A Critical Examination of Document-Level Machine Translation Systems

Prashanth Nayak

B.E., M.Tech

A dissertation submitted in fulfillment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Dublin City University
School of Computing

Supervisors:
Prof. Andy Way, Dublin City University
Dr. Rejwanul Haque, South East Technological University
Prof. John D. Kelleher, Maynooth University

January, 2024

I hereby certify that this material, which I now submit for assessment on the program of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID No.: 19213962
Date: 04/01/2024

ಪ್ರೀತಿಯ ಅಮ್ಮ, ಅಪ್ಪ ಮತ್ತು ತಂಗಿಗೆ

Contents

List of Tables

List of Figures

A Critical Examination of Document-Level Machine Translation Systems

Prashanth Nayak

Abstract

The need for accurate and effective translation cannot be overstated in an increasingly globalised world where communication is paramount. Bridging language barriers is important for promoting understanding and cooperation among diverse individuals and communities, making translation an indispensable tool. Over the past two decades, Machine Translation (MT) has undergone remarkable advancements, with significant progress attributed to the emergence of Neural Machine Translation (NMT), primarily the groundbreaking Transformer models. This rapid development in MT, which started with a focus on sentence-level translation, has not only bridged communication gaps but also brought MT systems close to delivering human-like performance on various translation tasks. While these advances are significant, focusing mainly on sentence-level modeling and evaluation, they miss the valuable contextual information around each sentence. Contextual information in document-level translation helps resolve language ambiguities and ensures consistency and coherence in the translated text, making the translation more accurate and readable. While considerable efforts have been made to incorporate context into NMT systems, the community has not reached a consensus on the most effective methods and the types of context to be integrated. In this thesis, my primary focus is on understanding document-level systems. Specifically, I explore how these systems incorporate context into their translation processes and investigate the span of context utilised by these systems. I also investigate the terminology translation mechanisms within these systems. Furthermore, with the emergence of modern-day powerful Large Language Models (LLMs), I examine their capabilities in terminology translation and propose new methodologies to improve terminology translation for these powerful models.

Acknowledgments

As I set off on this academic journey in a foreign land on September 14, 2019, I did not know what lay ahead. But I knew that the best was about to happen. These four years have shaped me academically and personally. I have missed home and family, met incredible people, made forever friendships, learned about life, and most of all, I was on a path to an incredible opportunity. As I reach the culmination of this amazing journey, there is no better time than now for me to think of and thank each one of those who supported me and contributed to my success and well-being during this endeavour.

First and foremost, I bow to my parents, Smt. Manorama and Sri Pandauranga, for their desire to see me in a good place and give me the best life they could. No amount of words can convey how deeply indebted I am to them for all their sacrifices, endless patience, and unconditional love. It was their motivation that raised me up again when I became weary. I am where they wanted me to be in life. This moment is theirs. Then, to my little sister, Shilpa, for holding the fort back home. Thank you for reminding me that my only job these years was to study, stay safe, and be happy.

As I approach this milestone in my academic pursuit, I would like to express my sincere gratitude to Prof. Andy Way, my primary supervisor, for his professional guidance and motivation throughout my research, enabling me to graduate successfully. His unwavering trust in me has been a significant factor in my achievement, and I aim to apply the knowledge gained from him in my future research endeavours. I would also like to extend my deepest appreciation to Dr. Rejwanul Haque, whose mentorship has been a key factor in my success. His invaluable and timely feedback on my work has enabled me to meet my project timelines. I am very grateful to Prof. John D. Kelleher for his thorough supervision and the numerous insightful discussions about the research. I would also like to recognise the invaluable support extended by Project Managers, Angela and Stephen.

I am immensely grateful to Rohit, a lifelong friend and a mentor, for standing by me through tough times. A big thank you to Bharat, my fellow Ph.D. student and housemate at DCU, for the laughs and the support we shared. I appreciate my DCU seniors, Ghanshyam, Joseph, and Pintu, for always encouraging and guiding me. A special shoutout goes to Ashok Kamath's family for their kindness and hospitality in Ireland, which made me feel welcome. Thanks to my dear friend Sam, whose friendship I will treasure for years. Lastly, the camaraderie and support from Nuki and Yasmin, my friends and Ph.D colleagues, and everyone in ADAPT have been a big part of my life and this journey.

DCU has been a wonderful place for me, and I will remain forever grateful to this outstanding institution for the opportunities it has provided. The financial support from Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, along with Microsoft Research Ireland was essential to my research success.

Chapter 1

Introduction

Language is essential for communication, which the internet has made effortless, opening doors to travel, work, study, and exploration of international content. Despite these advances, a challenge persists due to the many languages spoken worldwide. While this linguistic diversity is a huge asset that needs to be maintained and supported, it can lead to misunderstandings (Rehm and Way, 2023). Historically, human translators have bridged these gaps, especially during critical events like international business meetings or governmental talks.

In recent times, there has been an increasing demand for commercial translation services like those offered by Bing,[1] Google,[2] and DeepL.[3] For instance, Google Translate processes billions of words daily,[4] highlighting the growing need for translation services. Real-time translation is popular due to the growing demand for instant, cross-language communication in an increasingly globalised world. Its convenience, cost-effectiveness, and accessibility are driven by technological advancements, and with limited professional translators available, MT has become indispensable for businesses, travellers, and educators. This technology can provide quick and reliable translations from one natural language to another natural language, efficiently narrowing the communication gap. In other words, MT is not just

[1] https://www.bing.com/translator

[2] https://translate.google.com/

[3] https://www.deepl.com/en/translator

[4] https://blog.google/products/translate/ten-years-of-google-translate/

a convenience but a necessity for clear and effective communication in today's interconnected global society. It serves as a crucial tool for individuals and businesses, simplifying the understanding of various international languages. People from different linguistic backgrounds can better engage with and understand each other via MT, promoting a more inclusive and connected global community.

MT is a branch of Natural Language Processing (NLP) that utilises computer software to translate text from one language to another. Over time, MT has evolved significantly, giving rise to different methodologies. Initially, we had Rule-Based Machine Translation (RBMT) (Hutchins, 1986), which relied on manually crafted linguistic rules and dictionaries for translation. Then came Statistical MT (SMT) (Brown et al., 1993; Koehn et al., 2003), which used statistical models to translate text by analysing large volumes of bilingual/multilingual text corpora. The latest evolution in this field is NMT (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015), which utilises deep learning techniques to generate more fluent and contextually accurate translations. Each of the discussed methodologies represents significant advancements that led to an increase in the quality of automated translation.

Despite all the progress in NMT, most systems still operate primarily at the sentence level, focusing on translating individual sentences without considering the broader context of the entire document. Even models utilising state-of-the-art Transformer (Vaswani et al., 2017) architectures often neglect the context embedded within larger text documents. This approach often neglects essential aspects such as document-level coherence, context, and cross-sentence dependencies, which can significantly impact the overall quality and understanding of the translation. However, it is essential to note that efforts have been made to develop document-level MT systems (Wang et al., 2017; Maruf and Haffari, 2018; Miculicich et al., 2018; Voita et al., 2018; Zhang et al., 2022; Sun et al., 2022; Bao et al., 2023; Herold and Ney, 2023b,a; Zhang et al., 2023). While these existing document systems are still in their infancy, they represent a promising direction for the future of document-level

MT.

Based on this background, this thesis aims to contribute to a better understanding of document-level MT systems. I will focus on understanding their current methodologies, identifying their limitations, and exploring potential improvements. This detailed study aims to contribute to developing and refining more efficient and contextually accurate document-level MT systems.

1.1 Document-level translation

NMT systems primarily work at the sentence level, i.e., they do not consider document context while translating. Document-level MT systems overcome this limitation of sentence-level NMT by incorporating context from the document, improving the quality of translation. It also helps to resolve contextual ambiguities that depend on a context broader than a single sentence. It ensures cohesion across the document, manages inter-sentential relations for coherence, and adapts translations to fit the cultural context.

Source	He arrived late and found the bank closed, which complicated his plans for the evening.
Reference	वह देर से पहुंचा और पाया कि बैंक बंद है, जिससे शाम की उसकी योजनाएँ जटिल हो गईं। [*Vah der se pahuanchā aur pāyā ik baianka banda hai, jasase shām ka usaka yojanāe jiṭal ho gaī*]
NMT	वह देर से पहुंचा और पाया कि किनारा बंद है, जिससे शाम की उसकी योजनाएँ जटिल हो गईं। [*Vah der se pahuanchā aur pāyā ki baianka banda hai, jisase shām kī usakī yojanāe jaṭil ho gaīan*]

Table 1.1: An example showing the importance of context in translation.

For example, the English word **bank** can be translated in Hindi either to बैंक (Financial institution) or to किनारा (River bank) depending upon the context of the translation. I show a translation by my English-to-Hindi MT system in Table 1.1. We see from Table 1.1 that in the absence of contextual clues, the English term "bank" is incorrectly translated to Hindi by my NMT system as described in Section 3.5.

The document-level NMT systems described in the literature can be grouped into the following categories:

- **Window-based Models**: Window-based models in document-level MT offer a balanced approach when it comes to incorporating context into the translation process. Unlike sentence-level models, which translate sentences in isolation, window-based models consider a set of adjacent sentences, referred to as a “window”, when translating a given text. This window provides crucial contextual clues, helping resolve linguistic ambiguities and select appropriate terminology. For example, a term in the source language might have multiple potential translations in the target language. The surrounding sentences within the window help choose the translation that best fits the context. The window size is adjustable, with more bigger windows offering more context but requiring more computational resources. The critical advantage of window-based models is their ability to provide more coherent and contextually accurate translations than those by sentence-level NMT models.[5]

- **Cache-based Models**: Cache-based models in document-level MT utilise a specialised memory mechanism known as a cache to store and recall recently translated sentences. This approach improves the consistency and coherence of the translation output. When a new sentence is being translated, the model first refers to the cache to identify if similar or identical phrases have been translated before in the same document. If matches are found, the model uses these cached translations as a basis or reference, ensuring that terms and expressions are translated consistently throughout the text. This becomes important when translating documents with repeated technical terms, names, or specialised vocabulary, where consistency in translation is essential. By learning and adapting to the context within a document, cache-based models offer a dynamic and responsive translation process, effectively handling the

[5] Some commentators may wonder if global models can be considered a special case of window-based models with window size ∞. In my opinion, this premise is faulty, as window base models are designed for smaller window sizes, unlike global models.

challenges of long texts. The cache mechanism helps minimise errors, reduces redundancy in translation efforts, and significantly improves the quality and reliability of translated documents.

- **Global Models**: Global models in document-level MT offer an integrated approach to translation, where the entire document's context is considered to produce coherent and contextually appropriate translations. These models analyse the entire document, recognising themes, repeated phrases, and overall tone to create translations that are accurate on a sentence level and consistent and coherent across the entire document. For instance, if a specific term is used in a particular way in one section of the document, global models ensure that the term is translated consistently in subsequent sections, maintaining clarity and avoiding confusion for readers. This holistic approach works well when translating complex or lengthy documents, such as legal contracts, technical manuals, or literary works, where understanding the document's overall structure and content is crucial for producing high-quality translations. However, it is essential to note that while global models are powerful, they often require more computational resources and sophisticated algorithms to process and analyse entire documents effectively.

1.1.1 Objectives of document-level translation

- **Contextual Understanding**: It uses the document's context to translate accurately, understanding the relationship between sentences and paragraphs to keep the original meaning.

- **Consistency**: It ensures the exact words and style are used throughout the document for a translation that makes sense from start to finish.

- **Coherence**: It helps maintain the logical flow and structure of the original document, resulting in a translation that reads naturally and makes sense to readers.

- **Handling Ambiguities**: It effectively deals with words or phrases with multiple meanings by using the surrounding text to inform the correct interpretation.

- **Preserving Tone and Style**: It is necessary to retain the tone and style of the original document, whether it is formal, informal, persuasive, informative, or any other style.

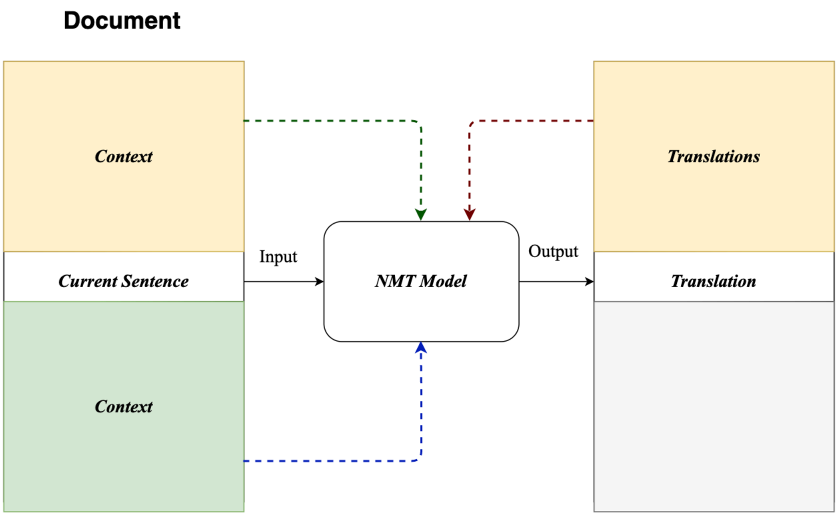


Figure 1.1: Context usage in Document-level NMT.

