,1}$ et $\boldsymbol{y}^{i,2}$:

$\boldsymbol{x}^i$ (EN): How are you today ?
$\boldsymbol{y}^{i,1}$ (FR): *Comment vas-tu aujourd'hui?*
$\boldsymbol{y}^{i,2}$ (FR): *Comment allez-vous aujourd'hui?*

Cependant, lorsque nous replaçons $\boldsymbol{x}^i$ dans son contexte en fournissant la phrase précédente $\boldsymbol{x}^{i-1}$, il devient clair que $\boldsymbol{y}^{i,2}$ est la seule option valide :

$\boldsymbol{x}^{i-1}$ (EN): Good morning Mr. President.
$\boldsymbol{x}^i$ (EN): How are you today ?
$\boldsymbol{y}^{i,1}$ (FR): *Comment vas-tu aujourd'hui?*
$\boldsymbol{y}^{i,2}$ (FR): *Comment allez-vous aujourd'hui?*

Le titre de l'interlocuteur ("M. le Président") exige l'utilisation du "vous" formel (distinction T-V) et, par conséquent, la conjugaison correcte du verbe correspondant.

Le paradigme NMT a évolué autour de la tâche de traduction agnostique au contexte. Ainsi, jusqu'à récemment, de nombreux systèmes NMT de pointe ne pouvaient pas contextualiser la phrase actuelle au-delà des limites de la phrase (Bojar et al., 2017a,b). Malheureusement, les systèmes NMT agnostiques par rapport au contexte produisent systématiquement des incohérences intersententielles, ce qui nuit à la qualité globale de la traduction (Läubli et al., 2018; Toral et al., 2018; Voita et al., 2019b). À titre d'exemple, jusqu'à fin décembre 2022, Google Translate, l'un des systèmes de traduction automatique les plus populaires au monde, traduisait toujours $\boldsymbol{x}^i$ par $\boldsymbol{y}^{i,1}$, même lorsque

$\boldsymbol{x}^{i-1}$ était fourni comme contexte. Pour remédier à cette limitation, des chercheurs se sont penchés sur la traduction neuronale sensible au contexte (CANMT) depuis début 2017 (Jean and Cho, 2019; Tiedemann and Scherrer, 2017; Wang et al., 2017). On peut classer la majorité des approche de CANMT en deux familles (Kim et al., 2019): les approches à encodage multiple et les approches à encodage unique, également appelées approches de concaténation. La première famille comprend toutes les approches qui utilisent l'architecture standard d'encodeur-décodeur pour produire des représentations latentes de la phrase courante et qui introduisent des modules supplémentaires pour encoder et intégrer son contexte, c'est-à-dire les phrases précédentes ou suivantes. Les approches par concaténation, au contraire, reposent entièrement sur l'architecture standard d'encodeur-décodeur, mais elles concatènent le contexte à la phrase actuelle avant de l'introduire dans le système. Aujourd'hui, cet axe de recherche est toujours en évolution, et aucun consensus n'a encore été atteint sur les meilleures approches contextuelles et méthodes d'évaluation.

Notre recherche vise à donner un aperçu de certains des défis auxquels sont confrontées les approches actuelles du CANMT et à proposer de nouvelles solutions.

D.2 Publications

La plupart des contributions présentées dans le chapitre 3 sont publiées dans :

> Lupo et al. (2022a) - Lupo, L., Dinarelli, M. and Besacier, L. (2022). Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557-4672, Dublin, Ireland.

La plupart des contributions présentées dans le chapitre 4 sont publiées dans :

> Lupo et al. (2022b) - Lupo, L., Dinarelli, M. and Besacier, L. (2022). Focused Concatenation for Context-Aware Neural Machine Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, December 7–8, 2022. Association for Computational Linguistics.

Les contributions présentées dans le chapitre 5 seront soumises au *Fourth Workshop on Insights from Negative Results in NLP*, co-localisé avec EACL 2023.

$\mathbf{x}^{i,1}$	He said that it was a project of peace
$\mathbf{x}^{i,2}$	and unity and that it brought people together .
$\mathbf{y}^{i,1}$	*Il disait que c' était un projet de paix*
$\mathbf{y}^{i,2}$	*et d' unité et qu' il réunissait les gens .*
$\mathbf{x}^{j,1}$	I think single-cell organisms are
$\mathbf{x}^{j,2}$	possible within two years .
$\mathbf{y}^{j,1}$	*Je pense que les organismes unicellulaires*
$\mathbf{y}^{j,2}$	*sont* possibles dans 2 ans .

