Document-wide Neural Machine Translation

Sameen Maruf

A thesis submitted for the degree of *Doctor of Philosophy at*

Monash University in 2019

Clayton School of Information Technology

This thesis is dedicated to my mother,

Dr. Attiya Maruf,

for always being the epitome of strength and wisdom

Copyright Notice

Abstract

Machine translation is an important task in natural language processing as it automates the translation process and reduces the reliance on human translators. For any machine translation task, we are given a set of sentences in the source language and the goal is to generate their translations in the target language. With the advent of neural networks, the translation quality surpasses that of the translations obtained using statistical techniques. However, most of the neural translation models still translate sentences independently, without incorporating any extra-sentential information. This research aims to build efficient neural models for document-level translation, which incorporate global contextual information when translating sentences. In this work, we focus on two use cases of document-level machine translation, monologue translation and dialogue translation, and endeavour to model them efficiently followed by rigorous evaluation and analysis.

For monologue translation, we start off by formulating the document-level machine translation problem as a structured prediction task to account for the correlations among the possible output translations and also between the source sentences and their corresponding translations. The resulting structured prediction problem is tackled with a neural translation model equipped with two memory components, one each for the source and target side, to capture the documental interdependencies. We train the model end-to-end and propose an iterative decoding algorithm based on block coordinate descent. After successfully formulating the problem, we narrow down our focus to better modelling of document context. We do this by proposing a novel and scalable top-down approach to hierarchical attention for document-context modelling which is able to selectively focus on relevant sentences in the document context and then attend to key words in those sentences. We further improve the model we proposed in the first phase and compare it to this approach. For both works mentioned here, we perform quantitative and qualitative evaluation.

Dialogue translation is another practical aspect of document translation but is underexplored. We propose the task of translating bilingual multi-speaker conversations and explore neural architectures that exploit both source and target-side conversation histories for this task. We introduce datasets for this task extracted from Europarl v7 and OpenSubtitles2016. Our experiments on public and in-house customer service data confirm the significance of leveraging conversation history, both in terms of automatic and manual evaluation.

Ours is the first work to look at these two aspects of document-level machine translation, and in general, the first work to use the document-wide context for improving machine translation.

Publications

Large portions of this thesis have been previously published in the form of conference papers. This applies to:

- Chapter 3 previously presented in:

  Sameen Maruf and Gholamreza Haffari (2018). Document context neural machine translation with memory networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2018), pages 1275–1284, Melbourne, Australia.

- Chapter 4 previously presented in:

  Sameen Maruf, André F. T. Martins and Gholamreza Haffari (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 (NAACL-HLT 2019), pages 3092–3102, Minneapolis, MN, USA.

- Chapter 5 previously presented in:

  Sameen Maruf, André F. T. Martins and Gholamreza Haffari (2018). Contextual Neural Model for Translating Bilingual Multi-Speaker Conversations. In Proceedings of the 3rd Conference on Machine Translation: Research Papers (WMT 2018), pages 101–112, Brussels, Belgium.

Declaration

I, Sameen Maruf, hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis. All main sources of help have been acknowledged where necessary.

Signature: **Date:** 12/11/2019

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my main supervisor, Assoc. Prof. Gholamreza Haffari, for taking me under his wing for the length of my PhD studies. I will always be thankful to him for introducing me to the field of NLP and deep learning, challenging me to develop the critical thinking necessary to be a good researcher, and motivating me to push myself when I needed it the most. If it were not for him, I would not have had the chance to get acquainted and work with my external supervisor, Dr. André Martins, who was added to my supervisory team in the second year of my PhD. André always lend a kind ear to my questions and helped me in improving my mathematical foundations. I also had the pleasure of working with him in person during the summer internship at Unbabel in Lisbon, Portugal. It was an incredible opportunity to witness first-hand, what goes into executing a commercial-grade MT system, and I will always cherish the time I spent there.

I would like to thank all members of my panel committee namely Mark Carman, Lan Du, Yuan-Fang Li, Bala Srinivasan, Chung-Hsing Yeh and John Betts for their feedback and support during my candidature. This thesis has benefited greatly from the thorough comments and suggestions made by my examiners, particularly Dr. George Foster (Google Research) and also Dr. Shafiq Joty (Nanyang Technological University).

I would also like to thank Monash University for providing me financial support during my PhD, in the form of Monash International Postgraduate Research Scholarship and Monash Graduate Scholarship, without which it would have been impossible for me to pursue this program. Furthermore, I would like to thank Danette Deriane and Helen Cridland from the former and current FIT GRS team for helping me navigate all the requirements and amenities available to me during my candidature. Also thanks to the support team at Monash Advanced Research Computing Hybrid (MonARCH) and Multi-modal Australian

ScienceS Imaging and Visualisation Environment (MASSIVE) for all their help with using the GPU nodes for running my experiments.

This thesis would not have been possible without the educational and moral support provided to me by my peers and colleagues in the NLP team at FIT in Monash. Firstly, I would like to thank Ehsan Shareghi for giving me feedback on the first paper I submitted in my PhD; Quan Tran and Poorya Zaremoodi for always answering my questions related to deep learning toolkits; Ming Liu, Najam Zaidi, Narjes Askarian, Fahimeh Saleh and Snow Situ for always listening to the troubles I was facing as a PhD student and teaching me to enjoy myself during the process, and last but not the least, Xuanli He, Trang Vu and Saeed Basirian, it was a pleasure to have known you. I also developed strong bonds with Jishan E Giti, Fatima Seeme and Ayesha Sadiq, who provided me with a sense of belonging and were there to listen to anything and everything I had to say whether it was about my studies or personal life.

Finally, I would like to thank my support system here and back home. My parents for shaping me into the person I am today; I owe all my achievements to you. My father, Syed Maruf Ahmed Ali, who gave me the freedom to develop as an individual, and my mother, Dr. Attiya Maruf, who was just a phone call away whenever I needed her – thank you for always sending your prayers my way. My sister, Maheen, and brother, Osama, it was comforting knowing that you were there to care for *ami* and *abu* when I could not. This list would be incomplete without the mention of my dearest husband, Irfan, who encouraged me to apply for PhD programs abroad and moved to Australia with me without any hesitation. If it were not for you, I would have never even considered the option of pursuing my studies overseas. I know that I do not say this enough (or at all) but thank you for all the emotional support, for always being there for me even when I was at my worst and for helping me navigate my life and career. I cannot imagine this journey without you being a part of it.

Contents

List of Tables

List of Figures

List of Notations

$\boldsymbol{x}$	The source/input sentence
$\boldsymbol{y}$	The target/output sentence
$\|\boldsymbol{d}\|$	The length of a bilingual document $\boldsymbol{d}$ equivalent to the number of sentences in that document
$\boldsymbol{X} = \{\boldsymbol{x}^j\}_{j=1}^{\|\boldsymbol{d}\|}$	The set of all source sentences in a document $\boldsymbol{d}$
$\boldsymbol{Y} = \{\boldsymbol{y}^j\}_{j=1}^{\|\boldsymbol{d}\|}$	The set of all target sentences in a document $\boldsymbol{d}$
$\boldsymbol{D}^{-j} = \{\boldsymbol{X}^{-j}, \boldsymbol{Y}^{-j}\}$	The collection of all sentences in a source and target document $\boldsymbol{d}$ except the j^{th} sentence
x_i^j, y_i^j	The i^{th} word in j^{th} source or target sentence
$\boldsymbol{h}_i^j$, $\boldsymbol{s}_i^j$	The hidden representation of i^{th} word in j^{th} source or target sentence from the encoder or decoder respectively
$\boldsymbol{W}, \boldsymbol{U}$	Parameter matrices
$\boldsymbol{b}$	Parameter vector

Chapter 1

Introduction

"Being that could be understood is language"

– Hans-Georg Gadamer

Machine translation (MT) is the process of automating translation between natural languages with the aid of computers. Translation, in itself, is a difficult task even for humans as it requires a thorough understanding of the source text and a good knowledge of the target language, hence requiring the human translators to have high degree of proficiency in both languages. Due to the dearth of professional translators and the rapid need of availability of multilingual digital content, for example, on the Internet, MT has grown immensely over the past few decades for the purposes of international communication.

Up until a few years ago, MT was mostly formalised through statistical techniques, hence very aptly named statistical machine translation (SMT), which involved meticulously crafting features to extract implicit information from corpora of bilingual sentence-pairs (Brown et al., 1993). These hand-engineered features were an intrinsic part of SMT and were one of the reasons behind its inflexibility. MT has come a long way from there to the state-of-the-art neural machine translation (NMT) systems (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) employed commercially today, which are based on neural network black-box models requiring little to no feature engineering. The results obtained by MT systems have seen rapid improvements in the past few years, and have added to their popularity among the general public (Metz, 2016; Lewis-Kraus, 2016) and the research community (Wu et al., 2016; Johnson et al., 2017; Dehghani et al., 2019).

Inspite of its success, MT has been based on strong independence and locality assumptions, that is either translating word-by-word or phrase-by-phrase (as done by SMT) or translating sentences in isolation (as done by NMT). Text, on the contrary, does not consist of isolated, unrelated elements, but of collocated and structured group of sentences bound together by complex linguistic elements, referred to as the discourse (Jurafsky and Martin, 2009). Ignoring the inter-relations among these discourse elements, results in translations which may be perfect at the sentence-level but lack crucial properties of the text hindering understanding. One way to address this issue is to exploit the underlying discourse structure of a text by utilising the information in the wider-sentential context. This is not a novel idea in itself and has been advocated by MT pioneers for decades (Bar-Hillel, 1960; Sennrich, 2018), but was mostly ignored in the era of SMT due to computational efficiency and tractability concerns by the MT community. Recently, with the increase in computational power available to us and the wide-scale application of neural networks to machine translation, we are finally in a position to forego the independence constraints that have impeded the progress in MT since long. The aim of this thesis is to explore different ways to leverage the wider extra-sentential (aka document-wide) context thus improving upon the state-of-the-art in NMT. We also endeavour to analyse the extent to which our context-dependent models enhance translation quality of the text in comparison to their context-agnostic counterparts.

1.1 Motivation and Goals

Neural machine translation has improved and in some cases even surpassed statistical MT to the extent that it is employed in all commercial systems today. To take it a step further, the next obvious action is to move from the sentence-based translation, which is plagued by independence assumptions, to a context-based translation that exploits the inter-sentential context. To illustrate the need for this and highlight the limitations of sentence-based translation, let us give a more concrete example of Chinese→English translation as provided by Sennrich (2018):

> Members of the public who find their cars obstructed by unfamiliar vehicles during their daily journeys can use the “Twitter Move Car” feature to address this distress when the driver of the unfamiliar vehicle cannot be reached.

At first glance, it is difficult to distinguish this MT output from a human translation. However, let us now provide a translation of the complete text as generated by an MT system:

> Members of the public who find their cars obstructed by unfamiliar vehicles during their daily journeys can use the "Twitter Move Car" feature to address this distress when the driver of the unfamiliar vehicle cannot be reached. On August 11, Xi'an traffic police WeChat service number "Xi'an traffic police" launched "WeChat mobile" service. With the launch of the service, members of the public can tackle such problems in their daily lives by using the "WeChat Move" feature when an unfamiliar vehicle obstructs the movement of their vehicle while the driver is not at the scene. [. . .]

An obvious problem with the translated text is the inconsistent translation of the name of the service "WeChat Move the Car". In other words, although it seemed that the sentence-based translation is adequate on its own, it still contained some ambiguous words which were inconsistent with the rest of the text. Let us now look at another example translation for Urdu→English generated by Google Translate.

> My grandfathers legs have failed because of the fluid. He had another visit today. Then his nature worsened. They can not speak for a few moments.

Even if we do not have access to the Urdu source text, we can concur that the English target text has some prominent issues, including inconsistent usage of pronouns (*he*, *they*) and ambiguous words (*fluid*, *visit*, *nature*), resulting in non-fluency and miscomprehension of the target text. From the previous two examples, it must be clear that despite the success of MT, it will never achieve human-level translation if it continues to be grounded on sentence-independence or locality assumptions.

Past works in SMT (discussed in Chapter 2) have tried to address the locality constraints of MT but they failed to produce significant improvements upon automatic evaluation. This was mostly due to the requirement of intensive feature engineering and the complexity of the SMT pipeline. With the successful application of neural networks to MT and the availability of powerful computational resources (e.g., GPUs), we are now in a favourable position to take advantage of its feature learning capacity and end-to-end training mechanism. This would allow us to build richer parameterised models with almost no dependence on explicit linguistic information. To this end, we formulate the primary goal of our research: *to explore effective methodologies that exploit the document-wide context information*[1] *to improve the quality of machine translation*.

To employ neural networks for any task, the first step is to *model* the problem itself, which involves representing the inputs and outputs (source and target text for NMT) with real-valued vectors through millions of parameters. After we have come up with a neural architecture, the next step involves *learning* the parameters of the model by optimising a training objective given the training data. Once the model is trained, the last step is to perform *decoding* (aka inference) to validate how well the model generalises to unseen testing data. Keeping this workflow and our primary research goal in mind, we come up with secondary goals for our research: we want (i) to model the document-wide context in NMT such that we have (ii) a context-dependent but end-to-end training and decoding framework not constrained by the sentence-independence assumption. The first part of this thesis addresses these goals with respect to *monologue translation*, where monologues are categorised as the discourse in which information flows in one language from a writer (or speaker) and the goal is to translate it into another language for the reader (or hearer). The second part of the thesis achieves the stated goals for the practical and under-explored problem of *dialogue translation*, which consists of the scenario in which we have speakers talking in their native language and the goal is to translate each speaker's utterance for the other non-native speakers to have an uninterrupted flow of information.

Our work in this thesis has been steered by a few desiderata, which we believe to be salient given the task of document-level machine translation.[2] The first of these is that, similar to all MT models, our models are data-driven, however, we do not require any explicit

[1] We will use the words 'document-wide' and 'global' interchangeably in this thesis.

[2] The terms 'document-level MT' and 'context-aware MT' are both used interchangeably by the MT community and by us for the purposes of this thesis.

linguistic annotation. Having such annotation could indeed be an interesting research direction but it does not fall within the scope of this thesis. Our models do require the provision of document boundaries in the datasets to be able to restrict the wider extra-sentential context to that within a document. For this purpose, we have endeavoured to extract the data used in this thesis from raw corpora available publicly and having document boundaries in the form of metadata. In other words, we have used aligned document-pairs instead of sentence-pairs (as done by classical MT) for training and decoding our models. The second desiderata is related to the extent of context information used in our models. Prior work mostly incorporates local context information coming from a few previous sentences, while we consider the global context information coming from all the sentences in the document whether they are on the source or target-side. This helps us in developing models which (i) are generalisable to documents of any length, (ii) do not devoid sentences at the beginning of a document of crucial information, and most importantly, (iii) can be thought of as having access to infinite context, unlike prior work. Using the document-wide context may also be crucial for the preservation of discourse phenomena from the source to the target text.

1.2 Thesis Statement and Contributions

Thesis Statement We claim that the performance of neural machine translation can be improved by utilising document-wide context information. First, we aim to model the correlations among the possible output translations and also between the source sentences and their corresponding translations. Second, using our document-level model, we aim to make joint predictions on the translations of a document given the source document for which we require end-to-end training and an efficient decoding framework. Finally, we show the efficacy of the previous three aspects by performing rigorous experimentation on data from a variety of language-pairs and domains.

Modelling:

1. We develop a model that leverages both the global source and target document contexts to improve the performance of neural machine translation. We do this by combining the generic sentence-based NMT model with external memory components and

employing coarse attention over the sentences in the context.

2. We propose an efficient and scalable top-down hierarchical attention approach for document-wide NMT which has the ability to focus on relevant sentences in the document context (using sparse attention) and then attend to key words in those sentences.

3. We design a contextual neural model for translating bilingual multi-speaker conversations, where we incorporate the source and target-side conversation histories into a sentence-based attentional model. We explore different ways of computing the source context representation for this task. Furthermore, we present an effective approach to leverage the target-side context, and also present an intuitive approach for incorporating both contexts simultaneously.

Learning:

4. We cast document MT as a structured prediction problem and introduce a pseudo-likelihood based training objective which allows the document-level NMT model to be trained end-to-end efficiently.

Decoding:

5. For decoding, we encumber the sentence-independence assumption by conditioning a sentence translation on both source and target-side document-wide context and proposing an iterative decoding algorithm based on block coordinate descent.

Experimental:

6. We experiment with our coarse attention approach using the proposed training and decoding strategies in an offline document MT setting and show that our model is effective in exploiting both source and target document contexts, and statistically significantly outperforms the previous work in terms of BLEU and METEOR.

7. We experiment with our hierarchical selective attention approach in both offline (past and future context) and online (only past) document MT settings and perform qualitative and quantitative analysis. Using sparse attention, instead of the standard soft attention, allows to dig deeper into the interpretability of the contextual NMT model.

8. We are the pioneer in dialogue translation and propose the task of translating bilingual multi-speaker conversations, a popular use-case in customer service chat, extract datasets for the said task and provide benchmark results. Our experiments on public datasets for four language-pairs confirm the significance of leveraging conversation history, both in terms of automatic and manual evaluation. Our models also achieve promising results on in-house customer service chat datasets for English-French and English-German.

1.3 Organisation of this Thesis

In this section, we provide an outline of the rest of the thesis. The primary contribution of this thesis are three content chapters: Chapters 3, 4 and 5, where the first two fall in the first part specific to monologue translation and the last falls in the second part for dialogue translation. The outline and summary of each chapter are as follows:

- **Chapter 2: Background** This chapter provides a thorough overview of the foundations for the research described in this thesis, including the state-of-the-art architectures for sentence-based neural machine translation and a detailed description of prior work in document-level machine translation. We also shed light on the evaluation approaches that are being used in this domain.

- **Chapter 3: Document Context Modelling with Coarse Attention** In this chapter, we present our document-level NMT model, which takes both source and target document context into account using memory networks, and a description of our training and decoding methodologies for the said model. Experimental results for three language-pairs are reported showing drastic improvements when incorporating both types of contexts.

- **Chapter 4: Document Context Modelling with Hierarchical Attention** We present our novel and scalable top-down approach to hierarchical attention for context-aware NMT and also present single-level attention approaches based on sentence or word-level information in the context. Our approach not only significantly outperforms context-agnostic baselines but also surpasses context-aware baselines in most cases.

- **Chapter 5: Translating Bilingual Multi-Speaker Conversations** We introduce the task of translating bilingual multi-speaker conversations, and explore neural architectures that exploit both source and target-side conversation histories for this task. We also introduce datasets extracted from Europarl v7 and OpenSubtitles2016. Our experiments on four language-pairs in the public domain and two language-pairs in the commercial domain confirm the significance of leveraging conversation history, both in terms of BLEU and manual evaluation.

- **Chapter 6: Conclusions** This chapter summarises our findings and contributions in this thesis and also highlights potential directions for future work.

Chapter 2

Background

This chapter provides a thorough overview of the foundations for the research described in this thesis, including sentence-based neural machine translation and document-level machine translation.

We start off by describing the evolution of sentence-based neural machine translation systems in Section 2.1, where we state the basics of statistical machine translation, followed by a detailed description of the state-of-the-art NMT architectures based on attention and the Transformer model. These will be the cornerstones of our models described in later chapters. We also describe how MT outputs are evaluated using automatic metrics.

Then, in Section 3.4, we shed light on prior and recent approaches that have tried to model document-level and discourse information in both SMT and NMT, followed by a detailed description of various evaluation strategies that have been proposed for document-level MT to-date.

2.1 Neural Machine Translation (NMT)

Machine translation has been around for a long time and various approaches have been proposed to make it on-par with human translation. The approach to MT which is strongly correlated with the current state-of-the-art NMT models and worth mentioning here is statistical machine translation (SMT). SMT models the probability of a sentence translation in one language given a source sentence in another language. This probability is determined automatically by training a statistical model using a parallel corpus containing source and target translation pairs. The advantages of SMT over its predecessors were that it was data-driven and language independent and was considered a state-of-the-art technique up until the advent of neural-based approaches.

Mathematically, the goal of SMT (and NMT) is to find the most probable target sequence $\hat{\boldsymbol{y}}$ given a source sentence, that is:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y} \mid \boldsymbol{x}) \tag{2.1}$$

Using Bayes' rule, this conditional probability can be reformulated as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y})P(\boldsymbol{x} \mid \boldsymbol{y}) \tag{2.2}$$

where $P(\boldsymbol{y})$, aka the *language model* (LM) usually based on trigram probabilities and estimated using monolingual corpora, assigns a higher probability to fluent, grammatical sentences and $P(\boldsymbol{x} \mid \boldsymbol{y})$, aka the *translation model*, assigns a higher probability to sentences that have corresponding meaning. The translation model is parameterised using an alignment function which represents how a source word is aligned to a target word (Brown et al., 1993). The more often two words occur together in different sentence-pairs, the more likely they are aligned to each other and have equivalent meaning. These word-based models were superseded by *phrase-based models* (Marcu and Wong, 2002; Koehn et al., 2003) which used many-to-many alignments between the source and target words stored in a phrase table (Och and Ney, 2004).

While SMT was successfully deployed in many commercial systems, it did not work very well and suffered from two major drawbacks. First, translation decisions were local as the translation was performed phrase-by-phrase and long-distance dependencies were often ignored. Secondly and more problematically, the entire MT pipeline became increasingly complex as many different components had to be tuned separately, e.g., translation models, language models, reordering models, etc., which made it difficult to combine them together and have a single end-to-end model. As a result, when the AI winter was over and neural networks resurfaced as the new approach to solve natural language processing (NLP) problems, it was seen as the next logical step to use them for machine translation as well. NMT, even though quite recent (after 2014), has opened a new era in MT for both research and commercial purposes.

NMT models, in general, are based on an encoder-decoder framework (Figure 2.1) where the encoder reads the source sentence to compute a real-valued representation and

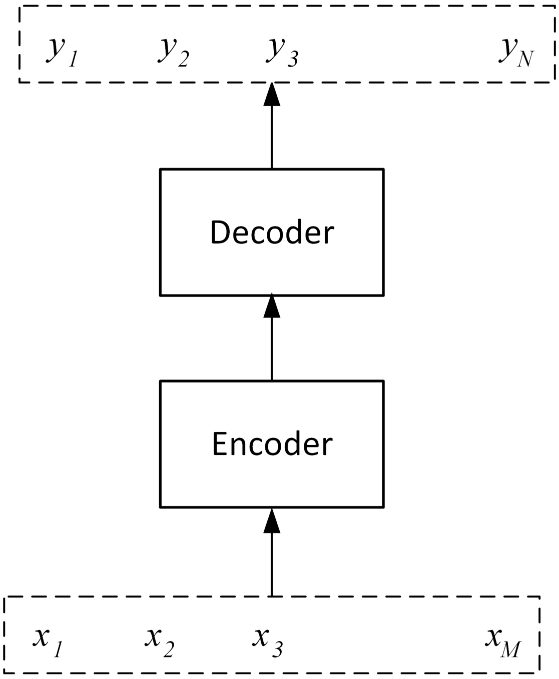


Figure 2.1: A general overview of an encoder-decoder model.

the decoder generates the target translation one word at a time given the previously computed representation. The initial model (Sutskever et al., 2014) used a fixed representation of the source sentence to generate the target sentence. It was quickly replaced by the attention-based encoder-decoder architecture which generated a dynamic context representation (Bahdanau et al., 2015). These models were based on recurrent neural networks (RNNs) (described shortly),[1] which use recurrent connections to exhibit temporal dynamic behavior over time, and were thus considerably suitable for modelling sequential information. However, the major drawback of such sequential computation was that it hindered parallelisation within training examples and became a bottleneck when processing longer sentences. Most recently, a new model architecture, the Transformer, was proposed which is based solely on attention mechanisms, dispensing with the recurrence entirely. It has proved to achieve state-of-the-art results on several language-pairs (Vaswani et al., 2017).

In the rest of this section, we will be detailing the RNN-based attentional encoder-decoder architecture (Bahdanau et al., 2015) (Section 2.1.1) and the Transformer architecture (Vaswani et al., 2017) (Section 2.1.2) as our document NMT models are grounded on

[1] We do not mention convolution-based NMT architectures (Gehring et al., 2017a,b) here as we have not used them in our research.

either one of them, followed by a description of the popular automatic evaluation metrics for MT (Section 2.1.5).

2.1.1 RNN-based Encoder-Decoder Architecture

Before we describe the attentional NMT model architecture, we will describe its key component: the recurrent neural network (RNN).

2.1.1.1 Recurrent Neural Networks

Recurrent neural network (RNN) (Elman, 1990) is a powerful neural architecture which has been used for a variety of sequence modelling tasks like language modelling (Mikolov et al., 2010, 2011; Mikolov and Zweig, 2012), text summarisation (Rush et al., 2015; Nallapati et al., 2016), and speech recognition (Graves and Jaitly, 2014), to name a few. Formally, an RNN (Figure 2.2) takes as input a sequence of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ one at a time, and for each input $\boldsymbol{x}_m$, the RNN updates its *hidden state* using the output from previous hidden states as additional inputs. The hidden state at timestep m can be thought of as a representation for the partial sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$. Mathematically, at timestep m, a *vanilla RNN* computes the hidden state $\boldsymbol{h}_m$ as:

$$\begin{aligned} \boldsymbol{h}_m &= \mathrm{RNN}(\boldsymbol{h}_{m-1}, \boldsymbol{x}_m) \\ &= f(\boldsymbol{W}_{\boldsymbol{hh}}\boldsymbol{h}_{m-1} + \boldsymbol{W}_{\boldsymbol{hx}}\boldsymbol{x}_m + \boldsymbol{b}_{\boldsymbol{h}}) \end{aligned} \tag{2.3}$$

where $\boldsymbol{x}_m$ is the input at timestep m, $\boldsymbol{h}_{m-1}$ is the previous hidden state, f is a non-linear activation function (Table 2.1) and $\{\boldsymbol{W}, \boldsymbol{b}\}$ are parameters of the RNN shared across timesteps. The initial state $\boldsymbol{h}_0$ is often set to the zero vector or is randomly initialised.

At each timestep m, the RNN (optionally) outputs a discrete symbol y_m which is sampled from a probability distribution via a softmax operation:

$$y_m \sim \mathrm{softmax}(\boldsymbol{W}_{\boldsymbol{y}}\boldsymbol{h}_m + \boldsymbol{b}_{\boldsymbol{y}}) \tag{2.4}$$

where the input (linear transformation of $\boldsymbol{h}_m$) is a score vector over the different output classes, and the softmax function $\left(\mathrm{softmax}(\boldsymbol{z}) = \frac{\exp(z_j)}{\sum_{k=1}^{K}\exp(z_k)}\right)$ converts the score vector into a probability vector.

The main advantage of using RNNs for sequence modelling tasks is that they have the ability to capture long-range dependencies in sequences due to their recurrent connections.

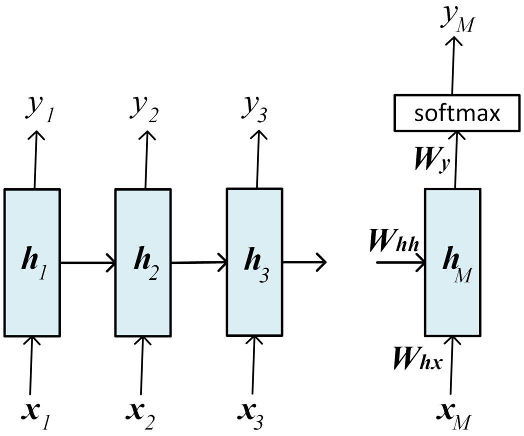


Figure 2.2: An overview of a recurrent neural network over timesteps.

Function	Formula
Sigmoid (σ)	$f(z) = \dfrac{1}{1 + \exp(-z)}$
Hyperbolic Tangent ($\tanh$)	$f(z) = \dfrac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)}$
Rectified Linear Unit (ReLU)	$f(z) = \max(0, z)$

Table 2.1: Commonly used non-linear activation functions.

Apart from running in a left-to-right fashion, vanilla RNNs suffer from two classic problems of *exploding and vanishing gradients* as described by Bengio et al. (1994), due to the gradient-based learning methods with backpropagation. In short, due to the multiplicative nature of the gradient updates, the gradients become exponentially large (exploding gradients) or move exponentially fast towards zero (vanishing gradients) as we backpropagate over time, making learning unstable and resulting in a model that, in practice, is unable to capture long-range dependencies in sequences.

To deal with the exploding gradients problem, gradient norm clipping (Pascanu et al., 2012) is usually used where the gradient γ is clipped based on its norm $||\gamma||$, that is if the norm is greater than a threshold η, then $\gamma \leftarrow \frac{\eta\gamma}{||\gamma||}$. This method prevents exponential increase in the norm of the gradients thus reducing the exploding gradient problem in practice.

For resolving the vanishing gradient problem, a variety of approaches have been proposed, however, the most widely adopted ones include replacing the recurrent unit with a long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) or a gated recurrent unit (GRU) (Cho et al., 2014a). The main idea is to use some or all of specific gating units that control how much an RNN wants to reuse memory from previous timestep (forget gates $\boldsymbol{g}^{\boldsymbol{f}}_m$), receive an input signal (input gates $\boldsymbol{g}^{\boldsymbol{i}}_m$), and extract information (output gates $\boldsymbol{g}^{\boldsymbol{o}}_m$) at each timestep. All gates are computed as linear transformations of current input $\boldsymbol{x}_m$ and previous hidden state $\boldsymbol{h}_{m-1}$ followed by a sigmoid activation function:

$$\begin{aligned}
\boldsymbol{g}^{\boldsymbol{i}}_m &= \sigma(\boldsymbol{W}_{\boldsymbol{ih}}\boldsymbol{h}_{m-1} + \boldsymbol{W}_{\boldsymbol{ix}}\boldsymbol{x}_m + \boldsymbol{b}_{\boldsymbol{i}}) \\
\boldsymbol{g}^{\boldsymbol{f}}_m &= \sigma(\boldsymbol{W}_{\boldsymbol{fh}}\boldsymbol{h}_{m-1} + \boldsymbol{W}_{\boldsymbol{fx}}\boldsymbol{x}_m + \boldsymbol{b}_{\boldsymbol{f}}) \\
\boldsymbol{g}^{\boldsymbol{o}}_m &= \sigma(\boldsymbol{W}_{\boldsymbol{oh}}\boldsymbol{h}_{m-1} + \boldsymbol{W}_{\boldsymbol{ox}}\boldsymbol{x}_m + \boldsymbol{b}_{\boldsymbol{o}})
\end{aligned}$$

and each gate has different parameters shared across timesteps.

The LSTM, in addition to using all the three gates, also defines a new cell state $\boldsymbol{c}_m$ to store the temporal information whose recurrence follows a linear scaling rather than a multiplicative and non-linear update, while the GRU is a simplification using only input and forget gates and does not have an intermediate cell state. The hidden state $\boldsymbol{h}_m$ is computed using LSTM as:

$$\begin{aligned}
\boldsymbol{g}_m &= \sigma(\boldsymbol{W}_{\boldsymbol{gh}}\boldsymbol{h}_{m-1} + \boldsymbol{W}_{\boldsymbol{gx}}\boldsymbol{x}_m + \boldsymbol{b}_{\boldsymbol{g}}) \\
\boldsymbol{c}_m &= \boldsymbol{g}^{\boldsymbol{f}}_m \odot \boldsymbol{c}_{m-1} + \boldsymbol{g}^{\boldsymbol{i}}_m \odot \boldsymbol{g}_m \\
\boldsymbol{h}_m &= \boldsymbol{g}^{\boldsymbol{o}}_m \odot \tanh(\boldsymbol{c}_m)
\end{aligned}$$

and for GRU:

$$\begin{aligned}
\tilde{\boldsymbol{h}}_m &= \tanh(\boldsymbol{W}_{\tilde{\boldsymbol{h}}\boldsymbol{h}}(\boldsymbol{g}^{\boldsymbol{f}}_m \odot \boldsymbol{h}_{m-1}) + \boldsymbol{W}_{\tilde{\boldsymbol{h}}\boldsymbol{x}}\boldsymbol{x}_m + \boldsymbol{b}_{\tilde{\boldsymbol{h}}}) \\
\boldsymbol{h}_m &= (\boldsymbol{1} - \boldsymbol{g}^{\boldsymbol{i}}_m) \odot \boldsymbol{h}_{m-1} + \boldsymbol{g}^{\boldsymbol{i}}_m \odot \tilde{\boldsymbol{h}}_m
\end{aligned}$$

where $\odot$ denotes element-wise multiplication.