1.1.2 Why Investigating Document-Level MT is Necessary

In recent times, a variety of document-level MT architectures have emerged. These new models primarily focus on integrating diverse contexts, such as local, global, limited, and contexts from the source and/or target languages. Figure 1.1 visually represents this, highlighting how context can be drawn from preceding and succeeding text on the source side or from the target side in document-level NMT.

A big challenge in document-level MT is how to include context. Different MT systems use different methods to use information from the document. Some focus on the context immediately preceding (Jean et al., 2017; Voita et al., 2018; Jiang

Modelling Global Document Context							
Context Type			**Approach**		**Lang. Pair**	**Targeted Evaluation**	**Reference**
Past	**Future**	**Amount**	**Context Encoding**	**Integration in NMT**			
s	-	$3^{\star\star}$	encoder	encoder	En → De	-	Chen et al. (2020b)
s	s		augmented input		De → En/Fr	WSD	Rios Gonzales et al. (2017)
		$\star\star$			En ↔ Fr, En → De	-	Macé and Servan (2019)
		all	encoder	encoder, decoder	Zh/De → En	Pronouns	Tan et al. (2019)
s, t	-	all	encoder w/attention	decoder	Fr/De/Et/Ru ↔ En	-	Maruf et al. (2018)
s, t	s, t	all	attention	encoder, decoder	En → De	Pronouns	Maruf et al. (2019)
			encoder w/attention	decoder, output	Fr/De/Et → En	-	Maruf and Haffari (2018)

Table 1.2: Overview of studies focusing on modeling global document-level context for improved NMT performance. Adapted from Table 2 in Maruf et al. (2021).

Modelling Local Document Context							
Context Type			Approach		Lang. Pair	Targeted Evaluation	Reference
past	future	amount	context encoding	integration in NMT			
s	-	1	encoder w/attention	encoder	En → Ru	Anaphora	Voita et al. (2018)
					Zh/Fr → En En → De/Ru	-	Li et al. (2020)
					Zh/Es → En	-	Jiang et al. (2019)
				decoder	En → Fr/De	Pronouns	Jean et al. (2017)
					Zh ↔ En	Coherence	Kuang and Xiong (2018)
		2		encoder	En → De/Tk/Ko	Pronouns	Yun et al. (2020)
				encoder, decoder	Zh/Fr → En	-	Zhang et al. (2018)
					Fr → En	-	Wang et al. (2019)
		3	capsule network [99]	encoder	En → De	-	Yang et al. (2019)
			encoder	decoder	Zh → En	-	Wang et al. (2017)
s	s	1	concatenated inputs, additional embeddings		En → De	-	Ma et al. (2020)
		2	attention	encoder	De/Pt ↔ En	Pronouns/Cataphora	Wong et al. (2020)
s, t	-	1	concatenated inputs		En → De	-	Tiedemann and Scherrer (2017)
			encoder w/attention	source context vector	En → Fr	Anaphora Cohesion Coherence	Bawden et al. (2018)
					De/Zh/Ja ↔ En	-	Yamagishi and Komachi (2020)
		3	attention	encoder, decoder	Zh/Es → En	Pronouns Cohesion Coherence	Miculicich et al. (2018)
			encoder w/attention		En → Ru	Deixis, Ellipsis Lexical cohesion	Xu et al. (2020a)
		variable	cache	decoder	Zh → En	-	Tu et al. (2018)
						Coherence	Kuang et al. (2018)
s, t	s	3	concatenated inputs		En → It	-	Agrawal et al. (2018)
		variable			En ↔ De	Pronouns	Scherrer et al. (2019)
		20	relative attention	encoder, decoder	Zh → En, En → De	Deixis, Ellipsis Lexical cohesion	Zheng et al. (2021)

Table 1.3: Overview of studies focusing on modeling local document-level context for improved NMT performance. Adapted from Table 2 in Maruf et al. (2021).

et al., 2019), on the assumption that it is essential for what comes next, while others look at the whole document to obtain its overall theme. Some also think certain parts, like the introduction or conclusion (Maruf et al., 2019; Zheng et al., 2021), are more important for understanding the document. A comprehensive study on document-level MT systems by Maruf et al. (2021) examined how context is integrated within NMT systems. The integration of context in different document-level MT systems is displayed in Tables 1.3 and 1.2, where *s* and *t* indicate whether the context originates from the source or target side, respectively. "Amount" refers to the maximum quantity of context utilised in the referenced work.

The tables show no unanimous agreement among researchers on defining "context" in document-level MT. I believe understanding document-level MT is necessary for improving translation quality, and context usage becomes necessary for this understanding. Hence, in this thesis, I aim to investigate document-level systems, focusing on understanding how context is used within them.

1.1.3 Challenges in Document-level translation

- **Understanding context usage:** One of the most significant challenges is effectively understanding and using the context provided by the rest of the document. This includes identifying which parts of the context are relevant for translating a given sentence and how to represent this context in a way the model can use.

- **Long-Range Dependencies:** In many documents, some dependencies and references span large text sections or even the entire document. Capturing and resolving these long-range dependencies is a significant challenge for document-level translation systems.

- **Maintaining Coherence and Consistency:** A document-level translation needs to be coherent and consistent, both within each sentence and across the whole document. This includes consistent translation of terms and phrases, maintaining the same style and tone, and preserving the overall flow and structure of the document.

- **Scalability and Efficiency:** Processing an entire document at once can be computationally expensive and time-consuming, particularly for long documents. Therefore, a major challenge is finding ways to scale the translation process and making it more efficient.

- **Handling Different Document Structures and Genres:** Documents can vary greatly in structure, style, genre, and content. Adapting the translation process to handle these variations is another significant challenge.

- **Evaluation of Translations:** Evaluating the quality of document-level translations is difficult because it requires considering not just the accuracy of each individual sentence, but also the coherence and consistency across the whole document. Most existing automatic evaluation metrics are primarily sentence-based and may not fully capture the quality of document-level translations.

As discussed above, document-level translation has many challenges, some more crucial than others for the growth of the field. I believe that understanding how context is used in document-level systems is a significant issue that demands attention. Addressing this concern can substantially contribute to advancing the field, and therefore, this thesis will primarily focus on investigating the context usage in document-level MT.

1.2 Research Questions

- **RQ1: How important is contextual information for improving translation in a document, and are there specific categories of sentences that demand contextual understanding more than others?**

  The main objective of this research question is to understand whether context impacts the translation quality of a sentence. I aim to investigate the sentences in a document, exploring how various contextual elements may influence translation. By doing so, I hope to obtain insights that can help improve the current context-aware MT systems.

- **RQ2: What is the ideal context span that can be incorporated into document-level translation systems to improve translation?**

  The main objective of this research question is to identify the optimal range of contextual information to consider during translation by analysing how various context spans affect translation quality. This analysis will help us understand how current document-level MT systems function and guide the development of improved document-level translation techniques.

- **RQ3: How effective are document-level translation systems and LLMs at translating domain-specific terminology, and to what extent can approaches based on terminology-aware mining improve the accuracy of domain-term translation in these systems ?**

  The main objective of this research question is to understand the effectiveness of document-level translation systems in translating terminology. This investigation includes a detailed analysis of current methods and their efficiency in handling domain-specific terms. With the emergence of LLMs, my study evaluates their capabilities for terminology translation. The research also discusses approaches based on terminology-aware mining in LLMs to improve the accuracy of terminology translation.

1.2.1 Thesis Outline

- **Chapter 2: Neural Machine Translation.**

  In this chapter, I discuss NMT, a method that has become the state-of-the-art methodology in the field of MT and on this robust foundation, all document-level MT systems are constructed. I explore how NMT is constructed (architecture), how it is trained (the training process), and how it generates translations (inference). I also discuss the significant benefits and challenges of using NMT. Finally, I review some of the tools I used to study and work with this technology.

- **Chapter 3: Investigating contextual influence in document-level translation.**

  This chapter investigates document-level MT systems by utilising the Hierarchical Attention Networks (HAN) (Miculicich et al., 2018) framework. The focus is on understanding why and when context helps. I conducted an in-depth qualitative analysis to understand the role of context in document-level MT. My investigations involved three morphologically distinct language pairs: Hindi-to-English, Spanish-to-English, and Chinese-to-English.

- **Chapter 4: Understanding the ideal context span in document-level translation.**

  This chapter discusses my experiments on understanding the ideal context span in document-level systems. Currently, there are many such systems, each working on different context spans. I investigate the ideal span for document-level systems and utilise this information to improve the existing systems.

- **Chapter 5: Terminology-aware mining for improving terminology translation.**

  In my previous chapters, I investigated document-level systems and tried to understand the context span of these systems. With advanced systems like LLMs emerging, I explore these models in this chapter. Specifically, I investigate their terminology translation capabilities and propose methods to help these systems improve terminology translation.

- **Chapter 6: Conclusion and Future Work.**

  In the final chapter of this thesis, I summarise my findings and reflect on the key outcomes and contributions of my research. I analyse the pros and cons of my approaches, measure their effectiveness, and consider their implications. I also discuss unexplored methodologies and future directions, identifying potential innovations and providing a roadmap for further inquiries. Finally, I explore opportunities to integrate emerging technologies to expand my research domain's understanding and application.

1.2.2 Publications

- **Prashanth Nayak**, Rejwanul Haque, and Andy Way. 2023. Instance-Based Domain Adaptation for Improving Terminology Translation. *In Proceedings of Machine Translation Summit XVII: Research Track*, pages 222-234, Macau SAR, China.

- **Prashanth Nayak**, Rejwanul Haque, and Andy Way. 2020. The ADAPT's Submissions to the WMT20 Biomedical Translation Task. *In Proceedings of the Fifth Conference on Machine Translation*, pages 841–848, Online. Association for Computational Linguistics. [Online].

- **Prashanth Nayak**, Rejwanul Haque, John D. Kelleher, and Andy Way. 2022. Investigating Contextual Influence in Document-Level Translation. *Information,* 13(5): Article number 249. ISSN 2078-2489.

- **Prashanth Nayak**, Rejwanul Haque, John Kelleher, Andy Way. 2023. Understanding Context Span in Document Level Translation, *Natural Language Engineering* [accepted, to appear online]

- **Prashanth Nayak**, Rejwanul Haque, and Andy Way. 2020. The ADAPT Centre's Participation in WAT 2020 English-to-Odia Translation Task. *In Proceedings of the 7th Workshop on Asian Translation*, pages 114–117, Suzhou, China. Association for Computational Linguistics.

Chapter 2

Neural Machine Translation

2.1 Introduction

Today's document-level translation systems are primarily built using NMT technology, so a thorough grounding in NMT is essential for understanding document-level MT. NMT systems are designed to handle context more effectively, a crucial aspect of accurately translating longer documents. We can better develop and refine document-level translation systems by understanding how NMT operates, particularly its ability to process and translate large chunks of text while maintaining context and meaning. So, in this chapter, I will give a brief overview of NMT and how it works.

NMT is an approach to MT that utilises neural networks, specifically deep learning algorithms, to automatically translate text from one natural language to another. NMT has become extremely popular in recent times due to its ability to handle complex linguistic structures and generate more fluent and contextually accurate translations over traditional rule-based and SMT systems. The NMT systems automatically translate a sentence from a source language, represented as $x_1, x_2, \ldots, x_n$ where each x_i corresponds to a word or token in the source sentence. This is translated into a corresponding sentence in a target language, denoted as $y_1, y_2, \ldots, y_m$ where each y_j corresponds to a word or token in the target language. The primary objective is to identify a target sequence with the highest likelihood given the input sequence. Formally, this involves the selection of a target sequence that results in the maximisation of the associated conditional probability, as in (1) and (2).

$$\hat{y} = \arg\max_{y} P_\theta(y|x) \tag{1}$$

$$P_\theta(y|x) = \sum_{n=1}^{N} P_\theta(y_n|y_{<n}, x) \tag{2}$$

In (1), $\hat{y}$ represents the predicted output sequence in the target language for a given input sequence x from the source language. The function $\arg\max_y$ seeks the output sequence y that maximises the conditional probability $P_\theta(y|x)$, where $P_\theta(y|x)$ denotes the conditional probability of the output sequence y given the input sequence x, as determined by the parameters θ of the neural network.

This conditional probability is further decomposed in (2). Here, the conditional probability of the output sequence y given x as the sum of the conditional probabilities of each output y_n given all previous outputs $y_{<n}$ and the input x. Essentially, this equation simplifies the problem by calculating the probability of each element in the sequence based on the previous elements and the input.

2.1.1 Architecture

Early models of NMT were based on Recurrent Neural Networks (RNNs), such as Recurrent Translation Models (Kalchbrenner and Blunsom, 2013), RNN Encoder-Decoder models (Cho et al., 2014), and Sequence-to-Sequence models (Sutskever et al., 2014). These models encode a source sentence to be translated into fixed-length vectors to encapsulate source sentences. This means that irrespective of the length of the source sentence, its representation was constrained to fit into a fixed-length vector. The RNNs were the preferred models for managing these fixed-length representations due to their ability to employ recurrent connections to demonstrate dynamic temporal behaviour. This characteristic gave them a specific proficiency in handling sequential data. However, these models encounter significant limitations when processing longer sentences, exposing a notable shortcoming in this approach. The fixed-length vector often failed to capture all the critical information from the source sentence, leading to a decline in translation quality, especially as the sentence structure became more complex. The emergence of attention mechanisms (Bahdanau et al., 2015) addressed this issue, marking a significant advancement in the field of NMT. These mechanisms permitted the model to focus on different parts of the input text during translation. Additionally, this approach introduced a method for generating variable-length representations, thereby improving the model's capacity to handle longer sentences and more complex structures.