Figure D.1 – Exemple de paire de phrases provenant de En→Fr IWSLT17, après avoir été jetonisées et divisées au milieu. Après la division, certaines relations syntaxiques s'étendent sur deux segments (soulignées).

D.3 Chapitre 3: diviser et régner

Les système è encodage multiple (ou multi-encodage) sont plus flexibles que les systèmes de concaténation et potentiellement plus efficaces, mais ils ont été critiqués pour agir comme de simples méthodes de régularisation (Kim et al., 2019; Li et al., 2020). Dans certains cas, ils se sont même avérés moins performants que les systèmes agnostiques au contexte sur des tests contrastifs pour la désambiguïsation de phénomènes discursifs (Lopes et al., 2020). Dans ce chapitre, nous répondons à cette critique en montrant que l'entraînement de modèles multi-encodage est difficile pour deux raisons :

- la rareté du *signal d'entraînement contextuel*, c'est-à-dire le signal qui pousse les systèmes à traduire en tenant compte du contexte, qui provient des mots qui ont besoin de contexte pour être correctement traduits ;
- la rareté des mots contextuels pertinents, ceux qui sont nécessaires pour désambiguïser la traduction.

Une façon triviale d'améliorer l'apprentissage contextuel est d'augmenter la quantité de documents d'entraînement. Les grands jeux de documents parallèles (dans les deux langues) ne sont pas toujours disponibles, mais certains travaux ont proposé des techniques d'augmentation des données pour remédier à ce manque : (Sugiyama and Yoshinaga, 2019; Stojanovski et al., 2020; Huo et al., 2020). Cependant, comme nous le montrerons dans notre section expérimentale, cette solution n'est pas efficiente et souvent sous-optimale.

Par conséquent, nous introduisons une nouvelle stratégie de pré-entraînement, *diviser et régner* (*d&r*), qui est basée sur une technique simple et pourtant puissante pour augmenter

le signal d'entraînement contextuel et faciliter l'apprentissage de manière efficace : la division de phrases parallèles en segments (voir Figure D.1). En termes simples, le fait d'alimenter un modèle sensible au contexte avec une séquence de segments incomplets, plus courts et consécutifs, l'oblige à rechercher le contexte (c'est-à-dire les segments environnants) plus fréquemment et facilite la récupération du contexte pertinent puisque les segments sont plus courts. Il en résulte un apprentissage efficient. Nous pré-entraînons des modèles d'encodage multiple sur des ensembles de données fractionnés et les évaluons de deux manières : Le score BLEU et une évaluation contrastive de la traduction de phénomènes discursifs. Nos principales contributions sont les suivantes :

- nous montrons que les modèles de multi-encodage tenant compte du contexte doivent être entraînés avec soin car le signal d'entraînement contextuel est clairsemé, ainsi que les éléments contextuels utiles à la contextualisation;
- nous proposons la stratégie de pré-entraînement *d&r*, qui facilite l'entraînement des paramètres contextuels en divisant les phrases en segments;
- nous proposons 4 méthodes alternatives pour la division des phrases;
- nous soutenons cette stratégie par une analyse de l'impact de la division des phrases sur la distribution des phénomènes discursifs;
- nous démontrons que cette stratégie est à la fois efficace et efficiente, car elle permet aux modèles multi-encodage d'apprendre mieux et plus rapidement qu'en augmentant simplement les données d'entraînement.

D.4 Chapitre 4: concaténation focalisée

Les approches de concaténation ont l'avantage de traiter la tâche de CANMT de la même manière que la NMT agnostique au contexte, ce qui facilite l'apprentissage car les paramètres apprenables responsables de la contextualisation au delà de la phrase sont les mêmes que ceux qui entreprennent la contextualisation dans la phrase. En effet, comme nous l'avons vu dans le chapitre précédent, l'apprentissage des paramètres responsables de la contextualisation au delà de la phrase dans les approches de multi-encodage (θ_C) s'est avéré difficile car le signal d'apprentissage est clairsemé et la tâche de récupération des éléments de contexte utiles est ardue. Malgré sa simplicité, il a été démontré que l'approche de concaténation permet d'obtenir des performances compétitives ou supérieures à celles de systèmes multi-encodage (Lopes et al., 2020; Ma et al., 2021b).