The stated modifications in LSTM and GRU make them more robust, capable of learning long-range dependencies and have superior performance in comparison to the vanilla RNN models (Cho et al., 2014a). Thus, these are the most popular choices today when using recurrent neural networks for sequence modelling. In our work as well, we use either of these two types of RNNs.

Now that we have become familiar with the basic working of an RNN, let us move on to describe how RNNs are employed in NMT.

2.1.1.2 Neural Machine Translation with RNNs

In NMT, we model the conditional probability $P(\boldsymbol{y} \mid \boldsymbol{x})$ using neural networks where $\boldsymbol{x} = (x_1, \ldots, x_M)$ is the input (source) sequence and $\boldsymbol{y} = (y_1, \ldots, y_N)$ is the output (target) sequence. To allow the model to cater to sequences with arbitrary lengths, special start-of-sentence (`<s>`) and end-of-sentence (`</s>`) tokens are added at the beginning and end of each sentence. All the source and target words in the parallel corpus constitute two fized-size vocabularies, denoted by V_S and V_T respectively. Out-of-vocabulary words are represented by a special token `<unk>` . The conditional probability of a target sentence $\boldsymbol{y}$ given the source sentence $\boldsymbol{x}$ is decomposed as:

$$P_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{n=1}^{N} P_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{y}_{<n}, \boldsymbol{x}) \tag{2.5}$$

where $\boldsymbol{\theta}$ denote the learnable parameters of the neural network, y_n is the current target word and $\boldsymbol{y}_{<n}$ are the previously generated words. Let us now describe the key components of an NMT model, illustrated in Figure 2.3, that yield this conditional probability.

Word Embeddings The first step in any neural model is to convert the discrete words in the input and output sequences to real-valued vector representations aka word embeddings. We have two embedding tables, one for the source language $\boldsymbol{E_S}$ (dimensions $H \times |V_S|$) and one for the target language $\boldsymbol{E_T}$ (dimensions $H \times |V_T|$) (where H is a pre-defined value, and $|V_S|$ and $|V_T|$ are source and target vocabulary sizes respectively), which are learned along with the other parameters in the model. For simplicity, we use the same H value for the dimensions of both source and target embeddings and the size of the hidden unit in the encoder and decoder.

Encoder It is a bi-directional (forward and backward) RNN whose hidden states represent individual words of the source sentence. These representations capture information not only of the corresponding word but also of other words in the sentence, i.e., the sentential context. The forward and backward RNNs run over the source sentence in a left-to-right and right-to-left direction:

$$\begin{aligned} \overrightarrow{\boldsymbol{h}_m} &= \overrightarrow{\text{RNN}}(\overrightarrow{\boldsymbol{h}}_{m-1}, \boldsymbol{E_S}[x_m]) \\ \overleftarrow{\boldsymbol{h}}_m &= \overleftarrow{\text{RNN}}(\overleftarrow{\boldsymbol{h}}_{m+1}, \boldsymbol{E_S}[x_m]) \end{aligned}$$

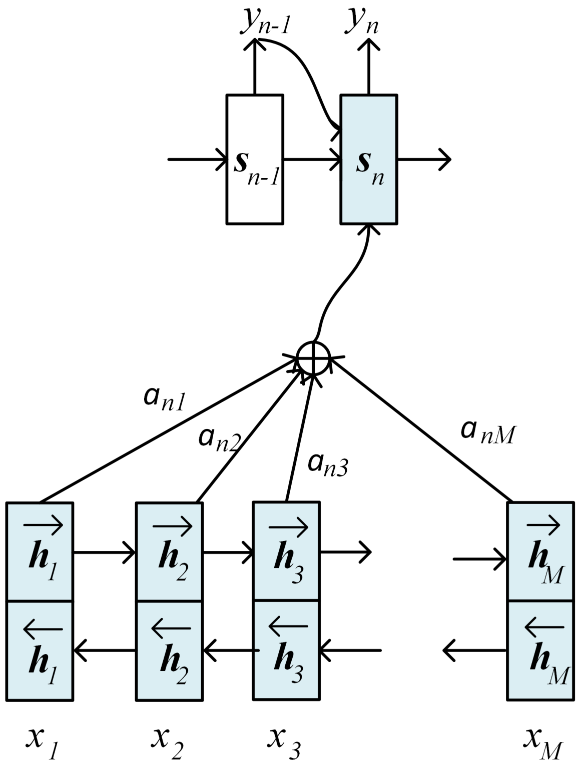


Figure 2.3: A detailed view of an RNN-based encoder-decoder model with attention when generating the n^{th} target word.

where $\boldsymbol{E_S}[x_m]$ is embedding of the word x_m from the embedding table $\boldsymbol{E_S}$ of the source language, and $\overrightarrow{\boldsymbol{h}}_m$ and $\overleftarrow{\boldsymbol{h}}_m$ are the hidden states of the forward and backward RNNs for current timestep m. Each word in the source sentence is then represented by the concatenation of the corresponding bidirectional hidden states, that is $\boldsymbol{h}_m = [\overrightarrow{\boldsymbol{h}}_m; \overleftarrow{\boldsymbol{h}}_m]$.

Attentional Decoder An integral component of the NMT architecture is the attention mechanism. This enables the decoder to dynamically *attend* to relevant parts of the source sentence at each step of generating the target sentence. The dynamic context vector $\boldsymbol{c}_n$ is computed as a weighted linear combination of the hidden states produced by the bidirectional RNNs in the encoder, that is:

$$\boldsymbol{c}_n = \sum_{m=1}^{M} \alpha_{nm} \boldsymbol{h}_m \tag{2.6}$$

The weight α_{nm} of each representation $\boldsymbol{h}_m$ is given by:

$$\begin{aligned}\alpha_{nm} &= \frac{\exp(e_{nm})}{\sum_{m'=1}^{M}\exp(e_{nm'})} \\ e_{nm} &= \boldsymbol{v}^\top \tanh(\boldsymbol{W_{ah}}\boldsymbol{h}_m + \boldsymbol{W_{as}}\boldsymbol{s}_{n-1})\end{aligned}$$

and can be thought of as the *alignment probability* between a target symbol at position n and a source symbol at position m. e_{nm} is the alignment score which tells how well the inputs around position m and the output at position n match. It is calculated based on the previous decoder state $\boldsymbol{s}_{n-1}$ and the representation $\boldsymbol{h}_m$ of source word at position *m* and referred to as *additive attention*.[2] The parameter vector $\boldsymbol{v}$ is of size H while the matrices $\boldsymbol{W_{ah}}$ and $\boldsymbol{W_{as}}$ are of size $H \times 2H$ and $H \times H$ respectively.

The backbone of the decoder is a uni-directional RNN which generates words of the target translation one-by-one in a left-to-right fashion. The decoder hidden state is computed as follows:

$$\boldsymbol{s}_n = \text{RNN}(\boldsymbol{s}_{n-1}, \boldsymbol{E_T}[y_{n-1}], \boldsymbol{c}_n)$$

where $\boldsymbol{s}_{n-1}$ is the previous decoder state, $\boldsymbol{E_T}[y_n]$ is embedding of the word y_n from the embedding table $\boldsymbol{E_T}$ of the target language, and $\boldsymbol{c}_n$ is the dynamic context vector formulated previously. The RNN in the decoder is similar to the one defined in Eq. 2.3 but with $\boldsymbol{c}_n$ as an additional input and a dedicated parameter matrix. The probability of generation of each word y_n is then conditioned on all of the previously generated words $\boldsymbol{y}_{<n}$ via the state of the RNN decoder $\boldsymbol{s}_n$, and the source sentence via $\boldsymbol{c}_n$ as follows:

$$\begin{aligned}\boldsymbol{u}_n &= \tanh(\boldsymbol{s}_n + \boldsymbol{W_{uc}}\boldsymbol{c}_n + \boldsymbol{W_{un}}\boldsymbol{E_T}[y_{n-1}]) \\ P_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{y}_{<n}, \boldsymbol{x}) &= \text{softmax}(\boldsymbol{W_y}\boldsymbol{u}_n + \boldsymbol{b_y}) \quad (2.7) \\ y_n &\sim P_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{y}_{<n}, \boldsymbol{x}) \quad (2.8)\end{aligned}$$

where $\boldsymbol{W}$ matrices and $\boldsymbol{b_y}$ vector are also parameters of the NMT model and the input to the softmax (linear transformation of $\boldsymbol{u}_n$) is a score vector over the target vocabulary. Hence, we have formulated Eq. 2.5 using an RNN-based NMT model.

[2] Another variant called the dot-product attention, proposed by Luong et al. (2015), is defined as $\boldsymbol{h}_m^\top \boldsymbol{s}_{n-1}$.

2.1.2 Transformer-based Encoder-Decoder Architecture

RNN-based encoder-decoder architectures are prevalent in various NLP tasks and were a popular approach for NMT up until two years ago. The limitations to NMT were mostly due to its grounding on RNNs. The first limitation is the sequential nature of RNNs, that is for processing each input token, the model has to wait until all previous input tokens have been processed, which proves to be a bottleneck when processing long sequences. The second limitation is learning long-range dependencies among the tokens within a sequence. The number of operations required to relate signals from two arbitrary input or output positions grows with the distance between positions, making it difficult to learn complex dependencies between distant positions. The recent Transformer architecture, proposed by Vaswani et al. (2017), circumvents these limitations by having a model that is still based on the philosophy of encoder-decoder, but instead of employing recurrence, uses stacked self-attention and point-wise, fully connected layers for both the encoder and decoder.

The model architecture is provided in Figure 2.4 and comprises the following components:

Embeddings Similar to the attentional encoder-decoder model described previously (Bahdanau et al., 2015), we have two word embedding tables, one each for the source and target language (denoted by $\boldsymbol{E}_{\boldsymbol{S}}^{\boldsymbol{w}}$ and $\boldsymbol{E}_{\boldsymbol{T}}^{\boldsymbol{w}}$ respectively) to convert the discrete source and target sequences to real-valued vectors. Now, the word embedding of a particular source sentence is given by the set of vectors $\boldsymbol{E}_{\boldsymbol{S}}[x_1], \ldots, \boldsymbol{E}_{\boldsymbol{S}}[x_M]$ and for a particular target sentence by the set of vectors $\boldsymbol{E}_{\boldsymbol{T}}[y_1], \ldots, \boldsymbol{E}_{\boldsymbol{T}}[y_N]$. However, since the Transformer does not use any form of recurrence, it needs to inject some positional information about the tokens in the sequence so that the model is aware of the word-order. This is done by using separate positional encodings for both the source and target sequences (denoted by $\boldsymbol{E}_{\boldsymbol{S}}^{\boldsymbol{p}}$ and $\boldsymbol{E}_{\boldsymbol{T}}^{\boldsymbol{p}}$ respectively) and adding them to the corresponding word embeddings. Hence, the total embeddings of a source and target sentence are given by:

$$\boldsymbol{E}[\boldsymbol{x}] \quad : \quad \boldsymbol{E}_{\boldsymbol{S}}^{\boldsymbol{w}}[x_1] + \boldsymbol{E}_{\boldsymbol{S}}^{\boldsymbol{p}}[1], \ldots, \boldsymbol{E}_{\boldsymbol{S}}^{\boldsymbol{w}}[x_M] + \boldsymbol{E}_{\boldsymbol{S}}^{\boldsymbol{p}}[M] \tag{2.9}$$

$$\boldsymbol{E}[\boldsymbol{y}] \quad : \quad \boldsymbol{E}_{\boldsymbol{T}}^{\boldsymbol{w}}[y_1] + \boldsymbol{E}_{\boldsymbol{T}}^{\boldsymbol{p}}[1], \ldots, \boldsymbol{E}_{\boldsymbol{T}}^{\boldsymbol{w}}[y_N] + \boldsymbol{E}_{\boldsymbol{S}}^{\boldsymbol{p}}[N] \tag{2.10}$$

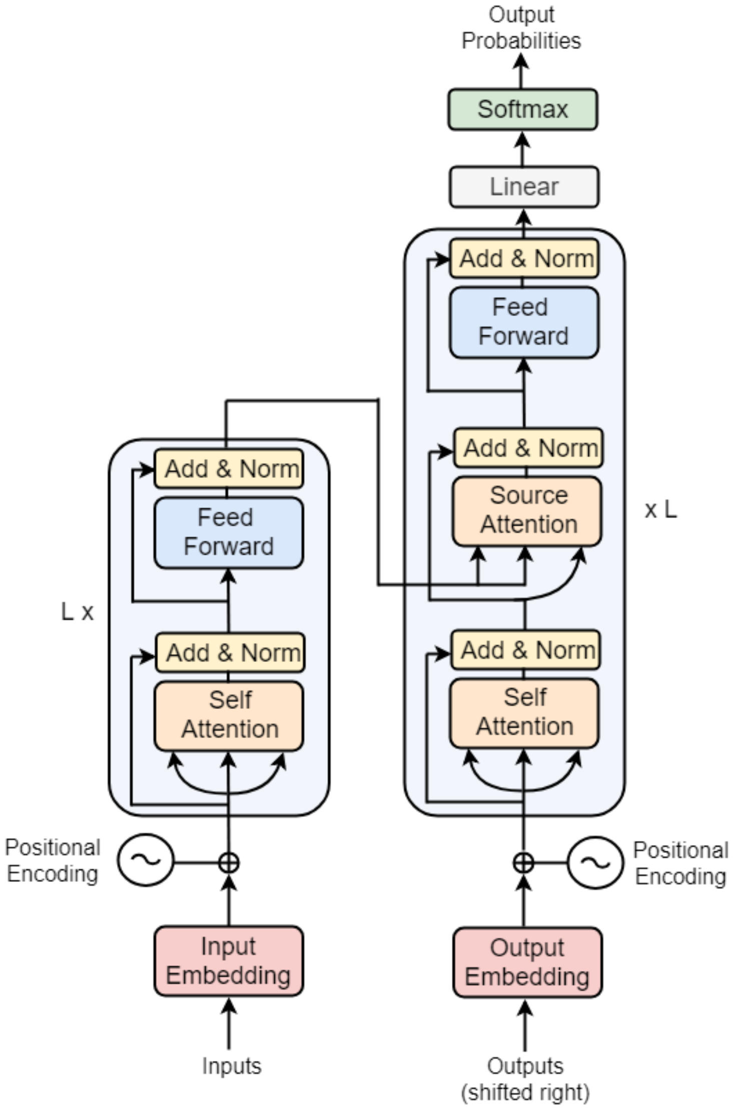


Figure 2.4: The Transformer - model architecture.

For the positional encodings, Vaswani et al. (2017) proposed to use a fixed sinusoidal encoding, that is for a particular source sentence the positional encoding is given by the set of vectors $\boldsymbol{f^p}(1), \dots, \boldsymbol{f^p}(M)$ and for a particular target sentence by the set of vectors $\boldsymbol{f^p}(1), \dots, \boldsymbol{f^p}(N)$, where $\boldsymbol{f^p}(pos)$ is a positional encoding vector at position pos in the sequence. These are computed based on sine and cosine functions of different frequencies forming wavelengths with a geometric progression from 2π to $10000\times 2\pi$:

$$\begin{aligned} \boldsymbol{f^p}(pos) &= \begin{bmatrix} \vdots \\ PE(pos, 2i) \\ PE(pos, 2i+1) \\ \vdots \end{bmatrix} \\ PE(pos, 2i) &= \sin\left(\frac{pos}{10000^{2i/H}}\right) \\ PE(pos, 2i+1) &= \cos\left(\frac{pos}{10000^{2i/H}}\right) \end{aligned}$$

where i is the dimension and belongs to $[0, \ldots, \frac{H-1}{2}]$. The advantages of sinusoidal positional encoding are that they do not add any parameters to the model and may handle sequences of lengths longer than the ones encountered during training.

Another method, proposed by Gehring et al. (2017b), defines positional encodings using additional weight matrices $\boldsymbol{W_S^p}$ and $\boldsymbol{W_T^p}$ of dimensions $H \times |x_{max}|$ and $H \times |y_{max}|$ respectively, where $|x_{max}|$ and $|y_{max}|$ are the maximum source and target sequence lengths to be chosen, and each column of the weight matrix at position pos corresponds to the encoding for the token at that position. To further elaborate, for a particular source sentence the positional encoding is now given by the set of vectors $\boldsymbol{W_S^p}[1], \ldots, \boldsymbol{W_S^p}[M]$ and for a particular target sentence by the set of vectors $\boldsymbol{W_T^p}[1], \ldots, \boldsymbol{W_T^p}[N]$, where M and N are the lengths of the particular sequences and these parameters are learned jointly with the model.

Encoder The encoder stack is composed of L identical layers, each containing two sub-layers. The first, a multi-head self-attention sub-layer (denoted by $\textsc{MultiHead}_{self}$), allows each position in the encoder to attend to all positions in the previous layer of the encoder, while the second sub-layer, a feed-forward network (denoted by FFN), uses two linear transformations with a ReLU activation. Both of these will be described in detail shortly. Residual connections (He et al., 2016) and layer normalisation (Ba et al., 2016) are employed around both sub-layers.[3] Hence, at the l^{th} layer of the encoder stack, the output is given by:

$$\begin{aligned} \boldsymbol{X}^l &= \text{LayerNorm}\Big(\boldsymbol{X}^l_{MHself} + \text{FFN}(\boldsymbol{X}^l_{MHself})\Big) \\ \boldsymbol{X}^l_{MHself} &= \text{LayerNorm}\Big(\boldsymbol{X}^{l-1} + \textsc{MultiHead}_{self}(\boldsymbol{X}^{l-1})\Big) \end{aligned} \tag{2.11}$$

where $\boldsymbol{X}^0$ is the output of the encoder embedding layer and $l \in [1, \ldots, L]$.

[3] Layer normalisation normalises the inputs across their neuron units within a hidden layer, thus stabilising the interactions between sub-layers in the Transformer encoder and decoder.

Decoder The decoder stack is also composed of L identical layers. In addition to the two sub-layers in the encoder, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack (denoted by $\textsc{MultiHead}_{src}$). Masking is used in the self-attention sub-layer to prevent positions from attending to subsequent positions thus avoiding leftward flow of information. Hence, at the l^{th} layer of the decoder stack, the output is given by:

$$\begin{aligned}
\boldsymbol{Y}^{l} &= \text{LayerNorm}\Big(\boldsymbol{Y}^{l}_{MHsrc} + \text{FFN}(\boldsymbol{Y}^{l}_{MHsrc})\Big) \\
\boldsymbol{Y}^{l}_{MHsrc} &= \text{LayerNorm}\Big(\boldsymbol{Y}^{l}_{MHself} + \textsc{MultiHead}_{src}(\boldsymbol{Y}^{l}_{MHself}, \boldsymbol{X}^{L})\Big) \\
\boldsymbol{Y}^{l}_{MHself} &= \text{LayerNorm}\Big(\boldsymbol{Y}^{l-1} + \textsc{MultiHead}_{self}(\boldsymbol{Y}^{l-1})\Big)
\end{aligned} \tag{2.12}$$

where $\boldsymbol{Y}^{0}$ is the output of the decoder embedding layer.

Similar to the RNN-based encoder-decoder architecture, the conditional probability of generating a target word y_n given the source sentence is computed as follows:

$$P_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{y}_{<n}, \boldsymbol{x}) = \text{softmax}(\boldsymbol{W}_{\boldsymbol{y}}\boldsymbol{Y}^{L} + \boldsymbol{b}_{\boldsymbol{y}}) \tag{2.13}$$

where $\boldsymbol{Y}^{L}$ is the final output from the decoder.

Multi-Head Attention (MultiHead) In general, an attention function can be described as the mapping of a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. The Transformer employs a variant of the dot-product attention function (Luong et al., 2015), that is it computes the dot products of the query with all keys, and divides each by a scaling factor $\sqrt{d_k}$, followed by a softmax function to obtain the weights on the values (Figure 2.5). This is referred to as scaled dot-product attention:

$$\textsc{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{V}\text{softmax}\Big(\frac{\boldsymbol{K}^{\top}\boldsymbol{Q}}{\sqrt{d_k}}\Big) \tag{2.14}$$

where the inputs are matrices and d_k is the dimensions of the keys taken to be H here. An advantage of using the dot-product attention, instead of the additive attention (Bahdanau et al., 2015), is that the former is much faster and more space-efficient in practice since it can be implemented using highly optimised matrix multiplications.

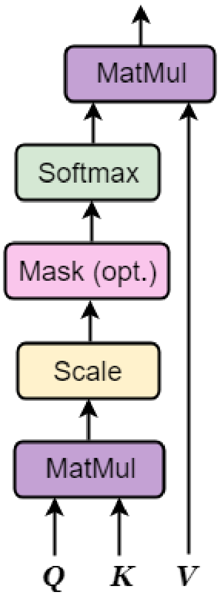


Figure 2.5: Scaled dot-product attention.

The main innovation of the Transformer, however, is that instead of employing a single attention function, the inputs are linearly projected $\mathcal{H}$ times. On each of these projected versions of the inputs, the attention is performed in parallel, yielding the outputs, which are then concatenated (row-wise) and again projected, resulting in the final values. This allows the model to jointly attend to information from different representation subspaces at different positions.

$$\textsc{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{W^O}\text{Concat}(\boldsymbol{head}_1; ...; \boldsymbol{head}_{\mathcal{H}}) \tag{2.15}$$

where $\boldsymbol{head}_h = \textsc{Attention}(\boldsymbol{W}_h^{\boldsymbol{Q}}\boldsymbol{Q}, \boldsymbol{W}_h^{\boldsymbol{K}}\boldsymbol{K}, \boldsymbol{W}_h^{\boldsymbol{V}}\boldsymbol{V})$, the projection matrices $\boldsymbol{W}_h^{\boldsymbol{Q}}$, $\boldsymbol{W}_h^{\boldsymbol{K}}$, $\boldsymbol{W}_h^{\boldsymbol{V}}$ are of size $\frac{d_k}{\mathcal{H}} \times H$, $\boldsymbol{W^O}$ is of size $H \times H$ and $\mathcal{H}$ is the total number of attention heads. Please note that $d_k = d_v = H$ and is the column-wise dimension for attention matrix of each head. Due to the reduced dimension of each head, the total computational cost of multi-head attention is similar to that of single attention with full dimensionality.

There are three ways in which multi-head attention is utilised in the Transformer:

- **Multi-Head Self-Attention in Encoder** This is the attention of the current encoder layer to its input, denoted by $\textsc{MultiHead}_{self}(\boldsymbol{X}^{l-1})$. Here, the keys, values, and queries come from the same place, that is, the output of the previous encoder layer:

$$\textsc{MultiHead}_{self}(\boldsymbol{X}^{l-1}) = \textsc{MultiHead}(\boldsymbol{X}^{l-1}, \boldsymbol{X}^{l-1}, \boldsymbol{X}^{l-1})$$

- **Multi-Head Self-Attention in Decoder** This is the attention of the current decoder layer to its input (denoted by $\textsc{MultiHead}_{self}(\boldsymbol{Y}^{l-1})$) by allowing each position in the decoder to attend to all positions in the decoder up to and including that position (all future inputs are masked out):

$$\textsc{MultiHead}_{self}(\boldsymbol{Y}^{l-1}) = \textsc{MultiHead}(\boldsymbol{Y}^{l-1}, \boldsymbol{Y}^{l-1}, \boldsymbol{Y}^{l-1})$$

- **Multi-Head Source Attention** This is the attention of the current decoder layer to the output of the encoder, mimicking the attention mechanism in RNN-based encoder-decoder architecture. Here, the queries come from the multi-head self-attention sub-layer in the current decoder layer and the keys, values come from the output of the encoder:

$$\textsc{MultiHead}_{src}(\boldsymbol{Y}^{l}_{MHself}, \boldsymbol{X}^{L}) = \textsc{MultiHead}(\boldsymbol{Y}^{l}_{MHself}, \boldsymbol{X}^{L}, \boldsymbol{X}^{L})$$

Feed-Forward Network (FFN) Both encoder and decoder layers have a sub-layer containing a position-wise fully connected feed-forward network (FFN), defined as two linear transformations with a ReLU activation in between:

$$\text{FFN}(\boldsymbol{X}) = \boldsymbol{W}_{\boldsymbol{ff_2}}\text{ReLU}(\boldsymbol{W}_{\boldsymbol{ff_1}}\boldsymbol{X} + \boldsymbol{b}_{\boldsymbol{ff_1}}) + \boldsymbol{b}_{\boldsymbol{ff_2}} \tag{2.16}$$

where $\{\boldsymbol{W}, \boldsymbol{b}\}$ are parameters defined specifically for each layer and ReLU is as defined in Table 2.1.

2.1.3 Training and Decoding

Training All parameters in the encoder-decoder architecture (RNN-based or Transformer) are jointly trained via backpropagation (LeCun, 1988; Rumelhart et al., 1986) to minimise the negative log-likelihood (conditional) over the training set. The conditional log-likelihood is defined as the sum of the log-probability of predicting a correct symbol in the output sequence y_n for each instance $\boldsymbol{x}$ in the training set $\mathcal{D}$. Thus, we want to find the optimum set of parameters $\boldsymbol{\theta}^*$ as follows:

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} -\log P_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}) &(2.17)\\ &= \arg\min_{\boldsymbol{\theta}} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \sum_{n=1}^{|\boldsymbol{y}|} -\log P_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{y}_{<n}, \boldsymbol{x}) &(2.18)\end{aligned}$$

The most common method to find $\boldsymbol{\theta}^*$ is the gradient descent (GD) algorithm which updates the parameters in the opposite direction of the gradient of the objective function with respect to the parameters (Ruder, 2016), that is, at each step i of the training:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{2.19}$$

where $J(\boldsymbol{\theta}) \triangleq \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}} -\log P_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x})$ is the objective function, and η is the learning rate which determines the size of the steps that are taken to reach a (local) minimum. Eq. 2.19 is referred to as the vanilla gradient descent or batch gradient descent (BGD). In practice, BGD can be very slow and intractable since we need to calculate the gradients for the whole dataset to perform just one update. To mitigate this issue, stochastic gradient descent (SGD) is used which computes the parameter update for each training example, and thus, is much faster. However, SGD performs frequent updates with a high variance that cause the objective function to fluctuate heavily. Another variant of the gradient descent algorithm, the mini-batch gradient descent, computes the training update for a small batch of training examples and is able to reduce the variance of the parameter updates leading to a more stable convergence. Mini-batch gradient descent is usually the algorithm of choice when training any neural network and the term SGD is employed even when mini-batches are used.

Vanilla mini-batch gradient descent, however, does not guarantee good convergence and offers a few challenges. Firstly, choosing a good learning rate can be difficult. A learning rate that is too small can lead to a very slow training process, while a learning rate that is too large may cause the loss function to fluctuate heavily around the minimum and may even result in divergence. Secondly, SGD is non-adaptive, that is, the learning rate is fixed throughout the training process. Usually learning rate annealing needs to be employed so as to reduce the learning rate according to a pre-defined schedule or when the change in objective between epochs (one pass over the training set) falls below a certain threshold. This is necessary to avoid getting trapped in suboptimal local minima.

Many GD-based methods have been proposed to address the shortcomings of SGD (Ruder, 2016). Out of these, the most popular method that works well for training neural sequence models is the Adaptive Moment Estimation (Adam) (Kingma and Ba, 2015). As will be mentioned in the subsequent chapters, we use SGD or Adam for training our NMT models.

Decoding Having trained an NMT model, we need to be able to use it to translate or decode unseen source sentences. The best output sequence for a given input sequence is produced by:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) \tag{2.20}$$

Solving this optimisation problem exactly is computationally hard, and hence an approximate solution is obtained using greedy decoding or beam search.

The basic idea of greedy decoding is to pick the most likely word (having the highest probability) at each decoding step until the end-of-sentence token is generated. Beam search, on the other hand, keeps a fixed number (b) of translation hypotheses with the highest log-probability at each timestep. When the end-of-sentence token is appended to a hypothesis, it is removed from the beam and added to the final candidate list. The algorithm then picks the translation with the highest log-probability (normalised by the number of target words) from the final candidate list. If the number of candidates at each timestep is chosen to be one, beam search reduces to greedy decoding. In practice, the translation quality obtained via beam search (size of 4) is significantly better than that obtained via greedy decoding (Chen et al., 2018b). However, beam search is computationally very expensive (25%-50% slower depending on the base architecture and the beam size) in comparison to greedy decoding. Thus, we resort to greedy decoding in this work.

2.1.4 Regularisation Techniques for Neural Architectures

Here we briefly describe the two regularisation techniques that have been employed in our work.

Dropout This simple regularisation technique prevents a neural sequence-to-sequence model from overfitting (Srivastava et al., 2014). The main idea is to ignore or drop (with a probability p) certain hidden units, chosen at random, during the training of the model. For the attentional RNN-based encoder-decoder model, dropout can be applied to either the embedding layers, within the RNNs or to the output layer. The Transformer employs four types of dropouts (Chen et al., 2018a): (i) input dropout applied to the sum of token embeddings and positional encodings, (ii) residual dropout applied to the output of each sub-layer before adding to the sub-layer input, (iii) feedforward dropout applied to output of the feed-forward sub-layer, and (iv) attention dropout applied to attention weights in each attention sub-layer.

Label Smoothing This is another effective regularisation technique that prevents the model from being over-confident on output labels (Szegedy et al., 2016). It means that for a training example with ground-truth label y for a specific token, we slightly lower its correctness from 1 to $1 - \epsilon$, where ϵ is a pre-defined small value. More formally, if we have a set of $|V_T|$ labels, then we replace the label distribution $\log p(y|\boldsymbol{x})$ with $(1 - \epsilon) \log p(y|\boldsymbol{x}) + \frac{\epsilon}{|V_T|} \sum_k \log p(y_k|\boldsymbol{x})$ by imposing the 1 and 0 target classification with targets of probabilities $1 - \epsilon$ and $\frac{\epsilon}{|V_T|}$, respectively.

2.1.5 Evaluation

Now that we have described some recent MT models, and their training and decoding strategies, the final topic of discussion is how to evaluate the quality of the generated translations. The first automatic evaluation metric we are going to describe is BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) which has been a de-facto standard for evaluating translation outputs since it was proposed in 2002. The core idea is to aggregate the count of words and phrases (n-grams) that overlap between machine and reference translations. Mathematically, BLEU is calculated as:

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

where N is the maximum length of n-grams (usually up to 4), w_n are positive weights summing to one (usually chosen uniformly), and BP is the brevity penalty used to penalise translation outputs longer than translation references and is defined as:

$$\text{BP} = \begin{cases} 1, & \text{if } t > r. \\ \exp^{(1-r/c)}, & \text{if } t \leq r. \end{cases}$$

where t and r are the lengths of the translation output and reference, respectively, and p_n is the geometric average of the n-gram precisions defined as:

$$p_n = \frac{\sum_{\mathcal{C}} \sum_{\text{n-gram} \in \mathcal{C}} Count_{clip}(\text{n-gram})}{\sum_{\mathcal{C}'} \sum_{\text{n-gram}' \in \mathcal{C}'} Count(\text{n-gram}')}$$

where $Count_{clip}(\text{n-gram}) = \min(\text{count, max\_ref\_count})$ and is used to truncate the total count of each word in the output translation with its maximum count in the references. The BLEU metric ranges from 0 to 1 where 1 means an identical output with the reference.