The primary drawback of sequential computation lies in its limitation on parallelisation within training examples, creating a bottleneck when processing lengthy sentences. Nonetheless, a novel model architecture called the Transformer was introduced by Vaswani et al. (2017) that addressed this limitation. This model overcomes the need for recurrence by relying mainly on attention mechanisms. While numerous other NMT models exist, my experiments have predominantly focused on the Transformer and its variants. This choice is made considering the Transformers state-of-the-art performance across various NLP tasks, its efficiency in handling complex dependencies, and the abundant support and resources available within the research community. Therefore, I will focus on a more in-depth discussion of the Transformer architecture in the following section.

Transformer Architecture

The Transformer architecture has become a cornerstone of modern NMT systems (Touvron et al., 2023). This neural network is designed explicitly for sequence-to-sequence tasks, leveraging self-attention mechanisms and positional encoding to effectively and efficiently model complex language structures. This architecture comprises an encoder and a decoder, each comprising multiple identical layers. I now describe each of its components as shown in Figure 2.1:

- **Encoder**: The encoder converts the input sequence into a continuous repre-

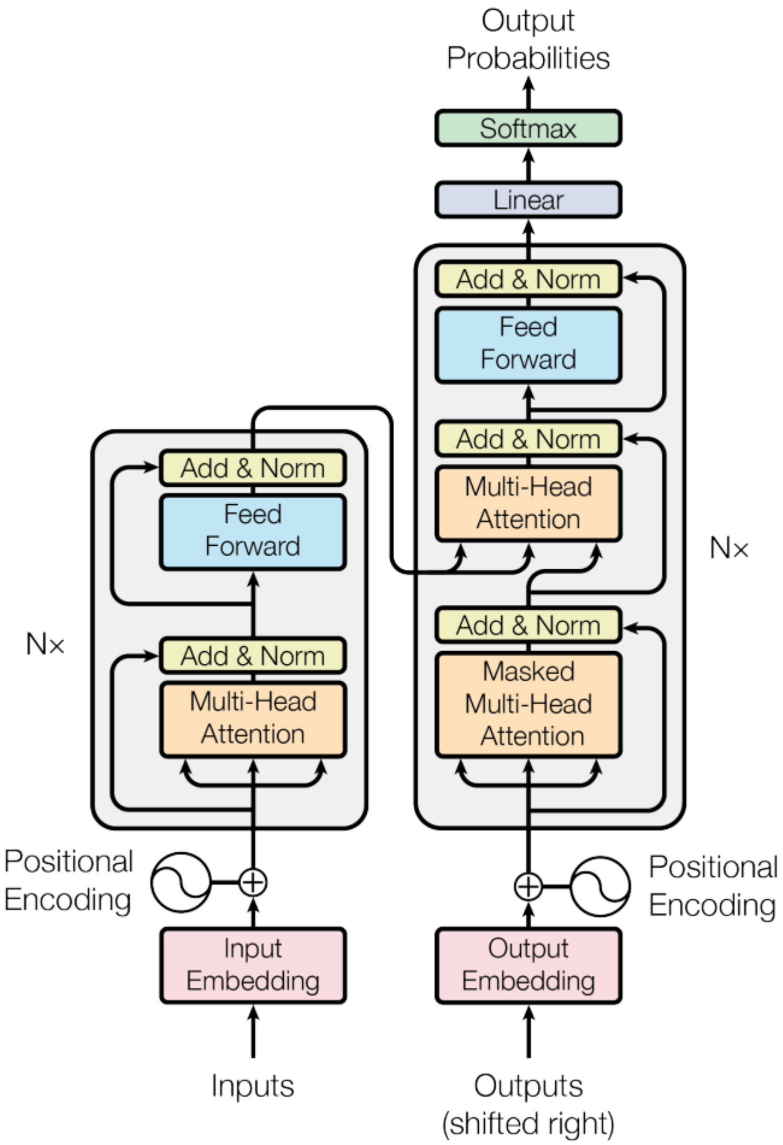


Figure 2.1: Transformer Architecture by Vaswani et al. (2017)

sentation. It comprises a stack of identical layers, each having two sub-layers: (i) a multi-head self-attention mechanism and (ii) a position-wise fully connected feed-forward network. Residual connections[1] and layer normalisation are utilised around each sub-layer to help with training.

- **Decoder**: The decoder generates the output sequence one token at a time. Like the encoder, it is composed of a stack of identical layers. It has three

[1] Residual connections help mitigate the vanishing gradient problem, a common issue in deep learning models, by creating shortcut paths for the gradient during back-propagation. They also ensure the preservation of information locality within the layers of the Transformer.

sub-layers: (i) a multi-head self-attention mechanism, (ii) a multi-head attention mechanism that attends to the encoder's output, and (iii) a position-wise fully connected feed-forward network. The decoder also employs residual connections and layer normalisation around each sub-layer.

- **Multi-head attention**: The Transformer employs multi-head self-attention mechanisms, which enable it to concentrate on distinct portions of the input sequence for each token in the output sequence. The self-attention mechanism computes a weighted sum of the input tokens' representations, with each token's relevance to the current context determining its weight. Multi-head attention enables the model to understand multiple relationships between input parts simultaneously.

- **Positional encoding**: Since the Transformer architecture has no recurrent or convolutional layers, it relies on positional encodings to incorporate information about the position of tokens in the input sequence.

- **Feed-forward neural networks**: The feed-forward (FF) networks within each layer of the encoder and decoder are crucially designed with a two-step process involving dimensionality manipulation. Initially, the first linear layer expands the dimensionality of the input, a important step that allows the network to explore a broader and more complex feature space. This is followed by applying a Rectified Linear Unit (ReLU) activation function, which introduces non-linearity and aids in complex pattern recognition. The subsequent linear layer then reduces this expanded dimensionality, condensing the information back into a more compact representation. Expanding and reducing the dimensionality in the FF component is critical to the Transformer model's ability to capture and process the complexities of the input data effectively.

2.1.2 Data Pre-processing

Pre-processing is a critical step in NMT for preparing data. It typically involves tokenisation, dividing text into smaller units (tokens), like words or subwords. The choice between words and subwords is significant: subword tokenisation can reduce out-of-vocabulary (OOV) issues by breaking down words into smaller, more manageable units. In contrast, using total words can worsen the OOV problem, especially with a limited vocabulary set. Vocabulary management, therefore, becomes a strategic decision where a fixed vocabulary is selected to represent best the text the model will encounter and learn from. This process is particularly effective with subword units, allowing the model to handle new or rare words more efficiently. Additionally, normalisation techniques are employed to standardise text variations, while cleaning methods are used to remove irrelevant or noisy data. These steps collectively ensure the creation of a consistent and coherent dataset, optimising the neural network' s learning efficiency and enhancing the accuracy of mapping between source and target languages.

2.1.3 Training

NMT models are typically trained on large datasets divided into training, validation (development), and testing. The training set, the largest of the three, is used to train the model to predict the target language from the source language. The more data in this set, the better the model can learn. The validation set is used during the model development process to fine-tune the model's hyperparameters and make decisions on the model architecture. It checks against overfitting and provides insights into how well the model generalises to unseen data. The testing set, kept distinct from the training and validation processes, is used after the model is finalised to provide an unbiased evaluation of the model's performance on completely unseen data.

Training in NMT involves optimising a model to translate sentences from a source to a target language. This optimisation is achieved by minimising the negative log-likelihood of the correct translation given the source sentence. The process involves forward propagation for prediction, loss calculation, and backward propagation for updating the parameters. This procedure is repeated across numerous epochs until the model's performance no longer improves. The effectiveness of the trained model is then evaluated on a test set.

In a typical NMT setting, the objective to minimise the negative log-likelihood of the correct translation, given the source sentence can be formally represented as in (3).

$$\theta^* = \arg\min_{\theta} \sum_{(x,y)\in D} \sum_{n=1}^{|y|} -\log P_{\theta}(y_n|y_{<n}, x) \tag{3}$$

In (3), the equation is modeled to find the best set of parameters, labeled as θ^*, which makes the model best at predicting each word in the translations. The symbol θ stands for the model's parameters and $\arg\min_{\theta}$ means that we are looking for θ that minimises what follows in the equation. The model goes through each pair of original and translated sentences (x, y) in the dataset D and each word n in the translation sentence y. It calculates $-\log P_{\theta}(y_n|y_{<n}, x)$, which is essentially a measure of how good the model is at predicting the *n-th* word in the translated sentence y, given the original sentence and the words in the translated sentence that came before, based on the current parameters θ.

2.1.4 Inference: Beam Search

Beam search (Och and Ney, 2004) has been the most commonly used decoding algorithm in NMT systems. Rather than keeping only the best sequence at each step, the beam search algorithm keeps the n best sequences at each step, where n is a user-defined parameter known as the beam width. The higher the beam width, the more sequences the algorithm explores, which typically leads to better results. The algorithm follows these general steps:

- In the initial step, the model generates the first token for all possible tokens in the vocabulary.

- Then, the probabilities of these initial tokens are computed, and only the top n sequences are retained.
- In each subsequent step, for every sequence currently in the beam, the model generates the next token for all possible tokens in the vocabulary.
- Next, the probabilities of these extended sequences are computed, and once again, only the top n sequences are retained. This process continues until a stopping condition is met, which could involve either reaching a maximum sequence length, or having all sequences in the beam end with an end-of-sequence token.
- Ultimately, the sequence with the highest cumulative score balanced by heuristics for different lengths is chosen as the output to ensure fair comparison and accuracy across varying hypothesis lengths.

In this manner, beam search effectively balances the trade-off between computational efficiency and high-quality translations.

2.1.5 Post Processing

Post-processing is the last stage of NMT, where the model's raw translation output is modified to improve its readability and precision. This stage involves detokenisation, which consists of merging tokens into comprehensible sentences. The translation is sometimes checked against a predetermined glossary, ensuring terminological consistency and accuracy. Error corrections are applied, and stylistic and grammatical adjustments are made to comply with the conventions of the target language. The main aim of post-processing step is to produce translations that are not only syntactically accurate but also rich in contextual and cultural appropriateness.

2.1.6 Advantages

NMT has several advantages over traditional rule-based and SMT methods (Sutskever et al., 2014), contributing to its widespread adoption and success in recent years.

Some of the key advantages include:

- Improved translation quality: NMT systems produce more fluent, accurate, and contextually appropriate translations than rule-based and statistical methods. This is due to their ability to learn complex language patterns and structures from the training data.

- Handling of long-range dependencies: NMT systems, particularly those using attention mechanisms and Transformer models, can effectively capture long-range dependencies within sentences, allowing them to handle complex linguistic structures and maintain translation coherence.

- End-to-end learning: Unlike rule-based or statistical methods that require extensive feature engineering and multiple processing steps, NMT systems learn to translate text from one language to another in an end-to-end manner. This simplifies the overall pipeline and allows for more efficient training and deployment.

- Scalability: NMT models can be scaled to handle multiple language pairs and domains by training them on diverse and large parallel corpora. This helps the development of Multilingual Translation Systems (MTS) that support many languages and can easily be adapted to different domains.

- Robustness to noisy data: NMT systems are particularly adept at being robust to noisy training data. Some practitioners have shown (Rarrick et al., 2011) that detecting and removing MT-ed data from training can improve SMT translation. Fast forward to NMT, again, filtering data was shown to be useful in the works by Junczys-Dowmunt et al. (2018) and Schamper et al. (2018). Interestingly, studies have shown that having a large volume of data contributes to the superiority of NMT systems over SMT (Koehn and Knowles, 2017). This raises the question: Where does this immense volume of data originate? It simply doesn't exist. Given its non-existence, practitioners have begun to augment the parallel data with synthetic data using methods such as back-translation (Sennrich et al., 2016a; Poncelas et al., 2018), to construct high-performing models, and these models have shown improvement in translation quality. Given the fact that NMT systems improve translation over MT-ed data, I believe they are robust to some amount of noise in the data.

- Adaptability: NMT systems can be fine-tuned to specific domains, styles, or genres by training them on specialised corpora, making it helpful to develop customised translation solutions for domains like legal and medical. While SMT models can also be adapted this way, NMT models generally yield higher-quality translations due to their ability to consider the full context of sentences. Moreover, NMT systems require less manual feature engineering than SMT systems, making them more flexible and easier to adapt.
- Limiting Vocabulary Size and Handling Rare and Unseen Words: NMT systems employ subword tokenization techniques such as Byte Pair Encoding (BPE) (Sennrich et al., 2016b)[2] to manage the source/target vocabulary size, helping to mitigate computational complexity. Most models incorporating these techniques reduce the possibility of encountering OOV tokens by retaining single characters as a fallback option. Consequently, OOV occurrences are rare and typically only arise in exceptional cases, such as when dealing with unusual character encodings or words from languages vastly different from those in the training dataset. Thus, while these techniques greatly improve the systems adaptability and capability to handle a wide range of words, there remain challenges in eliminating OOV issues.

2.1.7 Disadvantages

Although NMT provides numerous benefits over standard translation practices, it does have some drawbacks and difficulties that must be overcome.

- Data requirement: NMT systems typically require large parallel corpora (Koehn

[2] while this is an advantage to NMT systems, it also applies to SMT models

and Knowles, 2017) for training to achieve high-quality translations. Acquiring and preparing such data can be time-consuming and resource-intensive, particularly for low-resource languages or specialised domains where parallel corpora may be scarce or unavailable.