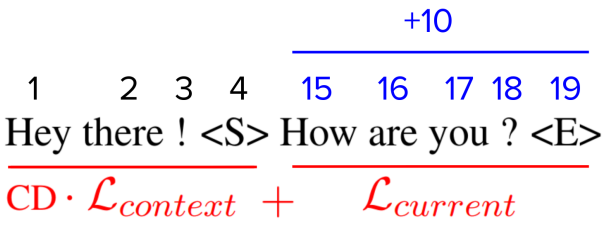

Figure D.2 – Exemple de l'approche proposée appliquée sur une fenêtre de deux phrases, avec une décote sur le contexte CD et des positions décalées par segment d'un facteur 10.

Néanmoins, les systèmes NMT basés sur des transformateurs ont du mal à apprendre les propriétés de localité (Rizzi, 2013) à la fois du langage lui-même et de l'alignement source-cible lorsque la séquence d'entrée augmente en longueur, comme dans le cas de la concaténation. Sans surprise, la présence du contexte rend l'apprentissage plus difficile pour les modèles de concaténation en distrayant l'attention. De plus, le chapitre 3 nous apprend que les systèmes NMT n'ont besoin de contexte que pour un ensemble clairsemé de phénomènes discursifs qui s'étalent sur plusieurs phrases. Par conséquent, il est souhaitable de rendre les modèles de concaténation plus axés sur les phénomènes linguistiques locaux afin d'améliorer les performances. Des travaux récents (Zhang et al., 2020a; Bao et al., 2021) ont démontré qu'une solution viable est l'introduction de biais inductifs forts sur la localité dans l'architecture NMT, comme le masquage partiel du contexte (voir la section 2.4.4 pour plus de détails). Sur la base de ces prémisses, nous proposons une approche de concaténation améliorée pour la CANMT, qui se concentre davantage sur la traduction de la phrase actuelle au moyen de deux solutions simples et qui ne rajoutent pas de paramètres apprenables à l'architecture :

- *la décote sur le contexte* : une simple modification de la fonction de perte qui améliore la traduction contextuelle d'une phrase en rendant le modèle moins distrait par son contexte;
- *les positions décalées par segment* : une modification du système d'encodage des positions des jetons, qui aide à la réalisation de l'objectif de traduction avec décote en soutenant l'apprentissage des propriétés locales des phrases concaténés.

Nous soutenons nos solutions par des expériences, des analyses et des évaluations comparatives approfondies.

D.5 Chapitre 5: encoder la position de la phrase dans les approches de concaténation

Dans le chapitre précédent, nous avons présenté les plongements de position décalés par segment comme un moyen d'aider les approches de concaténation avec décompte à discerner les phrases dans la fenêtre de concaténation. Nous pensons que le fait de fournir des informations explicites sur la position des phrases, au niveau des représentations vectorielles des jetons, aide les approches de concaténation à surmonter le défi d'apprentissage présenté dans le chapitre précédent. En effet, si les représentations latentes des jetons contiennent des informations sur la position des phrases auxquelles ils appartiennent, elles peuvent être traitées par la fonction d'attention en conséquence. Par exemple, l'attention peut reconnaître plus facilement les jetons appartenant à la même phrase, qui ont plus de chances d'être liés les uns aux autres, ainsi que la distance de leur contexte. Comme il est fondamental pour le modèle Transformer d'avoir une notion de la séquentialité des jetons, il est également précieux de connaître l'ordre des phrases dans la fenêtre de concaténation. La structure temporelle du document constitue une information essentielle pour sa compréhension et la désambiguïsation correcte des phénomènes discursifs qui s'étalent sur plusieurs phrases. On pourrait argumenter que l'information nécessaire pour déterminer à quelle phrase appartient un jeton, et sa position dans la fenêtre, est déjà disponible grâce aux plongements de position des jetons. Cependant, les plongements de position ne constituent pas une information directe sur leur appartenance à la phrase, car cette information ne peut être récupérée que par une comparaison avec la position des jetons séparateurs. Nous proposons d'injecter des informations sur l'appartenance à la phrase directement dans les représentations des jetons afin de contourner la nécessité d'apprendre et d'effectuer une telle comparaison.