Although BLEU correlates well with human judgment (Papineni et al., 2002), it relies on precision alone and does not take into account recall – the proportion of the matched n-grams out of the total number of n-grams in the reference translation. METEOR (Banerjee and Lavie, 2005; Lavie and Agarwal, 2007) was proposed to address the shortcomings of BLEU. It scores a translation output by performing a word-to-word alignment between the translation output and a given reference translation. If more than one reference translation is provided, the translation is scored against each reference independently, and the best scoring pair is used. The alignments are produced via a sequence of word-mapping modules: (i) the exact module maps two words if they are exactly the same, (ii) the stem module maps two words if they are the same after they are stemmed using the Porter stemmer, and (iii) the synonym module maps two words if they are synonyms of each other. After the word-mapping modules have identified an initial set of possible alignments, the largest subset of these word-mappings is identified such that the resulting set constitutes an alignment. If more than one such set is found, the alignment for which the word order in the two translations is most similar is selected. Now that a final alignment has been produced, the METEOR score is calculated as follows:

$$\text{METEOR} = F_{mean}(1 - Penalty)$$

where $Penalty = 0.5\Big(\dfrac{\#chunks}{m}\Big)^3$ and $F_{mean} = \dfrac{PR}{0.9P + 0.1R}$.

The $Penalty$ takes into account longer matches by dividing the sequence of matched words into the fewest possible number of "chunks" such that the matched words in each chunk are adjacent (in both output and reference) and have the same word-order. Thus, the longer the n-grams, the fewer the chunks, and in the extreme case, if the translation output exactly matches the reference translation, there is only one chunk. F_{mean} is just the parameterised harmonic mean of unigram precision and recall (Rijsbergen, 1979), where unigram precision P is the ratio of the number of mapped unigrams to the total number of unigrams in the output translation (m/t), while unigram recall R is the ratio of the number of mapped unigrams to the total number of unigrams in the reference translation (m/r). METEOR has also demonstrated to have a high level of correlation with human judgment, even outperforming that of BLEU (Banerjee and Lavie, 2005).

To make the results of the aforementioned MT evaluation metrics more reliable, a statistical significance test should be performed (Koehn, 2004) which indicates whether the difference in translation quality of two or more systems is due to a difference in true system quality. Although other MT evaluation metrics have been proposed, we only mention the most popular BLEU and METEOR, as these are sufficient for the purposes of this thesis.

2.2 Document-level Machine Translation

By this point, it must be clear to the reader that all MT models are built on strong independence assumptions whether it is based on locality assumptions within a sentence as done by phrase-based models or that outside the sentence as done by even the most advanced NMT models today. From a linguistics perspective, this assumption in practice is invalid, as any piece of text is much more than just a single sentence and making this assumption means ignoring the underlying discourse structure of the text (described shortly) and still hoping that the translation would not fall short. Although the problem of machine translation itself has been around for decades, the works which have tried to address the problem of discourse in MT are still just brushing the surface with more research yet to be undertaken.

In the rest of this section, I will describe the research on discourse in SMT (Section 2.2.1) and NMT (Section 2.2.2) followed by a description of how to evaluate translation outputs of larger pieces of text (Section 2.2.3).

2.2.1 Discourse in Statistical Machine Translation

A group of sentences that are contiguous, structured and exhibit coherency are regarded as a *discourse* (Jurafsky and Martin, 2009). In terms of SMT, we will be mentioning research which has tried to incorporate different aspects of discourse in SMT, beginning with the document-level discourse structure and moving on to specific discourse phenomena like pronominal anaphora, lexical cohesion and consistency, coherence, and discourse connectives.

2.2.1.1 Discourse and Document Structure

An initial work on discourse in MT by Marcu et al. (2000) (predating SMT) used a discourse transfer model to re-order the clauses and sentences of an input text (in Japanese) to make it closer to the natural discourse structure of text in a target language (English) and thus cater to the cross-lingual discourse shift. Almost a decade later, Foster et al. (2010) presented an SMT system for translating the Canadian Hansard corpus (parliamentary proceedings) in which they change the language model to incorporate structural features at the sentence-level (year, source language, speaker name, title, and section) without being explicity dependent on the content of the other sentences. Louis and Webber (2014) proposed a structured model for translating Wikipedia biography articles using a cache to encourage the use of article sub-structure (based on topics) by using words conforming to the smaller topic segments in the article.

It is a challenge to include document structure when training an SMT model, but a more challenging problem is to incorporate this information at the decoding stage. This is because the decoding of phrase-based SMT models not only relies on the sentence-independence assumption but is realised as a search for the highest-scoring translation in the space of exponentially possible translations that could be generated by the model (Koehn et al., 2003). A possible solution, proposed by Hardmeier et al. (2012, 2013a), is to start from an initial translation generated from a baseline decoder like Moses (Koehn, 2005) and make local changes to that translation via elementary operations (changing phrase translations or word-order, and resegmentation) and transform it into a better translation. This decoder, referred to as Docent, was followed up by Stymne et al. (2013) who incorporated readability constraints (including the ones to promote lexical consistency) into Docent to produce simplified translations; however, this resulted in deteriorated performance based on automatic evaluation.

2.2.1.2 Cohesion

Cohesion is a surface property of the text and refers to the way textual units are linked together grammatically or lexically (Halliday and Hasan, 1976). The first form, grammatical cohesion, is based on the logical and structural content, while the second, lexical cohesion, is based on the usage of semantically related words. Most research on discourse in SMT has focused on lexical cohesion while some has focused on grammatical cohesion in terms of pronominal anaphora.

Pronominal Anaphora Pronominal anaphora is the use of a pronoun to refer to someone or something mentioned previously in a text and is a challenging problem in MT due to the variation of the usage and distribution of pronouns across languages. This can only be dealt with access to inter-sentential context, specifically if the antecedent is not present in the same sentence. For example, a neutral pronoun in a source language (English) may have a gender-sensitive pronoun in the target language (German), requiring access to the antecedent to resolve the gender.

Initial attempts to exploiting anaphora information for the improvement of SMT systems, by Hardmeier and Federico (2010) using a word-dependency model to incorporate the output of a coreference resolution system in SMT, and by Le Nagard and Koehn (2010) using a two-pass approach, that includes annotations from a coreference system in the second pass, did not yield promising results. There have also been attempts to cross-lingual pronoun prediction by Novák and Žabokrtský (2014) and Hardmeier et al. (2013b) where the latter attempt to use anaphora links as latent variables in a neural network classifier.

Luong and Popescu-Belis (2016) proposed to use a pronoun-aware language model that determines a target pronoun based on the number and gender of preceding nouns or pronouns. Their method then re-ranks the translation hypotheses using the new LM and showed improvements over the baseline for English→French shared task in DiscoMT 2015. Luong and Popescu-Belis (2017) developed a fully probabilistic model that combines an additional translation model for pronouns, based on morphological and semantic features, with a Spanish→English SMT system to improve the translation of personal and possessive pronouns in Spanish to English. Miculicich Werlen and Popescu-Belis (2017a) presented a coreference-aware decoder for SMT based on similarity of coreference links in the source

(Spanish) and target (English) texts. Their post-editing scheme resulted in significant improvements in the accuracy of pronoun translation (Miculicich Werlen and Popescu-Belis, 2017b), while the BLEU scores remained constant.

Lexical Cohesion Lexical cohesion has two forms: repetition and collocation. The former is achieved through synonyms and hyponyms (sometimes also referred to as lexical consistency), while the latter uses related words that generally co-occur. There are three lines of work that try to incorporate lexical cohesion in SMT, by employing: (i) cache-based approaches, (ii) lexical chains, and (iii) two-pass approaches.

In terms of the first line of work, Tiedemann (2010) tried to promote lexical consistency in SMT by using adaptive language and translation models that use an exponentially decaying cache to carry over word preferences from one sentence to the next. Gong et al. (2011) also used a cache-based approach in which they employ three types of caches: (i) a dynamic cache (similar to Tiedemann (2010)) built using bilingual phrase pairs from the best translation hypotheses of previous sentences, (ii) a static cache which stores relevant bilingual phrase pairs extracted from similar bilingual documents, and (iii) a topic cache which stores the relevant target-side topic words. Their approach yielded significant improvements over the baseline in terms of BLEU score.

Falling into the second line of work, Xiong et al. (2013a) proposed a model that looks for lexical cohesion devices in the translation outputs of their MT system and then rewards the model for their appropriate usage based on conditional likelihood and mutual information. They reported significant improvements for Chinese→English SMT in terms of BLEU. Xiong et al. (2013b) presented a framework that attempts to incorporate lexical cohesion in the translations via lexical chains. The source document lexical chains are first identified and then projected to the target-side using maximum entropy classifiers. Then, a lexical cohesion based translation is generated from the target lexical chains by integrating their cohesion models into a hierarchical phrase-based SMT system. Instead of relying on lexical resources, the method proposed by Mascarell (2017) detects lexical chains in the source and tries to preserve the semantic similarity among the words in their corresponding lexical chains in the target via word embeddings. They integrated their model into Docent and through manual evaluation found that their model had a tendency to produce consistent translations of words in the chain.

The last line of work is based on incorporating document contexts into an initial translation obtained from a baseline MT system. Xiao et al. (2011) first identified ambiguous words in the source and then obtained a set of consistent translations for each word using the distribution of its translation over the target document, after which the phrase table is updated by removing inconsistent phrase-pairs and a second pass of decoding is performed. The semantic document language model in Hardmeier et al. (2012) rewarded the use of semantically related words (found based on latent semantic analysis) in the translation output thus promoting lexical cohesion. Garcia et al. (2014) proposed a two-pass approach to improve the translations already obtained by a sentence-level model. After the initial translation is obtained, they detect incorrect translations in the target document based on inconsistencies in meaning, gender and number disagreement among words, and suggest possible corrections. Their method did not yield improvement based on automatic evaluation which they claim to be due to the local changes made by their model. Garcia et al. (2015) designed a document-level scoring feature for lexical consistency by measuring the suitability of a word translation according to its context and its other possible translations in the document based on word embeddings. They also extended Docent to incorporate a new operation that guides the search process to yield consistent translations. Finally, Garcia et al. (2017) made use of bilingual word vector models as the semantic language model in Docent to enforce translation choices that are semantically similar to the context.

2.2.1.3 Coherence

As opposed to cohesion which is a surface property of the text, coherence refers to the underlying meaning relation between units of text and its continuity (Jurafsky and Martin, 2009). It is a stronger requirement for a piece of text to meet than is required by cohesion, and not only embodies cohesion, but other referential components like different parts of text referring to the same entities (entity-based coherence), and relational components like connections between utterances in a discourse via coherence relations (Hardmeier, 2014; Smith, 2018). Hence, coherence governs whether a text is semantically meaningful overall and how easily a reader can follow it.

Coherence has been explored for monolingual text but not much for bilingual text, like the one we deal with in MT. For SMT, the research in coherence mostly deals with studies that try to extend previously proposed coherence models for monolingual text to translation outputs (Smith and Specia, 2017; Smith, 2018). Smith et al. (2016) further extend these models by proposing a new method to learn the syntactic patterns in a text.

2.2.1.4 Discourse Connectives

Discourse connectives, also referred to as discourse markers or cue words, are the words that signal the existence of a specific discourse relation or discourse structure in the text. These are mostly domain-specific and may be implicit or explicit depending on the language. If implicit, these may be missed by the MT system in the translation although a human translator may be able to introduce them explicitly in the translation (Hatim and Mason, 1990). There have been studies that have tried to assess the ambiguity of discourse connectives for MT and have reported that the mismatches between implicit and explicit discourse connectives across languages result in deteriorated translation quality (Li et al., 2014a,b). Even explicitly annotating the discourse markers in the source text has a limited effect on translation quality for Chinese→English as reported by Yung et al. (2015) and Steele and Specia (2016).

Meyer et al. (2011) proposed to use an automatic scheme which annotates words with discourse sense by gathering informing from the different ways they are translated in their correct translations, also referred to as translation spotting (Cartoni et al., 2013). The impact of using this methodology was pretty low in terms of BLEU score for English-French (Meyer and Popescu-Belis, 2012).

2.2.1.5 Conclusion

After going through related work for discourse in SMT, it must be clear that incorporating discourse in SMT is a hard problem due to the various components in the SMT pipeline and the reliance on well-crafted and intuitive hand-engineered features for the various discourse phenomena. Furthermore, SMT is not very good at handling sentence-level phenomena such as syntactic reordering and long-distance agreement. Even if one can improve the discourse characteristics of a MT system output (that frequently contains local grammatical mistakes) via a post-editing step, noise from local errors make such improvements difficult to measure. These were the main reasons that for a long time the MT community was put off to pursue valuable research in this area, mostly resulting in studies which highlighted the importance of pursuing document-level MT but less hands-on work which actually attempted to do it.

2.2.2 Discourse in Neural Machine Translation

Up until two years ago, there was no work in NMT that tried to incorporate any type of discourse phenomena mentioned previously, but with most sentence-based NMT systems achieving state-of-the-art performance compared to their SMT counterparts, this area of research has finally started to gain the popularity it deserves. The main difference between the research on discourse in NMT and SMT, apart from the general building blocks, is that the works in NMT rarely try to model discourse phenomena explicitly. On the contrary, they use sentences in the context directly via different modelling techniques and show how they perform on automatic evaluation while sometimes measuring the performance on specific test sets.

The first work that we mention here is by Jean et al. (2017) who augment the attentional RNN-based NMT architecture with an additional attentional component over the previous source sentence. The context vector generated from this source-context attention is then added as an auxiliary input to the decoder hidden state. Through automatic evaluation and cross-lingual pronoun prediction, they found that although their approach yielded moderate improvements for a smaller training corpus, there was no improvement when the training set was much larger. Furthermore, their method suffered from an obvious limitation: an additional attention component meaning that their method could only incorporate limited context. Around the same time, Tiedemann and Scherrer (2017) conducted a pilot study in which they extend the translation units in two ways: (i) only extend the source sentence to include a single previous sentence, and (ii) extend both source and target sentences to include previous sentence in the corresponding context, without changing the underlying RNN-based NMT model. They again reported marginal improvements in terms of BLEU for German→English subtitle translation, but through further analysis and manual evaluation found output examples in which referential expressions across sentence boundaries could be handled properly.

The first work that yielded significant improvements over a sentence-based NMT model in terms of automatic evaluation was by Wang et al. (2017). They use a two-level hierarchical RNN to summarise the information in three previous source sentences, where the first-level RNNs are run over individual sentences and the second-level RNN is run over the single output vectors produced from the first-level RNN over each sentence. The final summary vector is then used to initialise the decoder, or added as an auxiliary input to the decoder state directly or after passing through a gate. Their approach showed promising results when using source-side context even though they found that considering target-side history inversely harmed translation performance. We will show in Chapter 3 an effective strategy to fruitfully incorporate the global target-side context in NMT.

There have also been two approaches that use cache to store relevant information from a document and then use this external memory to improve the translation quality (Tu et al., 2018; Kuang et al., 2018). The first of these approaches by Tu et al. (2018) uses a continuous cache to store recent hidden representations from the bilingual context, that is the key is designed to help match the query (current context vector produced via attention) to the source-side context, while the value is designed to help find the relevant target-side information to generate the next target word. The final context vector from the cache is then combined with the decoder hidden state via a gating mechanism. The cache has a finite length and is updated after generating a complete translation sentence. Their experiments on multi-domain Chinese→English datasets showed the effectiveness of their approach with negligible impact on the computational cost. The second approach by Kuang et al. (2018) uses dynamic and topic caches (similar to the ones in Gong et al. (2011)) to store target words from the preceding sentence translations and a set of target-side topical words semantically related to the source document, respectively. As opposed to the cache in Tu et al. (2018), their dynamic cache follows a first-in, first-out scheme and is updated after generating each target word. At each decoding step, the target words in the final cache are scored and a gating mechanism is used to combine the score from the cache and the one produced by the NMT model. Their experimental results on the NIST Chinese→English translation task revealed that the cache-based neural model achieved consistent and significant improvements in terms of translation quality.

More recent works in NMT have started to use the new state-of-the-art Transformer architecture (Vaswani et al., 2017) as the base model. Voita et al. (2018) change the encoder in the Transformer to a context-aware encoder which has two sets of encoders, a source encoder and a context encoder, with the first $L-1$ layers shared. The previous source sentence serves as input to the context encoder and its output is attended to by the L^{th} layer of the source encoder, and then combined with the source encoder output using a gate. The final output of the context-aware encoder is then fed into the decoder. Their experiments on English→Russian subtitles data and analysis on the effect of context information for translating pronouns revealed that their model implicitly learned anaphora resolution which is quite promising as the model used no specialised features. Along similar lines, Zhang et al. (2018) also use a context-aware encoder in the Transformer, however, instead of training their model from scratch, like Voita et al. (2018), they use pre-trained embeddings from the sentence-based Transformer as input to their context encoder. In the second stage of training, they only learn the document-level parameters and do not fine-tune the sentence-level parameters of their model similar to Tu et al. (2018). They experimented with NIST Chinese→English and IWSLT French→English translation tasks and reported significant gains over the baseline in terms of BLEU score.

Inspired from Yang et al. (2016), Miculicich et al. (2018) use three previous sentences as context by employing a hierarchical attention network (HAN) having two levels of abstraction: the word-level abstraction allows to focus on words in previous sentences, and the sentence-level abstraction allows access to relevant sentences in the context for each query word. They combine the contextual information with that from the current sentence using a gate. The context is used during encoding or decoding a word, and is taken from previous source sentences or previously decoded target sentences. Their experiments on Chinese→English and Spanish→English datasets demonstrated significant improvements in terms of BLEU. They further evaluated their model based on noun and pronoun translation and lexical cohesion and coherence, but did not report whether the gains achieved by their model were statistically significant with respect to the baseline.

Similar to the two-pass approaches in SMT, Xiong et al. (2019) use a two-pass decoder approach to encourage coherence in NMT. In the first pass, they generate locally coherent preliminary translations for each sentence using the Transformer architecture. In the second step, their decoder refines the initial translations with the aid of a reward teacher (Bosselut et al., 2018) which promotes coherent translations by minimising the similarity between a sequence encoded in its forward and reverse direction. Their model improved the translation quality in terms of sentence-level and document-level BLEU and METEOR scores where the document-level scores were measured by concatenating sentences in one document into one long sentence and then using the traditional metrics.

In conclusion, a lot of work remains to be done in the field of document-level NMT, even though there is more promising work now than until a few years ago. In this thesis, our contribution mainly falls in the category of modelling document-level context in NMT for both monologues and dialogue, in contrast to previous work which has only focused on monologue translation. We also present effective decoding methodologies for our models which use the document-wide context as opposed to local context used in most previous works.

2.2.3 Evaluation

MT outputs are almost always evaluated using metrics like BLEU and METEOR which use n-gram overlap between the translation and reference to judge translation quality; however, these metrics do not look for specific discourse phenomena in the translation, and thus may fail when it comes to evaluating the quality of longer pieces of generated text. There has been some work in terms of proposing new evaluation metrics for specific discourse phenomena (described shortly), which may seem promising but there is no consensus among the MT community about their usage. There are also those that suggest using evaluation test sets or better yet combining them with semi-automatic evaluation schemes (Guillou and Hardmeier, 2018). More recently, Stojanovski and Fraser (2018) propose to use oracle experiments for evaluating the effect of pronoun resolution and coherence in MT.

Automatic Evaluation for Specific Discourse Phenomena There have been a few works which have proposed reference-based automatic evaluation metrics for evaluating specific discourse phenomena. For pronoun translation, the first metric proposed by Hardmeier and Federico (2010) measures their precision and recall directly. Firstly, word alignments are produced between the source and translation output, and the source and the reference translation. For each pronoun in the source, a clipped count (described in Section 2.1.5 for BLEU (Papineni et al., 2002)) is computed, defined as the number of times the pronoun occurs in the translation output, limited by the number of times it occurs in the reference translation. The final metric is then the precision, recall or F-score based on these clipped counts. Miculicich Werlen and Popescu-Belis (2017b) proposed a metric that estimates the accuracy of pronoun translation (APT), that is for each source pronoun, it counts whether its translation can be considered correct. It first identifies triples of pronouns: (source pronoun, reference pronoun, candidate pronoun) based on word alignments which are improved through heuristics. Next, the translation of a source pronoun in the MT output and the reference are compared and the number of identical, equivalent, or different/incompatible translations in the output and reference, as well as cases where candidate translation is absent, reference translation is absent or both, are counted. Each of these cases is assigned a weight between 0 and 1 to determine the level of correctness of MT output given the reference. The weights and the counts are then used to compute the final score. Most recently, Jwalapuram et al. (2019) proposed a specialised evaluation measure for pronoun evaluation which is trained to distinguish a good translation from a bad one based on pairwise evaluations between two candidate translations (with or without past context). The measure performs the evaluation irrespective of the source language and is shown to be highly correlated with human judgments. They also present a targeted pronoun test suite that covers multiple source languages and various target pronouns in English. Both their test set and evaluation measure are based on actual MT system outputs.

For lexical cohesion, Wong and Kit (2012) extended the sentence-level evaluation metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Translation Edit Rate) (Snover et al., 2006) to incorporate a feature that scores lexical cohesion. To compute the new score, they identify lexical cohesion devices via clustering based on WordNet (Fellbaum, 1998) and repetition via stemming, and then combine this score with the sentence-level one through weighted average. They claimed that this new scoring feature increases the correlation of BLEU and TER with human judgments, but does not have any effect on the correlation of METEOR. Along similar lines, Gong et al. (2015) augmented a cohesion score, based on simplified lexical chain, and a gist consistency score, based on topic model, with document-level BLEU or METEOR (concatenating sentences in one document into one long sentence and applying the traditional metrics) using a weighted average. Their hybrid metrics could obtain significant improvements for BLEU but only slight improvements for METEOR.

For discourse connectives, Hajlaoui and Popescu-Belis (2013) proposed new automatic and semi-automatic metrics referred to as ACT (Accuracy of Connective Translation) (Meyer et al., 2012). For each connective in the source, ACT counts one point if the translations are the same and zero otherwise based on a dictionary of possible translations and word alignments. The insertion of connectives is counted manually. The final score is the total number of points divided by the number of source connectives. Guzmán et al. (2014) used discourse structure for improving MT evaluation. They developed two discourse-aware evaluation metrics, which first generate discourse trees for the translation output and reference using a discourse parser (lexicalised and un-lexicalised) followed by a similarity measure between the two. This is based on the assumption that good translations would have a similar discourse structure to that of the reference. Smith and Specia (2018) proposed a reference-independent metric that assesses the translation output based on the source text by measuring the extent to which the discourse connectives and relations are preserved in the translation. Their metric combines bilingual word embeddings pre-trained for discourse connectives with a score reflecting the correctness of the discourse relation match. However, their metric depends on other lexical elements like a parser which may miss some constituents or discourse relations.

Guillou and Hardmeier (2018) studied the performance of automatic metrics for pronouns, proposed by Hardmeier and Federico (2010) and Miculicich Werlen and Popescu-Belis (2017b), on the PROTEST test suite (Guillou and Hardmeier, 2016) of English→French translations and explored the extent to which automatic evaluation based on reference translations can provide useful information about an MT system's ability to handle pronouns. They found that such automatic evaluation can capture some linguistic patterns better than others and recommend emphasising high precision in the automatic metrics and referring negative cases to human evaluators. It has also been suggested that to take

MT to another level, “the outputs need to be evaluated not based on a single reference translation, but based on notions of fluency and of adequacy – ideally with reference to the source text” (Sim Smith, 2017).

Evaluation Test Sets Inspired by examples from OpenSubtitles2016 (Lison and Tiedemann, 2016), Bawden et al. (2018) hand-crafted two contrastive test sets for evaluating anaphoric pronoun translation and coherence and cohesion in English→French translation. Models are then assessed on their ability to rank the correct translation of a sentence in the test set higher than the incorrect translation. Müller et al. (2018) presented a contrastive test suite to evaluate the accuracy with which NMT models translate the English pronoun *it* to its German counterparts *es*, *sie* and *er*. Such an evaluation using test suites is feasible but has a restricted scope since it is for specific language-pairs.

In conclusion, there is no consensus in the MT community about how to evaluate documents. A recent study contrasting the evaluation of individual sentences and entire documents with the help of human raters found that they prefer human translations over machine translated ones when assessing adequacy and fluency of translations (Läubli et al., 2018). Hence, as translation quality improves, there is a dire need for document-level evaluation since errors related to discourse phenomena remain invisible in a sentence-level evaluation. For the purposed of this thesis, we still use automatic evaluation (BLEU and METEOR) following prior work, but also validate the performance of our models through extensive analysis as deemed necessary.

Part I

Monologue Translation

Chapter 3

Document Context Modelling with Coarse Attention

For many years, document-level machine translation, similar to sentence-based statistical machine translation (SMT), suffered from one major drawback: the use of hand-crafted features. This resulted in models which were restrictive and failed to achieve desirable results upon automatic evaluation. Thus, document-level MT was overlooked in MT research with the works in this field being few and far between. Neural machine translation (NMT) eliminates the need of hand-engineering complex features by having a single big neural network (having millions of parameters) designed to model the entire MT process (Sutskever et al., 2014; Cho et al., 2014b). Would the success of neural networks for sentence-level machine translation be enough to reinstigate the significance of document context for machine translation?

In this chapter,[1] we answer this question by augmenting the generic sentence-based NMT model with two external memory components to capture the documental interdependencies in an offline setting (past and future context) for the task of monologue translation. Our goal is to demonstrate that using both the source and target-side context is lucrative for enhancing NMT performance. While this thesis was in preparation, there have been a few works which use local context with promising results (Wang et al., 2017; Bawden et al., 2018; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018) but ours is the first work to successfully incorporate target context and global context, aka document-wide information, in general.

[1] First presented in Maruf and Haffari (2018).

This chapter serves a three-fold purpose: to demonstrate how global context information can be incorporated into an otherwise context-agnostic NMT model, how to train the model end-to-end and to propose an iterative decoding algorithm based on block coordinate descent for the model.

3.1 Introduction

With the resurgence of neural networks, neural machine translation (NMT) has proven to be powerful (Sutskever et al., 2014; Bahdanau et al., 2015). It is on-par, and in some cases, even surpasses the traditional statistical MT (Luong et al., 2015) while enjoying more flexibility and significantly less manual effort for feature engineering. Despite their flexibility, most neural MT models still translate sentences independently. Discourse phenomena such as pronominal anaphora and lexical consistency, depending on long-range dependencies going farther than a few previous sentences, are thus neglected in sentence-based translation (Bawden et al., 2018).

There are only a handful of attempts to document-wide machine translation in statistical and neural MT camps. Hardmeier and Federico (2010); Gong et al. (2011); Garcia et al. (2014) propose document translation models based on statistical MT but are restrictive in the way they incorporate the document-level information and fail to gain significant improvements. More recently, there have been a few attempts to incorporate source-side context into neural MT (Jean et al., 2017; Wang et al., 2017; Bawden et al., 2018); however, these works only consider a very local context including a few previous source/target sentences, ignoring the global source and target documental contexts. The latter two report deteriorated performance when using the target-side context.

In this chapter, we present a document-level machine translation model which combines sentence-based NMT (Bahdanau et al., 2015) with memory networks (Sukhbaatar et al., 2015). We capture the global source and target document contexts with two memory components, one each for the source and target-side, and incorporate it into the sentence-based NMT by changing the decoder to condition on it as the sentence translation is generated. We conduct experiments on three language-pairs: French-English, German-English and Estonian-English. The experimental results and analysis demonstrate that our model is effective in exploiting both source and target document context, and statistically significantly outperforms the previous work in terms of BLEU and METEOR.

3.2 Preliminaries

Memory Networks (MemNets) Memory networks (Weston et al., 2015) are a class of neural models that use external memories to perform inference based on long-range dependencies. A memory is a collection of vectors $\boldsymbol{M} = \{\boldsymbol{m}_1, .., \boldsymbol{m}_K\}$ constituting the memory cells, where each cell $\boldsymbol{m}_k$ may potentially correspond to a discrete object $\boldsymbol{x}_k$. The memory is equipped with a *read* and optionally a *write* operation. Given a query vector $\boldsymbol{q}$, the output vector generated by reading from the memory is $\sum_{k=1}^{K} p_k \boldsymbol{m}_k$, where p_k represents the relevance of the query to the k^{th} memory cell defined as $\boldsymbol{p} = \text{softmax}(\boldsymbol{q}^\top \boldsymbol{M})$. For the rest of this chapter, we denote the read operation by $\textsc{MemNet}(\boldsymbol{M}, \boldsymbol{q})$.

3.3 Document NMT as Structured Prediction

We formulate document-wide machine translation as a *structured* prediction problem. Given a set of sentences $\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^{|\boldsymbol{d}|}\}$ in a source document $\boldsymbol{d}$, we are interested in generating the collection of their translations $\{\boldsymbol{y}^1, \dots, \boldsymbol{y}^{|\boldsymbol{d}|}\}$ by taking into account *interdependencies* among them imposed by the document. We achieve this by the factor graph in Figure 3.1, which represents a conditional random field (CRF) (Lafferty et al., 2001) that models the probability of the target document given the source document.

Our model has two types of factors:

- $f_{\boldsymbol{\theta}}(\boldsymbol{y}^j; \boldsymbol{x}^j, \boldsymbol{X}^{-j})$ to capture the interdependencies between the translation $\boldsymbol{y}^j$, the corresponding source sentence $\boldsymbol{x}^j$ and all the other sentences in the source document $\boldsymbol{X}^{-j}$, and

- $g_{\boldsymbol{\theta}}(\boldsymbol{y}^j; \boldsymbol{Y}^{-j})$ to capture the interdependencies between the translation $\boldsymbol{y}^j$ and all the other translations in the document $\boldsymbol{Y}^{-j}$.

Hence, the probability of a document translation given the source document is given by

$$P(\boldsymbol{y}^1, \dots, \boldsymbol{y}^{|\boldsymbol{d}|} | \boldsymbol{x}^1, \dots, \boldsymbol{x}^{|\boldsymbol{d}|}) \propto \exp\Big(\sum_j f_{\boldsymbol{\theta}}(\boldsymbol{y}^j; \boldsymbol{x}^j, \boldsymbol{X}^{-j}) + g_{\boldsymbol{\theta}}(\boldsymbol{y}^j; \boldsymbol{Y}^{-j}) \Big) \tag{3.1}$$

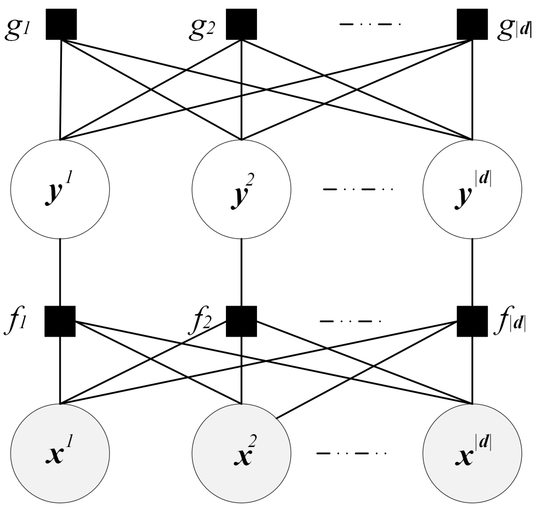


Figure 3.1: Factor graph for document-level MT

The factors $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\theta}}$ are realised by neural architectures (explained shortly) whose parameters are collectively denoted by $\boldsymbol{\theta}$.

Training It is challenging to train the model parameters by maximising the (regularised) likelihood since computing the partition function of Eq. 3.1 is hard.[2] This is due to the enormity of factors $g_{\boldsymbol{\theta}}(\boldsymbol{y}^j; \boldsymbol{Y}^{-j})$ over a large number of translation variables $\boldsymbol{y}^j$'s (i.e., the number of sentences in the document) as well as their unbounded domain (i.e., all sentences in the target language). Thus, we resort to maximising the *pseudo-likelihood* (Besag, 1975) for learning the parameters:

$$\arg\max_{\boldsymbol{\theta}} \prod_{d \in \mathcal{D}} \prod_{j=1}^{|\boldsymbol{d}|} P_{\boldsymbol{\theta}}(\boldsymbol{y}^j | \boldsymbol{x}^j, \boldsymbol{Y}^{-j}, \boldsymbol{X}^{-j}) \tag{3.2}$$

where $\mathcal{D}$ is the set of bilingual training documents. We directly model the document-conditioned NMT model $P_{\boldsymbol{\theta}}(\boldsymbol{y}^j | \boldsymbol{x}^j, \boldsymbol{Y}^{-j}, \boldsymbol{X}^{-j})$ using a neural architecture that subsumes both the $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\theta}}$ factors (covered in the next section).