- Computational resources: Training NMT models, especially large and deep architectures like Transformers, requires significant computational resources, such as powerful Graphical Processing Units (GPUs). This can make developing and deploying NMT systems expensive and challenging, particularly for smaller organisations or individual researchers, who lack access to such hardware.

    Model complexity: NMT models are complex (Forcada, 2017), which makes them challenging to understand and correct when errors and biases occur in translation. This lack of transparency can pose challenges in industries such as law or medicine, where the ability to explain and trust the system is critical for its adoption. Similarly, SMT models are also complex (Way and Hearne, 2011; Hearne and Way, 2011) they may contain multiple submodels, making them hard to interpret. This complexity in NMT and SMT models shows the difficulty in achieving transparency and reliability in MT, especially in fields where precision and accountability are paramount.

- Bias and fairness: NMT systems can unintentionally acquire and propagate biases in the training data, resulting in unfair or biased translations. In order to address these issues, it is necessary to carefully consider the training data and devise techniques to mitigate bias in the generated translations.

- Context issues: Because NMT systems typically work on one sentence at a time, they can struggle to understand the broader context from the rest of the text, leading to incorrect translation.

2.1.8 Evaluating Machine Translation

Evaluating machine translation poses a challenge due to the subjective nature of assessing the quality of the translated text. Several different metrics have been proposed for evaluating MT. I will discuss some of the metrics that we utilised for my experiments:

- Bilingual Evaluation Understudy (BLEU): BLEU (Papineni et al., 2002) is a widely-used metric for MT. It measures the overlap of n-grams between the MT and a set of reference translations. The scores range from 0 to 1, with 1 indicating a perfect match with the reference(s).

- Metric for Evaluation of Translation with Explicit Ordering (METEOR): METEOR (Banerjee and Lavie, 2005) is another evaluation metric that considers precision, recall, synonymy, stemming, and word order when comparing MT and reference translation. The scores range from 0 to 1, with higher scores indicating better translation quality.

- Translation Edit Rate (TER): TER (Snover et al., 2006) measures the number of edits required to change a system output into one of the references. The edits can include word shifting, insertion, deletion, and substitution based on Levenshtein distance. A lower score is an indicator of better translation quality.

- character *n*-gram F-score (chrF): chrF (Popović, 2015) is another automatic metric for MT evaluation. Unlike BLEU, which operates at the word level, chrF measures the precision and recall of character *n*-grams in MT translations compared to the reference translations. It can be beneficial when evaluating translations into languages where word segmentation is challenging or character-level errors are more prominent and are important to be identified. chrF focuses exclusively on character-level analysis, providing a score with a higher value, indicating better translation quality.

- Crosslingual Optimised Metric for Evaluation of Translation (COMET): COMET (Rei et al., 2020) leverages large-scale multilingual pre-training, followed by fine-tuning on translation ranking tasks. In contrast to traditional metrics like BLEU or chrF, COMET's deep learning-based approach enables it to capture more complex and subtle aspects of translation quality that align closely with human judgement. As a result, a higher COMET score corresponds to a translation that more closely resembles a human-evaluated, high-quality translation, offering a more comprehensive assessment of translation quality.

2.1.9 Human Evaluation

Human evaluation in translation, particularly in MT, is important for assessing the quality and fluency of translated texts. While automated metrics like BLEU, ChrF, TER and METEOR can provide a quick and easy way to measure translation accuracy, they often fail to evaluate aspects like fluency, idiomatic usage, and cultural appropriateness, where a human evaluator's perspective is irreplaceable. Human evaluation (Way, 2018) involves examining translations by linguists or native speakers who assess the output based on specific criteria such as:

- Fluency: The measure to check if the translation reads as if it were written by a native speaker by looking for natural phrasing and grammatical correctness.
- Adequacy: The measure to check whether the translation conveys all the information from the source text accurately without adding, omitting, or distorting the meaning.
- Comprehensibility: This measures how easily the translated text can be understood.
- Coherence and Cohesion: This measures how well the translated text has flow, ensuring that it is logically organised and connected.

This process gives a detailed understanding of how well an MT system works and, importantly, areas where it can improve. Feedback from human evaluators serves as an essential resource for training and refining MT systems, ensuring that they align more closely with human expectations for quality translation.

2.2 Toolkits Used

In this section, I discuss the various tools I used for my experiments.

2.2.1 OpenNMT

OpenNMT (Klein et al., 2017) is an open-source ecosystem that offers implementations of NMT models. OpenNMT provides robust, flexible, and user-friendly implementations of sequence-to-sequence models, making it an excellent choice for researchers. It supports various models, such as Transformers, LSTMs, and more, along with various features necessary for training and deploying models for machine translation and other NLP tasks. The toolkit offers high flexibility and control, allowing users to experiment with different network architectures, training procedures, and other parameters. These characteristics and its open-source nature make OpenNMT a valuable tool for those working in machine translation and related fields.

2.2.2 Hugging Face

Hugging Face[3] created the popular open-source Transformers library for NLP. The Transformers library offers thousands of pre-trained models for various NLP tasks, like text classification, named entity recognition, text generation, translation, and more. These models include well-known architectures such as BERT (Devlin et al., 2019), Generative Pretrained Transformer (GPT) (Brown et al., 2020), T5 (Raffel et al., 2020), and DistilBERT (Sanh et al., 2019). One key feature of the Transformers library is its user-friendly design, making it easy for developers, researchers, and businesses to use cutting-edge NLP models without needing specialised knowledge of

[3] https://huggingface.co/

technology. Hugging Face also maintains a model hub where community members can share their pre-trained models, promoting collaboration and helping the field to grow.

2.2.3 Moses

Moses[4] (Koehn et al., 2007) is a renowned SMT toolkit that offers a flexible and extensive platform for MT and language processing research. The toolkit provides many features, including word alignment, language modeling, and translation model training, allowing users to modify the system to specific needs and languages. With support for various algorithms and a robust framework well-suited for experimentation, Moses has been widely adopted in academia and industry. Its open-source nature encourages collaboration and development within the community, maintaining Moses's reputation as a fundamental tool in MT.

[4] https://github.com/moses-smt/mosesdecoder

Chapter 3

Document Level Translation

3.1 Introduction

NMT (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017) is currently the prominent approach in the field of MT. However, interestingly, even the best-performing NMT systems (such as Vaswani et al. (2017)) do not consider the context of the sentence being translated, which means that the translation happens in isolation without considering the context of the document. However, isolating sentences for translation may not be ideal, as the semantics of a source sentence are often more accurately interpreted when viewed in the specific context of the document. For example, human translators typically use Computer-Aided Translation tools that present the sentence to be translated alongside surrounding sentences for contextual reference. Recently, there has been a growing trend among researchers to incorporate document-level context into NMT systems (Wang et al., 2017; Maruf and Haffari, 2018; Miculicich et al., 2018; Voita et al., 2018; Sun et al., 2022; Zhang et al., 2022; Bao et al., 2023; Herold and Ney, 2023b,a; Zhang et al., 2023). The results of these studies show promising signs that this approach indeed has the potential to improve the translation quality of today's NMT systems. This chapter discusses my work on document-level MT, where I investigate contextual influence in document-level translation. In this chapter, I try to understand why and when context helps translation in document-level MT system.

3.2 Related Work

Incorporating document-level context into MT systems helps resolve linguistic ambiguities and inconsistencies that may arise when translated in isolation. For example, phenomena such as anaphoric pronoun resolution, maintaining lexical cohesion, and preserving the document's overall theme are better handled in a document-level approach (Bawden et al., 2018). Castilho et al. (2020) in their analysis on document-level evaluation found that a context window of 10 sentences both preceding and following is sufficient to handle major linguistic issues.

The efforts to effectively incorporate document-level context into NMT systems have seen several innovative approaches. For instance, Wang et al. (2017) introduced a context-aware MT architecture using a hierarchical Recurrent Neural Network (RNN) that synthesised the context from the preceding n sentences of a source sentence to be translated. Tiedemann and Scherrer (2017) adopted a similar approach to capturing the context and implemented an RNN-based MT model. This document-level MT system used the preceding sentence as the context window for both the source and target sides. Subsequently, Bawden et al. (2018) used multi-encoder NMT models that harness the context from the prior source sentence. Finally, Maruf and Haffari (2018) proposed a unique document-level NMT architecture employing memory networks to track global context, with separate memory components for both source and target sides.

Further studies saw the Transformer architecture (Vaswani et al., 2017) being employed by Voita et al. (2018) to investigate document-level MT. The approach added an extra encoder to incorporate document-level context. In addition, they used a single sentence, either preceding or succeeding, as the context for translation. The sentences are concatenated using a seperator to indicate context usage. Unlike most works, in an effort to balance local and global context,Tan et al. (2019) proposed a hierarchical model that captures local dependencies with a sentence encoder and global dependencies with a document encoder. This approach was intended to minimise mistranslations and attain context-specific translations.

Breaking away from the dual-encoder architecture, standard in document-level MT, Ma et al. (2020) introduced a flat-structured Transformer model with a unified encoder that attends to local and global contexts. Similarly, Zhang et al. (2021) proposed a novel Multi-Hop Transformer architecture which refines sentence-level translations iteratively using context clues from the previous source and target sentences.Yin et al. (2021) found that regularising attention with Supporting Context for Ambiguous Translations enhanced anaphoric pronoun translation, suggesting the potential for further attention supervision with context.

Additionally, in the development of context encoders, Yun et al. (2020) introduced a Hierarchical Context Encoder that extracts sentence-level information from preceding sentences and hierarchically encodes context-level information using a hierarchical attentional network. Kim et al. (2019) examined advances in document-level MT using general (non-targeted) datasets, attributing the observed improvements not to context utilisation but to the effects of regularisation (promoting models to pay more attention to words that humans use to resolve linguistic issues). Lopes et al. (2020) systematically compared various document-level MT systems based on large pre-trained language models, introducing a variant of the Star Transformer (Guo et al., 2019) that incorporates document-level context. Exploring the application of contextual information for zero-resource domain adaptation, Stojanovski and Fraser (2021) proposed two variants of the Transformer model to handle exceedingly large contexts (10 previous sentences).

While the experiments by Miculicich et al. (2018) demonstrate that context improves translation quality, it remains unclear why their context-aware models outperform those disregarding context. I aimed to investigate why and when context improves translation quality in document-level MT. This chapter discusses my work on document-level MT, utilising the HAN (Miculicich et al., 2018) (cf. Section 3.3.4) framework for my experiments. My investigations involved three morphologically distinct language pairs: Hindi-to-English, Spanish-to-English, and Chinese-to-English. I conducted an in-depth qualitative analysis to understand the role of context in document-level MT. At the time of my experiments, HAN was the state-of-the-art for document-level MT, which is why I used it; additionally, the availability of models and source code made it user-friendly.

3.3 Dataset Used

In this section, I detail the datasets that I used for my experiments for three language pairs.

3.3.1 Hindi-to-English

My NMT systems were trained using the IIT-Bombay[1] parallel corpus (Kunchukuttan et al., 2018). I randomly extracted 1000 judicial domain sentences from the training data for development purposes. My test data came from the judicial domain and contained domain-specific term annotations[2] (Haque et al., 2019). Hindi sentences in the parallel corpus were tokenised using the IndicNLP[3] library, and for English, I used the Moses toolkit[4] (Koehn et al., 2007). Detailed data statistics are shown in Table 3.1.

Table 3.1: Corpus statistics for Hindi-to-English.

Hindi-to-English			
	Sentences	**English (Words)**	**Hindi (Words)**
Train	1,049,198	18,132,805	18,907,775
Dev	1000	26,106	28,535
Test	1270	26,284	27,414

3.3.2 Spanish-to-English

I used the same Spanish-to-English dataset for my experiment as the one referenced in Miculicich et al. (2018). Data for my experiments was sourced from the TED talks

[1] https://www.cfilt.iitb.ac.in/ parallelcorp/iitb_en_hi_parallel/

[2] https://github.com/rejwanul-adapt/EnHiTerminologyData

[3] https://anoopkunchukuttan.github.io/indic_nlp_library/

[4] https://github.com/moses-smt/mosesdecoder

dataset.[5] As part of my evaluation setup, I used datasets proposed by Cettolo et al. (2012, 2015). Furthermore, following the methodology suggested by Miculicich et al. (2018), I utilised the *dev2010* dataset for my development stage and a combination of *tst2010*, *tst2011*, and *tst2012* test sets for my evaluation stage. Finally, I used the tokeniser scripts provided within the Moses toolkit to tokenise English and Spanish sentences. Detailed statistics of the data are exhibited in Table 3.2. The "No. Discourses" column in the table represents the discourse boundaries provided in the dataset.

Table 3.2: Corpus statistics for Spanish-to-English.

	Spanish-to-English			
	Sentences	**English (Words)**	**Spanish (Words)**	**No. Discourses**
Train	187,958	3,190,760	308,6205	1421
Dev	887	17,454	16,944	8
Test	4706	90,288	83,526	42

3.3.3 Chinese-to-English

I used the same Chinese-to-English dataset for my experiment as the one referenced in Miculicich et al. (2018). Similar to the Spanish-to-English dataset, I used TED talks data for the Chinese-to-English translation task, as provided by Cettolo et al. (2012, 2015). Following the guidelines from Miculicich et al. (2018), I utilised *dev2010* for validation and a combination of *tst2010*, *tst2011*, *tst2012*, and *tst2013* for comparative evaluation against existing work. The Moses toolkit was applied to the English sentences for the tokenisation process, whereas the Jieba segmentation toolkit[6] was used for Chinese. Table 3.3 provides detailed data statistics. The "No. Discourses" column in the table represents the discourse boundaries provided in the dataset.