Ce chapitre présente une étude comparative de différentes approches d'encodage de l'appartenance à la phrase et de sa position, pour les approches de concaténation. En plus des plongements de position décalées par segment, nous évaluerons trois différents types de plongements de segment : one-hot, sinusoïdaux (Vaswani et al., 2017) et appris (Devlin et al., 2019). Inspirés par la littérature sur les plongements de position (Chen et al., 2021; Liu and Zhang, 2020), nous proposons de rendre les plongements de segment persistants sur plusieurs couches, en les ajoutant à l'entrée de chaque couche en plus de la première. De plus, nous proposons de fusionner les plongements de position et de segment en un seul vecteur où les positions des jetons et des segments sont encodées dans deux ensembles orthogonaux de dimensions, ce qui permet une distinction plus claire entre eux, ainsi que des économies de mémoire. À notre connaissance, il s'agit de la première étude comparative sur l'emploi d'encodages de position de phrase pour les approches de concaténation de

CANMT. Bien que quelques études dans la littérature aient adopté des plongements de segments fixes ou appris pour des approches de multi-encodage (Voita et al., 2018; Zheng et al., 2020; Bao et al., 2021), une comparaison complète manque dans ce cas également.

Toutes les combinaisons d'encodages de la position du segment et d'intégration de cette information dans le Transformer ont été évaluées empiriquement sur la paire de langues Anglais-Russe. Nous avons constaté que le modèle de concaténation standard ne bénéficie pas de ces approches, alors que le modèle de concaténation avec décote du contexte en bénéficie. Les plongements persistants de position avec décalage et les plongements de segment persistants se sont avérés particulièrement efficaces pour améliorer les performances sur la désambiguïsation des phénomènes discursifs. Par la suite, nous avons entraîné et évalué les approches les plus prometteuses sur un jeu Anglais-Allemand. Malheureusement, nous n'avons pas observé d'amélioration par rapport à l'approche de concaténation simple ou à l'approche de concaténation avec décote. Nous n'avons donc pas pu tirer de conclusions générales sur les approches proposées en ce chapitre.

D.6 Conclusions

Dans cette thèse de doctorat, nous avons identifié certains des principaux défis auxquels sont confrontées les architectures actuelles d'apprentissage profond pour la traduction sensible au contexte, et proposé des solutions pour les relever. En particulier, nous nous sommes concentrés sur l'aspect apprentissage des approches de multi-encodage et de concaténation, en proposant une solution de pré-entraînement dans le premier cas, et une fonction de perte modifiée dans le second. Les approches proposées se sont avérées efficaces pour améliorer la traduction de documents. Ensuite, nous avons poursuivi le travail sur les approches de concaténation en proposant d'étendre l'architecture standard de Transformer avec des plongements de position décalées par segments ou des plongements de segments. Nous avons entrepris une étude comparative des ces approches, qui n'a malheureusement pas donné de résultats concluants.

Au cours de cette recherche, plusieurs nouvelles questions ont émergé et méritent d'être approfondies. Voici une liste de pistes potentielles pour de futures recherches qui, à notre avis, méritent d'être entreprises.

D.6.1 Traduction en ligne avec concaténation

L'approche de concaténation pourrait être adaptée aux scénarios de traduction en ligne et sensible au contexte (e.g., traduction des chats à chaque fois qu'un nouveau message

est envoyé). L'objectif serait de produire des traductions tenant compte du contexte de manière plus efficace et cohérente. L'approche SlidingKtoK traduit à la fois la phrase actuelle et son contexte, puis rejette la traduction du contexte. Dans les cas où le document est traduit de manière séquentielle, le contexte passé a déjà été traduit et peut être utilisé pour guider la génération de la phrase actuelle. Au lieu de générer à nouveau le contexte cible passé, on pourrait le fournir au décodeur pour conditionner la génération de la phrase actuelle. Dans le cas d'un chat, par exemple, le système peut traduire le message actuel en conditionnant les messages passés qui ont déjà été traduits, en les utilisant comme contexte source et cible. Limiter la génération à la phrase actuelle rendrait l'inférence plus efficiente. De plus, le fait de conditionner sur ce qui a déjà été traduit devrait améliorer la cohérence du texte cible.