Decoding To generate the best translation for a document according to our model, we need to solve the following optimisation problem:

$$\arg\max_{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^{|\boldsymbol{d}|}} \prod_{j=1}^{|\boldsymbol{d}|} P_{\boldsymbol{\theta}}(\boldsymbol{y}^j | \boldsymbol{x}^j, \boldsymbol{Y}^{-j}, \boldsymbol{X}^{-j})$$

[2] The partition function is given by: $\sum_{\boldsymbol{y}^j} \exp\left(\sum_j f_{\boldsymbol{\theta}}(\boldsymbol{y}^j; \boldsymbol{x}^j, \boldsymbol{X}^{-j}) + g_{\boldsymbol{\theta}}(\boldsymbol{y}^j; \boldsymbol{Y}^{-j})\right)$

which is hard (due to similar reasons as mentioned earlier). We hence resort to a block coordinate descent optimisation algorithm. More specifically, we initialise the translation of each sentence using the base neural MT model $P(\boldsymbol{y}^j|\boldsymbol{x}^j)$. We then repeatedly visit each sentence in the document and update its translation using our document-context dependent NMT model $P(\boldsymbol{y}^j|\boldsymbol{x}^j, \boldsymbol{Y}^{-j}, \boldsymbol{X}^{-j})$ while the translations of other sentences are kept fixed.

3.4 Context-Dependent NMT with MemNets

We augment the sentence-level attentional NMT model (RNN-based) by incorporating the document context (both source and target) using memory networks when generating the translation of a sentence, as shown in Figure 3.2.

Our model generates the target translation word-by-word from left to right, similar to the vanilla attentional neural translation model. However, it conditions the generation of a target word not only on the previously generated words and the current source sentence (as in the vanilla NMT model), but also on all the other source sentences in the document $\boldsymbol{X}^{-j}$ and their translations $\boldsymbol{Y}^{-j}$. That is, the generation process is as follows:

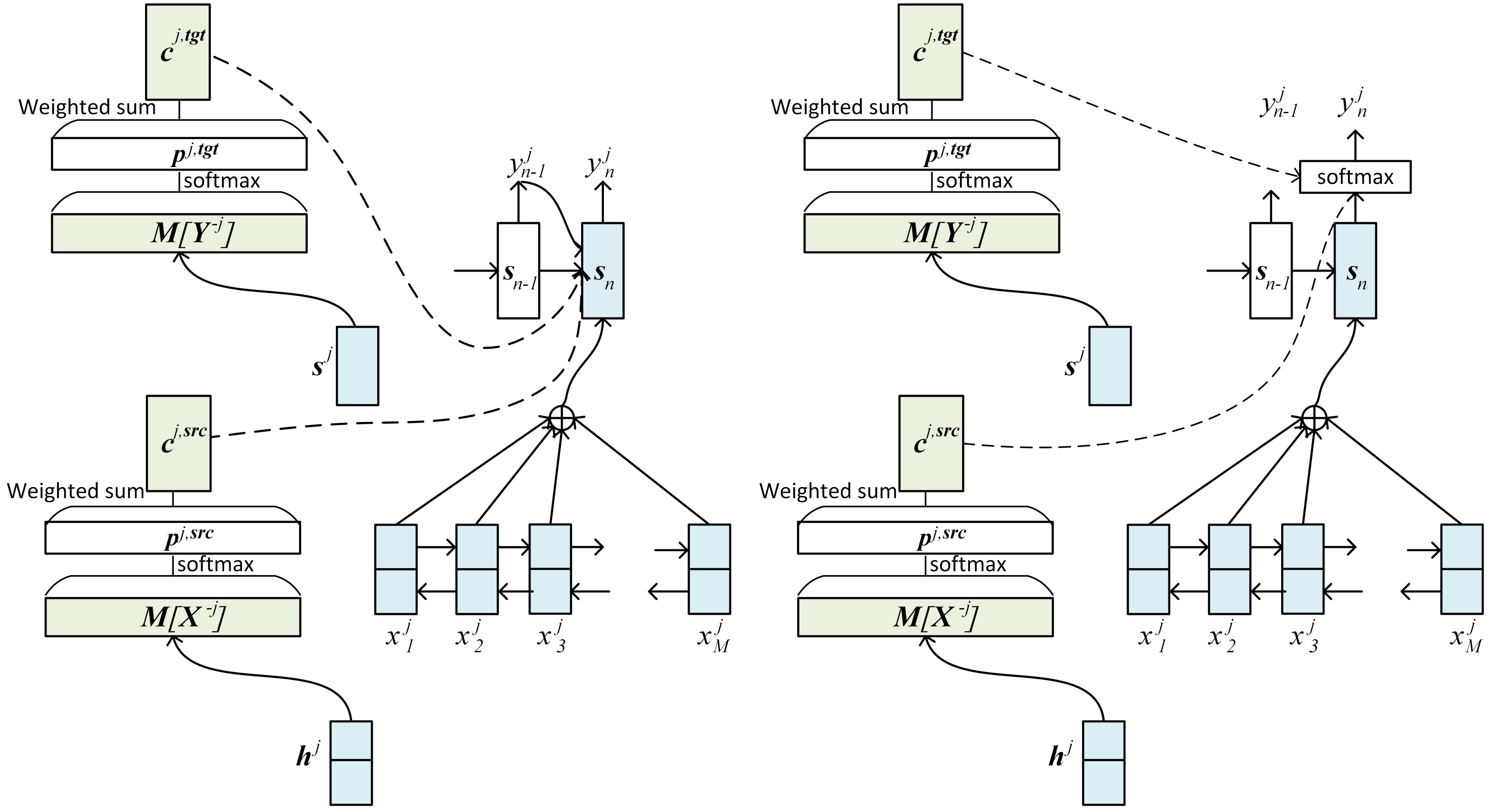
$$P_{\boldsymbol{\theta}}(\boldsymbol{y}^j|\boldsymbol{x}^j, \boldsymbol{Y}^{-j}, \boldsymbol{X}^{-j}) = \prod_{n=1}^{N} P_{\boldsymbol{\theta}}(y_n^j|\boldsymbol{y}_{<n}^j, \boldsymbol{x}^j, \boldsymbol{Y}^{-j}, \boldsymbol{X}^{-j}) \tag{3.3}$$

where y_n^j is the n^{th} word of the j^{th} target sentence, $\boldsymbol{y}_{<n}^j$ are the previously generated words, and $\boldsymbol{X}^{-j}$ and $\boldsymbol{Y}^{-j}$ are as introduced in the List of Notations.

Our model represents the source and target document contexts as external memories and *attends* to relevant parts of these external memories when generating the translation of a sentence. Let $\boldsymbol{M}[\boldsymbol{X}^{-j}]$ and $\boldsymbol{M}[\boldsymbol{Y}^{-j}]$ denote external memories representing the source and target document context, respectively. These contain memory cells corresponding to all sentences in the document except the j^{th} sentence (described shortly). Let $\boldsymbol{h}^j$ and $\boldsymbol{s}^j$ be representations of the j^{th} source sentence and its current translation, from the encoder and decoder respectively. We make use of $\boldsymbol{h}^j$ as the query to get the relevant *context* from the source external memory:

$$\boldsymbol{c}^{j,src} = \textsc{MemNet}(\boldsymbol{M}[\boldsymbol{X}^{-j}], \boldsymbol{h}^j)$$

where MEMNET(.) is as defined in Section 3.2 and the attention is *coarse* since the values are abstract representations of sentences in the document rather than of words. Furthermore,



(a) Memory-to-Context

(b) Memory-to-Output

Figure 3.2: Our document NMT model consisting of a sentence-based NMT model with source and target external memories.

for the j^{th} sentence, we get the relevant information from the target context as follows:

$$\boldsymbol{c}^{j,\boldsymbol{tgt}} = \textsc{MemNet}(\boldsymbol{M}[\boldsymbol{Y}^{-j}], \boldsymbol{s}^j + \boldsymbol{W_{at}} \cdot \boldsymbol{h}^j)$$

where the query consists of the representation of the translation $\boldsymbol{s}^j$ from the decoder endowed with that of the source sentence $\boldsymbol{h}^j$ from the encoder to make the query robust to potential noises in the current translation and circumvent error propagation, and $\boldsymbol{W_{at}}$ projects the source representation into the hidden state space.

Now that we have representations of the relevant source and target document contexts, Eq. 3.3 can be re-written as:

$$P_{\boldsymbol{\theta}}(\boldsymbol{y}^j|\boldsymbol{x}^j, \boldsymbol{Y}^{-j}, \boldsymbol{X}^{-j}) = \prod_{n=1}^{N} P_{\boldsymbol{\theta}}(y_n^j|\boldsymbol{y}_{<n}^j, \boldsymbol{x}^j, \boldsymbol{c}^{j,\boldsymbol{tgt}}, \boldsymbol{c}^{j,\boldsymbol{src}}) \tag{3.4}$$

More specifically, the memory contexts $\boldsymbol{c}^{j,\boldsymbol{src}}$ and $\boldsymbol{c}^{j,\boldsymbol{tgt}}$ are incorporated into the NMT decoder as:

- **Memory-to-Context** in which the memory contexts are incorporated when computing the next decoder hidden state (Figure 3.2a):

$$\boldsymbol{s}_n^j = \text{RNN}(\boldsymbol{s}_{n-1}^j, \boldsymbol{E_T}[y_{n-1}^j], \boldsymbol{c}_n^j, \boldsymbol{c}^{j,\boldsymbol{src}}, \boldsymbol{c}^{j,\boldsymbol{tgt}})$$

- **Memory-to-Output** in which the memory contexts are incorporated in the output layer (Figure 3.2b):

$$y_n^j \sim \text{softmax}(\boldsymbol{W_y}\boldsymbol{u}_n^j + \boldsymbol{W_{ys}}\boldsymbol{c}^{j,\boldsymbol{src}} + \boldsymbol{W_{yt}}\boldsymbol{c}^{j,\boldsymbol{tgt}} + \boldsymbol{b_y})$$

where $\boldsymbol{W_{ys}}$, and $\boldsymbol{W_{yt}}$ are the new parameter matrices. We use only the source, only the target, or both external memories as the additional conditioning contexts. Furthermore, we use either the Memory-to-Context or Memory-to-Output architectures for incorporating the document contexts. In the experiments, we will explore these different options to investigate the most effective combination. We now turn our attention to the construction of the external memories for the source and target sides, $\boldsymbol{M}[\boldsymbol{X}^{-j}]$ and $\boldsymbol{M}[\boldsymbol{Y}^{-j}]$ respectively, of a document.

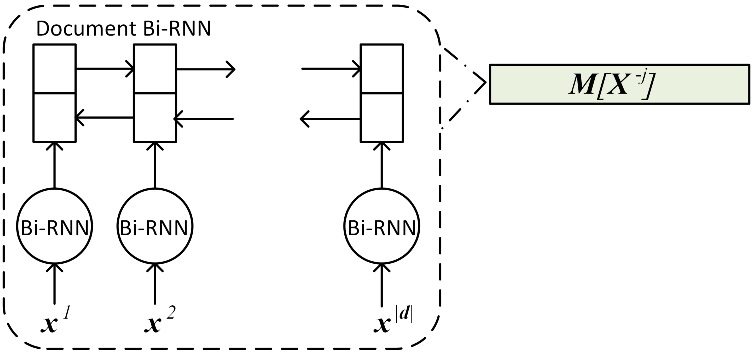


Figure 3.3: Hierarchical RNNs for source memory.

The Source Memory We make use of a hierarchical two-level RNN architecture to construct the external memory of the source document. More specifically, for the first level, we pass each sentence of the document through a sentence-level bi-directional RNN to get the representation of the sentence, i.e., we run two RNNs - one in the forward and one in the backward direction, and get the sentence representation by concatenating the last hidden states of the forward and backward RNNs. We then pass these sentence representations through a document-level bi-directional RNN to propagate sentences' information across the document as shown in Figure 3.3. We take the hidden states of the document-level bi-directional RNNs as the memory cells of the source external memory.

The source external memory is built once for each mini-batch and does not change throughout the document translation. To be able to fit the computational graph of the document NMT model within GPU memory limits, we pre-train the sentence-level bi-directional RNN using the language modelling training objective $\left(\prod_{m=1}^{M} P_{\phi}(x_m|\boldsymbol{x}_{<m})\right)$ on the original and reverse sentence independently; in other words, we train them as RNNLMs. However, the document-level bi-directional RNN is trained together with other parameters of the document NMT model by backpropagating the document translation training objective (Eq. 3.4).

The Target Memory The memory cells of the target external memory represent the current translations of the document. Recall from the previous section that we use coordinate descent iteratively to update these translations. Let $\{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^{|\boldsymbol{d}|}\}$ be the current translations, and let $\{\boldsymbol{s}^{|\boldsymbol{y}^1|}, \ldots, \boldsymbol{s}^{|\boldsymbol{y}^{|\boldsymbol{d}|}|}\}$ be the last states of the decoder when these translations were generated. We use these last decoder states as the cells of the external target memory.

	#Documents	#Sentences	Document length	Src/Tgt Vocab
French-English	1000/120/153	123K/15K/19K	123/128/124	25.1K/21K
Estonian-English	15K/1000/1776	209K/14K/25K	14/14/14	48.6K/24.9K
German-English	4871/87/110/160	191K/2K/3K/3K	39/23/27/19	45.1K/34.7K

Table 3.1: Training/development/test corpora statistics: number of documents and sentences (K stands for thousands), average document length (in sentences) and source/target vocabulary size (in thousands). For German-English, we report statistics of the two test sets `news-test2011` and `news-test2016`.

We could make use of hierarchical sentence-document RNNs to transform the document translations into memory cells (similar to what we do for the source memory); however, it would have been computationally expensive and may have resulted in error propagation. We will show in the experiments that our efficient target memory construction is indeed effective.

3.5 Experiments

3.5.1 Setup

Datasets We conducted experiments on three language-pairs: French-English, German-English and Estonian-English. Table 3.1 shows the statistics of the datasets used in our experiments. The French-English dataset is based on the TED talks corpus[3] (Cettolo et al., 2012) where each talk is considered a document. The Estonian-English data comes from the Europarl v7 corpus[4] (Koehn, 2005). Following Smith et al. (2013), we split the speeches based on the `SPEAKER` tag and treat them as documents. The French-English and Estonian-English corpora were randomly split into training, development and test sets. For German-English, we use the News Commentary v9 corpus for training,[5] `news-dev2009` for development, and `news-test2011` and `news-test2016` as the test sets. This corpus already has document boundaries provided.

We pre-processed all corpora to remove very short documents and those with missing translations. Out-of-vocabulary and rare words (frequency less than 5) are replaced by the `<unk>` token, following Cohn et al. (2016).[6]

[3] `https://wit3.fbk.eu/`

[4] `http://www.statmt.org/europarl/`

[5] `http://statmt.org/wmt14/news-commentary-v9-by-document.tgz`

[6] For this work, we did not split words into subwords using byte-pair encoding (BPE) (Sennrich et al., 2016). However, we will show later (Section 4.4) that our model can be extended to do that with minimum effort.

Evaluation Measures We use BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) scores to measure the quality of the generated translations. We use bootstrap resampling (Clark et al., 2011) to measure statistical significance, $p < 0.05$, upon comparison to the baselines.

Implementation and Hyperparameters We implement our document-level neural machine translation model in C++ using the `DyNet` library (Neubig et al., 2017), on top of the RNN-based sentence-level NMT implementation in `mantis` (Cohn et al., 2016). For the source memory, the sentence and document-level bi-directional RNNs use LSTM and GRU units, respectively. The translation model uses GRU units for the bi-directional RNN encoder and the two-layer RNN decoder. GRUs are used instead of LSTMs to reduce the number of parameters in the main model. The RNN hidden dimensions and word embedding sizes are set to 512 in the translation and memory components, and the alignment dimension is set to 256 in the translation model.

Training We use a stage-wise method to train the variants of our document-context NMT model. Firstly, we pre-train the Memory-to-Context/Memory-to-Output models, setting their *readings* from the source and target memories to the zero vector. This effectively learns parameters associated with the underlying sentence-based NMT model, which is then used as initialisation when training *all* parameters in the second stage (including fine-tuning the ones from the first stage). For the first stage, we make use of stochastic gradient descent (SGD)[7] with an initial learning rate of 0.1 and a decay factor of 0.5 after the fourth epoch for a total of ten epochs. The convergence occurs in 6-8 epochs. For the second stage, we use SGD with an initial learning rate of 0.08 and a decay factor of 0.9 after the first epoch for a total of 15 epochs.[8] The best model is picked based on minimum perplexity on development set. To avoid overfitting, we employ dropout with a rate of 0.2 for the single memory model. For the dual memory model, we set dropout for document RNN in source memory to 0.2 and for the encoder and decoder to 0.5. Mini-batching is used in both stages

[7] In our initial experiments, we found SGD to be more effective than Adam/Adagrad; an observation also made by Bahar et al. (2017).

[8] For the document NMT model training, we did some preliminary experiments using different learning rates and used the scheme which converged to the best perplexity in the least number of epochs while for sentence-level training we follow Cohn et al. (2016).

to speed up training. For the largest dataset, the document NMT model takes about 4.5 hours per epoch to train on a single P100 GPU, while the sentence-level model takes about 3 hours per epoch for the same settings.

When training the document NMT model in the second stage, we need the target memory. One option would be to use the ground-truth translations for building the memory. However, this may result in inferior training, since at test-time, the decoder iteratively updates the translation of sentences based on the noisy translations of other sentences (accessed via the target memory). Hence, while training the document NMT model, we construct the target memory from the translations *generated* by the pre-trained sentence-level model.[9] This effectively exposes the model to its potential test-time mistakes during the training, resulting in more robust learned parameters.

3.5.2 Main Results

We have three variants of our model, using: (i) only the source memory (S-NMT+SRC MEM), (ii) only the target memory (S-NMT+TGT MEM), or (iii) both the source and target memories (S-NMT+BOTH MEMS). We compare these variants against the RNN-based sentence-level NMT model (S-NMT). We also compare the source-memory variants of our model to the local-context NMT models of Jean et al. (2017) and Wang et al. (2017),[10] which use a few previous source sentences to generate a context representation, which is augmented to the decoder hidden state (similar to our Memory-to-Context model).

Memory-to-Context We consistently observe +1.15/+1.13 BLEU/METEOR score improvements across the three language-pairs upon comparing our best model to S-NMT (see Table 3.2), with the maximum improvement for Estonian→English (+1.9 BLEU and +1.69 METEOR). Overall, our document NMT model with both source and target memories has been the most effective variant for all of the three language-pairs.

We further experiment to train the target memory variants using *gold* translations instead of the generated ones for German→English. This led to -0.16 and -0.25 decrease[11]

[9] We report results for two-pass decoding, i.e., we only update the translations once using the initial translations generated from the base model. We tried multiple passes of decoding at test-time but it was not helpful.

[10] We implemented and trained the baseline local-context models using the same hyperparameters and training procedure that we used for training our memory models.

[11] Latter is statistically significant decrease with respect to the S-NMT+BOTH MEMS model trained on generated target translations.

	Memory-to-Context							
	BLEU				METEOR			
	Fr→En	De→En		Et→En	Fr→En	De→En		Et→En
		NC-11	NC-16			NC-11	NC-16	
S-NMT	20.85	5.24	9.18	20.42	23.27	10.90	14.35	24.65
+Src Mem	21.91†	6.26†	10.20†	22.10†	24.04†	11.52†	15.45†	25.92†
+Tgt Mem	21.74†	6.24†	9.97†	21.94†	23.98†	11.58†	15.32†	25.89†
+Both Mems	**22.00**†	**6.57**†	**10.54**†	**22.32**†	**24.40**†	**12.24**†	**16.18**†	**26.34**†
	Memory-to-Output							
	BLEU				METEOR			
	Fr→En	De→En		Et→En	Fr→En	De→En		Et→En
		NC-11	NC-16			NC-11	NC-16	
S-NMT	20.85	5.24	9.18	20.42	23.27	10.90	14.35	24.65
+Src Mem	**21.80**†	6.10†	9.98†	21.50†	23.99†	11.53†	15.29†	25.44†
+Tgt Mem	21.76†	**6.31**†	10.04†	21.82†	24.06†	**12.10**†	15.75†	25.93†
+Both Mems	21.77†	6.20†	**10.23**†	**22.20**†	**24.27**†	11.84†	**15.82**†	**26.10**†

Table 3.2: BLEU and METEOR scores for the RNN-based sentence-level NMT baseline (S-NMT) vs. variants of our document NMT model. **bold**: Best performance, †: Statistically significantly better than the baseline.

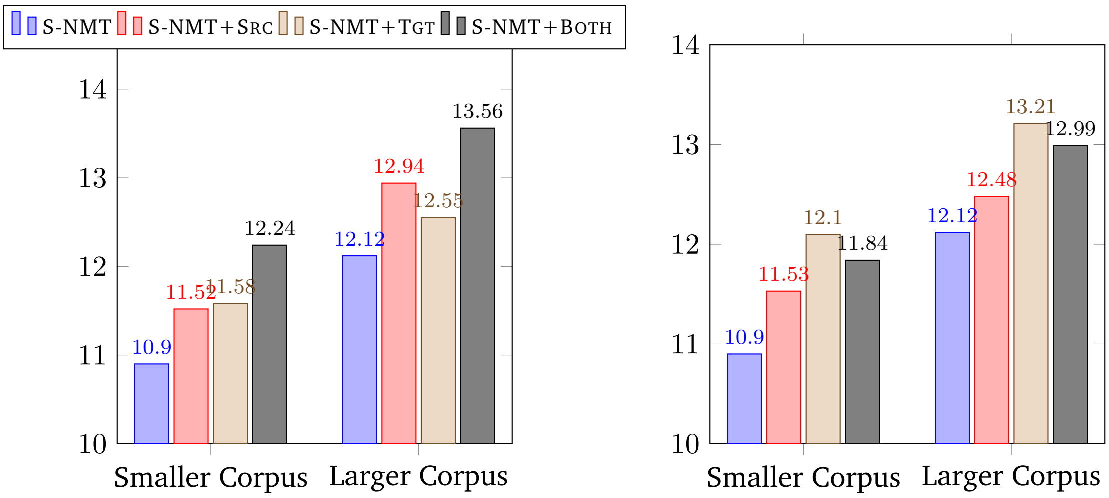
in the BLEU scores for the S-NMT+Tgt Mem and S-NMT+Both Mems variants, which confirms the intuition of constructing the target memory by exposing the model to its noises during training time.

Memory-to-Output From Table 3.2, we consistently see +.95/+1.00 BLEU/METEOR improvements between the best variants of our model and the sentence-level baseline across the three language-pairs. For French→English, all variants of document NMT model show comparable performance when using BLEU; however, when evaluated using METEOR, the dual memory model is the best. For German→English, the target memory variants give comparable results, whereas, for Estonian→English, the dual memory variant proves to be the best. Overall, the Memory-to-Context model variants perform better than their Memory-to-Output counterparts. We attribute this to a large number of parameters in the latter architecture (Table 3.3) and limited amount of data.

Large datasets with document boundaries are hard to obtain, however, one can use freely available sentence-level corpora for pre-training the document-level model. Hence, we further experiment with more data when training the sentence-based NMT model to investigate the extent to which document context is useful in this setting and to show that

	Memory-to-Context			**Memory-to-Output**		
Lang. Pair	Fr→En	De→En	Et→En	Fr→En	De→En	Et→En
S-NMT	42.5	66.8	58.4	42.5	66.8	58.5
+Src Mem	48.8	73.1	64.8	68.7	107.1	88.7
+Tgt Mem	43.8	68.1	59.8	53.8	85.1	71.8
+Both Mems	50.1	74.4	66.1	80	125.4	102

Table 3.3: Number of model parameters (in millions).



(a) Memory-to-Context

(b) Memory-to-Output

Figure 3.4: METEOR scores on German→English (NC-11) while training S-NMT with smaller vs. larger corpus.

the improvements obtained with our model are not due to data size bias. We randomly choose an additional 300K German-English sentence-pairs from WMT'14 data[12] to train the base NMT model in stage 1. In stage 2, we use the same document corpus as before to train the document-level models. As seen from Figure 3.4, the document MT variants still benefit from the document context even when the base model is trained on a larger bilingual corpus. For the Memory-to-Context model, we see massive improvements of +0.72 and +1.44 METEOR scores for the source memory and dual memory model respectively, when compared to the baseline. On the other hand, for the Memory-to-Output model, the target memory model's METEOR score increases significantly by +1.09 compared to the baseline, slightly differing from the corresponding model using the smaller corpus (+1.2).

[12] https://nlp.stanford.edu/projects/nmt/

	BLEU				METEOR			
	Fr→En	De→En		Et→En	Fr→En	De→En		Et→En
		NC-11	NC-16			NC-11	NC-16	
Jean et al. (2017)	21.95	6.04	10.26	21.67	24.10	11.61	15.56	25.77
Wang et al. (2017)	21.87	5.49	10.14	22.06	24.13	11.05	15.20	26.00
S-NMT	20.85	5.24	9.18	20.42	23.27	10.90	14.35	24.65
+SRC MEM	21.91†	6.26$^{\clubsuit}$	10.20	22.10$^{\spadesuit}$	24.04†	11.52$^{\clubsuit}$	15.45$^{\clubsuit}$	25.92$^{\spadesuit}$
+BOTH MEMS	**22.00**†	**6.57**$^{\diamondsuit}$	**10.54**$^{\clubsuit}$	**22.32**$^{\diamondsuit}$	**24.40**$^{\diamondsuit}$	**12.24**$^{\diamondsuit}$	**16.18**$^{\diamondsuit}$	**26.34**$^{\diamondsuit}$

Table 3.4: Our Memory-to-Context source-memory NMT variants vs. S-NMT and source-context NMT baselines. **bold**: Best performance, †, ♠, ♣, ♢: Statistically significantly better than only S-NMT, S-NMT & Jean et al. (2017), S-NMT & Wang et al. (2017), all baselines, respectively.

Local Source Context Models Table 3.4 shows a comparison of our Memory-to-Context model variants to local source-context NMT models (Jean et al., 2017; Wang et al., 2017). For French→English, our source memory model is comparable to both baselines. For German→English, our S-NMT+SRC MEM model is comparable to Jean et al. (2017) but outperforms Wang et al. (2017) for one test set with respect to BLEU, and for both test sets with respect to METEOR. For Estonian→English, our model outperforms Jean et al. (2017). Our global source-context model has only surface-level sentence information and is oblivious to the individual words in the context since we do offline training to obtain the sentence representations (as previously mentioned in Section 3.4). However, the other two context baselines have access to that information, yet our model's performance is either better or quite close to those models. We also look into the unigram BLEU scores to see how much our global source-memory variants lead to improvement at the word-level. From Table 3.5, it can be seen that our model's performance is better than the baselines for the majority of cases. The S-NMT+BOTH MEMS model gives the best results for all three language pairs, showing that leveraging both source and target document context is indeed beneficial for improving MT performance.

3.5.3 Analysis

Using Global/Local Target Context We first investigate whether using a local target context would have been equally sufficient in comparison to our global target memory model for the three datasets. We condition the decoder on the previous target sentence representation (obtained from the last hidden state of the decoder) by adding it as an additional input

	BLEU-1			
	Fr→En	De→En		Et→En
		NC-11	NC-16	
Jean et al. (2017)	52.8	30.6	39.2	51.9
Wang et al. (2017)	52.6	28.2	38.3	52.3
S-NMT	51.4	28.7	36.9	50.4
+Src Mem	53.0	30.5	39.1	52.6
+Both Mems	**53.5**	**33.1**	**41.3**	**53.2**

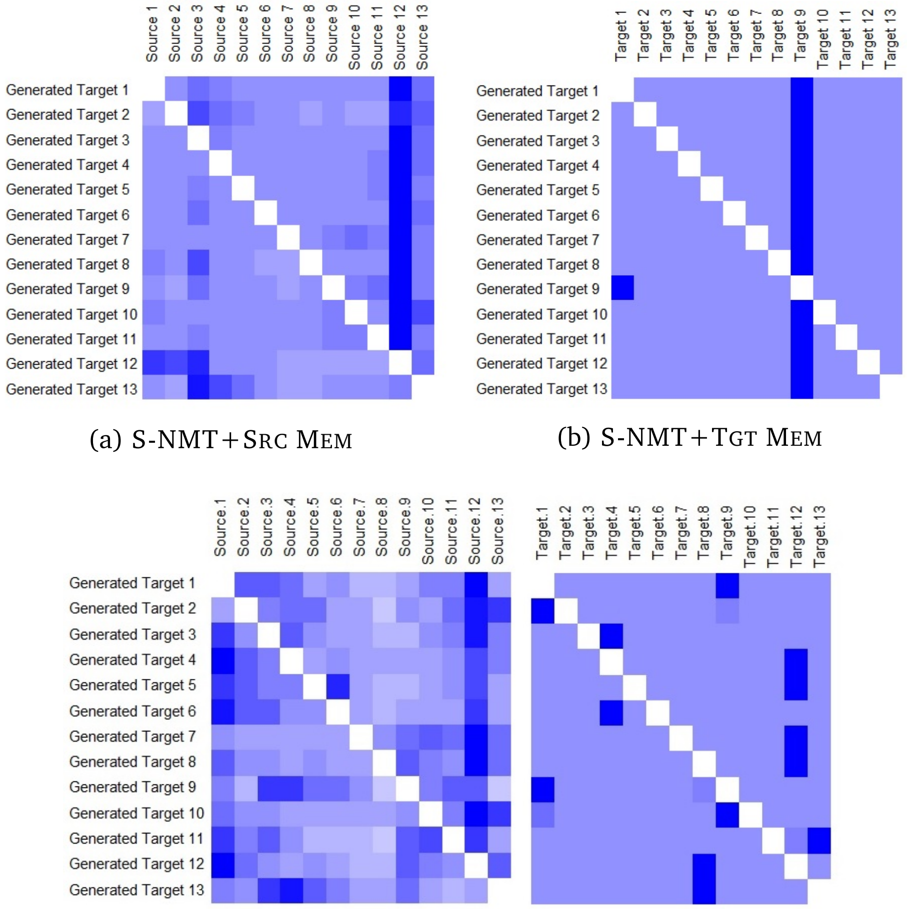
Table 3.5: Unigram BLEU for our Memory-to-Context document NMT models vs. S-NMT and source-context NMT baselines. **bold**: Best performance.

to all decoder states (Prev Tgt) similar to our Memory-to-Context model. From Table 3.6, we observe that for French→English and Estonian→English, using all sentences in the target context or just the previous target sentence gives comparable results. We may attribute this to these specific datasets, that is documents from TED talks or European Parliament Proceedings may depend more on the local than on the global context when using coarse context information. However, for German→English (NC-11), the target memory model performs the best showing that for documents with richer context (e.g., news articles) we do need the global target document context to improve MT performance.

Qualitative Analysis Figure 3.5 illustrates the attention matrices for an example test document in Estonian→English (provided in Appendix A), inferred by the variants of Memory-to-Context model. The attention matrices for the single memory models (Figures 3.5a, 3.5b) are quite different, majorly because both focus on different sentences in the source and target documents. Upon further inspection, it was found that these sentences were key in delivering the gist of the text. For the dual memory model (Figure 3.5c), we see that the attention matrix for the source document is still mostly focused around the same sentences as the source-only one, except that now it also attends to the first sentence which introduces

	BLEU			METEOR		
Lang. Pair	Fr→En	De→En	Et→En	Fr→En	De→En	Et→En
S-NMT	20.85	5.24	20.42	23.27	10.90	24.65
+Prev Tgt	**21.75**	5.93	**22.08**	**24.03**	11.40	**25.94**
+Tgt Mem	21.74	**6.24**	21.94	23.98	**11.58**	25.89

Table 3.6: Analysis of target context model.



(a) S-NMT+SRC MEM

(b) S-NMT+TGT MEM

(c) S-NMT+BOTH MEMS

Figure 3.5: Inferred attention weights by Memory-to-Context models for an Et→En test document. The horizontal axis gives the position of the sentence being generated and the vertical axis gives the position of the sentence in the source or target documents. Darker shades denote higher values.

the topic; while the attention matrix for the target memory has a more granular and guided attention spread out over the sentences.