[5] https://www.ted.com/talks

[6] https://github.com/fxsjy/jieba

Table 3.3: Corpus statistics for Chinese-to-English.

	Chinese-to-English			
	Sentences	**English (Words)**	**Chinese (Words)**	**No. Discourses**
Train	223,685	3,756,209	545,708	1718
Dev	887	17,454	2348	8
Test	5473	108,937	12,897	56

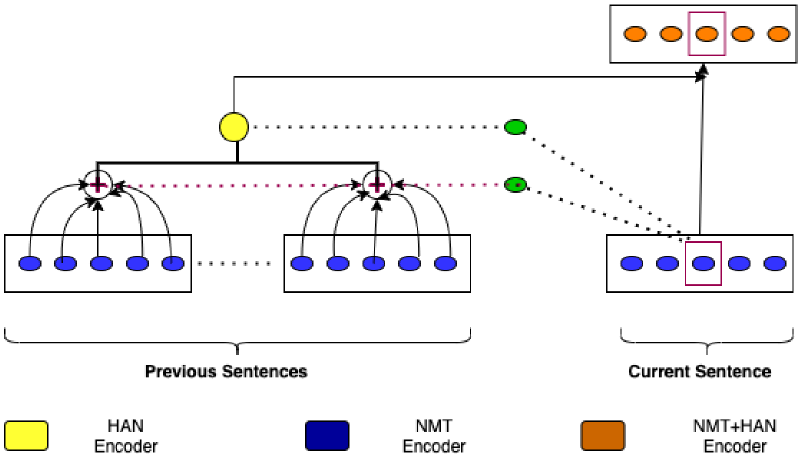


Figure 3.1: Illustration of the HAN architecture based on Figure 1 in Miculicich et al. (2018)

3.3.4 Context-Aware HAN Model

HAN is a context-aware NMT model that utilises hierarchical attention (varying importance to different parts of data, focusing on keywords in sentences and significant sentences in documents) to incorporate prior context. It systematically structures contextual and source sentence information by leveraging word- and sentence-level abstractions. For each predicted word, the hierarchical attention provides dynamic access to the context by selectively focusing on different sentences and words. Specifically, HAN considers the preceding n sentences as context from the source and target sentences. Figure 3.1 depicts how context integration occurs in HAN. The process involves combining hidden representations from the encoder and decoder of past translations and then feeding this unified information into both the encoder and decoder during the translation of the current sentence. This method of integration enables the model to optimise across multiple sentences simultaneously. My context-aware systems were built using HAN, with the context window of the previous three sentences as in Miculicich et al. (2018).

3.4 Evaluation Methodology

The progression of sentences within a document creates context (for instance, the sentences that have appeared previously), which is likely to be beneficial for document-level translation. However, shuffling the sentences in a document typically interferes with this progression. In such a scenario, the advantages that context typically brings to document-level translation are likely to be lost. Shuffling context in a document-level MT system aims to investigate how changing the context affects translation performance. This approach tests HAN's ability to handle information that is provided out of sequence. This mixing up of context from various parts of the document allows me to understand the context usage by HAN. For this, I examined the performance of HAN under two separate evaluation setups:

1. Original test set sentences: These sentences preserve the original contextual order of the document. In my future references, I will refer to this as OrigTestset.

2. Shuffled test set sentences: I generated this test set by randomly shuffling the OrigTestset, thereby rearranging the order of the sentences in the document. Subsequently, I will refer to this test set as ShuffleTestset.

In order to understand the influence of context on translation quality, I translated both the test sets discussed above (i.e., OrigTestset and ShuffleTestset) using HAN. I employed four distinct evaluation metrics to assess the translations produced: BLEU, chrF, TER, and METEOR.

3.5 Experiment and Results

3.5.1 Results

I assessed my MT systems, Transformer and HAN models, utilising the OrigTestset for Hindi-to-English, Spanish-to-English, and Chinese-to-English translation tasks. The corresponding BLEU, chrF, TER, and METEOR scores are tabulated in Table 3.4. As seen in Table 3.4, HAN surpasses the Transformer across all evaluation metrics. Further, upon conducting statistical significance tests through bootstrap resampling Koehn (2004), I discovered that these scores are statistically significant. This underlines the advantages of integrating context into NMT models.

Table 3.4: Baseline scores of NMT systems (HAN).

	Hindi-to-English			
	BLEU	**chrF**	**TER**	**METEOR**
Transformer	31.78	0.535	48.53	0.658
HAN	33.27	0.543	46.78	0.665
	Spanish-to-English			
	BLEU	**chrF**	**TER**	**METEOR**
Transformer	36.19	0.558	40.95	0.707
HAN	39.08	0.579	38.58	0.714
	Chinese-to-English			
	BLEU	**chrF**	**TER**	**METEOR**
Transformer	15.60	0.375	67.75	0.484
HAN	18.14	0.388	64.09	0.519

In Table 3.4, I see that the BLEU score for Hindi-to-English improved by 4.68%, showing that translations are more accurate. Hindi-to-English also saw minor improvements in other metrics, with a 1.49% increase in CHRF, 3.74% in TER, and 1.06% in Meteor, making translations slightly better than the baseline overall. Spanish-to-English translations improved more, with a increase in BLEU, 3.76% in CHRF, 6.14% in TER, and 0.99% in Meteor. This means Spanish-to-English translations are moderately better and more accurate than the baseline Transformer. Lastly, Chinese-to-English translations improved the most, with a significant 16%

increase in BLEU and improvements in CHRF (3.46%), TER (5.63%), and Meteor (7.23%). This indicates that Chinese translations are more accurate and have seen an excellent overall improvement in quality. To be noted that all the improvements suggested above are relative to their respective baseline scores.

To further understand how context affects HAN's translations, I performed five random shuffles of the OrigTestset. This procedure yielded five distinct test sets, which I shall henceforth call ShuffleTestsets. Subsequently, I conducted evaluations of my MT systems on these ShuffleTestsets and documented the BLEU, chrF, TER, and METEOR scores in Table 3.5. Observations from Table 3.5 indicate that the context-aware NMT model generates roughly consistent BLEU, chrF, TER, and METEOR scores across all ShuffleTestsets. Despite the improved translation quality of HAN seen from the scores in Table 3.4, which were seemingly influenced by context, the scores in Table 3.5 challenge the notion of the positive impact of context in HAN. At first glance, it might seem as if the results from Table 3.5 are challenging the notion of the positive impact of context on the HAN model, but a deeper look offers a different perspective. Despite the context being shuffled, the translation scores are still significantly better than those of the baseline system. This suggests that the presence of some context – even though it is somewhat randomly generated from the document – might be beneficial, providing some relevant information for the task of translation. This observation indicates that HAN might be effectively utilising information from context in ways that are not solely dependent on its original, sequential order. Subsequently, I conducted statistical significance tests using bootstrap resampling to compare the BLEU metric scores between the baseline and the context-aware system. The results of these tests revealed that the differences in BLEU scores were statistically significant.

I also examined the translation scores (BLEU, chrF, TER, and METEOR) generated by HAN. I discovered that 14%, 16%, and 17% of sentence translations significantly fluctuated across the five shuffles (i.e., five ShuffleTestsets) for Hindi-to-English, Spanish-to-English, and Chinese-to-English, respectively. Furthermore,

Table 3.5: Performance of NMT systems (HAN) on shuffled data.

	Hindi-to-English			
	BLEU	**chrF**	**TER**	**METEOR**
shuffle 1	33.06	0.542	46.78	0.664
shuffle 2	33.19	0.544	46.78	0.663
shuffle 3	33.07	0.544	46.87	0.665
shuffle 4	32.93	0.540	47.24	0.663
shuffle 5	33.34	0.543	46.69	0.665
Mean	33.11	0.542	46.87	0.664
	Spanish-to-English			
	BLEU	**chrF**	**TER**	**METEOR**
shuffle 1	38.31	0.577	38.77	0.716
shuffle 2	39.00	0.578	38.48	0.712
shuffle 3	38.84	0.578	38.77	0.713
shuffle 4	38.59	0.577	39.17	0.714
shuffle 5	38.39	0.577	38.87	0.715
Mean	38.62	0.577	38.81	0.714
	Chinese-to-English			
	BLEU	**chrF**	**TER**	**METEOR**
shuffle 1	17.36	0.392	65.18	0.519
shuffle 2	16.99	0.387	64.79	0.518
shuffle 3	16.50	0.387	65.48	0.514
shuffle 4	16.96	0.385	65.28	0.516
shuffle 5	16.49	0.386	64.99	0.520
Mean	16.86	0.387	65.14	0.517

I noted that 58%, 64%, and 61% of sentence translations remained unchanged or consistent across the five shuffles (i.e., five ShuffleTestsets) for Hindi-to-English, Spanish-to-English, and Chinese-to-English, respectively.

These findings encouraged me to scale up my experiments, so I increased the samples to obtain further insights. For this, I shuffled my test data fifty times, providing us with fifty ShuffleTestSets. Finally, I computed the mean variances of the obtained translation scores for each sentence in the discourse over the fifty ShuffleTestSets. From now on, I call this measure MV (mean of the variance). This resulted in a single MV score for each sentence in the test set. I then calculated the sample mean ($\overline{x}$) and standard deviation (s) from the sampling distribution i.e. the MV scores, and the 95% confidence interval of the population mean (μ) using the formula: $\overline{x} \pm Z(\sigma_{\overline{x}}) = \overline{x} \pm Z(\sigma/\sqrt{n}) = \overline{x} \pm Z(s/\sqrt{n})$.[7]

Table 3.6: Mean variances of two sentences across fifty shuffles. They were selected from the test set of the Spanish-to-English task.

	BLEU	**chrF**	**TER**	**METEOR**
Sent 1	0.26	0.04	0.20	0.31
Sent 2	21.79	2.93	6.72	5.40
Confidence Interval	1.07–4.86	0.46–1.56	0.61–2.20	0.81–4.06

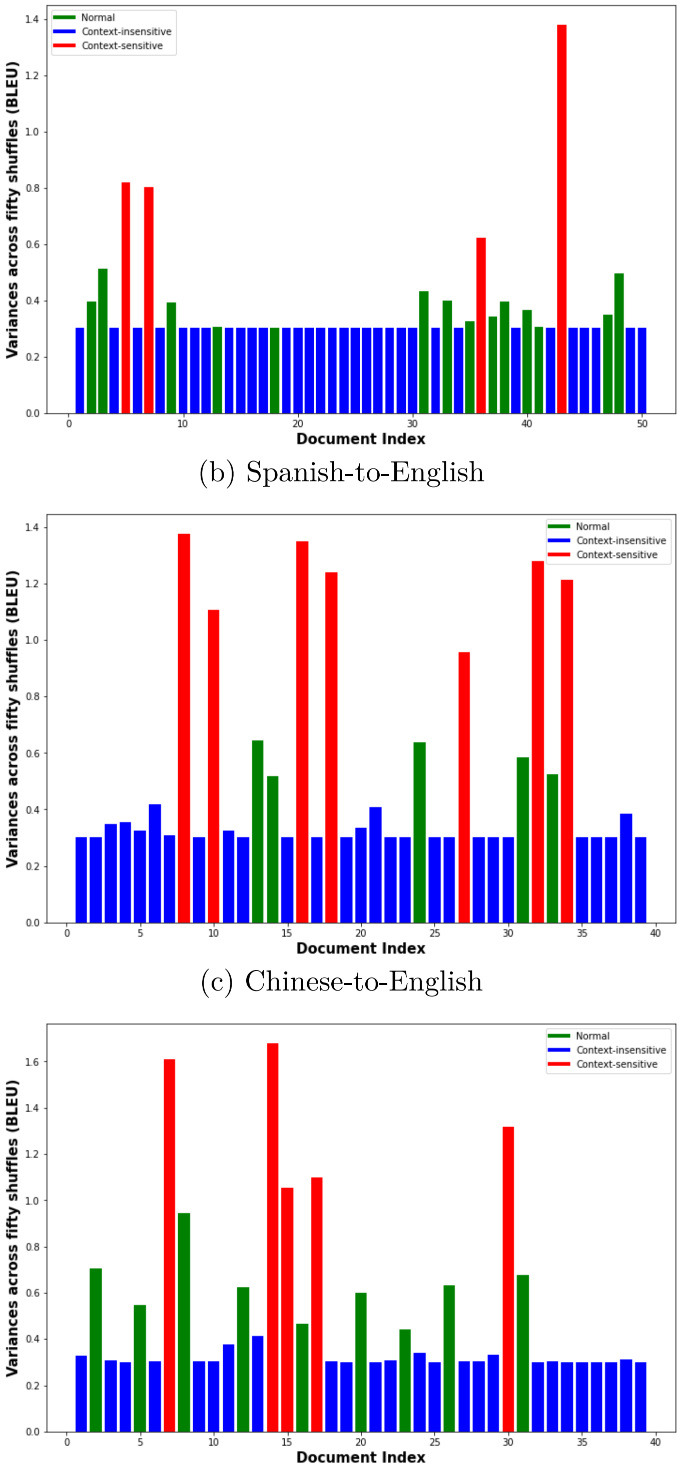
The final row of Table 3.6 displays the 95% confidence intervals for sentence-level BLEU (Post, 2018), chrF, TER, and METEOR, derived from the sampling distribution of the MV scores for sentences in the test set. This methodology enables me to categorise test set sentences into three groups: (i) *context-sensitive*, (ii) *context-insensitive*, and (iii) *normal* (MV scores that fall within the 95% confidence interval). The normal category of sentences neither show the high sensitivity of the "context-sensitive" category nor the low sensitivity of the "non-context-insensitive" group, representing a standard or average context responsiveness in translation. Hence, for the purpose of my study, I primarily focus on analysing sentences from the two extremes of this classification, namely, the context-sensitive and context-insensitive categories.