D.6.2 Attention efficace pour un contexte long

Ces dernières années, plusieurs études ont porté sur la modification de l'*auto-attention* (*self-attention* en anglais) afin de réduire sa complexité à une échelle sous-quadratique par rapport à la longueur de la séquence (Lin et al., 2022; Tay et al., 2022). Des travaux récents ont également étudié l'emploi d'alternatives d'attention efficientes pour la traduction en contexte long avec concaténation (Petrick et al., 2022; Wu et al., 2022). Petrick et al. (2022) a adopté l'approche de hachage sensible à la localité de Kitaev et al. (2020) pour les couches d'attention du Transformer, tandis que Wu et al. (2022) a utilisé l'attention de caractéristiques aléatoires de Peng et al. (2021). Les deux travaux ont permis d'observer des accélérations pertinentes pendant l'inférence par rapport au Transformer avec l'attention standard, mais au détriment de la précision sur des ensembles de tests contrastés pour la désambiguïsation de phénomènes discursifs (sur ContraPro (Müller et al., 2018), et Voita (Voita et al., 2019b), respectivement). (Petrick et al., 2022) a également constaté que l'attention efficiente était moins performante que l'attention standard en termes de BLEU sur les fenêtres de concaténation courtes et longues, tandis que (Wu et al., 2022) a constaté que les performances étaient comparables sur les fenêtres courtes. Dans certaines expériences préliminaires, nous avons également trouvé des performances décourageantes sur la paire de langues anglais-français lorsque l'on remplace l'attention standard par l'attention efficiente "Luna" de Kitaev et al. (2020). Les approximations de l'attention efficiente pourraient être vouées à sous-performer sur la tâche de traduction. Cependant, l'attention efficiente peut résoudre efficientement les dépendances à longue distance actuellement inaccessibles par l'attention standard.

À notre connaissance, les alternatives efficientes d'auto-attention n'ont pas encore été étudiées en complément des approches de CANMT par multi-encodage. Nous pensons

qu'une architecture multi-encodage qui utilise à la fois l'attention standard et l'auto-attention efficiente peut exploiter les avantages des deux. Suivant une approche multi-résolution, l'attention standard pourrait être utilisée pour traiter la phrase courante seule, ce qui requiert la plus grande précision, tandis que la concaténation des phrases courante et contextuelles pourrait être traitée avec une alternative d'attention efficiente afin de produire une représentation contextualisée qui sera fusionnée avec celle agnostique au contexte.

D.6.3 Évaluation de systèmes sensibles au contexte long

À mesure que la recherche progresse dans le domaine de l'apprentissage profond et de la CANMT, il sera possible de modéliser avec précision de plus en plus de contexte, dans l'espoir d'améliorer toujours plus les performances. Cependant, il n'est pas encore clair dans quelle mesure le contexte long (plus de quatre phrases) est utile pour améliorer les performances des modèles existants. En fait, bien que certaines tentatives de développement d'architectures sensibles aux contexte long aient déjà été publiées (Zheng et al., 2020; Sun et al., 2020; Petrick et al., 2022; Wu et al., 2022), il y a un manque de méthodes d'évaluation pour mesurer les améliorations obtenues avec l'inclusion du contexte long. Leurs résultats ont montré une dégradation des performances dans la désambiguïsation des phénomènes discursifs s'étendant sur quelques phrases, mais les résultats sur les phénomènes discursifs qui s'étendent sur beaucoup de phrases reste inconnus. Par conséquent, nous estimons qu'il est utile d'analyser la performance marginale réalisable avec l'inclusion du contexte long, en suivant les méthodologies adoptées par Bawden et al. (2018); Müller et al. (2018); Voita et al. (2019b). Nous pensons qu'une certaine amélioration est possible, bien que marginale. Nous pensons également qu'à mesure que les machines se rapprochent de la qualité humaine dans la traduction des documents, de telles améliorations deviendront nécessaires pour combler l'écart. Si nos attentes se confirment, le développement de nouveaux ensembles de test contrastifs pour l'évaluation sur les phénomènes discursifs de longue portée serait utile pour la communauté des chercheurs.

Bibliography

Agrawal, R. R., Turchi, M., and Negri, M. (2018). Contextual Handling in Neural Machine Translation: Look Behind, Ahead and on Both Sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 11–20, Alacant, Spain.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer Normalization. *ArXiv preprint*, abs/1607.06450.