To better understand the dual memory model, we look at the first sentence example in Table 3.7. It can be seen that the source sentence has the noun '*Qimonda*' but the sentence-level NMT model fails to attend to it when generating the translation. On the other hand, the single memory models are better in delivering some, if not all, of the underlying information in the source sentence but the dual memory model's translation quality surpasses them. This is because the word '*Qimonda*' was being repeated in this specific document, providing a strong contextual signal to our global document-context model while the local-context model by Wang et al. (2017) is still unable to correctly translate the noun even

Source	qimonda täidab lissaboni strateegia eesmärke.
Target	qimonda meets the objectives of the lisbon strategy.
S-NMT	<unk> is the objectives of the lisbon strategy.
+Src Mem	the millennium development goals are fulfilling the millennium goals of the lisbon strategy.
+Tgt Mem	in writing. - (ro) the lisbon strategy is fulfilling the objectives of the lisbon strategy.
+Both Mems	qimonda fulfils the aims of the lisbon strategy.
Wang et al. (2017)	<unk> fulfils the objectives of the lisbon strategy.
Source	... et riigis kehtib endiselt lukašenka diktatuur, mis rikub inim- ning etnilise vähemuse õigusi.
Target	... this country is still under the dictatorship of lukashenko, breaching human rights and the rights of ethnic minorities.
S-NMT	... the country still remains in a position of lukashenko to violate human rights and ethnic minorities.
+Src Mem	... the country still applies to the brutal dictatorship of human and ethnic minority rights.
+Tgt Mem	... the country still keeps the <unk> dictatorship that violates human rights and ethnic rights.
+Both Mems	... the country still persists in lukashenko's dictatorship that violate human rights and ethnic minority rights.
Wang et al. (2017)	... there is still a regime in the country that is violating the rights of human and ethnic minority in the country.

Table 3.7: Example Estonian→English sentence translations (Memory-to-Context) from two test documents.

when it has access to the word-level information of previous sentences.

We resort to manual evaluation as there is no standard metric that evaluates document-level discourse information like consistency or pronominal anaphora. By manual inspection, we observe that our models can identify nouns in the source sentence to resolve coreferent pronouns, as shown in the second example of Table 3.7. Here the topic of the sentence is '*the country under the dictatorship of Lukashenko*' and our target and dual memory models are able to generate the appropriate pronoun/determiner as well as accurately translate the word '*diktatuur*', hence producing much better translation as compared to both baselines. Apart from these improvements, our models are better in improving the readability of sentences by generating more context appropriate grammatical structures such as verbs and adverbs.

Furthermore, to validate that our model improves the consistency of translations, we look at five documents (roughly 70 sentences) from the test set of Estonian→English, each of which had a word being repeated in the gold translation. Our model is able to resolve the consistency in 22 out of 32 cases as compared to the sentence-based model which only accurately translates 16 of those. Following Wang et al. (2017), we also investigate the extent to which our model can correct errors made by the baseline system. We randomly choose five documents from the test set. Out of the 20 words/phrases which were incorrectly translated by the sentence-based model, our model corrects 85% of them while also generating 10% new errors.

3.6 Related Work

At the time of this research, most of the works in document-level MT were based on the conventional SMT approaches relying on hand-engineered features. These have been extensively covered in Section 2.2.1. Here, we will only briefly mention the works which have been used as baselines in this chapter.

Jean et al. (2017) extend the vanilla attention-based neural MT model (Bahdanau et al., 2015) by conditioning the decoder on the previous sentence via an additional attention over its words. Extending their model to consider the global source document-context is challenging due to large size of the computation graph as a result of having different attentional components for the individual sentences in the source document. Wang et al. (2017) employ a two-level hierarchical RNN to summarise three previous source sentences, and then feed the summary vector as an additional input to the decoder hidden state. Both these works consider a very local source context and completely ignore the global source and target document contexts.

3.7 Summary

In this chapter, we have presented a document-level neural MT model that captures global source and target document context via coarse attention over the sentences in the source and target documents. Our model augments the vanilla RNN-based sentence-level NMT model with external memories to incorporate documental interdependencies on both source and target sides. We train the model end-to-end and propose an iterative decoding algorithm based on block coordinate descent. We show statistically significant improvements in the translation quality over the context-agnostic baseline for three language-pairs. We also compare our model to recent local source-context baselines, where our model outperforms them in terms of automatic evaluation metrics BLEU and METEOR.

Chapter 4

Document Context Modelling with Hierarchical Selective Attention

In the previous chapter, we demonstrated the significance of using document-wide context for neural machine translation via a simple approach (coarse attention). Despite not having explicit extra-sentential word-level information, our model was able to outperform other context-aware baselines on three language-pairs. Having proposed and tested efficient training and decoding strategies for document-wide neural machine translation, now we can focus on developing better modelling strategies for the document context. One effective and scalable approach that uses explicit word-level information is hierarchical attention, which has gained popularity in NLP tasks of text summarisation (Nallapati et al., 2016), document/sentiment classification (Yang et al., 2016; Li et al., 2018) and reading comprehension (Zhu et al., 2018), to name a few. Miculicich et al. (2018) successfully apply a bottom-up hierarchical attention mechanism to model word-level and sentence-level abstractions for local-context NMT.

Keeping scalability and efficiency in mind, this chapter[1] presents a novel and scalable top-down approach to hierarchical selective attention for document-wide NMT which uses sparse attention to selectively focus on relevant sentences in the document context and then attends to key words in those sentences. For completeness, single-level attention approaches based on sentence and word-level information in the context are also proposed. Our experiments on English→German monologue datasets in both offline (past and future) and online (only past) settings show that our selective attention approach not only significantly outperforms context-agnostic baselines but also surpasses context-aware baselines

[1] First presented in Maruf et al. (2019).

in most cases. We also conduct an extensive analysis to evaluate our models in terms of its ability to translate pronouns, adequacy and fluency of generated translations, model complexity, and interpretability of hierarchical selective attention.

4.1 Introduction

Neural machine translation has grown immensely in the past few years, from the simplistic RNN-based encoder-decoder models (Sutskever et al., 2014; Bahdanau et al., 2015) to the state-of-the-art Transformer architecture (Vaswani et al., 2017). Most of these models rely on the attention mechanism as a major component, which involves focusing on different parts of a sequence to compute new representations, and has proven to be quite effective in improving the translation quality (Vaswani et al., 2017). However, all of these models share the same inherent problem: the translation is still performed on a sentence-by-sentence basis, thus ignoring the long-range dependencies which may be useful when it comes to translating discourse phenomena.

More recently, context-aware NMT has been gaining significant traction from the MT community with the majority of works coming out in the past two years. Most of these focus on using a few previous sentences as context (Jean et al., 2017; Wang et al., 2017; Tu et al., 2018; Voita et al., 2018; Zhang et al., 2018; Miculicich et al., 2018) and neglect the rest of the document. Our work (Maruf and Haffari, 2018) in the previous chapter is the only one to have considered the full document context, thus proposing a more generalised approach to document-level NMT. However, the model is restrictive as the document-level attention computed is sentence-based and static (computed only once for the sentence being translated). A more recent work (Miculicich et al., 2018) proposes to use a hierarchical attention network (HAN) (Yang et al., 2016) to model the contextual information in a structured manner using word-level and sentence-level abstractions; yet, it uses a limited number of past source and target sentences as context and is not scalable to the entire document.

In this work, we propose a *selective attention* approach to first selectively focus on relevant sentences in the global document-context and then attend to key words in those sentences, while ignoring the rest.[2] Towards this goal, we use *sparse attention*, enabling an efficient and scalable use of the context. The intuition behind this is the way humans translate a sentence containing ambiguous words. They may look for sentences in the whole document which contain similar words and just focus on those for the translation. This attention, which we refer to as hierarchical attention, is computed dynamically for each query word. Furthermore, we propose a flat attention approach that is based on either sentence or word-level information in the context. We integrate the document-level context representation, produced from these attention modules, into the encoder or decoder of the Transformer model depending on whether we consider monolingual (source-side) or bilingual (both source and target-side) context.

Our contributions are as follows: (i) we propose a novel and efficient top-down approach to hierarchical attention for context-aware NMT, (ii) we compare variants of selective attention with both context-agnostic and context-aware baselines, and (iii) we run experiments in both online (only past context) and offline (both past and future context) settings on three English→German datasets. Experiments show that our approach improves upon the Transformer by an overall +1.3, +2.1 and +1.2 BLEU for TED talks, News-Commentary and Europarl, respectively. It also outperforms two recent context-aware baselines (Zhang et al., 2018; Miculicich et al., 2018) in the majority of cases.

4.2 Preliminaries

Sparsemax Transformation The softmax function, we previously described, is strictly positive; meaning that it forces all of its inputs to receive *some* probability mass, thus discouraging sparsity. It is also convenient to use as it is simple to compute and differentiate. Sparsemax, on the other hand, has the distinctive property that it can return sparse posterior distributions, that is, it may assign exactly zero probability to some of its output variables (Martins and Astudillo, 2016), and also preserves the appealing properties of the softmax. Mathematically, the sparsemax transformation is defined as:

$$\text{sparsemax}(\boldsymbol{z}) = \underset{\boldsymbol{\alpha} \in \Delta^{d_z - 1}}{\text{arg min}} \, ||\boldsymbol{\alpha} - \boldsymbol{z}||^2 \tag{4.1}$$

[2] The term "selective attention" comes from cognitive science and is defined as the act of focusing on a particular object for a period of time while simultaneously ignoring irrelevant information that is also occurring (Dayan et al., 2000).

where $\Delta^{d_z-1} := \{\boldsymbol{\alpha} \in \mathbb{R}^{d_z} | \boldsymbol{\alpha} \geq \mathbf{0}, \sum_{d_z} \alpha_{d_z} = 1\}$ is the $d_z - 1$ dimensional probability simplex. In simpler terms, sparsemax is the Euclidean projection of the scores $\boldsymbol{z}$ onto the probability simplex. Since these projections are likely to hit the boundary of the simplex, this yields a sparse probability distribution. We rely on this transformation to identify the relevant sentences and words in a document as required by our hierarchical attention model.

4.3 Proposed Approach

The main goal of this work is to have a document-level NMT model which is memory-efficient, scalable, and capable of listening to the entire document. To achieve this, we augment a sentence-level NMT model (the Transformer (Vaswani et al., 2017)) with an efficient hierarchical attention mechanism which has the ability to identify the key sentences in the document context and then attend to the key words within those sentences. As mentioned previously (Section 3.3), we want to maximise the probability of a document translation given the source document, that is $P_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{j=1}^{|\boldsymbol{d}|} P_{\boldsymbol{\theta}}(\boldsymbol{y}^j|\boldsymbol{x}^j, \boldsymbol{D}^{-j})$, where $\boldsymbol{y}^j$ and $\boldsymbol{x}^j$ denote the j^{th} target and source sentence, respectively, and $\boldsymbol{D}^{-j} = \{\boldsymbol{X}^{-j}, \boldsymbol{Y}^{-j}\}$ is the collection of all other sentences in the source and target documents. In this chapter, we take $\boldsymbol{D}^{-j}$ to be either the monolingual source or bilingual source and target-side context in two settings: *offline*—the context comes from both past and future, and *online*—the context comes from only the past.

In this section, we show how to represent the document-level context using our Context Layer, how to regulate the information at the sentence and document-level using context gating and finally we present our integrated model.

4.3.1 Document-level Context Layer

The context $\boldsymbol{D}^{-j}$ is modeled via a single Document-level Context Layer (Figure 4.1) comprising of two sub-layers: (i) a multi-head context attention sub-layer, and (ii) a feed-forward sub-layer, where the former consists of either a top-down hierarchical attention module or a flat attention module (both explained shortly), and the latter is similar to the feed-forward network (FFN) in the original Transformer architecture. Each sub-layer is

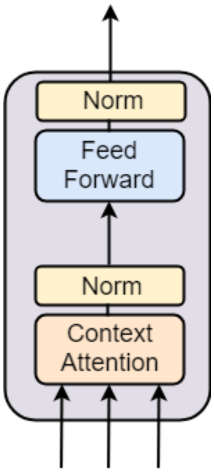


Figure 4.1: Document-level Context Layer.

followed by layer normalisation.[3]

Let us now describe the attention modules which independently form the multi-head context attention sub-layer.

4.3.1.1 Hierarchical Attention

Our hierachical attention module H-ATTENTION($\boldsymbol{Q_s}$, $\boldsymbol{Q_w}$, $\boldsymbol{K_s}$, $\boldsymbol{K_w}$, $\boldsymbol{V_w}$) (Figure 4.2) is a reformulation of the scaled dot-product attention by Vaswani et al. (2017) described in Section 2.1.2. For our module, we have five inputs consisting of two types of keys and queries, one each for the sentences and the words, while the values are based only on words in the context. The hierarchical attention module has four operations:

1. **Sentence-level Key Matching:** This is performed on a set of queries simultaneously, packed together into a matrix $\boldsymbol{Q_s}$. The sentence-level keys are also packed into a matrix $\boldsymbol{K_s}$. We will describe in Section 4.3.3 how $\boldsymbol{Q_s}$ and $\boldsymbol{K_s}$ are computed. The sentence-level attention weights are computed as:

$$\boldsymbol{\alpha_s} = \text{sparsemax}\Big(\frac{\boldsymbol{K_s}^\top \boldsymbol{Q_s}}{\sqrt{d_k}}\Big) \tag{4.2}$$

where d_k is the dimension of the keys and $\boldsymbol{\alpha_s}$ has dimensions equal to the total number of sentences in the document. We propose to use *sparsemax* (Martins and Astudillo, 2016), instead of softmax, as this gives us the intended selective attention

[3] We do not have residual connections after sub-layers in our Document-level Context Layer as we found them to have a deteriorating effect on the translation scores (also reported by Zhang et al. (2018)).

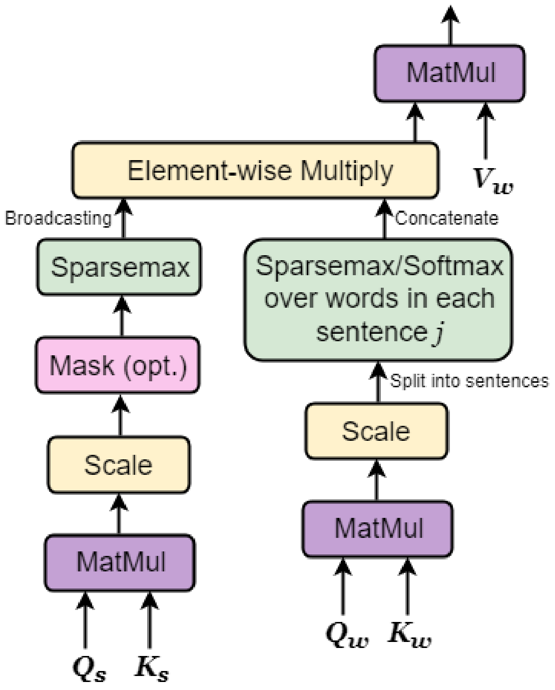


Figure 4.2: Hierarchical context attention module.

behavior, that is identifying the key sentences that may potentially be relevant to the current sentence, hence making the model more efficient in compressing its memory. A softmax attention, on the other hand, can still assign low probability to sentences, forming a long-tail and absorbing significant probability mass, and it cannot fully *ignore* those sentences. An additive mask is used (before the *sparsemax* operation) based on whether we train for offline or online setting by masking out only the current sentence or current and future sentences, respectively.

2. **Word-level Key Matching:** Here the query and key matrices, $\boldsymbol{Q_w}$ and $\boldsymbol{K_w}$, are word-level. We perform a word-level key matching for each sentence j in the document:

$$\boldsymbol{\alpha}_{\boldsymbol{w}}^{j} = \text{sparsemax}\Big(\frac{\boldsymbol{K}_{\boldsymbol{w}}^{j\top}\boldsymbol{Q_w}}{\sqrt{d_k}}\Big) \qquad (4.3)$$

where $\boldsymbol{\alpha}_{\boldsymbol{w}}^{j}$ is the word-level attention vector for j^{th} sentence.[4] We can also use softmax, instead of sparsemax, for a coarser key matching. We explore the two variants in our experiments.

[4] This can be done for only the sentences with non-zero probabilities (obtained from the sentence-level key matching), however, we found it to be computationally expensive, as it required breaking down the batched matrices.

3. **Re-scaling attention weights:** The word-level attention is further re-weighted by the corresponding sentence-level attention (Nallapati et al., 2016) such that the probability of j^{th} sentence in a document is given by:

$$\boldsymbol{\alpha}^j_{\boldsymbol{hier}} = \boldsymbol{\alpha_s}(j)\boldsymbol{\alpha}^j_{\boldsymbol{w}} \tag{4.4}$$

   where $\boldsymbol{\alpha_s}(j)$ is the attention weight for the j^{th} sentence obtained via Eq. 4.2 and $\boldsymbol{\alpha}^j_{\boldsymbol{w}}$ is as in Eq. 4.3. The re-weighting, thus, produces a scaled attention vector $\boldsymbol{\alpha_{hier}} = \text{Concat}(\boldsymbol{\alpha}^1_{\boldsymbol{hier}}; \ldots; \boldsymbol{\alpha}^{|\boldsymbol{d}|}_{\boldsymbol{hier}})$, each entry of which corresponds to the attention weight of a specific word in the document.

4. **Value Reading:** The set of word-level values is packed together into a matrix $\boldsymbol{V_w}$ and the matrix of outputs is given by $\boldsymbol{V_w}\boldsymbol{\alpha_{hier}}$. This multiplication, combined with sparsemax attention, allows to *prune* the hierarchy.

We further extend the MULTIHEAD attention function proposed by Vaswani et al. (2017) for our hierarchical attention module as:

$$\text{H-MULTIHEAD}(\boldsymbol{Q_s}, \boldsymbol{K_s}, \boldsymbol{Q_w}, \boldsymbol{K_w}, \boldsymbol{V_w}) = \boldsymbol{W^O}\text{Concat}(\boldsymbol{head}_1; ...; \boldsymbol{head}_{\mathcal{H}})$$

where $\boldsymbol{head}_h = \text{H-ATTENTION}(\boldsymbol{W}_h^{\boldsymbol{Q_s}}\boldsymbol{Q_s}, \boldsymbol{W}_h^{\boldsymbol{Q_w}}\boldsymbol{Q_w}, \boldsymbol{W}_h^{\boldsymbol{K_s}}\boldsymbol{K_s}, \boldsymbol{W}_h^{\boldsymbol{K_w}}\boldsymbol{K_w}, \boldsymbol{W}_h^{\boldsymbol{V_w}}\boldsymbol{V_w})$, $\boldsymbol{W}$'s are parameter matrices and all (five) inputs[5] are transformed using separate linear layers.

4.3.1.2 Flat Attention

Another way to model the context $\boldsymbol{D}^{-j}$ is via single-level attention by directly re-using the scaled dot-product attention in Vaswani et al. (2017):

$$\text{ATTENTION}(\boldsymbol{Q}_*, \boldsymbol{K}_*, \boldsymbol{V}_*) = \boldsymbol{V}_*\text{softmax}\Big(\frac{\boldsymbol{K}_*^\top \boldsymbol{Q}_*}{\sqrt{d_k}}\Big) \tag{4.5}$$

where the subscript $*$ corresponds to $\boldsymbol{s}$ or $\boldsymbol{w}$ depending on whether the attention is on sentence or word-level. The attention[6] is *sentence-level* if $\boldsymbol{K_s}$, $\boldsymbol{V_s}$ are computed for sentences in the document, and *word-level* if $\boldsymbol{K_w}$, $\boldsymbol{V_w}$ are computed for words in the document.[7] The former module is similar to our Memory Networks architecture in Chapter 3 (Maruf

[5] To re-emphasise, the only difference between $\boldsymbol{Q_s}$ and $\boldsymbol{Q_w}$ is that the queries (from words in current source or target sentence) are transformed using separate linear layers.

[6] Investigation into sparse flat attention is left for future work.

[7] Here $\boldsymbol{Q_s}$, $\boldsymbol{K_s}$, $\boldsymbol{Q_w}$, $\boldsymbol{K_w}$ and $\boldsymbol{V_w}$ are equivalent to the ones computed for hierarchical attention.

and Haffari, 2018) in that it uses sentence-level information. However, there are two key differences, here: (i) we use MultiHead attention as in the Transformer architecture, and (ii) our context attention is dynamic such that we have a separate attention for each query word.

4.3.2 Context Gating

As mentioned previously, the multi-head context attention sub-layer (employing either the hierarchical or flat attention module) is part of the Document-level Context Layer (Figure 4.1), the output of which is fed into the Transformer architecture through context gating (Tu et al., 2018). For i^{th} word in the j^{th} source or target sentence:

$$\gamma_i^j = \sigma(\boldsymbol{W_r} \boldsymbol{r}_i^j + \boldsymbol{W_d} \boldsymbol{c}_i^{j,\boldsymbol{d}}) \tag{4.6}$$

$$\tilde{\boldsymbol{r}}_i^j = \gamma_i^j \odot \boldsymbol{r}^j{}_i + (\boldsymbol{1} - \gamma_i^j) \odot \boldsymbol{c}_i^{j,\boldsymbol{d}} \tag{4.7}$$

where $\boldsymbol{W}$'s are parameter matrices, $\boldsymbol{r}_i^j$ is the output of encoder or decoder stack for i^{th} word in j^{th} sentence, $\boldsymbol{c}_i^{\boldsymbol{d}}$ is the output from the Document-level Context Layer for i^{th} word in j^{th} sentence and $\tilde{\boldsymbol{r}}_i^j$ is the final hidden representation for the same.

4.3.3 Integrated Model

The context can be integrated into the encoder or decoder of the NMT model depending on if it is monolingual or bilingual.[8]

Monolingual context integration in Encoder We add the Document-level Context Layer alongside the encoder stack as shown in Figure 4.3. The Encoder Context Encoding block stores the sentence and word-level keys and values produced from the pre-trained sentence-level NMT model. The word-level keys $\boldsymbol{K_w}$ and values $\boldsymbol{V_w}$ are composed of vector representations of source words (from last encoder layer) in the document, while the sentence-level keys $\boldsymbol{K_s}$ and values $\boldsymbol{V_s}$ are composed of vector representations of sentences in the document where the vector representation of each sentence is an average of the word representations in that sentence.[9] The queries $\boldsymbol{Q_w}$, $\boldsymbol{Q_s}$ are linear transformations of the output of the

[8] We do not integrate context into both encoder and decoder as it would have redundant information from the source (the context incorporated in the decoder is bilingual), in addition to increasing the complexity of the model.

[9] $\boldsymbol{V_s}$ is only used in the sentence-level flat attention.

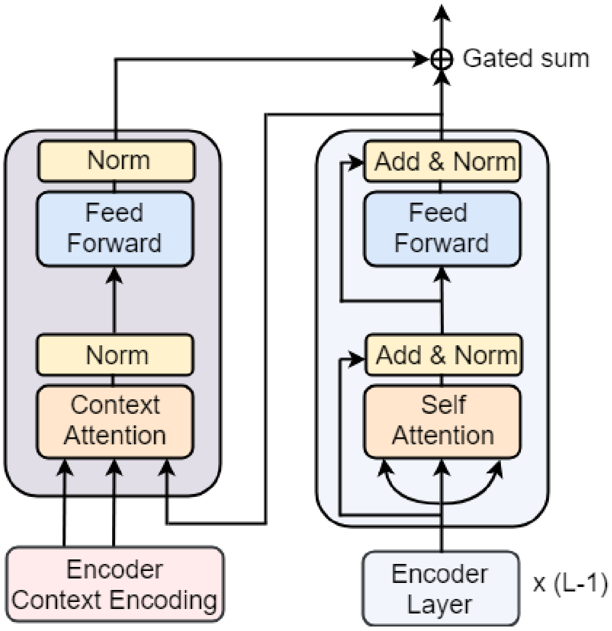


Figure 4.3: Encoder-side context integration.

L^{th} encoder layer (for each query word in the current sentence) which are then matched with the corresponding keys and values stored in the Encoder Context Encoding block just described.

Bilingual context integration in Decoder We again add the Document-level Context Layer alongside the decoder stack as in Figure 4.4. However, instead of choosing the keys and values to be computed from the encoder, we follow Tu et al. (2018) in choosing the keys to match to the source-side context, while designing the values to match to the target-side context. To elaborate, the keys (in the Decoder Context Encoding block) are composed of context vectors from the Source Attention sub-layer in the last decoder layer, while the values are composed of the hidden representations of the target words output from the last decoder layer. Again the keys $\boldsymbol{K_w}$ and $\boldsymbol{K_s}$ are either for individual target words or target sentences, and the same goes for $\boldsymbol{V_w}$ and $\boldsymbol{V_s}$. The queries $\boldsymbol{Q_w}$, $\boldsymbol{Q_s}$ for the Context Layer are linear transformations of the output from the Source Attention sub-layer (for each query word) in the L^{th} layer of the decoder (Figure 4.4).

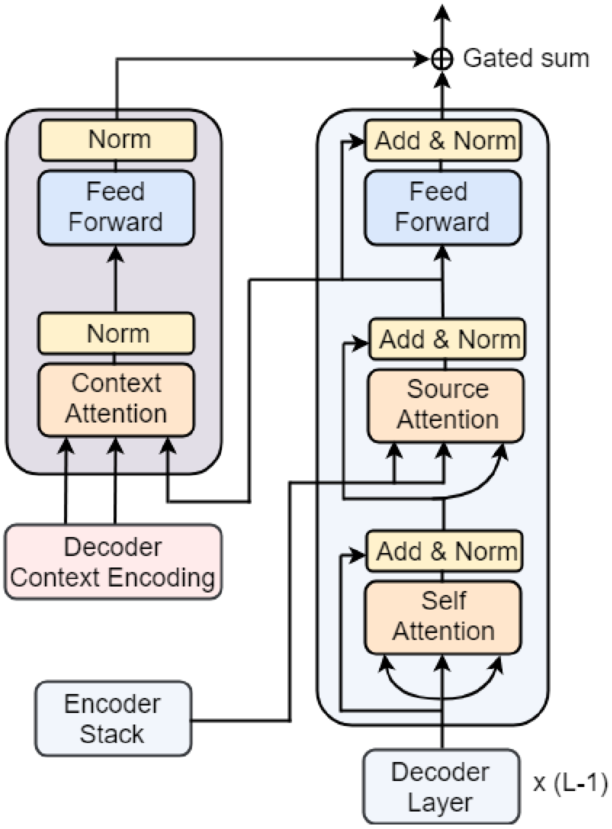


Figure 4.4: Decoder-side context integration.

4.4 Experiments

4.4.1 Setup

Datasets We conduct experiments for English→German translation on three different domains: TED talks, News Commentary and Europarl. These datasets are chosen based on their variance in genre, style and level of formality:

- **TED** This corpus is from the IWSLT 2017 MT track (Cettolo et al., 2012) and contains transcripts of TED talks aligned at sentence-level. Each talk is considered to be a document. We combine `tst2016-2017` into the test set and the rest is used for development.

- **News Commentary** We obtain the sentence-aligned document-delimited News Commentary v11 corpus for training.[10] The WMT'16 `news-test2015` and `news-test2016` are used for development and testing, respectively.

[10] www.casmacat.eu/corpus/news-commentary.html

Domain	#Documents	#Sentences	Document length	Sentence length
TED	1698/93/23	0.21M/9K/2.3K	120.89/96.42/98.74	20.3/19.7/19.6
News	6069/81/155	0.24M/2K/3K	38.93/26.78/19.35	25/21.6/21.5
Europarl	118K/240/360	1.67M/3.6K/5.1K	14.14/14.95/14.06	27.7/28/28

Table 4.1: Training/development/test corpora statistics: number of documents and sentences (K stands for thousands and M for millions), average document length (in sentences) and average sentence length for English (in tokens).

- **Europarl** This dataset is extracted from Europarl v7 (Koehn, 2005). The source and target sentences in each session are aligned using the links provided by Tiedemann (2012). Following our previous work (Maruf and Haffari, 2018), we use the SPEAKER tag as the document delimiter. Documents longer than 5 sentences are kept and the resulting corpus is randomly split into training, dev and test sets.

The corpora statistics are provided in Table 4.1. All datasets[11] are tokenised and truecased using the Moses toolkit (Koehn et al., 2007), and split into subword units using a joint BPE model with 30K merge operations (Sennrich et al., 2016).

Models and Baselines For offline document MT, we have two context-agnostic baselines: (i) a modified version of RNNSearch (attentional RNN-based NMT model) (Bahdanau et al., 2015), which incorporates dropout on the output layer and improves the attention model by feeding the previously generated word, and (ii) the state-of-the-art Transformer architecture. We also compare to our document-wide NMT model with coarse attention (Maruf and Haffari, 2018). For the online case, we again have the Transformer as a context-agnostic baseline and two recent context-aware baselines (Zhang et al., 2018; Miculicich et al., 2018).

All models are implemented in C++ using DyNet (Neubig et al., 2017). For RNNSearch, we modify the sentence-based NMT implementation in mantis (Cohn et al., 2016). The encoder is a single layer bi-directional GRU (Cho et al., 2014a) and the decoder is a two-layer GRU with embeddings and hidden dimensions set to 512. The dropout rate for the output layer is set to 0.2. For the Transformer, we use Transformer-DyNet implementation[12] and extend it for our context-aware NMT model.The hidden dimensions and feed-forward layer

[11] The data is available at https://github.com/sameenmaruf/selective-attn.

[12] https://github.com/duyvuleo/Transformer-DyNet

size is set to 512 and 2048 respectively. We use 4 layers[13] each in the encoder and decoder with 8 attention heads and employ label smoothing with a value of 0.1. We also employ all four types of dropouts as in the original Transformer with a rate of 0.1 for the sentence-based model and 0.2 for our context-aware model. For all models, we use separate source and target embeddings.

For training all models, we use the default Adam optimiser (Kingma and Ba, 2015) with an initial learning rate of 0.0001 and employ early stopping. For our context-aware NMT model, we use a two-stage training strategy, similar to the one described in Section 3.5, that is we pre-train the sentence-level NMT model[14] followed by optimising the parameters of the whole model, i.e., both the document-level and the sentence-level parameters. For inference, we use iterative decoding only when using the bilingual context. All experiments are run on a single Nvidia P100 GPU with 16GBs of memory.[15]

Evaluation Metrics For evaluation, we use BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) scores on tokenised text, and measure statistical significance with respect to the baselines, $p < 0.05$ (Clark et al., 2011).

4.4.2 Main Results

We divide our experiments into two parts: offline and online document MT, and report results for our models (and the context-aware baselines) depending on whether the context is integrated into encoder or decoder.

Offline Document MT From the scores of the two context-agnostic baselines in Table 4.2, we can see that the Transformer beats the RNNSearch model in all cases by at least +2.5 BLEU and +2.1 METEOR scores showing that our hyperparameter choice for the Transformer is indeed effective. Our document-wide NMT model with coarse attention outperforms RNNSearch in majority cases but is unable to beat the Transformer as it still extends the RNN-based encoder-decoder architecture.

[13] We found this configuration to be much more stable than using 6 layers with almost no difference in performance as reported by Xia et al. (2018).

[14] We have used the same sentence-level parameters as warm-start for all context-aware models and baselines.

[15] The experiments can also be run on GPUs with 10-12GBs of memory by reducing the batch size at the expense of increased computational cost.