To illustrate my classification process, I selected two sentences from the Spanish-to-English translation task's test set. The variances calculated from the distribution of the BLEU, chrF, TER, and METEOR scores for these sentences are displayed in the first two rows of Table 3.6. As evident from Table 3.6, the variances for both sentences fall outside the confidence interval (CI). Sentences with a variance greater than the CI are classified as context-sensitive, while those with a variance lower than the CI are classified as context-insensitive.

To gain a clearer understanding of the three categories of sentences, a detailed visualization is provided in Figure 3.2a–c. Due to the impracticality of visualising the entire test set, I selected specific discourses from the dataset. The lengths of

[7] The mean of the sampling distribution of $\overline{x}$ equals the mean of the sampled population. Since the sample size is large ($n = 50$), I will use the sample standard deviation, s, as an estimate for μ in the confidence interval formula.

Figure 3.2: Variances of the test set sentences for BLEU and their corresponding classes (green: *normal*, blue: *context-insensitive* and red: *context-sensitive*).



(a) Hindi-to-English

(b) Spanish-to-English

(c) Chinese-to-English

these discourses in Hindi, Spanish, and Chinese are 50, 39, and 39, respectively. Figure 3.2a–c illustrates the variances obtained for the BLEU score in the Hindi-to-English, Spanish-to-English, and Chinese-to-English tasks. Here, the green, blue, and red bars correspond to *normal*, *context-insensitive*, and *context-sensitive* sentences, respectively. These figures clearly describe the sentence distributions across the three classes.

I also manually reviewed the translations for context-sensitive and context-insensitive sentence categories. I found that the quality of translations for the context-sensitive category is indeed affected by contextual information. Meanwhile, for the context-insensitive category, the quality of translations largely remained consistent across different shuffles. In the following sections, I discuss the context-sensitive and context-insensitive categories of sentences in detail.

Table 3.7: Evaluation scores for the set of context-sensitive sentences.

		Hindi-to-English		
	BLEU	**chrF**	**TER**	**METEOR**
Max	25.13	0.47	57.89	0.55
Mean	16.39	0.45	54.93	0.52
Min	12.71	0.42	50.00	0.50
		Spanish-to-English		
	BLEU	**chrF**	**TER**	**METEOR**
Max	39.24	0.56	43.79	0.63
Mean	35.98	0.55	41.06	0.61
Min	30.40	0.53	39.22	0.58
		Chinese-to-English		
	BLEU	**chrF**	**TER**	**METEOR**
Max	29.62	0.43	66.51	0.51
Mean	18.88	0.39	62.50	0.48
Min	13.69	0.37	55.81	0.44

3.5.2 Context-Sensitive Sentences

Sentences defined as “context-sensitive” are highly impacted by contextual changes. The translation of these sentences often fluctuates significantly with alterations in the preceding context, which could either enhance or degrade their translations. In Table 3.7, I detail the highest, average, and lowest scores of the context-sensitive sentences within a test set. These statistics are calculated across all fifty ShuffleTestsets. As is evident from Table 3.7, the translation quality of context-sensitive sentences is significantly influenced by their surrounding context.

Table 3.8: Context-sensitive sentence example for the three language pairs.

	Hindi-to-English	**Spanish-to-English**	**Chinese-to-English**
Source	इसके अतिरिक्त , जिस चिकित्सक ने शल्य चिकित्सा लेखों को तैयार किया था उससे एक गवाह के रूप में पूछ – ताछ नहीं की ग	hablaba de una forma muy jovial y sociable acerca de Yo @-@ Yo Ma y de Hillary Clinton y de cómo los Dodgers nunca llegarían a la Serie Mundial , todo debido a la traicionera ejecución del pasaje del primer violín en el último movimiento de la cuarta sinfonía de Beethoven .	他才听了贝多芬第一，第四交响乐到后天来自我介绍
Target	Furthermore, the doctor who prepared the surgery notes was not was not **examined** as a witness .	I was talking about a very <unk> and social way about Yo @-@ Yo @-@ Ma and Hillary Clinton , and how the Dodgers never came to the World Series , all because of the <unk> execution of the first violin on the final movement of **Beethoven** .	he had just heard a performance of Beethoven &apos;s First and Fourth **symphonies** , and came backstage and introduced himself .
shuffle1	Moreover, the doctor who had prepared the surgery article was not **examined** as a witness .	I was talking about a very <unk> and social way about Yo @-@ Yo @-@ Ma and Hillary Clinton , and how the Dodgers never would get to the World Series , all because of the <unk> of the first violin on the final movement of the fourth **symphony** .	and he listened to the first , the fourth **symphony** to himself .
shuffle2	Moreover, the doctor who had prepared the surgery articles was not **questioned** as a witness .	I was talking about a very <unk> and social way about &quot; Yo @-@ Yo @-@ Yo <unk> and Hillary Clinton , and how the Dodgers never would come to the World Series , because of the <unk> <unk> of the first violin on the final movement of the fourth	and he was listening to Beethoven &apos;s first , and he was about to introduce himself .
shuffle3	Further, a witness from the doctor who had prepared the surgery article was not **questioned** .	now , I &apos;ve got to mention that Nathaniel is denied treatment , because when he was treated <unk> , <unk> and wives , and , that scar has remained in it all of their life .	and he listened to Beethoven first , and he was about to himself .
Gloss	पूछताछ - Examined/Enquiry	Beethoven -Beethoven	**交响乐** - Symphony

Table 3.8 provides examples of context-sensitive sentences for the Hindi-to-

English, Spanish-to-English, [8] and Chinese-to-English language pairs for the source, target, and various shuffled iterations. For instance, the Hindi word "पूछताछ" (examined) is translated as "examined", "questioned", and again "questioned" in the shuffle1, shuffle2, and shuffle3 test sets, respectively. For the Spanish-to-English pair, the word "**Beethoven**" in the Spanish sentence is incorrectly translated to "symphony" in the shuffle2 set. In contrast, it is correctly translated in the shuffle1 set and omitted entirely in the shuffle3 set. For the Chinese-to-English pair, the Chinese word **交响乐** translates to "Symphony" in the target set. Interestingly, the MT system correctly translates this Chinese word in the shuffle1 set. However, in both shuffle2 and shuffle3 sets, the translations do not include the target equivalent for the Chinese word **交响乐**.

3.5.3 Characteristics of context-sensitive sentences

In my work to date, I have discovered a class of sentences that are highly sensitive to context. These sentences yield significantly different translations when exposed to varied prior contexts. However, it would be good to try and discover whether there are any linguistic characteristics of these context-sensitive sentences. There appear to be at least three ways of investigating this issue:

- Comparison with a human-based labelling performed independently of the proposed method.
- A translation experiment where it is shown that context-insensitive sentences do not lose translation quality using a context-independent system, while context-sensitive ones improve when using a context-aware system.
- Applying the methodology of researchers who have developed discourse-specific test sets for translation.

We leave options (i) and (ii) for future work, and concentrate here on option (iii). To understand the characteristics of context-sensitive sentences, I used the methodology

[8] After producing this table, in consultation with a native Spanish speaker, it transpires that the Spanish examples are suboptimal

proposed in Castilho et al. (2021). In this paper, the authors produce the DELA testset, a document-level corpus annotated in English with context-aware issues that arise when translating from English into Brazilian Portuguese, namely ellipsis, gender, lexical ambiguity, number, reference, and terminology, with six domains.

Taking their method as a starting point, I manually checked the context-sensitive class of sentences to find different linguistic issues related to context-aware MT. I evaluated the Hindi-to-English language pair due to my native proficiency in Hindi. In the future, I plan to extend this manual evaluation to Spanish and Chinese, employing the expertise of professional human translators for these languages. Table 3.9 display the results from my manual evaluation.

Category	**Occurrence**
Gender	18%
Ellipsis	8%
Reference	27%
Lexical Ambiguity	6%
Terminology	9%

Table 3.9: Characteristics of context-sensitive sentences

Castilho et al. (2021) use six linguistic criteria to generate their DELA testset. In Table 3.9, we use five of these criteria; we omitted Numbers as it was a little unclear as to what this referred to. For the Hindi-to-English examples, Table 3.9 shows that for the 85 sentences out of the entire testset of 1273 sentences, 18% contained gender-sensitive material, 8% contained elliptical material, 27% contained reference material, 6% contained lexical ambiguity and 9% contained terminology.

Perhaps it was not entirely unexpected that Reference proved to be the most prominent of the DELA criteria in the context-sensitive subset. In future work, we aim to investigate this in more detail, not just for Hindi, but for the other language pairs too, as well as conducting a more in-depth study of this whole issue using options (i) and (ii) above.

3.5.4 Context-Insensitive Sentences

Sentences classified as “context-insensitive” are those least impacted by the influence of the surrounding context. Such sentences maintain a consistent quality of translation, regardless of the contextual backdrop. Table 3.10 presents the highest, average, and lowest scores for this category of sentences within a test set. I observe from Table 3.10 that contextual information changes impact context-insensitive sentences less than their context-sensitive counterparts. It was found that the BLEU, TER, METEOR and chrF scores demonstrated a consistent pattern across various shuffles despite changes in context. Thus, we can say that context has little to no effect on the translation quality of context-insensitive sentences.

Table 3.10: Evaluation scores for the set of context-insensitive sentences.

			Hindi-to-English	
	BLEU	**chrF**	**TER**	**METEOR**
Max	33.64	0.55	45.22	0.70
Mean	33.64	0.55	45.21	0.69
Min	33.54	0.55	45.07	0.69
			Spanish-to-English	
	BLEU	**chrF**	**TER**	**METEOR**
Max	40.72	0.61	36.48	0.76
Mean	40.37	0.61	36.32	0.75
Min	40.02	0.60	36.12	0.75
			Chinese-to-English	
	BLEU	**chrF**	**TER**	**METEOR**
Max	16.72	0.39	66.96	0.56
Mean	16.55	0.39	66.45	0.56
Min	16.39	0.38	65.54	0.56

3.6 Conclusion

In this chapter, I discussed my work on document-level MT, investigating the influence of context in NMT. I conducted experiments for Hindi-to-English, Spanish-to-English, and Chinese-to-English language pairs. My results confirm the already established sensitivity of the HAN model to context. I found that the context-aware NMT system significantly outperforms the context-agnostic NMT system in terms of BLEU, chrF, TER and METEOR.

I observed that at the discourse level, the BLEU, chrF, TER, and METEOR scores of the context-aware NMT model across different shuffles are nearly identical for the three language pairs (cf. Table 3.5). My investigation revealed that this similarity is mainly due to the context-sensitive class of sentences having the most significant impact on translation quality. This led us to classify the test set sentences into three categories:(i) *context-sensitive* (ii) *normal* and (iii) *context-insensitive* sentences. The quality of translation for context-sensitive sentences is influenced by the presence or absence of correct contextual information, while context-insensitive sentences remain unaffected. I am convinced that investigating this issue, specifically identifying the correct context for context-sensitive sentences, could significantly impact discourse-level MT research.

In the next chapter, I further investigate the class of context-sensitive sentences. I aim to not only understand their inherent nature and behaviour but also uncover the source and extent of context they require. Specifically, I will focus on identifying the "context span" these MT systems can effectively utilise during translation. This exploration is fundamental for improving our MT systems translation accuracy and relevance.

Chapter 4

Understanding the Context Span in Document Level Translation

The previous chapter investigated the influence of context on document-level translation using the HAN model. I found that HAN performs better than baseline NMT models, disregarding document-level context. Furthermore, the results of my experiments led me to categorise sentences into two primary classes based on their sensitivity to context: those that are context-sensitive and those that are context-insensitive.

This chapter further examines the experiments conducted in the previous chapter on the HAN. My main objective in this study is to better understand the context-sensitive class of sentences. Considering the impact this class of sentences can have on the quality of translation, I aim to determine the appropriate context span to be considered during the translation process. I conducted experiments involving three morphologically diverse language pairs: Hindi-to-English, Spanish-to-English, and Chinese-to-English. The primary contributions of this study can be summarised as follows:

- I discuss the categorisation of context based on its impact on the translation quality of context-sensitive sentences.

- I provide recommendations to be considered when integrating document-level context into NMT models and also provide more information on the acceptable context span.

4.1 Related Work

NMT systems currently operate under locality assumptions, which indicate that these systems overlook the broader context of a document during the translation process. Nonetheless, considering the entire document could result in more accurate translations. This is hypothesised on the understanding that a document is not a random compilation of sentences but a well-structured narrative where context is extremely important. Furthermore, context is essential for addressing linguistic complexities such as deixis and coherence. Acknowledging the significance of context, there has been a recent surge of research dedicated to integrating context into NMT systems (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Voita et al., 2018; Miculicich et al., 2018; Maruf and Haffari, 2018; Tan et al., 2019; Maruf et al., 2019; Stojanovski and Fraser, 2021; Zhang et al., 2021). These studies have highlighted context's critical role in improving the quality of translations and addressing specific linguistic challenges. As of today, most document-level frameworks have been developed by incorporating context from various document portions into the model. I further try to understand the context integration into document-level NMT frameworks based on the locality assumptions: (i) whether they use the local, (ii) and/or the global document-level context in their MT systems.