Bahdanau, D., Cho, K., and Bengio, Y. (2015a). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bahdanau, D., Cho, K., and Bengio, Y. (2015b). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Bao, G., Zhang, Y., Teng, Z., Chen, B., and Luo, W. (2021). G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

Bawden, R. (2018). *Going beyond the sentence : Contextual Machine Translation of Dialogue*. These de doctorat, Université Paris-Saclay (ComUE).

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*, abs/2004.05150.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017a). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017b). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Bridle, J. S. (1990). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In Soulié, F. F. and Hérault, J., editors, *Neurocomputing*, NATO ASI Series, pages 227–236, Berlin, Heidelberg. Springer.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Cai, X. and Xiong, D. (2020). A test suite for evaluating discourse phenomena in document-level neural machine translation. In *Proceedings of the Second International Workshop of Discourse Processing*, pages 13–17, Suzhou, China. Association for Computational Linguistics.

Carver, R. (2012). The Case Against Statistical Significance Testing. *Harvard Educational Review*, 48(3):378–399.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Chen, L., Li, J., Gong, Z., Chen, B., Luo, W., Zhang, M., and Zhou, G. (2021). Breaking the corpus bottleneck for context-aware neural machine translation with cross-task pre-training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2851–2861, Online. Association for Computational Linguistics.

Chierchia, G. and McConnell-Ginet, S. (2000). *Meaning and Grammar: An Introduction to Semantics*. MIT Press. Google-Books-ID: pxJGet3pKdoC.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). 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.

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlós, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J., and Weller, A. (2021). Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Denk, T. (2019). Linear Relationships in the Transformer's Positional Encoding.

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.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of*

the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Dryer, M. S. (2013a). Order of Adjective and Noun. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Dryer, M. S. (2013b). Order of Adverbial Subordinator and Clause. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Dryer, M. S. (2013c). Order of subject, object and verb. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Dryer, M. S. and Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Dylgjeri, A. and Kazazi, L. (2013). Deixis in Modern Linguistics and Outside. *Academic Journal of Interdisciplinary Studies*.

Eberhard, D. M., Simons, G. F., and Fenning, C. D. (2021). Ethnologue: Languages of the World. Library Catalog: www.ethnologue.com.

Fellbaum, C. (1998). A Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities*, 32(2):209–220. 00194.

Fernandes, P., Yin, K., Neubig, G., and Martins, A. F. T. (2021). Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Freitag, M., Grangier, D., and Caswell, I. (2020). BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Foster, G., Lavie, A., and Bojar, O. (2021). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Graves, A. (2012). Sequence Transduction with Recurrent Neural Networks. arXiv:1211.3711 [cs, stat].

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.

Guo, J., Chen, X., Liu, Z., Yuan, W., Zhang, J., and Liu, G. (2022). Context Modeling with Hierarchical Shallow Attention Structure for Document-Level NMT. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ISSN: 2161-4407.

Hajlaoui, N. and Popescu-Belis, A. (2013). Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 236–247, Berlin, Heidelberg. Springer. 00000.

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. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, pages 283–289, Paris, France.

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 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. *ArXiv preprint*, abs/1712.00409.

Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2003). Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. *A Field Guide to Dynamical Recurrent Neural Networks*.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Huo, J., Herold, C., Gao, Y., Dahlmann, L., Khadivi, S., and Ney, H. (2020). Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Hwang, Y., Yun, H., and Jung, K. (2021). Contrastive learning for context-aware neural machine translation using coreference information. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1135–1144, Online. Association for Computational Linguistics.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Jean, S. and Cho, K. (2019). Context-Aware Learning for Neural Machine Translation. *ArXiv preprint*, abs/1903.04715.

Jiang, Y., Liu, T., Ma, S., Zhang, D., Yang, J., Huang, H., Sennrich, R., Cotterell, R., Sachan, M., and Zhou, M. (2022). BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202. Publisher: Royal Society.

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.

Jwalapuram, P., Joty, S., Temnikova, I., and Nakov, P. (2019). Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975, Hong Kong, China. Association for Computational Linguistics.

Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A. v. d., Graves, A., and Kavukcuoglu, K. (2016). Neural Machine Translation in Linear Time. *ArXiv preprint*, abs/1610.10099.