	Integration into Encoder						Integration into Decoder					
	TED		News		Europarl		TED		News		Europarl	
Model	**BLEU**	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**
RNNSearch	19.2	40.8	16.5	36.8	26.3	44.1	19.2	40.8	16.5	36.8	26.3	44.1
Maruf and Haffari (2018)	-	-	-	-	-	-	20.4	42.1	17.4	37.8	27.1	44.9
Transformer	23.3	44.2	22.8	42.2	28.7	46.2	23.3	44.2	22.8	42.2	28.7	46.2
+Attention, sentence	24.5	**45.2**	**24.8**	43.9	29.6	47.0	**24.4**	44.8	24.7	43.8	29.7	**47.0**
word	**24.6**	44.9	24.6	43.8	29.6	46.9	24.3	45.0	24.2	43.4	29.7	46.9
+H-Attention, sparse-soft	24.2	44.8	24.8	44.1	**29.7**	47.0	24.2	44.9	**24.7**	**43.9**	**29.7**	47.0
sparse-sparse	24.3	45.1	24.7	**44.2**	29.6	**47.0**	24.1	**45.3**	24.5	43.5	29.6	47.0

Table 4.2: BLEU and METEOR scores for variants of our model, two context-agnostic baselines and a context-aware baseline (Maruf and Haffari, 2018) for offline document MT. **bold**: Best performance w.r.t. two decimal places. All reported results for our model are significantly better than all baselines.

	Integration into Encoder						Integration into Decoder					
	TED		News		Europarl		TED		News		Europarl	
Model	**BLEU**	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**
Zhang et al. (2018)	24.0	44.7	23.1	42.4	29.3	46.7	23.8	44.5	22.8	42.2	29.4	46.7
Miculicich et al. (2018)	**24.6**	**45.5**	**25.0**	44.0	28.6	46.1	24.4	45.2	24.4	43.6	29.6	46.9
Transformer	23.3	44.2	22.8	42.2	28.7	46.2	23.3	44.2	22.8	42.2	28.7	46.2
+Attention, sentence	24.4	45.0	24.5$^{\bigstar}$	43.5$^{\bigstar}$	29.6$^{\clubsuit}$	47.0$^{\clubsuit}$	24.3$^{\bigstar}$	45.1$^{\bigstar}$	**24.8**$^{\clubsuit}$	**44.0**$^{\clubsuit}$	29.6	46.8
word	24.2	45.0$^{\bigstar}$	24.8$^{\bigstar}$	**44.3**$^{\bigstar}$	29.7$^{\clubsuit}$	47.0$^{\clubsuit}$	24.0	44.8	24.2$^{\bigstar}$	43.5$^{\bigstar}$	**29.9**$^{\clubsuit}$	47.1$^{\clubsuit}$
+H-Attention, sparse-soft	24.3	45.0$^{\bigstar}$	24.5$^{\bigstar}$	43.7$^{\bigstar}$	**29.8**$^{\clubsuit}$	**47.2**$^{\clubsuit}$	**24.6**$^{\bigstar}$	**45.3**$^{\bigstar}$	24.4$^{\bigstar}$	43.7$^{\bigstar}$	29.8$^{\bigstar}$	**47.1**$^{\clubsuit}$
sparse-sparse	24.4	45.4$^{\bigstar}$	24.7$^{\bigstar}$	44.1$^{\bigstar}$	29.4$^{\diamondsuit}$	46.8$^{\diamondsuit}$	24.4$^{\bigstar}$	45.1$^{\bigstar}$	24.6$^{\bigstar}$	43.8$^{\bigstar}$	29.6$^{\bigstar}$	46.9$^{\bigstar}$

Table 4.3: BLEU and METEOR scores for variants of our model, one context-agnostic and two context-aware baselines for online document MT. **bold**: Best performance w.r.t. two decimal places. $\bigstar$, $\diamondsuit$, $\clubsuit$: Statistically significantly better than our implementations of Zhang et al. (2018), Miculicich et al. (2018), or both. All reported results for our model are significantly better than the Transformer.

For the Encoder Context integration with monolingual context, our hierarchical attention models perform the (near) best for News and Europarl datasets with +2 and +1 BLEU and +2 and +0.8 METEOR improvements with respect to the Transformer. For TED talks, however, we find the flat attention based models (sentence and word-level) to be the best with +1.3 BLEU and +1.1 METEOR improvements.[16] For Decoder Context integration with bilingual context, we find the hierarchical attention to be the best in the majority of cases both in terms of BLEU and METEOR.

Online Document MT From Table 4.3, all our models significantly outperform the context-agnostic baseline and are significantly better than Zhang et al. (2018) in majority cases.[17] For Encoder Context integration, the HAN encoder[18] Miculicich et al. (2018) is the best for TED and News datasets, however, the results are statistically insignificant with respect to our best model. For Europarl, our hierarchical attention model performs significantly better than Miculicich et al. (2018) with a gain of +1.2 BLEU and +1.1 METEOR. For Decoder Context integration, our hierarchical attention models are the winner in majority cases and our best models beat Miculicich et al. (2018)'s HAN decoder[19] for all datasets based on BLEU and METEOR. The main conclusion we draw from these results is that efficiently using the context information at hand is crucial when it comes to improving the performance of context-aware NMT. Furthermore, shorter pieces of text (e.g., the ones in Europarl) benefit more from using global context because their sentences may exhibit higher interdependency than those in a longer piece of text.

Offline vs. Online Document MT Let us compare the overall results for the offline and online document MT settings. For all datasets and model variants, we find the best BLEU and METEOR scores in Tables 4.2 and 4.3 (highlighted in bold) to be quite close to each

[16] Please note that the sentence-level flat attention model is similar to the model presented in Chapter 3 but is different in that it computes a separate attention for each query word (as previously mentioned in Section 4.3.1.2).

[17] The major difference between our models and Zhang et al. (2018) is that they use the context representation from only two previous source sentences to integrate into the encoder or decoder, where the embeddings of these context sentences are the original embeddings (word+positional) while we use their abstract representations (as described in Section 4.3.3) in our models. Furthermore, they fix the sentence-level paramaters when training their context-aware model while we fine-tune the sentence-level parameters as well.

[18] This refers to their model which integrates context from three previous source sentences into the encoder.

[19] This refers to their model which integrates context from three previously decoded target sentences into the decoder. They do not propose a model which integrates bilingual context into the decoder.

other with those for the online setting slightly better. This is quite self-explanatory, because, in essence, the experimental datasets comprise of talks, speeches or commentaries, which are in fact produced in an online manner and hence we do not see drastic improvements in terms of BLEU and METEOR when conditioning on the future context. This, in our opinion, does not mean that we should never look into the future, but just that NMT models, in general, are highly subjective to data, and whether context-aware models benefit from future context is also dependant on that.

To summarise our experiments, if one wants to choose an optimal configuration between the flat and hierarchical attention models, then the latter is a clear winner provided the base model is trained on large datasets. This is also evident from our follow-up work (Maruf and Haffari, 2019), where we demonstrate the hierarchical attention to outperform the sentence-level attention (based on ensembling multiple independent runs) in the majority of cases for both translation directions of the English-German language pair.

4.4.3 Analysis

Evaluation on Contrastive Pronoun Test Set It has been argued that evaluation metrics that quantify the overall translation quality are somewhat ill-equipped to assess how well models translate inter-sentential phenomena such as pronouns. Hence, we use a test suite of contrastive translations designed to measure the accuracy of translating the English pronoun *it* to its German counterparts *es*, *er* and *sie* (Müller et al., 2018). The test set is automatically created from the OpenSubtitles corpus (Lison and Tiedemann, 2016) using the publicly available scripts.[20] It contains 4000 randomly sampled instances of each of the three translations of *it* under consideration. For each sentence pair in the resulting test set, a contrastive translation is introduced where the correct pronoun is replaced with an incorrect one, such that on its own the contrastive translation is grammatically correct if the antecedent is outside the current sentence. The trained model is then used to provide a score (the negative log-probability) for the reference and contrastive translation. The number of times the model scores the reference translation higher than the contrastive one is reported as accuracy. The test set is used to evaluate our models trained on TED talks. We

[20] https://github.com/ZurichNLP/ContraPro

Model	**antecedent distance**				
	0	1	2	3	>3
	Offline document MT				
RNNSearch	0.415	0.310	0.424	0.440	0.647
Maruf and Haffari (2018)	0.424	0.302	0.418	0.443	0.665
Transformer	0.586	0.308	0.437	0.48	0.642
+ATTENTION, sentence	0.677	0.314	0.439	0.478	0.697
word	**0.686**	**0.347**	**0.464**	**0.511**	0.679
+H-ATTENTION, sparse-soft	0.676	0.308	0.440	0.480	0.686
sparse-sparse	0.652	0.303	0.435	0.471	**0.701**
	Online document MT				
Zhang et al. (2018)	0.622	0.321	0.450	0.485	0.658
Miculicich et al. (2018)	0.722	0.326	0.451	0.471	0.661
Transformer	0.586	0.308	0.437	0.48	0.642
+ATTENTION, sentence	**0.732**	**0.340**	**0.460**	0.485	0.661
word	0.690	0.317	0.444	0.487	0.683
+H-ATTENTION, sparse-soft	0.692	0.329	0.446	0.464	0.656
sparse-sparse	0.711	0.317	0.437	**0.489**	**0.692**

Table 4.4: Accuracy on the contrastive test set with regard to antecedent distance (in sentences) on TED talks. Antecedent distance 0 means the pronoun occurs in the same sentence as the antecedent.

are interested to see if our global document-context models surpass the local context-aware baselines.

The first obvious thing we notice from Table 4.4 is that both context-agnostic baselines perform roughly the same when the antecedent is in a different sentence than the reference. Interestingly, for antecedent distance greater than three, we see our previous architecture with coarse attention (Maruf and Haffari, 2018) to outperform both context-agnostic baselines. In general, Table 4.4 shows that not only our global-context models are quite effective but our hierarchical attention model is most useful when the antecedent is farther than three previous sentences. We also conclude that models for offline MT perform better when antecedent distance is greater than two.

Subjective Evaluation We also conduct a subjective evaluation to validate the benefit of exploiting document-level context. Three native German speakers were asked to choose the better (with ties allowed) of two translations for each of 18 documents (randomly sampled from Europarl test set). The two translations, one produced by the Transformer and the other by our hierarchical attention model, were evaluated in terms of: *adequacy* (Which

Model	**#Params**	**Speed (words/sec.)**	
		Training	Decoding
Zhang et al. (2018)	59.5M	3300	84.94
Miculicich et al. (2018)	54.8M	1650	76.90
Transformer	50M	5100	86.33
+ATTENTION, sentence	53.7M	3750	83.84
word	53.7M	3100	81.38
+H-ATTENTION	54.2M	2600	74.11

Table 4.5: Model complexity of Encoder Context integration models for News Commentary dataset. The training speed is for batched training with 900 tokens in each mini-batch and decoding speed is for greedy decoding without batching. It should be mentioned that although the presented speed is for the online setting, the offline setting exhibits comparable performance.

translation expresses the meaning of the source text more adequately?) and *fluency* (Which text has better German?) (Läubli et al., 2018). Let a, b be number of ratings in favour of Transformer or our model, respectively, and t be number of ties, then number of successes $x = b + 0.5t$ and trials $n = a + b + t$. We test for statistically significant preference of our model over the Transformer by means of two-sided Sign Tests and find that our model is better than the Transformer both in terms of document-level adequacy (x = 39, n = 54, p = 0.0015) and fluency (x = 38, n = 54, p = 0.0038).

Model Complexity Model complexity is reported in Table 4.5. Our context-aware models introduce only 8% more parameters to the original Transformer model. In comparison to the Transformer, our hierarchical attention model is slow in training, dropping the speed by almost 50%,[21] but it is still almost 40% faster than Miculicich et al. (2018). At decoding time, our hierarchical attention model is almost equivalent to Miculicich et al. (2018) and only 13% slower than Zhang et al. (2018). Hence, attending to the whole document (instead of a few previous sentences) does not add to the time complexity of the model on average. It should be noted that the speed of our hierarchical attention model is slower than the word-level attention model because it has two levels of abstraction, but unlike the word-level attention model which may fail due to the large size of computation graph, our hierarchical attention model has the capability to scale to long documents.

[21] `DyNet` implementation of *sparsemax* is CPU-based and only operates on column vectors. We believe a GPU-based matrix implementation would bring the speed much closer to our word-level attention model.

Src: my **thoughts** are also with the victims. Tgt: meine **Gedanken** sind auch bei den Opfern.
Transformer: ich **denke** auch an die Opfer. Zhang et al. (2018): ich **denke** auch an die Opfer. Miculicich et al. (2018): ich **denke** auch an die Opfer.
Our Model: meine **Gedanken** sind auch bei den Opfern.

Head 2: Attention to related words *sympathy, support, hope*
s^{j-2}: (FR) Madam President, many things have already been said , but I would like to echo all the words of sympathy and support that have already been addressed to the peoples of Tunisia and Egypt .
s^{j+4}: it must implement a strong strategy towards these countries .
s^{j-1}: they are a symbol of hope for all those who defend freedom .

Table 4.6: Example of noun disambiguation. Source context sentences are ordered in decreasing probability mass. The intensity of color corresponds to the attention given to a specific word before rescaling.

Qualitative Analysis To analyse the effect of using sparse attention on both the sentence and word-level, we looked at the attention weights computed by *sparsemax*. Table 4.6 shows an example where our model helped generate a correct translation of the noun “thoughts” (highlighted in bold). The context sentences shown in the bottom box had the highest attention weights as assigned by sparsemax. It seems that this particular attention head focuses more on phrases like “words of sympathy”, “support’, “symbol of hope” which are related to the query “thoughts”. Another example in Table 4.7 shows how our model correctly translates the pronoun “their”. Upon looking at the words in the context sentences, it seems that this particular attention head focuses on words related to the antecedent “Croatia’s Serbian population” with most of the weight concentrated around neighbouring words in sentence s^{j-1}. It is evident from both examples that word-level sparsity is more prevalent in longer sentences in the context; the same holds for sparsity at sentence-level.

4.5 Related Work

Our work builds upon the research in document-level MT, broadly classified into conventional MT (refer to Section 2.2.1 for details) and neural MT, and the research in sparse attention for NLP.

Src: Croatia is **their** homeland, too. Tgt: Kroatien ist auch **ihre** Heimat.
Transformer: Kroatien ist auch **seine** Heimat. Our Model: Kroatien ist auch **ihr** Heimatland.

Head 8: Attention to words related to the antecedent.
s^{j-1}: to name but a few, these include cooperation with the Hague Tribunal , efforts made so far in prosecuting corruption , restructuring the economy and finances and greater commitment and sincerity in eliminating the obstacles to the return of Croatia 's Serbian population .
s^{j-4}: by signing a border arbitration agreement with its neighbour Slovenia , the new Croatian Government has not only eliminated an obstacle to the negotiating process , but has also paved the way for the resolution of other issues .

Table 4.7: Example of pronoun disambiguation. Context sentences are ordered in decreasing probability mass.

Document-level Neural MT We look at previous works from the perspective of the type of context they use, that is *online*—use previous context only, or *offline*—use both past and future contexts. Most works fall into the former category, with those that (i) use only a single previous sentence in the source by having a separate attention over it (Jean et al., 2017; Voita et al., 2018) or concatenating it with the current source sentence (Tiedemann and Scherrer, 2017); (ii) use a single previous sentence both in source and target via a multi-encoder model (Bawden et al., 2018); (iii) use more than one previous source sentence by having a two-level hierarchical RNN over three previous source sentences (Wang et al., 2017) or having a separate context encoder over concatenation of two previous source sentences (Zhang et al., 2018); and (iv) use a few previous source and target sentences by having a hierarchical attention network over three previous sentences (Miculicich et al., 2018). Apart from fixing the context length, there are few works that use cache-based memories to store contextual information from preceding sentences (Tu et al., 2018) and also store topical words in a cache (Kuang et al., 2018) to improve the MT system performance. A recent work (Maruf et al., 2018) reports promising results when using the complete history for translating online conversations.

For the offline setting, however, there is only one work that effectively uses the full document-context on both source and target-side using memory networks (Maruf and Haffari, 2018). The debate in document-level NMT today is mostly about how much of the previous context to use and there has been no comparison between the online and offline setting except using only one previous and following sentence (Voita et al., 2018). Our hierarchical selective attention approach is most similar in concept to the one by Miculicich et al. (2018) but outperforms it in terms of translation performances across different domains (Table 4.3) and is more efficient in training (Table 4.5).

Sparse Attention Sparse attention and its constrained variants have been used to address the coverage problem in NMT (Malaviya et al., 2018) by limiting the amount of attention that each source word can receive. Apart from NMT, sparse attention has been shown to yield promising results for NLP tasks of textual entailment (Martins and Astudillo, 2016) and summarisation (Niculae and Blondel, 2017).

4.6 Summary

In this chapter, we have presented a novel, scalable and efficient approach to hierarchical attention for context-aware NMT, which uses sparse attention to selectively focus on relevant sentences in the document context and then attend to key words in those sentences. We also present single-level attention approaches based on sentence or word-level information in the context. The document-level context representation, produced from these attention modules, is integrated into the encoder or decoder of the Transformer architecture depending on whether we use monolingual or bilingual context. Experiments and evaluation on three English→German datasets in offline and online document MT settings show that our approach surpasses context-agnostic and recent context-aware baselines. The qualitative analysis shows that the sparsity at sentence-level allows our model to identify key sentences in the document context and the sparsity at word-level allows it to focus on key words in those sentences allowing for efficient compression of memory and is a step towards better interpretation of document-level NMT models.

Part II

Dialogue Translation

Chapter 5

Translating Bilingual Multi-Speaker Conversations

5.1 Introduction

In the previous chapters, we presented two approaches to model the document-wide context for monologue translation. We demonstrated that using document-wide source and target context is crucial for achieving performance gains over context-agnostic NMT models and most context-aware NMT baselines for several language-pairs. Looking at current and past work in the field, we made a salient observation: all previous research in context-aware NMT has focused on improving the performance of translation models for the traditional task of monologue translation. This prompted us to consider other practical avenues of MT that could greatly benefit from context information. One application that instantly comes to mind is dialogue where context is crucial to achieve an uninterrupted and eloquent exchange of information.

The ultimate aim of all machine translation systems for dialogue is to enable a multilingual conversation between multiple speakers. However, translation of such conversations is not well-explored in the literature, even though translating such conversations online is ubiquitous in real life, e.g., in the European Parliament, United Nations, and customer service chats. This scenario involves leveraging the conversation history in multiple languages. The goal of this chapter[1] is to propose and explore a simplified version of such a setting, referred to as bilingual multi-speaker machine translation (Bi-MSMT), where speakers' turns in the conversation switch the source and target language. To illustrate the significance of

[1] First presented in Maruf et al. (2018).

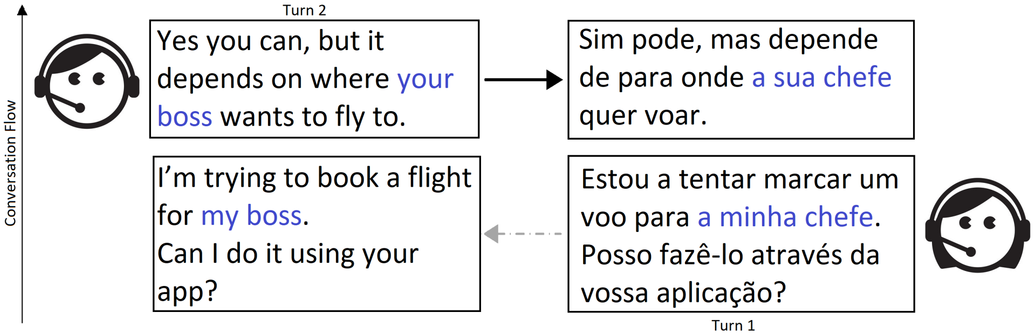

Turn 2
Yes you can, but it depends on where your boss wants to fly to.
Sim pode, mas depende de para onde a sua chefe quer voar.
I'm trying to book a flight for my boss.
Can I do it using your app?
Estou a tentar marcar um voo para a minha chefe.
Posso fazê-lo através da vossa aplicação?
Turn 1
Conversation Flow

Figure 5.1: An example English-Portuguese conversation in a customer service chat. Turn denotes the set of sentences from a single speaker.

using conversation history in this scenario, let us consider the example in Figure 5.1 showing a customer service chat. Assume that we want to translate the response by the agent in English to Portuguese for the client. It should be noted that the referent determiner of *boss* (highlighted in blue) in English is gender-insensitive and it is impossible to disambiguate the gender given the English source sentence alone. Hence, we require the Portuguese source sentence (first sentence in Turn 1), which has explicit mention of the gender of the boss (*a minha chefe*), to accurately translate the English response by the agent. In general, we may also require the English source or translations to capture anaphoric information or discourse connectives, and the Portuguese translations to maintain lexical coherence. The Bi-MSMT task is thus challenging as the conversation history consists of utterances in both languages.

As previously mentioned, there has been work focusing on using the discourse or document context to improve NMT, in an online setting, by using the past context (Jean et al., 2017; Wang et al., 2017; Bawden et al., 2018; Voita et al., 2018), and in an offline setting, using the past and future context (Maruf and Haffari, 2018). In this chapter, we design and evaluate a conversational Bi-MSMT model, where we incorporate the source and target-side conversation histories into a sentence-based attentional model (Bahdanau et al., 2015). [2] Here, the source history comprises of sentences in the original language (either English or Foreign), and the target history consists of their corresponding translations. We experiment with different ways of computing the source-context representation

[2] The Transformer architecture (Vaswani et al., 2017) was not a well-established baseline when this work was initiated. The ideas presented here are not specific to any base model and can easily be extended to the Transformer architecture.

for this task. Furthermore, we present an effective approach to leverage the target-side context and also present an intuitive approach for incorporating both contexts simultaneously. To evaluate this task, we introduce datasets extracted from Europarl v7 and OpenSubtitles2016, containing speaker information. Our experiments on English-French, English-Estonian, English-German and English-Russian language-pairs show improvements of +1.4, +1.2, +1.8 and +0.3 BLEU, respectively, for our best model over the context-free baseline. We also perform experiments on English-French and English-German customer service chat data yielding promising results. The results show the impact of conversation history on the translation of bilingual multi-speaker conversations and can be used as a benchmark for future work on this task.

5.2 Preliminaries

5.2.1 Problem Formulation

We are given a dataset that comprises parallel conversations, and each conversation consists of *turns*. Each turn is constituted by sentences spoken by a single speaker, where the sentences are denoted by $\boldsymbol{x}$ or $\boldsymbol{y}$ if in English or Foreign language, respectively. The goal is to learn a model that is able to leverage the mixed-language conversation history in order to produce high-quality translations.

5.2.2 Data

Standard machine translation datasets are inappropriate for the Bi-MSMT task since they are not composed of conversations or the speaker annotations are missing. In this section, we describe how we extract data from raw Europarl v7 (Koehn, 2005) and OpenSubtitles2016[3] (Lison and Tiedemann, 2016) for this task.[4]

Europarl The raw Europarl v7 corpus (Koehn, 2005) contains `SPEAKER` and `LANGUAGE` tags where the latter indicates the original language of utterance used by the speaker. The individual files (referred to as episodes in the corpus) are first split into conversations (referred to as chapters). The data is tokenised (using scripts by Koehn (2005)), and cleaned

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

[4] The data is publicly available at https://github.com/sameenmaruf/Bi-MSMT.git.

	Europarl			Subtitles
	En-Fr	En-Et	En-De	En-Ru
#Conversations	6997/140/209	4394/88/132	3582/70/108	23126/462/694
#Sentences	246.5K/4.9K/7.8K	174K/3.2K/5.2K	109K/2.1K/3.3K	291.5K/5.9K/9K

Table 5.1: General statistics for training/development/test sets (K stands for thousands).

	Europarl			Subtitles
	En-Fr	En-Et	En-De	En-Ru
#Sentences (English)	139.8K	130.6K	55.5K	157.9K
Mean Statistics per Conversation				
#Sentences	36.24	40.65	31.50	13.60
#Turns	4.77	4.85	4.79	7.12
Turn Length	7.12	7.92	6.16	1.68

Table 5.2: Statistics for training set.

(headings and single token sentences removed). Conversations are divided into smaller ones if the number of speakers is greater than 5.[5] The corpus is then randomly split into train/dev/test sets with respect to conversations in ratio 100:2:3. The English-side of the corpus is set as reference, and if the language tag is absent, the source language is English, otherwise Foreign. The sentences in the source-side of the corpus are kept or swapped with those in the target-side based on this tag.

We perform the aforementioned steps for English-French, English-Estonian and English-German, and obtain the bilingual multi-speaker corpora for the three language-pairs. Before splitting into train/dev/test sets, we remove conversations with sentences having more than 100 tokens for English-French, English-German and more than 80 tokens for English-Estonian respectively,[6] to limit the sentence-length for using subwords with BPE (Sennrich et al., 2016). The data statistics are given in Tables 5.1 and 5.2.[7]

Subtitles There has been recent work to obtain speaker labels via automatic turn segmentation for the OpenSubtitles2016 corpus (Lison and Meena, 2016; van der Wees et al., 2016; Wang et al., 2016). We obtain the English-side of OpenSubtitles2016 corpus anno-

[5] Using the conversations as is or setting a higher threshold further reduces the data due to inconsistencies in conversation/turn lengths in the source and target-side.

[6] Sentence-lengths of 100 tokens result in longer sentences than what we get for the other two language-pairs.

[7] Although the extracted dataset is small but we believe it to be a realistic setting for a real-world conversation task, where reference translations are usually not readily available and expensive to obtain.

tated with speaker information by Lison and Meena (2016).[8] To obtain the parallel corpus, we use the OpenSubtitles alignment links to align foreign subtitles to the annotated English ones. For each subtitle, we extract individual conversations with more than 5 sentences and at least two turns. Conversations with more than 30 turns are discarded. Finally, since subtitles are in a single language, we assign language tags such that the same language occurs in alternating turns. We thus obtain the Bi-MSMT corpus for English-Russian, which is then divided into training, development and test sets.

5.3 Conversational Bi-MSMT Model

Before we delve into the details of how to leverage the conversation history, we identify the three types of context we may encounter in an ongoing bilingual multi-speaker conversation, as shown in Figure 5.2. It comprises: (i) the previously completed English turns, (ii) the previously completed Foreign turns, and (iii) the ongoing turn (English or Foreign).

We propose a conversational Bi-MSMT model that is able to incorporate all three types of context using source, target or dual conversation histories into a context-agnostic base model. The base model caters to the speaker's language transition by having one sentence-based NMT model (Bahdanau et al., 2015) for each translation direction, English→Foreign and Foreign→English. We now describe our approach for extracting relevant information from the source and target bilingual conversation history.

5.3.1 Computing Representations of Source and Target-side Histories

Suppose we are translating an ongoing conversation having alternating turns of English and Foreign. We are currently in the $2k+1^{th}$ turn (in English) and want to translate its j^{th} sentence using the source and target-side conversation histories, represented by context vectors $\mathbf{o}^{\boldsymbol{src}}$ (dimensions H) and $\mathbf{o}^{\boldsymbol{tgt}}$ respectively (also dimensions H).

The simplest way to utilise the dual conversation history is to incorporate both context vectors $\boldsymbol{o}^{\boldsymbol{src}}$ and $\boldsymbol{o}^{\boldsymbol{tgt}}$ as additional inputs into the base model, referred to as Dual Context Src-Tgt. Let us now describe how we compute the source and target-side histories.

[8] The majority of sentences still have missing annotations (Lison and Meena, 2016) due to changes between the original script and the actual movie or alignment problems between scripts and subtitles. As for Wang et al. (2016), their publicly released data is even smaller than our En-De dataset extracted from Europarl.

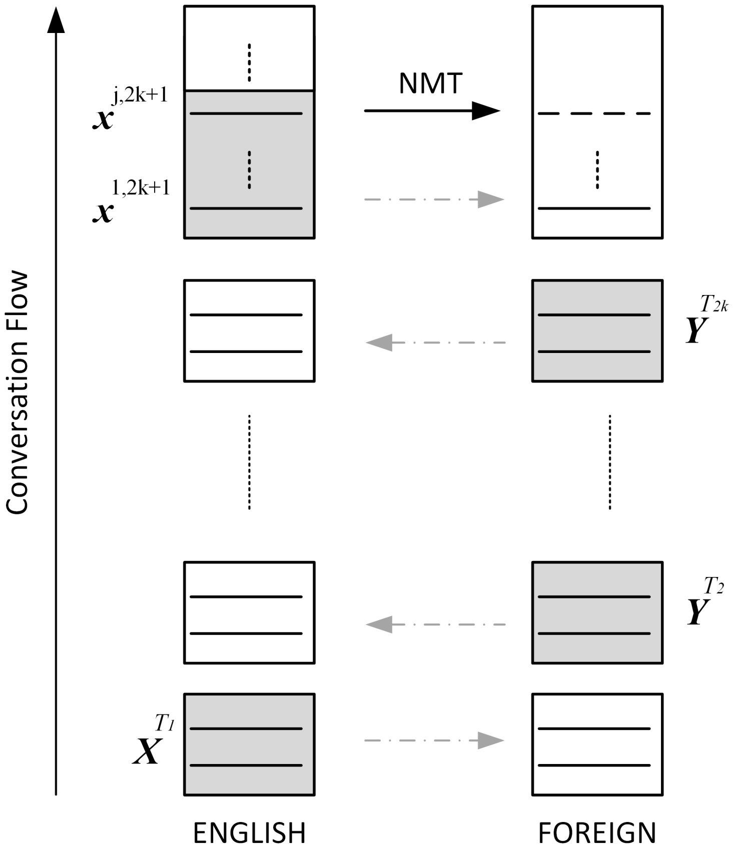


Figure 5.2: Illustration of an ongoing conversation while translating j^{th} sentence in $2k+1^{th}$ turn. $\boldsymbol{X}^{T_i}$ and $\boldsymbol{Y}^{T_i}$ contain the sentences in previously completed English and Foreign turn respectively, and $\mathbf{x}^{j,i}$ denotes the j^{th} sentence in ongoing English turn i. The shaded turns are observed, i.e., source (the speaker utterances), while the rest are unobserved, i.e., the target translations or the unuttered source sentences for the current turn.

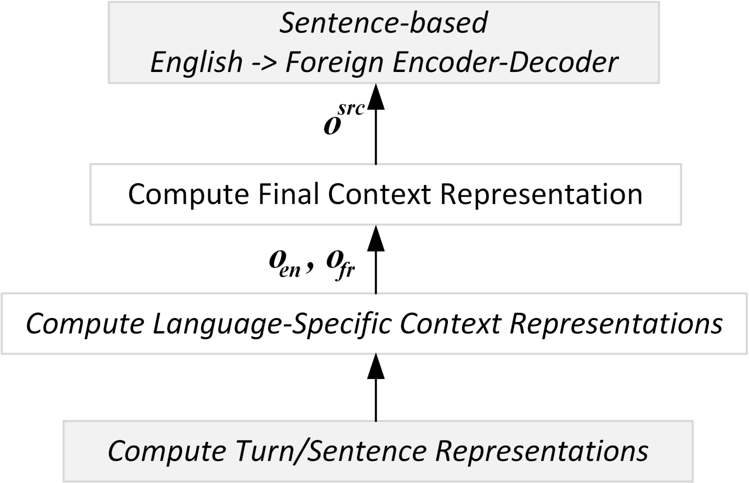

Sentence-based
English -> Foreign Encoder-Decoder
o^{src}
Compute Final Context Representation
o_{en}, o_{fr}
Compute Language-Specific Context Representations
Compute Turn/Sentence Representations

Figure 5.3: Operational overview of the model when using source-side conversation history.

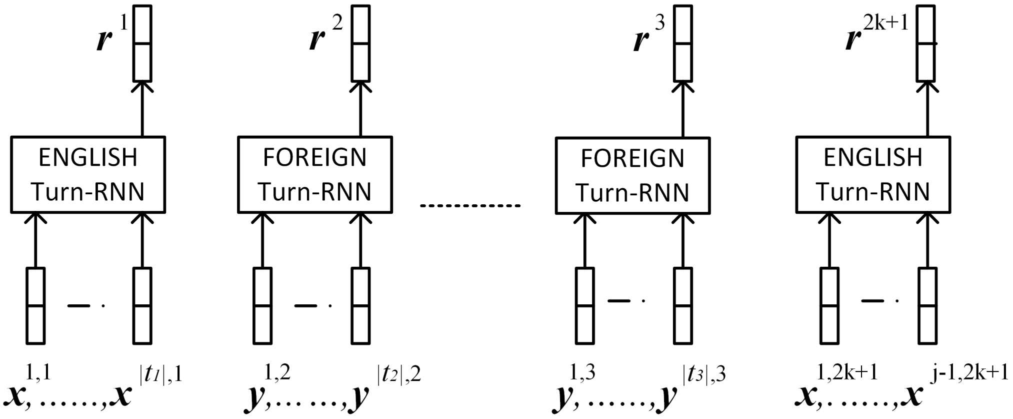

r¹
r²
r³
r²ᵏ⁺¹
ENGLISH Turn-RNN
FOREIGN Turn-RNN
FOREIGN Turn-RNN
ENGLISH Turn-RNN
x^{1,1},......,x^{|t1|,1}
y^{1,2},......,y^{|t2|,2}
y^{1,3},......,y^{|t3|,3}
x^{1,2k+1},......,x^{j-1,2k+1}

Figure 5.4: English and Foreign language Turn-RNNs to produce representations of completed and ongoing turns.