The local context in a document typically refers to the immediately surrounding text of the specific sentence or phrase being translated. This could include the previous sentence, the following sentence, or a block of text surrounding the current sentence. The global context refers to the entire document or body of text. This context allows a MT system to understand the document's broader themes, topics, or narrative.

4.1.1 Using local context

This section explores various context-aware NMT systems, focusing on utilising local context and its span. In their work, Voita et al. (2018); Li et al. (2020) and Jiang et al. (2021) adapted the encoder to incorporate context into their respective MT models, operating with a fixed context window sourced from the previous sentence. In contrast, Jean et al. (2017) and Kuang and Xiong (2018) integrated context into the decoder, drawing from the preceding sentence with a context window of one. Building on these approaches, Yun et al. (2020) modified their encoder to include context but opted for a larger context window by incorporating the two preceding sentences. Interestingly, Yang et al. (2019) implemented capsule networks for context encoding, comparing their proposed method to other traditional encoding approaches, both with and without attention. Putting aside the typical practice of drawing context from preceding sentences, Ma et al. (2020) and Wong et al. (2020) considered context from both preceding and following sentences, using context windows of one and two, respectively.

All the studies discussed so far have integrated context into their NMT systems using context derived solely from the source sentences. I will explore studies incorporating context drawn from source and target sentences into their NMT systems. Tiedemann and Scherrer (2017) were the first to propose the inclusion of target sentences in the context. However, instead of altering the architecture, they opted for a data concatenation method, utilising one previous sentence as context for their experiment. Bawden et al. (2018) and Yamagishi and Komachi (2020) integrated context into their NMT systems by merging data from the context with the current source sentence to be translated by using various concatenation methods. Miculicich et al. (2018) and Xu et al. (2020b) used a context window of three to structure their context-aware MT systems. Additionally, they incorporated context into both the encoder and decoder side of the NMT systems.

A few studies considered context from the following sentences of the source language in addition to the source and target side contexts from the previous sentences

(Agrawal et al., 2018; Scherrer et al., 2019; Zheng et al., 2021). Agrawal et al. (2018) considered up to three previous sentences and one following sentence from the source side. For the target language, they used up to two previous sentences as the context. Scherrer et al. (2019) carried out experiments with variable or random context. Finally, Zheng et al. (2021) considered a context span of twenty sentences while integrating context into their models.

4.1.2 Using global context

This section discusses some studies that have attempted to incorporate global context into their MT systems. In contrast to the local context, only a few studies have tried to integrate global context into their models. All the sentences within the document are considered for context in the global context.

Macé and Servan (2019) proposed an innovative approach for incorporating contextual information into NMT. They introduced a unique method integrating the source language context into their models. Their method effectively encapsulates the entire document, accurately defining its boundaries and giving attention to each word.

Tan et al. (2019) implemented a hierarchical architecture for incorporating global document context into document-level MT. Their approach uses a sentence encoder and a document encoder. The sentence encoder is designed to capture local dependencies, while the document encoder manages global dependencies. This methodology was successful in reducing mistranslations and was able to generate context-specific translations by considering the context of each word.

Maruf and Haffari (2018) and Maruf et al. (2019) introduced an NMT architecture designed explicitly for document-level translation. This architecture incorporated memory networks, a neural network that uses external memories to capture the global context. The architecture featured two memory components: one for the source language and one for the target language. This design allowed the system to consider contextual information from both languages. Experimental results suggested that this approach was effective in leveraging the document-level context.

Though the experiments conducted by Miculicich et al. (2018) demonstrate that context can help to improve translation quality, the reasons for the better performance of their context-aware model over context-independent ones remain unclear. In my previous chapter, I conducted a study to uncover the situations and underlying reasons for the improvement in translation quality when the context is incorporated in document-level NMT, specifically through HAN. I discovered that some sentences are particularly sensitive to the context. Despite my findings indicating that these context-sensitive sentences help improve translation quality, further examination is necessary to understand the role of the context span in translation. To date, few studies have explored the role of context span in document-level MT. An attempt to explore this topic was made by Castilho et al. (2020). They conducted a research study examining 300 sentences drawn from three distinct domains: reviews, subtitles, and literature. The primary aim of the study was to determine the extent of context needed for accurate translation. Through an in-depth analysis of the translated sentences, the researchers deduced that an extensive context was crucial to understanding and evaluating multiple sentences within a document. In this chapter, I further discuss the context-sensitive class of sentences using HAN. In addition to this, I detail the context span that the context-sensitive class of sentences could consider before performing the translation. As pointed out above, I again carried out my experiment for three pairs of morphologically diverse languages: Hindi-to-English, Spanish-to-English, and Chinese-to-English.

4.2 Experiments and Results

As we saw in Section 5 in Chapter 3, sentences classified as context-sensitive exhibit the highest degree of sensitivity to context and demonstrate a significant divergence in their translations when the previous context is altered. My research findings

Table 4.1: Example for context-sensitive Sentences

	Hindi-to-English	Spanish-to-English	Chinese-to-English
Source	प्रतिवादी , यह अभिकथन करते हैं कि वादी , सिनेमा पाराडिसो के नाम और अभिनाम कि संस्था का स्वत्वधारी / भागीदार है .	ahora , debo mencionar que Nathaniel se niega al tratamiento , porque cuando fue tratado emplearon terapia de choque , Toracina y esposas , y esa cicatriz ha permanecido en él toda su vida .	这正是我们创造音乐的原因我们用我们每个人都拥有的一种内在我们的最根本的核心我们的感情通过我们的艺术的镜头通过我们的创造力，将我们的情感塑造成现实
Target	The Defendant , allege the **Plaintiffs** , is the Proprietor / partner of the concern by the name and style of Cinema Paradiso .	now , I should mention that Nathaniel refuses treatment because when he was treated it was with shock therapy and thorazine and **handcuffs** , and that scar has stayed with him for his entire life .	this was the very reason why we made music , that we take something that exists within all of us at our very fundamental core , our emotions , and through our **artistic lens** , through our creativity , we &apos;re able to shape those emotions into reality .
shuffle1	<unk> , the <unk> ensure that the **plaintiff** , the cinema is the partner / partner of the institute.	now , I &apos;ve got to mention that Nathaniel is denied treatment , because when he was treated <unk> , <unk> and **wives** , and , that scar has remained in his entire life .	and that &apos;s why we create music , and we use each of us to have an internal inner core of ourselves , the core of our emotions **through our art** .
shuffle2	<unk> , assure that the **litigants** , the name of the cinema company and the <unk> / partner of the institute .	now , I &apos;ve got to mention that Nathaniel is denied treatment , because when he was treated <unk> , <unk> and **wife** , and that scar has remained in his entire life .	and that &apos;s why we create music , and we use each of us to have an internal inner core of ourselves , the core of our emotions through our **artistic lens** , to make our emotions , to make our feelings of reality .
shuffle3	<unk> , the <unk> ensure that the <unk> , the cinema is the <unk> / partner of the institution.	now , I &apos;ve got to mention that Nathaniel is denied treatment , because when he was treated <unk> , <unk> and **cuffs** , and , that scar has remained in it all of his life .	and that &apos;s why we create music , and we use each of us to have an internal inner core of ourselves , our feelings , through our art , through our **vision of our art** , to make our feelings of reality ..
Gloss	वादी - plaintiff	esposas - handcuffs	**艺术镜头** - artistic lens

indicate that context can have a profound impact on the quality of translations, either positively or negatively, depending on the specific circumstances. Therefore, providing appropriate context to these context-sensitive sentences can prove beneficial and result in an overall improvement in the translation quality of the entire document. In Table 4.1, we can observe how changes in context affect the quality of translation of context-sensitive sentences. I use these context-sensitive sentences further to understand the role of context span in document-level translation.

From my experiments in the previous chapter, I observed that a set of specific context-sensitive sentences can improve or deteriorate the overall quality of the translation of a document. Therefore, I try to understand where the context comes from in the document (i.e. context span) so that I can provide the right context to the context-sensitive class of sentences, thereby improving its translation quality. My experimental setups for HAN use the previous three sentences as context. Additionally, the shuffling of test sets can provide context to a sentence from different parts of the document. I use this idea to better understand the role of context span in translation. The outcome of this investigation can be crucial for improving the current state-of-the-art NMT models. I report my observations on translations of the Hindi-to-English, Spanish-to-English and Chinese-to-English tasks.

4.2.1 Context Span Score

In MT, “context span” is the term usually used to describe the number of words or tokens the model takes from the source sentence to generate the target sentence. It can be seen as the “window” that the MT model uses to interpret the source sentence. The length of this context span can considerably impact the performance of NMT models. When the context span is too short, the model may overlook important information, resulting in incomplete or translations with errors. On the contrary, when the context span is too long, the model may be burdened with irrelevant information, resulting in complex or nonsensical translations. An appropriate context span selection primarily depends on the specific NMT architecture and the particular task. For instance, models that leverage attention mechanisms (Bahdanau et al., 2015; Sukhbaatar et al., 2019) for context might be more effective with longer context spans, as they can dynamically focus on relevant sections of the source side data. Conversely, models using fixed context spans (Scherrer et al., 2019) may perform better with shorter spans. Therefore, a clear understanding of the context is crucial when conducting document-level translation.

In order to better understand context span usage in document-level translation, I propose a distance-based metric to quantify the context span and define it in (4). Context Span Score (CSS) is a measure based on the proximity (in terms of document position) of the context sentence to the translated sentence in the document. A particular context sentence with a high CSS implies that it is closely located to the translated sentence. Conversely, a context sentence with a lower CSS suggests that it is positioned farther away from the translated sentence.

$$\mathrm{CSS}_j = \sum_{n=1}^{N} \frac{1}{|\mathrm{D}_i - \mathrm{D}_j|} \quad \text{where } i \neq j \tag{4}$$

CSS_j represents context span score (CSS) that I use in my analysis to quantify context,

D_i represents the original document index of the sentence being translated,

D_j represents the original document index of the context sentences,

i represents the document index of the sentence being translated,

j represents the document index of the context sentences,

N context span of the document-level model.

Let us understand CSS with the help of an example. Suppose I have a document consisting of ten sentences numbered from 1 to 10. If my task is to translate sentence number 5, then the context spans would include sentences 1 to 4, which precede it, and sentences 6 to 10, which follow it. Let us compute CSS for sentence 5 with context sentences (3,1) and (7,9). The rate of change of the CSS score is hyperbolic as I move away from the target sentence. Now, according to the formula (4):

```
CSS_1 = 1 / (5 - 1)

      = 0.25

CSS_3 = 1 / (5 - 3)

      = 0.5
```

```
CSS_7 = 1 / (5 - 7)
      = 0.5

CSS_9 = 1 / (5 - 9)
      = 0.25
```

From these calculations, We can see that sentences 3 and 7 (with CSS=0.5) are closer to the target sentence (sentence 5) than sentences 1 and 9 (with CSS=0.25), which is farther away.

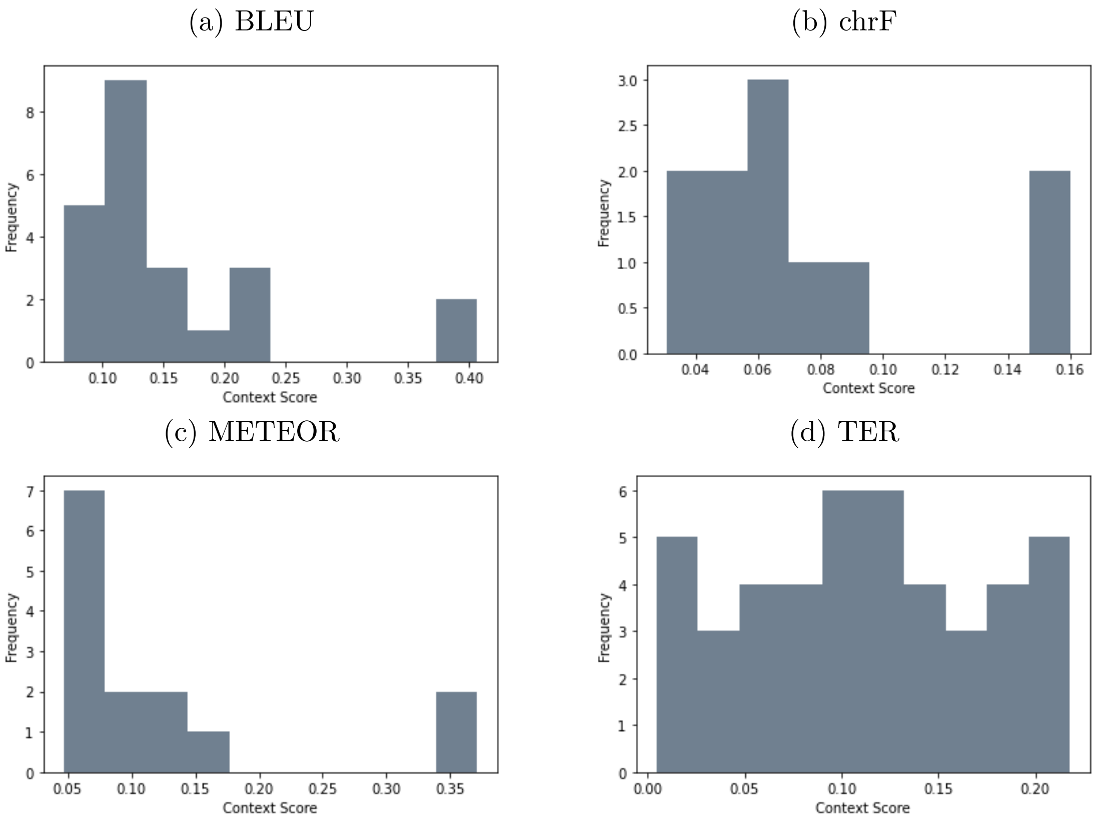
4.2.2 Context span for document-level MT

In my previous chapter, I discussed how contextual information impacts the quality of translation either positively or negatively. I have already classified these as context-sensitive sentences. In this study, my primary focus is to examine the effect of context span on the translation of these context-sensitive sentences. Specifically, I explored two types of contexts and their origins: (i) those that improve translation quality and (ii) those that cause deterioration. My experimental setup remained consistent as I continued to my investigation for the Hindi-to-English, Spanish-to-English, and Chinese-to-English tasks. This section discusses in detail my experimental results and includes related discussions.