Kang, X., Zhao, Y., Zhang, J., and Zong, C. (2020). Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models. *ArXiv preprint*, abs/2001.08361.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR.

Kim, Y., Tran, D. T., and Ney, H. (2019). When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Kitaev, N., Kaiser, L., and Levskaya, A. (2020). Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

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.

Kuang, S., Xiong, D., Luo, W., and Zhou, G. (2018). Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA. 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*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Li, B., Liu, H., Wang, Z., Jiang, Y., Xiao, T., Zhu, J., Liu, T., and Li, C. (2020). Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, 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.

Li, L., Jiang, X., and Liu, Q. (2019). Pretrained Language Models for Document-Level Neural Machine Translation. *ArXiv preprint*, abs/1911.03110.

Likhomanenko, T., Xu, Q., Synnaeve, G., Collobert, R., and Rogozhnikov, A. (2021). CAPE: Encoding Relative Positions with Continuous Augmented Positional Embeddings.

Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open*, 3:111–132.

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Liu, S. and Zhang, X. (2020). Corpora for document-level neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3775–3781, Marseille, France. European Language Resources Association.

Liu, X., Yu, H., Dhillon, I. S., and Hsieh, C. (2020a). Learning to encode position for transformer with continuous dynamical model. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6327–6335. PMLR.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020b). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lopes, A., Farajian, M. A., Bawden, R., Zhang, M., and Martins, A. F. T. (2020). Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Luong, M., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016). Multi-task sequence to sequence learning. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Luong, T., Sutskever, I., Le, Q., Vinyals, O., and Zaremba, W. (2015). Addressing the rare word problem in neural machine translation. 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 11–19, Beijing, China. Association for Computational Linguistics.

Lupo, L., Dinarelli, M., and Besacier, L. (2022a). Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Lupo, L., Dinarelli, M., and Besacier, L. (2022b). Focused Concatenation for Context-Aware Neural Machine Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi. Association for Computational Linguistics.

Ma, S., Zhang, D., and Zhou, M. (2020). A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Ma, X., Kong, X., Wang, S., Zhou, C., May, J., Ma, H., and Zettlemoyer, L. (2021a). Luna: Linear Unified Nested Attention. *ArXiv preprint*, abs/2106.01540.

Ma, Z., Edunov, S., and Auli, M. (2021b). A Comparison of Approaches to Document-level Machine Translation. *ArXiv preprint*, abs/2101.11040.

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.

Martínez Garcia, E., Creus, C., and España-Bonet, C. (2019). Context-aware neural machine translation decoding. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 13–23, Hong Kong, China. Association for Computational Linguistics.

Maruf, S., Martins, A. F. T., and Haffari, G. (2018). Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium. 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.

Maruf, S., Saleh, F., and Haffari, G. (2021). A Survey on Document-level Neural Machine Translation: Methods and Evaluation. *ACM Computing Surveys*, 54(2):45:1–45:36.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157. 03511.

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. and Popescu-Belis, A. (2017). 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.

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.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N. A., and Kong, L. (2021). Random feature attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., and Hinton, G. (2017). Regularizing Neural Networks by Penalizing Confident Output Distributions. *ArXiv preprint*, abs/1701.06548.

Petrick, F., Rosendahl, J., Herold, C., and Ney, H. (2022). Locality-sensitive hashing for long context neural machine translation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 32–42, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Popel, M. and Bojar, O. (2018). Training Tips for the Transformer Model. *ArXiv preprint*, abs/1804.00247.

Popescu-Belis, A. (2019). Context in Neural Machine Translation: A Review of Models and Evaluations. *ArXiv preprint*, abs/1901.09115.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 40(3):211–218. 10830.

Rae, J. W., Potapenko, A., Jayakumar, S. M., Hillier, C., and Lillicrap, T. P. (2020). Compressive transformers for long-range sequence modelling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016). Sequence level training with recurrent neural networks. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Richards, J. C. and Schmidt, R. W. (2013). *Longman Dictionary of Language Teaching and Applied Linguistics*. Routledge. Google-Books-ID: ziSsAgAAQBAJ.

Riezler, S. and Maxwell, J. T. (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Rios Gonzales, A., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Rizzi, L. (2013). Locality. *Lingua*, 130:169–186.