5.3.1.1 Source-Side History

A high-level view of the individual operations that are involved in computing the source (or target) conversation history is shown in Figure 5.3. Let us assume that we already have the representations of previous source sentences in the conversation. We pass the source sentence representations through Turn-RNNs, as shown in Figure 5.4, where the Turn-RNNs are composed of language-specific bi-directional RNNs irrespective of the speaker (one each for English and Foreign). We concatenate the last hidden states of the forward and backward Turn-RNNs to get the final turn representation $\mathbf{r}^i$, where i denotes the turn index.

The individual turn (or sentence) representations from the Turn-RNNs[9] are then combined in several possible ways (described shortly) to obtain context vectors $\mathbf{o}_{en}$ and $\mathbf{o}_{fr}$, which are further amalgamated using a gating mechanism so as to give differing importance to each element of the context vector:

$$\begin{aligned} \boldsymbol{o}_{en,fr} &= \boldsymbol{\alpha} \odot \boldsymbol{o}_{en} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \boldsymbol{o}_{fr} \\ \boldsymbol{\alpha} &= \sigma(\boldsymbol{U}_{en}\boldsymbol{o}_{en} + \boldsymbol{U}_{fr}\boldsymbol{o}_{fr} + \boldsymbol{b}_g) \end{aligned} \tag{5.1}$$

where σ is the logistic sigmoid function defined in Table 2.1, $\boldsymbol{U}$'s are parameter matrices and $\boldsymbol{b}_g$ is a parameter vector. Finally, since $\boldsymbol{o}_{en,fr}$ is of dimensions *2H*, we perform a

[9] For this work, we define the turns based on language and do not use the speaker information as for real-world chat scenarios (e.g., agent-client in a customer service chat), we do not have multiple speakers based on language. We leave this for future exploration.

dimensionality reduction to obtain:

$$\boldsymbol{o}^{src} = \tanh(\boldsymbol{W}_{ot}\boldsymbol{o}_{en,fr} + \boldsymbol{b}_o) \tag{5.2}$$

This final source-context vector $\boldsymbol{o}^{src}$ can then be incorporated into the corresponding encoder-decoder module in the base model (English→Foreign in this case).

Let us now provide a description of how the language-specific context vectors $\mathbf{o}_{en}$ and $\mathbf{o}_{fr}$ are computed.[10] In the remainder of this section, $\{\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{b}\}$ are language-specific learned parameters.

Direct Transformation The simplest approach is to combine turn representations using a language-specific dimensionality reduction transformation:

$$\begin{aligned}
\boldsymbol{o}_{en} &= \tanh\left([\boldsymbol{W}_{en}, \ldots, \boldsymbol{W}_{en}]\begin{bmatrix}\boldsymbol{r}^1\\ \vdots \\ \boldsymbol{r}^{2k+1}\end{bmatrix} + \boldsymbol{b}_{en}\right)\\
\boldsymbol{o}_{fr} &= \tanh\left([\boldsymbol{W}_{fr}, \ldots, \boldsymbol{W}_{fr}]\begin{bmatrix}\boldsymbol{r}^2\\ \vdots \\ \boldsymbol{r}^{2k}\end{bmatrix} + \boldsymbol{b}_{fr}\right)
\end{aligned}$$

Here $\mathbf{r}^i$'s are concatenated to form a column vector, $\boldsymbol{W}$'s are matrices of dimensions $H \times 2H$ and $\boldsymbol{b}$'s are vectors of size H. This is followed by Eq. 5.1 to get a single context representation. Since output vectors are already of size H, there is no need to perform dimensionality reduction (Eq. 5.2).

Hierarchical Gating We propose a language-specific exponential decay gating based on the intuition that the farther the previous turns are from the current one, the lesser their impact may be on the translation of a sentence in the ongoing turn, similar in spirit to the caching mechanism by Tu et al. (2018):

$$\boldsymbol{o}_{en} = \boldsymbol{g}_{en}(\boldsymbol{g}_{en}(\ldots \boldsymbol{g}_{en}(\boldsymbol{g}_{en}(\boldsymbol{r}^1, \boldsymbol{r}^3), \boldsymbol{r}^5)\ldots), \boldsymbol{r}^{2k-1}), \boldsymbol{r}^{2k+1})$$

where

$$\begin{aligned}
\boldsymbol{g}_{en}(\mathbf{a}, \mathbf{b}) &= \boldsymbol{\gamma} \odot \mathbf{a} + (\mathbf{1} - \boldsymbol{\gamma}) \odot \mathbf{b}\\
\boldsymbol{\gamma} &= \sigma(\boldsymbol{U}_{\mathbf{1},en}\mathbf{a} + \boldsymbol{U}_{\mathbf{2},en}\mathbf{b} + \boldsymbol{b}_{en})
\end{aligned}$$

$\boldsymbol{o}_{fr}$ is computed in a similar manner, followed by Eqs. 5.1 and 5.2.

[10] It should be mentioned that $\mathbf{o}_{en}$ and $\boldsymbol{o}_{fr}$ are computed so as to lie in the target language space (assumed to be Foreign here).

Language-Specific Attention The strongest of our approaches is to combine the English and Foreign turn representations (obtained from the Turn-RNNs) separately via attention so as to allow the model to focus on relevant turns in the English and the Foreign context. Computing the attention over the English turns is straightforward, however, the same cannot be said for the Foreign turns as they come from a different language space than the current source sentence being translated (assumed to be in English). To circumvent this issue, we perform a cross-language non-linear projection on the query (from the current English sentence) prior to computing attention weights over the Foreign turn representations (shown pictorially in Figure 5.5). We perform a similar projection when computing the attention-weighted output for the English history $\mathbf{o_{en}}$. Mathematically:

$$\begin{aligned}
\boldsymbol{p_{en}} &= \text{softmax}\Big([\boldsymbol{r}^1, \ldots, \boldsymbol{r}^{2k+1}]^\top \boldsymbol{h}^j\Big) \\
\boldsymbol{p_{fr}} &= \text{softmax}\Big([\boldsymbol{r}^2, \ldots, \boldsymbol{r}^{2k}]^\top (\tanh(\boldsymbol{W_{en}} \boldsymbol{h}^j + \boldsymbol{b_{en}}))\Big) \\
\boldsymbol{o_{en}} &= \tanh\Big(\boldsymbol{W_{en}}([\boldsymbol{r}^1, \ldots, \boldsymbol{r}^{2k+1}]\boldsymbol{p_{en}}) + \boldsymbol{b_{en}}\Big) \\
\boldsymbol{o_{fr}} &= [\boldsymbol{r}^2, \ldots, \boldsymbol{r}^{2k}]\boldsymbol{p_{fr}}
\end{aligned} \tag{5.3}$$

Here $\boldsymbol{r}^i$'s are concatenated column-wise, $\boldsymbol{h}^j$ is the concatenation of last hidden state of forward and backward RNNs in the encoder (dimensions *2H*) for current sentence j in turn $2k+1$ and $\{\boldsymbol{W_{en}}, \boldsymbol{b_{en}}\}$ transform the language space to that of the target language.

To evaluate the significance of a more fine-grained context, we also propose to use the sentence information explicitly via a sentence-level attention, that is, we replace the turn representations with their sentence-level counterparts. This variant is referred to as **Language-Specific Sentence-level Attention**. The sentence-level representations are obtained by concatenating the corresponding hidden states of forward and backward Turn-RNNs and getting a matrix $[\boldsymbol{r}^{1,1}, \ldots, \boldsymbol{r}^{|t_1|,1}, \ldots, \boldsymbol{r}^{1,2k+1}, \ldots, \boldsymbol{r}^{j-1,2k+1}]$ for all the previous English sentences, and another matrix $[\boldsymbol{r}^{1,2}, \ldots, \boldsymbol{r}^{|t_2|,2}, \ldots, \boldsymbol{r}^{1,2k}, \ldots, \boldsymbol{r}^{|t_{2k}|,2k}]$ for all the previous Foreign sentences.[11]

Another special case of the Language-Specific Attention is to project all turn representations into the target language space prior to computing the attention, referred to as **Combined Attention**. Since this is a language-independent attention, it is not followed by

[11] $\boldsymbol{r}^{j,i}$ is the representation of source sentence j in turn i computed by the bi-directional Turn-RNN.

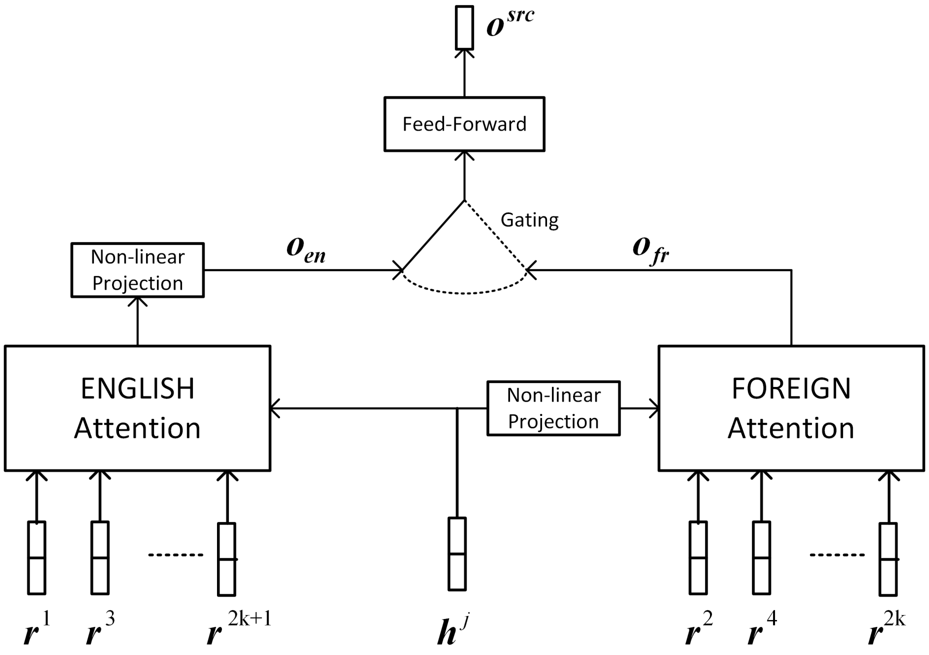


Figure 5.5: Computing source-side context representation $\boldsymbol{o^{src}}$ using Language-Specific Attention.

any context gating. The hypothesis here is to verify if the model actually benefits from Language-Specific attention or not.

$$\begin{aligned}\boldsymbol{p_{en,fr}} &= \text{softmax}\Big([\boldsymbol{r^1_{en}}, \boldsymbol{r}^2, \dots, \boldsymbol{r^{2k+1}_{en}}]^\top (\tanh(\boldsymbol{W_{en}h}^j + \boldsymbol{b_{en}}))\Big) \\ \boldsymbol{o_{en,fr}} &= [\boldsymbol{r^1_{en}}, \boldsymbol{r}^2, \dots, \boldsymbol{r^{2k+1}_{en}}]\boldsymbol{p_{en,fr}}\end{aligned}$$

Here $\boldsymbol{r^{2k+1}_{en}} = \tanh(\boldsymbol{W_{en}r}^{2k+1} + \boldsymbol{b_{en}})$.

5.3.1.2 Target-Side History

Using target-side conversation history is as important as that of the source-side since it helps in making the translation more faithful to the target language. This becomes crucial for translating conversations where the previous turns are all in the same language.

Let us again assume that we already have the representations of previous target sentences in the conversation. Instead of using Turn-RNNs, we directly combine the target representations using the *Language-Specific Sentence-level Attention* approach to obtain context vectors $\mathbf{o_{en}}$ and $\mathbf{o_{fr}}$. Specifically, let $\boldsymbol{R_{en}}$ be the matrix comprising of the target sentence representations in Foreign language for the English source sentences and $\boldsymbol{R_{fr}}$ be the matrix of target sentence representations (in English) for the previous Foreign turns. Here each target sentence representation has dimensions *H*. Then,

$$\begin{aligned}
\boldsymbol{p_{en}} &= \text{softmax}(\boldsymbol{R}_{\boldsymbol{en}}^{\top} \tanh(\boldsymbol{W_{tr,en}} \boldsymbol{h}^{j} + \boldsymbol{b_{tr,en}})) \\
\boldsymbol{p_{fr}} &= \text{softmax}(\boldsymbol{R}_{\boldsymbol{fr}}^{\top} (\boldsymbol{W_{dim,en}} \boldsymbol{h}^{j} + \boldsymbol{b_{dim,en}})) \\
\boldsymbol{o_{en}} &= \boldsymbol{R_{en}} \boldsymbol{p_{en}} \\
\boldsymbol{o_{fr}} &= \tanh(\boldsymbol{W_{tr,en}} (\boldsymbol{R_{fr}} \boldsymbol{p_{fr}}) + \boldsymbol{b_{tr,en}})
\end{aligned}$$

where $\{\boldsymbol{W_{tr,en}}, \boldsymbol{b_{tr,en}}\}$ are for both dimensionality reduction and changing the language space of the query vector $\boldsymbol{h}_j$ and the context vector, while $\{\boldsymbol{W_{dim,en}}, \boldsymbol{b_{dim,en}}\}$ are only for dimensionality reduction.

The context vectors $\boldsymbol{o_{en}}$ and $\boldsymbol{o_{fr}}$ are further combined using a gating mechanism as in Eq. 5.1 to obtain the final target context vector $\boldsymbol{o^{tgt}}$ (dimensions *H*).

5.3.2 Mixing Source and Target-side Histories along the Language axis

As already mentioned, the Dual Context Src-Tgt adds both context vectors $\boldsymbol{o^{src}}$ and $\boldsymbol{o^{tgt}}$ as auxiliary inputs into the base model. Another intuitive approach, as evident from Figure 5.2, is to separately model English and Foreign sentences using two context vectors $\boldsymbol{o^{en,m}}$ and $\boldsymbol{o^{fr,m}}$, where each is constructed from a mixture of the original source or target translations, is language-specific and possibly contain less noise. We refer to this as the Dual Context Src-Tgt-Mix. Suppose $\boldsymbol{R_{en,m}}$ contains the mixed source/target representations for English (the dimensions for source representations have been reduced to *H*) and $\boldsymbol{R_{fr,m}}$ contains the mixed source/target representations for Foreign language. Then,

$$\begin{aligned}
\boldsymbol{p_{en,m}} &= \text{softmax}(\boldsymbol{R}_{\boldsymbol{en,m}}^{\top} (\boldsymbol{W_{dim,en}} \boldsymbol{h}^{j} + \boldsymbol{b_{dim,en}})) \\
\boldsymbol{p_{fr,m}} &= \text{softmax}(\boldsymbol{R}_{\boldsymbol{fr,m}}^{\top} \tanh(\boldsymbol{W_{tr,en}} \boldsymbol{h}^{j} + \boldsymbol{b_{tr,en}})) \\
\boldsymbol{o^{en,m}} &= \tanh(\boldsymbol{W_{tl,en}} (\boldsymbol{R_{en,m}} \boldsymbol{p_{en,m}}) + \boldsymbol{b_{tl,en}}) \\
\boldsymbol{o^{fr,m}} &= \boldsymbol{R_{fr,m}} \boldsymbol{p_{fr,m}}
\end{aligned}$$

where $\boldsymbol{W_{dim,en}}$, $\boldsymbol{W_{tl,en}}$ and $\boldsymbol{W_{tr,en}}$ are for dimensionality reduction, changing the language space and both, respectively, and $\boldsymbol{o^{en,m}}$, $\boldsymbol{o^{fr,m}}$ are used as auxiliary inputs in the base model instead of $\boldsymbol{o^{src}}$, $\boldsymbol{o^{tgt}}$.

5.3.3 Incorporating Context into Base Model

The final representations $\boldsymbol{o}^{\boldsymbol{src}}$ and $\boldsymbol{o}^{\boldsymbol{tgt}}$ (or $\boldsymbol{o}^{\boldsymbol{en},\boldsymbol{m}}$ and $\boldsymbol{o}^{\boldsymbol{fr},\boldsymbol{m}}$), can be incorporated together or individually in the base model by:

- **InitDec** Using a non-linear transformation to initialise the decoder, similar to Wang et al. (2017): $\boldsymbol{s}_0^j = \tanh(\boldsymbol{U}\boldsymbol{o}^j + \boldsymbol{b_s})$, where j is the sentence index in current turn $2k+1$, $\{\boldsymbol{U},\ \boldsymbol{b_s}\}$ are encoder-decoder specific parameters and $\boldsymbol{o}^j$ is either a single context vector or a concatenation (followed by non-linear transformation) of the two.

- **AddDec** As auxiliary inputs to the decoder (similar to Jean et al. (2017); Wang et al. (2017); Maruf and Haffari (2018)):

$$\boldsymbol{s}_n^j = \text{RNN}(\boldsymbol{s}_{n-1}^j, \boldsymbol{E_T}[y_{n-1}^j], \boldsymbol{c}_n^j, \boldsymbol{o}^{j,\boldsymbol{src}}, \boldsymbol{o}^{j,\boldsymbol{tgt}})$$

- **InitDec+AddDec** Combination of previous two approaches.

5.3.4 Training and Decoding

The model parameters are trained end-to-end by maximising the sum of log-likelihood of the bilingual conversations in training set $\mathcal{D}$. That is, for a conversation having turns of both English and Foreign language, the log-likelihood is the sum of log of the conditional probability of producing the target translations given the source and conversation history:

$$\sum_{i \in odd} \sum_{j=1}^{|t_i|} \log P_{\boldsymbol{\theta}}(\boldsymbol{y}^{j,i} | \boldsymbol{x}^{j,i}, \boldsymbol{o}^{j,i}) + \sum_{i' \in even} \sum_{j'=1}^{|t_{i'}|} \log P_{\boldsymbol{\theta}}(\boldsymbol{x}^{j',i'} | \boldsymbol{y}^{j',i'}, \boldsymbol{o}^{j',i'})$$

where $j,\ j'$ denote sentences belonging to odd or even turns respectively, and $\boldsymbol{o}^{(\cdot)}$ is a representation of the conversation history (single or dual).

The best output sequence for a given input sequence for the j^{th} sentence at test time, aka decoding, is produced by:

$$\underset{\boldsymbol{y}^j}{\text{arg max}}\ P_{\boldsymbol{\theta}}(\boldsymbol{y}^j | \boldsymbol{x}^j, \boldsymbol{o}^j)$$

5.4 Experiments on Public Data

Implementation and Hyperparameters We implement our conversational Bi-MSMT model in C++ using the `DyNet` library (Neubig et al., 2017). The base model is built using `mantis` (Cohn et al., 2016) which is an implementation of the generic sentence-level attentional NMT model in `DyNet`.

The base model has single layer bi-directional GRUs in both encoders and two-layer GRUs in the decoders.[12] The hidden dimensions and word embedding sizes are set to 256, and the alignment dimension (for the attention mechanism in the decoder) is set to 128.

Models and Training As already mentioned, the base model consists of two encoder-decoder architectures, one translating from English and the other to English. We do a stage-wise training for this model, i.e., we train the English→Foreign architecture followed by the Foreign→English architecture, using the full sentence-level parallel corpus. Both architectures have the same vocabulary[13] but separate parameters to avoid biasing the embeddings towards the architecture trained last. The contextual model is pre-trained similar to training the base model. The best model is chosen based on the minimum overall perplexity on the bilingual dev set.

For computing the source-context representations, we use sentence representations generated by two bi-directional RNNLMs (one each for English and Foreign) trained offline, which are then fed as input to the Turn-RNNs in our source-context models. For the target sentence representations, we use the last hidden states of the decoder generated from the pre-trained base model.[14] At decoding time, however, we use the last hidden state of the decoder computed by our conversational model (not the base) as the target sentence representations.

For the base model, we make use of stochastic gradient descent (SGD)[15] with an initial learning rate of 0.1 and a decay factor of 0.5 after the fifth epoch for a total of 15 epochs.

[12] We follow Cohn et al. (2016) and Britz et al. (2017) in choosing hyperparameters for our model.

[13] For each language-pair, we use BPE (Sennrich et al., 2016) to obtain a joint vocabulary of size ≈30k.

[14] Even though the parameters of the base model are updated, the target sentence representations are fixed throughout training. We experimented with a scheduled updating scheme in preliminary experiments but it did not yield significant improvement over the current strategy.

[15] In our preliminary experiments, we tried SGD, Adam and Adagrad as optimisers, and found SGD to achieve better perplexities in lesser number of epochs (Bahar et al., 2017).

For the contextual model, we use SGD with an initial learning rate of 0.08 and a decay factor of 0.9 after the first epoch for a total of 30 epochs. To avoid overfitting, we employ dropout and set its rate to 0.2. To reduce the training time of our contextual model, we perform the computation of one turn at a time, for instance, when using the source context, we run the Turn-RNNs for previous turns once and re-run the Turn-RNN only for sentences in the current turn.

Evaluation We evaluate generated translations from all neural models using BLEU (Papineni et al., 2002). We also apply bootstrap resampling (Clark et al., 2011) to measure statistical significance, $p < 0.05$, of our models compared to the base model.

5.4.1 Results

Firstly, we evaluate the three strategies for incorporating context: InitDec, AddDec, InitDec+AddDec, and report the results for source context using *Language-Specific Attention* in Table 5.3. For the Europarl data, we see decent improvements with InitDec for En-Et (+1.1 BLEU) and En-De (+1.6 BLEU), and with InitDec+AddDec for En-Fr (+1.2 BLEU). We also observe that, for all language-pairs, both translation directions benefit from context, showing that our training methodology was indeed effective. On the other hand, for the Subtitles data, we see a maximum improvement of +0.3 BLEU for InitDec+AddDec.[16] We narrow down to three major reasons: (i) the data is noisier when compared to Europarl, (ii) the sentences are short and generic with only 1% having more than 27 tokens, and finally (iii) the turns in OpenSubtitles2016 are short compared to those in Europarl (see Table 5.1), and we show later (Section 5.4.2) that the context from current turn is the most important.

The next set of experiments evaluates the five different approaches for computing the source-side context. It is evident from Table 5.3 that for English-Estonian and English-German, our model indeed benefits from using the fine-grained sentence-level information (*Language-Specific Sentence-level Attention*) as opposed to just the turn-level one.

Finally, our results with source, target and dual contexts are reported. Interestingly, just using the source context is sufficient for English-Estonian and English-German. For

[16] We saw almost double the improvement if we used separate BPE vocabularies/embeddings for English and Russian. However, we don't report those results since the overall BLEU score for the base model was lower (18.2) than that of the current one.

	Europarl									Subtitles		
	En-Fr			En-Et			En-De			En-Ru		
	Overall	En→Fr	Fr→En	Overall	En→Et	Et→En	Overall	En→De	De→En	Overall	En→Ru	Ru→En
Base Model	37.4	38.1	36.0	20.7	18.6	26.6	24.7	21.8	27.7	19.0	14.9	23.0
+Source Context as Lang-Specific Attention via												
InitDec	38.4†	39.2†	36.9†	**21.8**†	19.5†	**28.3**†	**26.3**†	**23.3**†	29.4†	18.9	14.9	22.6
AddDec	38.5†	**39.4**†	37.0†	21.6†	**19.7**†	27.5†	26.3†	23.1†	**29.5**†	19.3	15.2	23.1
InitDec+AddDec	**38.6**†	39.3†	**37.1**†	21.5†	19.4†	27.6†	26.2†	23.2†	29.3†	**19.4**	**15.2**	**23.1**
+Source Context via												
Direct Transformation	38.4†	39.1†	37.0†	21.8†	**19.6**†	28.1†	26.3†	23.3†	29.2†	19.1	14.9	22.8
Hierarchical Gating	38.3†	39.1†	36.9†	21.6†	19.6†	27.6†	26.3†	23.2†	29.4†	19.2	15.1	22.7
Lang-Specific Attention	38.4†	39.2†	36.9†	21.8†	19.5†	28.3†	26.3†	23.3†	29.4†	**19.4**	**15.2**	**23.1**
Lang-Specific S-Attention	38.5†	39.2†	**37.1**†	**21.8**†	19.6†	**28.4**†	**26.5**†	**23.5**†	**29.5**†	19.1	14.6	23.0
Combined Attention	**38.5**†	**39.4**†	36.9†	21.7†	19.5†	27.9†	26.4†	23.3†	29.4†	19.0	14.8	22.9
+Lang-Specific S-Attention using												
Source Context	38.5†	39.2†	37.1†	**21.8**†	19.6†	**28.4**†	**26.5**†	**23.5**†	29.5†	19.1	14.6	23.0
Target Context	38.8†	**39.6**†	37.4†	21.8†	**19.7**†	27.9†	26.2†	23.2†	29.3†	19.2	14.8	**23.2**
Dual Context Src-Tgt	**38.8**†	39.5†	**37.5**†	21.7†	19.6†	28.0†	26.4†	23.3†	**29.5**†	18.9	14.5	23.1
Dual Context Src-Tgt-Mix	38.8†	39.5†	37.4†	21.7†	19.6†	27.7†	26.4†	23.3†	29.5†	**19.3**	**14.9**	23.0

Table 5.3: BLEU scores for the bilingual test sets. Here all contexts are incorporated as InitDec for Europarl and InitDec+AddDec for Subtitles unless otherwise specified. **bold**: Best performance w.r.t. two decimal places, †: Statistically significantly better than the base model, based on bootstrap resampling (Clark et al., 2011) with $p < 0.05$.

	Europarl			Subtitles
	En-Fr	En-Et	En-De	En-Ru
Prev Sent	38.2	21.7	26.1	**19.1**
Our Model	**38.5**†	**21.8**	**26.5**†	19.1

Table 5.4: BLEU scores for the bilingual test sets. **bold**: Best performance w.r.t. two decimal places, †: Statistically significantly better than the contextual baseline.

English-French, on the other hand, we see significant improvements for models using the target-side conversation history over using only the source-side. We attribute this to the base model being more efficient and able to generate better translations for En-Fr as it had been trained on a larger corpus as opposed to the other two language-pairs. Unlike Europarl, for Subtitles, we see improvements for our Dual Context Src-Tgt-Mix over the Src-Tgt one for En→Ru, showing this to be an effective approach when the target representations are noisier.

To summarise, for the majority of cases our model with *Language-Specific Sentence-level Attention* is a winner or a close second. Using the target context is useful when the base model generates reasonable-quality translations; otherwise, using the source context should suffice.

Local Source Context Model Most of the previous works for online context-based NMT consider only a single previous sentence as context (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2018). Drawing inspiration from these works, we evaluate our model (trained with *Language-Specific Sentence-Level Attention*) on the same test set but using only the previous source sentence as context. This evaluation allows us to hypothesise what proportion of the overall gain can be attributed to only the previous sentence. From Table 5.4, it can be seen that our model surpasses the local-context baseline for Europarl showing that the wider context is indeed beneficial if the turn lengths are longer. For Subtitles (En-Ru), it can be seen that using the previous sentence is sufficient due to short turns (see Table 5.1).

5.4.2 Analysis

Ablation Study We conduct an ablation study to validate our hypothesis of using the complete context versus using only one of the three types of contexts in a bilingual multi-speaker conversation: (i) current turn, (ii) previous turns in the current language, and

Type of Context	**BLEU**
No context (*Base Model*)	24.7
Current Turn	26.4
Current Language from Previous Turns	26.2
Other Language from Previous Turns	26.3
Complete Context (*Our Model*)	**26.5**

Table 5.5: BLEU scores for English-German bilingual test set.

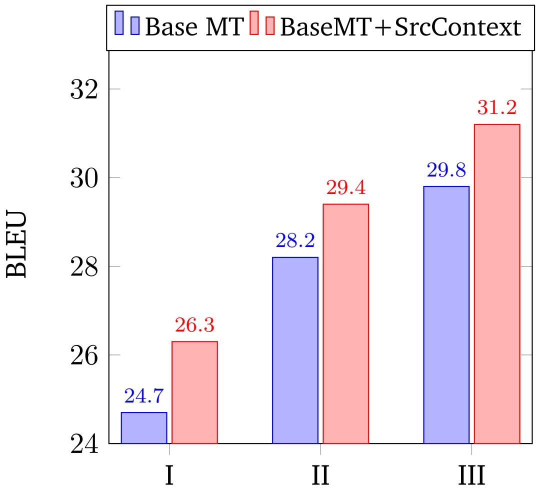


Figure 5.6: BLEU scores on En-De test set while training (I) smaller base model with smaller corpus (previous experiment), (II) smaller base model with larger corpus, and (III) a larger base model with larger corpus.

(iii) previous turns in the other language. The results for English-German are reported in Table 5.5. We see a decrease in BLEU for all types of contexts with significant decrease when considering only the current language from previous turns. The results show that the current turn has the most influence on the overall translation of a sentence, and we conclude that since our model is able to capture the complete context, it is generalisable to any conversational scenario.

Training base model with more data To analyse if the context is beneficial even when using more data, we perform an experiment for English-German where we train the base model with additional sentence-pairs from the full WMT'14 corpus[17] (excluding our dev/test sets and filtering sentences with more than 100 tokens). For training the contextual model, we still use the same bilingual multi-speaker corpus. We observe a significant improvement

[17] https://nlp.stanford.edu/projects/nmt/

En→Fr	les; par; est; a; dans; le; en; j’; un; afin; question; entre; qu’; être; ces; également; y; depuis; c’; ou
Fr→En	this; of; we; issue; europe; by; up; make; united; does; what; regard; s; must; however; such; whose; share; like; been
En→Et	eest; vahel; üle; nimel; ja; aastal; aasta; neid; ainult seepärast; nagu; kes; komisjoni; tehtud; küsimuses; sisserände; liikmesriigi; mulla; liibanoni; dawit
Et→En	for; this; of; is; political; important; culture; also; as; order; are; each; their; only; gender; were; its; economy; one; market
En→De	daß; auf; und; werden; nicht; müssen; aus; mehr; können; einem; rates; eines; insbesondere; wurden; habe; mitgliedstaaten; ist; sondern; europa; gemeinsamen
De→En	that; its; say; must; some; therefore; more; countries; an; favour; public; will; without; particularly; hankiss; much; increase; eu; them; parliamentary

Table 5.6: Most frequent tokens correctly generated by our model when compared to the base model.

of +1.1 BLEU for the context-based conversational model (Figure 5.6 II) over the base MT model, showing the significance of conversation history in this experiment condition.[18]

We perform another experiment where we use a larger base model, having almost double the number of parameters than our previous base model (hidden units and word embedding sizes set to 512, and alignment dimension set to 256), to test if the model parameters are being overestimated due to the additional context. We use the same WMT’14 corpus to train this larger base model and achieve significant improvement of +1.5 BLEU for our context-based model over the larger baseline (Figure 5.6 III).

How is the context helping? The underlying hypothesis for this work is that discourse phenomena in a conversation may depend on long-range dependency and these may be ignored by the sentence-based NMT models. To analyse if our contextual model is able to accurately translate such linguistic phenomena, we come up with our own evaluation procedure. We aggregate the tokens correctly generated by our model and those correctly generated by the baseline over the entire test set. We then take the difference of these counts and sort them.[19] Table 5.6 reports the top 20 tokens where our model is better than the baseline for the Europarl dataset. Figure 5.7 gives the density of counts obtained using

[18] It should be noted that the BLEU score for the base model trained with WMT does not match the published results exactly as the test set contains both English and German sentences. It does, however, fall between the scores usually obtained on WMT’14 for En→De and De→En.

[19] We do not normalise the counts with the background frequency as it favours rare words. Thus, obscuring the main reasons of improving the translation performance (BLEU score).

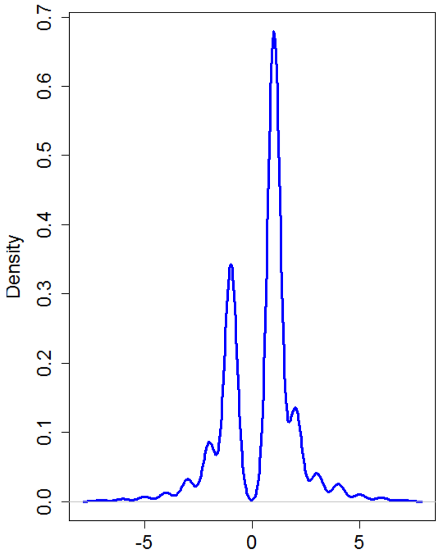


Figure 5.7: Density of token counts for En→Fr illustrating where our model is better (+ve x-axis) and where the base model is better (-ve x-axis).

our evaluation for English→French.[20] Positive counts correspond to correct translations by our model while the negative counts correspond to where the base model was better. It can be seen that, for the majority of cases, our model supersedes the base model. We observed a similar trend for other translation directions. In general, the correctly generated tokens by our model include pronouns (that, this, its, their, them), discourse connectives (e.g., 'however', 'therefore', 'also') and prepositions (of, for, by).

Table 5.7 reports an example where our model is able to generate the correct discourse connective '*however*' using the context. If we look at the context of the source sentence in French, we come to the conclusion that 'however' is indeed a perfect fit in this case, whereas the base model is at a disadvantage and completely changes the underlying meaning of the sentence by generating the inappropriate connective 'nevertheless'.

Table 5.8 gives an instance where our model is able to generate the correct pronoun '*their*'. It should be noted that in this case, the current source sentence does not contain the antecedent and thus the context-free baseline is unable to generate the appropriate pronoun. On the other hand, our contextual model is able to do so by giving the highest attention weights to sentences containing the antecedent (observed from the attention map in Figure 5.8).[21] Figure 5.8 also shows that for translating majority of the sentences, the

[20] Outliers and tokens with equal counts, for our model and the baseline, were removed.

[21] For this particular conversation, all previous turns were in Estonian.

Context	nous sommes également favorables au principe d'un système de collecte des miles commun pour le parlement européen, pour que celui-ci puisse bénéficier de billets d'avion moins chers, même si nous voyons difficilement comment ce système pourrait être déployé en pratique. enfin, nous ne sommes pas opposés à l'attribution de prix culturels par le parlement européen.
Source *Target*	néanmoins, nous sommes particulièrement critiques à l'égard du prix pour le journalisme du parlement européen et nous ne pensons pas que celui-ci puisse décerner des prix aux journalistes ayant pour mission de soumettre le parlement européen à un regard critique. however, we are highly critical of parliament's prize for journalism, and do not believe that it is appropriate for parliament to award prizes to journalists whose task it is to critically examine the european parliament.
Base Model *Our Model*	nevertheless, we are particularly critical of the price for the european union's european alism and we do not believe that it would be able to make a price to the journalists who have been made available to the european parliament to a critical view. however, we are particularly critical of the price for the european union's democratic alism and we do not believe that it can give rise to the prices for journalists who have been tabled to submit the european parliament to a critical view.

Table 5.7: Example En-Fr sentence translation showing how the context helps our model in generating the appropriate discourse connective.

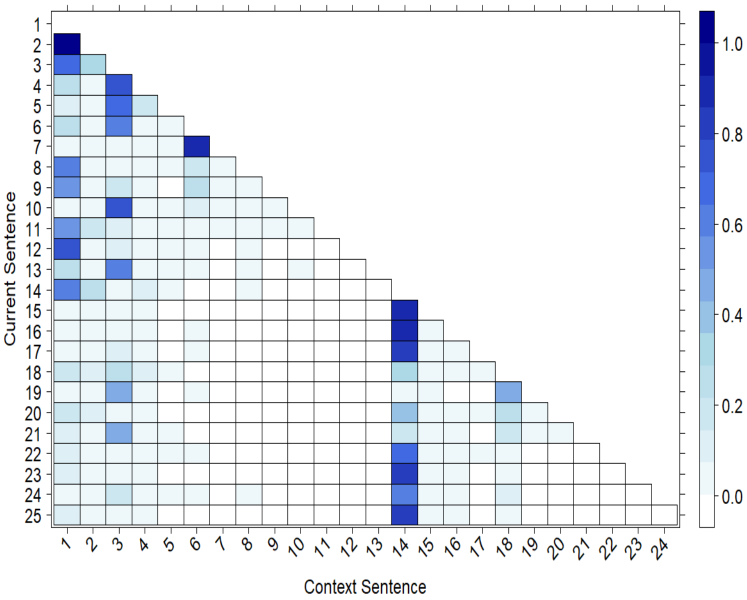


Figure 5.8: Attention map when translating a conversation from the En-Et test set.

Context	*(1)* see raport juhib tähelepanu krediidireitingute sektori äärmiselt olulistele probleemidele, sealhulgas konkurentsi puudumisele, oligopolidele tüüpilistele struktuuridele ning vastutuse ja läbipaistvuse puudumisele , eriti riigi võlakohustuse hindamisel. *(3)* oleks hea, kui reitinguagentuurid vastutaksid tulevikus enda tegevuse eest rohkem. ... *(14)* kirjalikult. - (it) kiites heaks wolf klinzi raporti, mille eesmärk on reitinguagentuuride tõhus reguleerimine, võtab parlament järjekordse sammu finantsturgude suurema läbipaistvuse suunas. ... *(18)* mul oli selle dokumendi üle hea meel, sest krediidireitingute valdkonnal on palju probleeme, millest kõige suuremad on oligopolidele tüpilised struktuurid ning konkurentsi, vastutuse ja läbipaistvuse puudumine.
Source	*(24)* selles suhtes tuleb rõhutada nende tegevuse suuremal äbipaistvuse põhirolli.
Target	in this respect, it is necessary to highlight the central role of increased transparency in their activities.
Base Model	in this regard it must be emphasised in the major role of transparency in which these activities are to be strengthened.
Our Model	in this regard, it must be stressed in the key role of greater transparency in their activities.

Table 5.8: Example English-Estonian translation showing how the wide-range context helps in generating the correct pronoun. The antecedent mentions and correct pronouns are highlighted in blue. The numbers at the start of the Estonian sentences give their position in the conversation.

model attends to wide-range context rather than just the previous sentence, hence strengthening the premise of using the complete context for our conversational Bi-MSMT model.

5.5 Experiments on Real-world Customer Service Data

Dialogue translation is an important task for companies providing multilingual services to customers. I had the opportunity to test my models on the in-house customer service chat data of an airline company provided to me during an internship at Unbabel.[22]

Datasets The datasets consist of conversations between agents and clients for two language-pairs: English-French and English-German. As is clear from the data statistics in Table 5.9, the size of the datasets is even smaller than the data we extracted from public resources and is one of the challenges that we face in the real-world setting.

[22] https://unbabel.com/

	En-Fr	**En-De**
#Conversations	2116/70/80	1717/96/81
#Sentences	46.5K/1.5K/1.7K	38.9K/2.1K/2.1K

Table 5.9: General statistics for training/development/test sets for customer service data.

Models and Training In a commercial setting, there is usually a generic model trained on an extremely large dataset, followed by a step of domain adaptation to adapt the model given the data from a specific domain. Keeping the time constraints in mind, we trained a generic model on the sentence-parallel Europarl v7 corpus (Koehn, 2005) which has approximately 2M sentence-pairs, followed by domain adaptation using the sentence-level chat data.

The target sentence representations are computed in same manner as in the previous set of experiments in Section 5.4. For the source-context representations, we use the sentence representations from the encoder of the NMT model instead of separate RNNLMs. This allows us to re-use those learned parameters thus saving time and resources which are both critical in a commercial setting.

The base model has single layer bi-directional GRUs in both encoders and two-layer GRUs in the decoders while the hidden dimensions and word embedding sizes are set to 256, and the alignment dimension (for the attention mechanism in the decoder) is set to 128 (similar to our base model in the previous section). For the generic model, we again make use of stochastic gradient descent (SGD) with an initial learning rate of 0.1 and a decay factor of 0.5 after the fifth epoch for a total of 15 epochs. For the domain adaptation and contextual model, we use SGD with an initial learning rate of 0.08 and a decay factor of 0.9 and use early stopping. To avoid overfitting, we employ dropout and set its rate to 0.2.

5.5.1 Results

The main results are reported in Table 5.10. In our preliminary experiments, we found the configuration of AddDec to work best for these datasets and thus we report results only for this specific configuration. We only use the *Language-Specific Sentence-level Attention* model as it was previously found to be the best among the different approaches for computing the context.

	En-Fr			**En-De**		
	Overall	En→Fr	Fr→En	Overall	En→De	De→En
Base Model w/ Domain Adaptation	55.6	65.5	36.5	46.7	60.1	30.4
+Source Context	57.6†	68.3†	36.8	51.2†	66.5†	32.6†
+Target Context	57.4†	68.1†	36.6	50.5†	**67.1**†	30.9†
+Dual Context Src-Tgt	**57.7**†	**68.3**†	**37.3**	**51.5**†	66.8†	**32.9**†
+Dual Context Src-Tgt-Mix	57.2†	67.6†	36.9	50.4†	66.3†	31.4†
+*Source Context from*						
Current Turn	57.6†	68.3†	**37.1**	50.9†	66.5†	32.0†
Current Language from Previous Turns	57.5†	68.2†	37.0	**51.4**†	66.1†	**32.9**†
Other Language from Previous Turns	**57.8**†	**68.6**†	37.0	50.2†	66.0†	32.9†
Complete Context	57.6†	68.3†	36.8	51.2†	**66.5**†	32.6†

Table 5.10: BLEU scores for the bilingual test sets. Here all contexts are incorporated as AddDec. **bold**: Best performance w.r.t. two decimal places, †: Statistically significantly better than the base model, based on bootstrap resampling (Clark et al., 2011) with $p < 0.05$.

For these experiments, we are more interested to see how the different types of context, that is source, target and dual, compare to each other. Across the board, we see the Dual Context Src-Tgt to outperform its counterpart and the models which use only a single type of context, reiterating the importance of using both source and target-side conversation history for this task. Our second set of experiments is similar to the ablation study we conducted in Section 5.4. For English-French, the overall and direction-specific BLEU scores are quite similar for the different conversational contexts. For English-German, however, we find the context from the same language (current and previous turn) to be the main contributing factors for improving the BLEU scores.

To summarise, we experimented with in-house customer service chat data for two language-pairs and found that translating multi-speaker conversations is an interesting avenue of research to pursue and that it greatly benefits from using source and target-side conversational histories.

5.6 Related Work

Our research builds upon prior work in the field of context-based language modelling and context-based machine translation.

Language Modelling There have been few works on leveraging context information for language modelling. Ji et al. (2016) introduced a document context language model (DCLM) which incorporates inter and intra-sentential contexts. Hoang et al. (2016) made use of side-information, for instance, metadata, and Tran et al. (2016) utilised inter-document context to boost the performance of RNN language models.

For conversational language modelling, Ji and Bilmes (2004) proposed a statistical multi-speaker language model (MSLM) that considers words from other speakers when predicting words from the current one. By taking the inter-speaker dependency into account using a normal trigram context, they reported significant reduction in perplexity.

Machine Translation For related work on context-based MT, the reader is encouraged to look at Section 3.4, where this topic has been covered in depth. To the best of our knowledge, there has been no work on dialogue translation or its variation to date.

5.7 Summary

In this chapter, we have investigated the challenges associated with translating multilingual multi-speaker conversations by exploring a simpler task referred to as bilingual multi-speaker conversation MT. We processed Europarl v7 and OpenSubtitles2016 to obtain an introductory dataset for this task. Compared to models developed for similar tasks, our work is different in two aspects: (i) the history captured by our model contains multiple languages, and (ii) our model captures 'global' history as opposed to 'local' history captured in most previous works. Our experiments on both public and real-world customer service chat data demonstrate the significance of leveraging the bilingual conversation history in such scenarios, in terms of BLEU and manual evaluation. Furthermore, the analysis shows that using wide-range context, our model generates appropriate pronouns and discourse connectives in some cases. We hope this work to be a first step towards translating multilingual multi-speaker conversations. A natural extension of this work is employing our hierarchical attention model, introduced in the previous chapter (chronologically performed after this work), to the turns/sentences in the conversation history. We leave this for future exploration.

Chapter 6

Conclusions

6.1 Summary of the Thesis

The primary contribution of this thesis is using global context information to build efficient neural models for document-level machine translation.

Chapter 3 We presented the first work which views document-level translation as a structured prediction problem with interdependencies among the observed and hidden variables, i.e., the source sentences and their unobserved target translations in the document. The resulting structured prediction problem was tackled with a neural translation model equipped with two memory components, one each for the source and target-side, to capture the documental interdependencies. We trained the model end-to-end using a pseudo-likelihood based training objective and proposed an iterative decoding algorithm based on block coordinate descent. We demonstrate improvements in the translation quality of three language-pairs with respect to context-free and local context-aware baselines.

Chapter 4 We proposed a novel approach based on hierarchical attention for document-level NMT using sparse attention, which is both scalable and efficient. Experiments and evaluation on three English→German datasets in offline and online document MT settings show that our approach surpasses context-agnostic and two recent context-aware baselines. The qualitative analysis indicates that the sparsity at sentence-level allows our model to identify key sentences in the document context and the sparsity at word-level allows it to focus on key words in those sentences allowing for efficient compression of memory. Using sparse attention may lead to better interpretability of the context-aware NMT models in general.

Chapter 5 We looked into the problem of dialogue translation by investigating the challenges associated with translating bilingual multi-speaker conversations. To initiate an evaluation for this task, we introduced datasets extracted from Europarl v7 and OpenSubtitles2016. The history captured by our models contains multiple languages and is global. Our experiments demonstrate the significance of leveraging the bilingual conversation history in such scenarios. Furthermore, the analysis reveals that, using wide-range context, our model can generate appropriate pronouns and discourse connectives in some cases.

6.2 Future Directions

This section briefly mentions a few of the possible research directions and insights gained from this thesis.

Document-aligned Datasets While there are many popular datasets for MT, all of them consist of aligned sentence-pairs without any metadata. Hence, the first problem that we and other researchers working on the problem of document-level machine translation encounter is to curate datasets for this purpose. Furthermore, it is not necessary that the discourse phenomena we aim to observe actually exist in the current public datasets. This problem further exacerbates when one tries to translate dialogues since datasets, like subtitles, lack speaker annotations. It is high time that the MT community starts investing their efforts in creating such resources so that the research process can be standardised with respect to the datasets used.

Explicit Linguistic Annotation We mentioned in Chapter 1 that using linguistic annotation is outside the scope of this thesis. However, if this process could be automated and we could obtain annotations of, for instance, entities in the discourse, it could directly impact the translation of their mentions thus improving lexical cohesion. The translation could also be conditioned on the evolution of entities as they are introduced in the source and target text (Ji et al., 2017). We believe annotation of discourse phenomena, for example, coreference or discourse markers, could be beneficial by having better quality translation outputs more faithful to the source text.

Document-level MT Evaluation From Chapter 2, it is evident that there is no consensus among the MT community when it comes to the evaluation of document-level MT. Reference-based automatic evaluation metrics, like BLEU and METEOR, which look at the overlap of MT output with a reference, are insensitive to the underlying discourse structure of the text (Läubli et al., 2018). These are still being used to evaluate MT outputs as they have been the de-facto standard in the community for more than a decade. The proposed document-level automatic metrics (detailed in Section 2.2.3) have their own flaws and are not widely accepted. A middle ground should be found between automatic and manual evaluation for MT that could make the process of manual evaluation cheaper and would still be better than the current automatic metrics at evaluating discourse phenomena. Evaluation test sets only resolve a part of the problem as they are mostly hand-engineered for specific language-pairs. Comparison to a single reference translation is also not a good way to evaluate translation output as it has its own shortcomings. To actually progress in document-level MT, we not only need models that address it but also evaluation schemes that have the ability to correctly gauge their performance.

To conclude, the contributions made in this thesis have tried to eliminate the sentence-independence assumption made by the state-of-the-art MT systems. The previous works in document-level NMT rely on context information in only a few local sentences, where the deciding factor for choosing the number of sentences is usually the model's BLEU score on a validation set. This assumption makes the model subjective to the dataset and lacking in generalisability. Hence, these works do not conform to our definition of incorporating global document context. Furthermore, this thesis also shows that conditioning on the global context is not computationally expensive in comparison to local-context models provided the training and decoding algorithms are efficient. We hope this work invigorates research in this domain with a greater inclination towards designing better training and decoding schemes, in addition to modelling, for monologue and dialogue translation.

Appendix A

Estonian→English Test Document

Source Document	**Target Document**
qimonda on praeguse ülemaailmse finants- ja majanduskriisi kontekstis paradigmaatiline juhtum.	qimonda is a paradigm case in the current context of global financial and economic crisis.
see on ettevõte, mis kasutab tipptehnoloogiat, on palganud kõrge kvalifikatsiooniga töötajaid ning edendab teadusuuringuid.	it is a company that uses cutting-edge technology, employs highly qualified workers and promotes research.
qimonda täidab lissaboni strateegia eesmärke.	qimonda meets the objectives of the lisbon strategy.
portugali valitsus on teinud kõik võimaliku, et leida lahendus asjaomase ettevõtte elujõulisena hoidmiseks, kuid lahendus sõltub saksamaa föderaalvalitsuse ning baierimaa ja saksimaa riiklike valitsuste sekkumisest.	the portuguese government has been doing everything to find a solution that makes this company viable, but the solution is also dependent on the involvement of the german federal government and the state governments of bavaria and saxony.
portugali valitsus on juba otsustanud eraldada sellel eesmärgil 100 miljonit eurot.	the portuguese government has already decided to make eur 100 million available for this purpose.
nagu ma juba ütlesin, on portugali valitsus juba praeguseks teinud ning kavatseb ka edaspidi teha kõik endast oleneva, nagu võisid tegelikult ka qimonda saksamaa töötajad portugali presidendi hiljutise ametliku visiidi ajal tõdeda.	as i said, it has been doing and will continue to do everything it can, as was, in fact, recognised by qimonda's german workers during the recent official visit by the president of the portuguese republic.
euroopa komisjon ja liikmesriigid on võtnud õigustatult kasutusele meetmeid mitmete pankade päästmiseks ja teatud tööstusharude, nagu näiteks autotööstuse toetamiseks.	the european commission and the member states have been taking steps - and rightly so - to save many banks and to support certain industries such as, for example, the automotive industry.
miks siis mitte toetada ka qimondat?	why not also support qimonda?
qimonda saatuse hooleks jätmisel saavad olema väga tõsised tagajärjed.	leaving qimonda to its fate will have extremely serious consequences.
lisaks sellele, et saksamaal ja portugalis kaotavad tuhanded töötajad oma töö, läheb kaduma ka euroopa intellektuaalomand ning palju qimondasse investeeritud ühenduse rahalisi ressursse.	not only will thousands of workers in germany and portugal lose their jobs, but invaluable european intellectual property and a lot of community funds that were invested in qimonda will also be lost.
qimonda tegevuses hoidmine saksamaal ja portugalis on euroopa jaoks nii suure strateegilise tähtsusega, et euroopa liidu toetus oleks igati õigustatud.	keeping qimonda going in germany and in portugal is of such strategic importance for europe that european union support is well justified.
volinik, me peame olema sihikindlad, sest nii teeme me kõik võimaliku, et qimondat päästa.	commissioner, we must be consistent and, if we are to be consistent, we will do everything to save qimonda.
qimonda ei ole lihtsalt mingi tavaline ettevõte!	qimonda is not just any company!

Bibliography

Ba, J., Kiros, R., and Hinton, G. E. (2016). Layer normalization. *CoRR*, abs/1607.06450.

Bahar, P., Alkhouli, T., Peter, J.-T., Brix, C. J.-S., and Ney, H. (2017). Empirical investigation of optimization algorithms in neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):13 – 25.

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

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, USA. Association for Computational Linguistics.

Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers*, 1:91–163.

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, USA. Association for Computational Linguistics.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195.

Bosselut, A., Celikyilmaz, A., He, X., Gao, J., Huang, P.-S., and Choi, Y. (2018). Discourse-aware neural rewards for coherent text generation. 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 173–184, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Britz, D., Goldie, A., Luong, M.-T., and Le, Q. (2017). Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451, Copenhagen, Denmark. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Cartoni, B., Zufferey, S., and Meyer, T. (2013). Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue & Discourse*, 4:65–86.

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

Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Schuster, M., Shazeer, N., Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Chen, Z., Wu, Y., and Hughes, M. (2018a). The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia. Association for Computational Linguistics.

Chen, Y., Li, V. O., Cho, K., and Bowman, S. (2018b). A stable and effective learning strategy for trainable greedy decoding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 380–390, Brussels, Belgium. Association for Computational Linguistics.

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

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

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 176–181, Portland, Oregon, USA. Association for Computational Linguistics.

Cohn, T., Hoang, C. D. V., Vymolova, E., Yao, K., Dyer, C., and Haffari, G. (2016). Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California, USA. Association for Computational Linguistics.

Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, 3:1218–1223.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, L. (2019). Universal transformers. In *International Conference on Learning Representations*.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Fellbaum, C., editor (1998). *WordNet: An electronic lexical database*. MIT Press.

Foster, G., Isabelle, P., and Kuhn, R. (2010). Translating structured documents. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, USA.

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

Garcia, E. M., España-Bonet, C., and Màrquez, L. (2014). Document-level machine translation as a re-translation process. *Procesamiento del Lenguaje Natural*, 53:103–110.

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

Gehring, J., Auli, M., Grangier, D., and Dauphin, Y. (2017a). A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

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

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

Gong, Z., Zhang, M., and Zhou, G. (2015). Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40, Lisbon, Portugal. Association for Computational Linguistics.

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1764–1772, Bejing, China. PMLR.

Guillou, L. and Hardmeier, C. (2016). PROTEST: A test suite for evaluating pronouns in machine translation. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 636–643, Portorož, Slovenia. European Language Resources Association.

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.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2014). Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hajlaoui, N. and Popescu-Belis, A. (2013). Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, pages 236–247, Samos, Greece. Springer Berlin Heidelberg.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

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

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

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

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

Hardmeier, C., Tiedemann, J., and Nivre, J. (2013b). Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, Washington, USA. Association for Computational Linguistics.

Hatim, B. and Mason, I. (1990). *Discourse and the Translator*. Longman, London.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, Nevada, USA.

Hoang, C. D. V., Cohn, T., and Haffari, G. (2016). Incorporating side information into recurrent neural network language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1255, San Diego, California, USA. Association for Computational Linguistics.

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

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017). Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

Ji, G. and Bilmes, J. (2004). Multi-speaker language modeling. In *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Short Papers*, pages 133–136, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ji, Y., Cohn, T., Kong, L., Dyer, C., and Eisenstein, J. (2016). Document context language models. In *Workshop track - International Conference on Learning Representations*.

Ji, Y., Tan, C., Martschat, S., Choi, Y., and Smith, N. A. (2017). Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839, Copenhagen, Denmark. Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2nd edition.

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 Ninth International Joint Conference on Natural Language Processing*, pages 2957–2966, Hong Kong, China. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations*.

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 the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. Asia-Pacific Association for Machine Translation.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54, Edmonton, Canada. 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.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc.

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.

Lavie, A. and Agarwal, A. (2007). METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Le Nagard, R. and Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden. Association for Computational Linguistics.

LeCun, Y. (1988). A theoretical framework for back-propagation. In Touretzky, D., Hinton, G., and Sejnowski, T., editors, *Proceedings of the 1988 Connectionist Models Summer School*, pages 21–28, Pittsburg, Pennsylvania, USA. Morgan Kaufmann.

Lewis-Kraus, G. (2016). The great A.I. awakening. *The New York Times Magazine*.

Li, J. J., Carpuat, M., and Nenkova, A. (2014a). 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, USA. Association for Computational Linguistics.

Li, J. J., Carpuat, M., and Nenkova, A. (2014b). Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 577–587, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Li, L., Liu, Y., and Zhou, A. (2018). Hierarchical attention based position-aware network for aspect-level sentiment analysis. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 181–189, Brussels, Belgium. Association for Computational Linguistics.

Lison, P. and Meena, R. (2016). Automatic turn segmentation of Movie & TV subtitles. In *Proceedings of the 2016 Spoken Language Technology Workshop*, pages 245–252, San Diego, California, USA.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from Movie and TV subtitles. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 923–929, Portorož, Slovenia. European Language Resources Association.

Louis, A. and Webber, B. (2014). Structured and unstructured cache models for SMT domain adaptation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 155–163, Gothenburg, Sweden. Association for Computational Linguistics.

Luong, N.-Q. and Popescu-Belis, A. (2016). A contextual language model to improve machine translation of pronouns by re-ranking translation hypotheses. In *Proceedings of the 19th Conference of the European Association for Machine Translation*, pages 292–304, Riga, Latvia.

Luong, N.-Q. and Popescu-Belis, A. (2017). Machine translation of Spanish personal and possessive pronouns using anaphora probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 631–636, Valencia, Spain. Association for Computational Linguistics.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Malaviya, C., Ferreira, P., and Martins, A. F. T. (2018). Sparse and constrained attention for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–376, Melbourne, Australia. Association for Computational Linguistics.

Marcu, D., Carlson, L., and Watanabe, M. (2000). The automatic translation of discourse structures. In *Proceedings of the First North American Chapter of the Association for Computational Linguistics Conference*, pages 9–17, Seattle, Washington, USA. Association for Computational Linguistics.

Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 133–139. Association for Computational Linguistics.

Martins, A. F. T. and Astudillo, R. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1614–1623, New York, New York, USA. PMLR.

Maruf, S. and Haffari, G. (2018). Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Maruf, S. and Haffari, G. (2019). Monash University's submissions to the WNGT 2019 document translation task. In *Proceedings of the Third Workshop on Neural Generation and Translation*, pages 256–261, 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, USA. Association for Computational Linguistics.

Mascarell, L. (2017). Lexical chains meet word embeddings in document-level statistical machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 99–109, Copenhagen, Denmark. Association for Computational Linguistics.

Metz, C. (2016). An infusion of AI makes google translate more powerful than ever. *Wired*.

Meyer, T. and Popescu-Belis, A. (2012). Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138, Avignon, France. Association for Computational Linguistics.

Meyer, T., Popescu-Belis, A., Hajlaoui, N., and Gesmundo, A. (2012). Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, San Diego, California, USA.

Meyer, T., Popescu-Belis, A., Zufferey, S., and Cartoni, B. (2011). Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Proceedings of the SIGDIAL 2011 Conference*, pages 194–203, Portland, Oregon, USA. Association for Computational Linguistics.

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

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

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

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Kobayashi, T., Hirose, K., and Nakamura, S., editors, *INTERSPEECH*, pages 1045–1048, Makuhari, Japan. International Speech Communication Association.

Mikolov, T., Kombrink, S., Burget, L., Cernocký, J., and Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5528–5531, Prague, Czech Republic.

Mikolov, T. and Zweig, G. (2012). Context dependent recurrent neural network language model. In *Proceedings of the 2012 IEEE Spoken Language Technology Workshop*, pages 234–239.

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.

Nallapati, R., Zhou, B., dos Santos, C. N., Gülçehre, Ç., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Neubig, G., Dyer, C., Goldberg, Y., Matthews, A., Ammar, W., Anastasopoulos, A., Ballesteros, M., Chiang, D., Clothiaux, D., Cohn, T., Duh, K., Faruqui, M., Gan, C., Garrette, D., Ji, Y., Kong, L., Kuncoro, A., Kumar, G., Malaviya, C., Michel, P., Oda, Y., Richardson, M., Saphra, N., Swayamdipta, S., and Yin, P. (2017). DyNet: The dynamic neural network toolkit. *CoRR*, abs/1701.03980.

Niculae, V. and Blondel, M. (2017). A regularized framework for sparse and structured neural attention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 3338–3348. Curran Associates, Inc.

Novák, M. and Žabokrtský, Z. (2014). Cross-lingual coreference resolution of pronouns. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24, Dublin, Ireland. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

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 on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012). Understanding the exploding gradient problem. *CoRR*, abs/1211.5063.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pages 318–362. MIT Press, Cambridge, Massachusetts, USA.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Sennrich, R. (2018). Why the time is ripe for discourse in machine translation? Presented at the Second Workshop on Neural Machine Translation and Generation.

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.

Sim Smith, K. (2017). On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark. Association for Computational Linguistics.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Smith, K. S. (2018). *Coherence in Machine Translation*. PhD thesis, University of Sheffield.

Smith, K. S., Aziz, W., and Specia, L. (2016). The trouble with machine translation coherence. In *Proceedings of the 19th Conference of the European Association for Machine Translation*, pages 178–189, Riga, Latvia.

Smith, K. S. and Specia, L. (2017). Examining lexical coherence in a multilingual setting. In Menzel, K., Lapshinova-Koltunski, E., and Kunz, K., editors, *New perspectives on cohesion and coherence*, pages 131–150. Language Science Press, Berlin.

Smith, K. S. and Specia, L. (2018). Assessing crosslingual discourse relations in machine translation. *CoRR*, abs/1810.03148.

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 Seventh Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Steele, D. and Specia, L. (2016). Predicting and using implicit discourse elements in Chinese-English translation. In *Proceedings of the 19th Conference of the European Association for Machine Translation*, pages 305–317, Riga, Latvia.

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, Belgium, Brussels. Association for Computational Linguistics.

Stymne, S., Tiedemann, J., Hardmeier, C., and Nivre, J. (2013). Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pages 375–386, Oslo, Norway. Linköping University Electronic Press, Sweden.

Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. (2015). End-to-end memory networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, Las Vegas, Nevada, USA.

Tiedemann, J. (2010). Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Calzolari, N., Choukri, K., Declerck, T., Doan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Tran, Q. H., Zuckerman, I., and Haffari, G. (2016). Inter-document contextual language model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–766, San Diego, California. 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.

van der Wees, M., Bisazza, A., and Monz, C. (2016). Measuring the effect of conversational aspects on machine translation quality. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2571–2581, Osaka, Japan. The COLING 2016 Organizing Committee.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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, L., Zhang, X., Tu, Z., Way, A., and Liu, Q. (2016). Automatic construction of discourse corpora for dialogue translation. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia. European Language Resources Association.

Weston, J., Chopra, S., and Bordes, A. (2015). Memory networks. In *Proceedings of the Third International Conference on Learning Representations*.

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.

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., ukasz Kaiser, 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. *CoRR*, abs/1609.08144.

Xia, Y., Tan, X., Tian, F., Qin, T., Yu, N., and Liu, T.-Y. (2018). Model-level dual learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5383–5392, Stockholm, Sweden. PMLR.

Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). Document-level consistency verification in machine translation. In *Proceedings of the 13th Machine Translation Summit*, pages 131–138, Xiamen, China. Asia-Pacific Association for Machine Translation.

Xiong, D., Ben, G., Zhang, M., Lü, Y., and Liu, Q. (2013a). Modeling lexical cohesion for document-level machine translation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2183–2189, Beijing, China.

Xiong, D., Ding, Y., Zhang, M., and Tan, C. L. (2013b). Lexical chain based cohesion models for document-level statistical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1563–1573, Seattle, Washington, USA. Association for Computational Linguistics.

Xiong, H., He, Z., Wu, H., and Wang, H. (2019). Modeling coherence for discourse neural machine translation. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, USA. Association for the Advancement of Artificial Intelligence.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, USA. Association for Computational Linguistics.

Yung, F., Duh, K., and Matsumoto, Y. (2015). Crosslingual annotation and analysis of implicit discourse connectives for machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 142–152, Lisbon, Portugal. Association for Computational Linguistics.

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.

Zhu, H., Wei, F., Qin, B., and Liu, T. (2018). Hierarchical attention flow for multiple-choice reading comprehension. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 6077–6085, New Orleans, Louisiana, USA. Association for the Advancement of Artificial Intelligence.