Rosendahl, J., Tran, V. A. K., Wang, W., and Ney, H. (2019). Analysis of positional encodings for neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Rysová, K., Rysová, M., Musil, T., Poláková, L., and Bojar, O. (2019). A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.

Saunders, D., Stahlberg, F., and Byrne, B. (2020). Using context in neural machine translation training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online. Association for Computational Linguistics.

Scherrer, Y., Tiedemann, J., and Loáiciga, S. (2019a). Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Scherrer, Y., Tiedemann, J., and Loáiciga, S. (2019b). Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

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. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with relative position representations. 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 464–468, New Orleans, Louisiana. 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 the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Stojanovski, D. and Fraser, A. (2018). Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium. Association for Computational Linguistics.

Stojanovski, D. and Fraser, A. (2019). Improving anaphora resolution in neural machine translation using curriculum learning. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 140–150, Dublin, Ireland. European Association for Machine Translation.

Stojanovski, D., Krojer, B., Peskov, D., and Fraser, A. (2020). ContraCAT: Contrastive coreference analytical templates for machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4732–4749, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sugiyama, A. and Yoshinaga, N. (2019). Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Sun, Z., Wang, M., Zhou, H., Zhao, C., Huang, S., Chen, J., and Li, L. (2020). Capturing Longer Context for Document-level Neural Machine Translation: A Multi-resolutional Approach. *ArXiv preprint*, abs/2010.08961.

Sun, Z., Wang, M., Zhou, H., Zhao, C., Huang, S., Chen, J., and Li, L. (2022). Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

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.

Tan, X., Zhang, L., and Zhou, G. (2022). Discourse Cohesion Evaluation for Document-Level Neural Machine Translation.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2020). Efficient Transformers: A Survey. *ArXiv preprint*, abs/2009.06732.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2022). Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6):109:1–109:28.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

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.

Tomlin, R. S. (2014). *Basic Word Order (RLE Linguistics B: Grammar): Functional Principles*. Routledge. Google-Books-ID: OlPIAgAAQBAJ.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

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*, 6:407–420.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Vernikos, G., Thompson, B., Mathur, P., and Federico, M. (2022). Embarrassingly Easy Document-Level MT Metrics: How to Convert Any Pretrained Metric Into a Document-Level Metric.

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, Melbourne, Australia. Association for Computational Linguistics.

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, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. (2020). Linformer: Self-Attention with Linear Complexity. *ArXiv preprint*, abs/2006.04768.

Wilcoxon, F. (1946). Individual Comparisons of Grouped Data by Ranking Methods. *Journal of Economic Entomology*, 39(2):269–270.

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.

Wong, K., Maruf, S., and Haffari, G. (2020). Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online. Association for Computational Linguistics.

Wu, F., Fan, A., Baevski, A., Dauphin, Y. N., and Auli, M. (2019). Pay less attention with lightweight and dynamic convolutions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv preprint*, abs/1609.08144.

Wu, Z., Peng, H., Pappas, N., and Smith, N. A. (2022). Modeling Context With Linear Attention for Scalable Document-Level Translation.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Yin, K., Fernandes, P., Pruthi, D., Chaudhary, A., Martins, A. F. T., and Neubig, G. (2021). Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801, Online. Association for Computational Linguistics.

Yu, L., Sartran, L., Stokowiec, W., Ling, W., Kong, L., Blunsom, P., and Dyer, C. (2020). Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big bird: Transformers for longer sequences. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018). Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Zhang, L., Zhang, T., Zhang, H., Yang, B., Ye, W., and Zhang, S. (2021). Multi-hop transformer for document-level machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 3953–3963, Online. Association for Computational Linguistics.

Zhang, L., Zhang, Z., Chen, B., Luo, W., and Si, L. (2022). Context-Adaptive Document-Level Neural Machine Translation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6232–6236. ISSN: 2379-190X.

Zhang, P., Chen, B., Ge, N., and Fan, K. (2020a). Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zheng, Z., Yue, X., Huang, S., Chen, J., and Birch, A. (2020). Towards making the most of context in neural machine translation. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3983–3989. ijcai.org.

Zhou, C., Ma, X., Hu, J., and Neubig, G. (2019). Handling syntactic divergence in low-resource 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 1388–1394, Hong Kong, China. Association for Computational Linguistics.

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

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).