Hindi-to-English

In this section, I discuss the impact of context span on the translation performance for the Hindi-to-English task. We first took those sentences of documents (context) into account for the investigation that caused improvement in translation quality. Accordingly, I measured CSSs for such context-sensitive sentences (cf. (4)). Note that the improvements in translation quality are measured using automatic evaluation metrics, i.e. BLEU, chrF, TER and METEOR. I then produced histogram distributions of the context-sensitive sentences over the CSSs for the metrics, which are shown in Figure 4.1. Similarly, I performed the same analysis for context-sensitive sentences that showed deterioration in translation quality. The histogram distribution for this is shown in Figure 4.2.

Figure 4.1: Distribution of context-sensitive sentences over context span scores in the Hindi-to-English task for improvement in translation quality.



(a) BLEU

(b) chrF

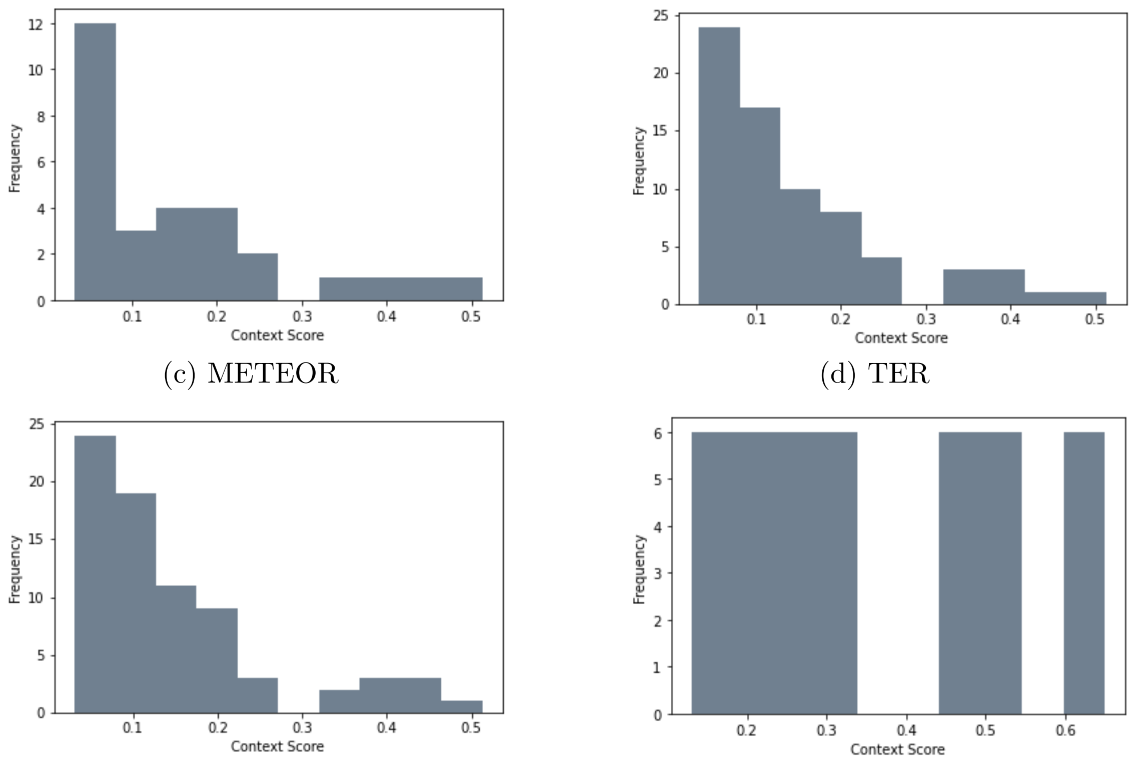
(c) METEOR

(d) TER

In Figures 4.1 and 4.2, the x-axis of the histogram plot represents context scores, which measure the relevance of the provided context to the sentence being translated, while the y-axis represents the frequency of occurrence of the context-sensitive sentences for each CSS, illustrating how often a particular score appears in the data. I examined the two distributions for context-span of sentences in Hindi-to-English translation tasks, comparing those that improve translation quality (Figure 4.1) and those that worsen it (Figure 4.2). I obtained a p-value of less than 0.05 upon implementing the Kolmogorov-Smirnov test for my analysis. This low p-value indicates that the observed differences between the two distributions are statistically significant.

As can be seen from the histogram plot in Figure 4.1, translation quality tends to improve when the context is farther away from the sentence being translated.

Figure 4.2: Distribution of context-sensitive sentences over context span scores in the Hindi-to-English task for deterioration in translation quality.


(a) BLEU

(b) chrF

(c) METEOR

(d) TER

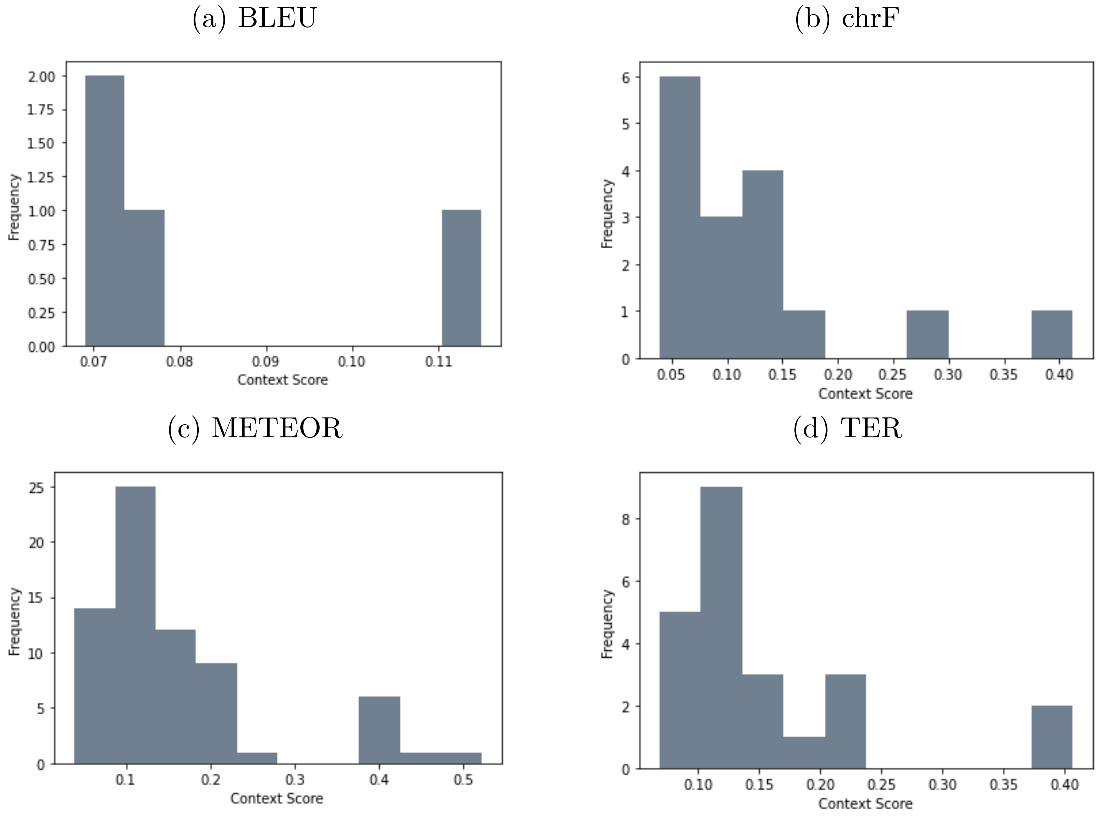
This is indicated by low CSSs, suggesting that a less directly related context may still contain valuable information for the translation process. This finding shows us the importance of considering the broader context span when translating context-sensitive sentences in order to achieve improved translation quality.

Interestingly, in Figure 4.2, I observe that the translation quality deteriorates when the context is farther away from the sentence being translated (TER does not hold this observation). This is indicated by low context scores, suggesting that a distant context would unlikely contain valuable information for the translation process. My findings suggest that a more distant context may not always be beneficial, and in some cases, local context could be equally important as the more distant one.

Spanish-to-English

This section presents the results and a discussion of the Spanish-to-English translation task. Like the Hindi-to-English task, I calculated CSSs for the context-sensitive sentences of the Spanish-to-English translation task when I see an improvement in translation quality. I produce the histogram distributions of the context-sensitive sentences over the CSSs. The histogram plots across metrics (BLEU, METEOR, TER and chrF) are shown in Figure 4.3. As above, I produced histogram distributions of the context-sensitive sentences over CSSs when the translation scores were measured in BLEU, METEOR, TER and chrF drops. The histogram plots are shown in Figure 4.4. Similar to the Hindi-to-English experiments, I conducted statistical significance tests and found that the differences in distributions are statistically significant.

Figure 4.3: Distribution of context-sensitive sentences over context span scores in the Spanish-to-English translation task for improvement in translation quality.


(a) BLEU

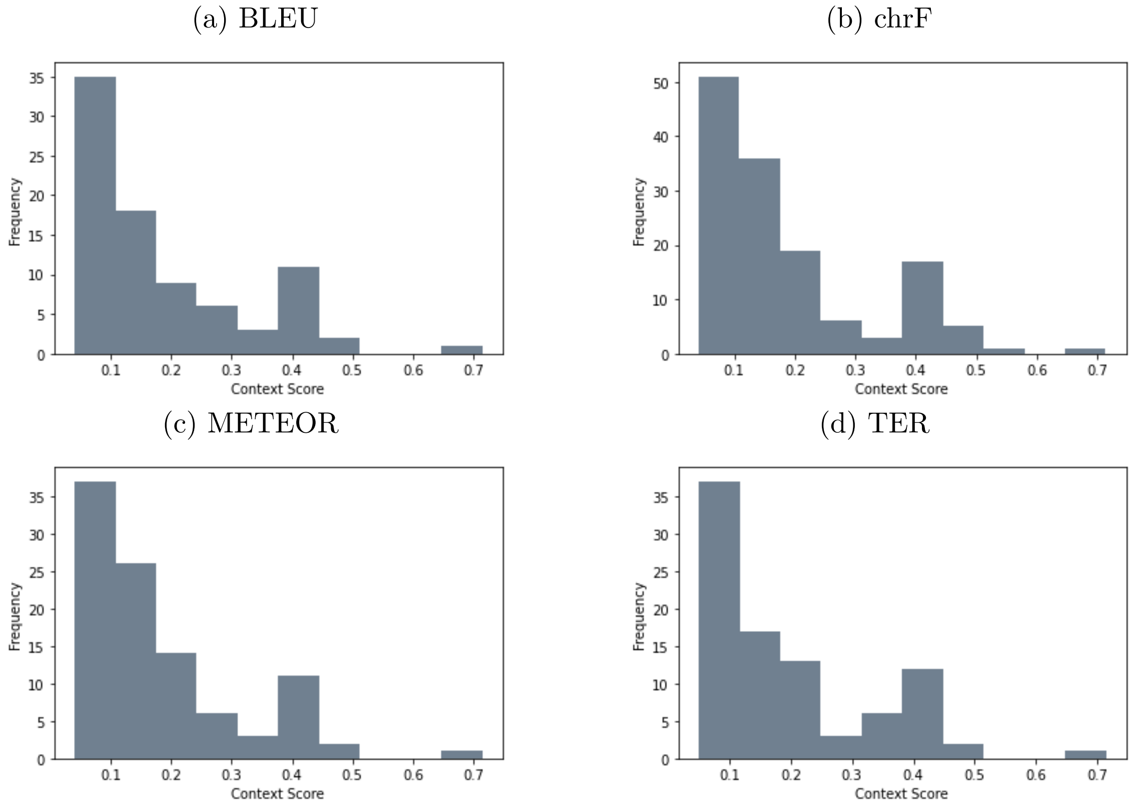
(b) chrF

(c) METEOR

(d) TER

I see from the plots in Figure 4.3 that translation quality tends to improve more when the context is farther away from the sentence being translated. This finding is similar to that of the Hindi-to-English translation task (cf. Figure 4.1). Similarly, I see from the plot in Figure 4.4 that translation quality tends to drop more when the context is farther away from the sentence being translated. This finding is similar

Figure 4.4: Distribution of context-sensitive sentences over context span scores in the Spanish-to-English task for deterioration in translation quality.



(a) BLEU

(b) chrF

(c) METEOR

(d) TER

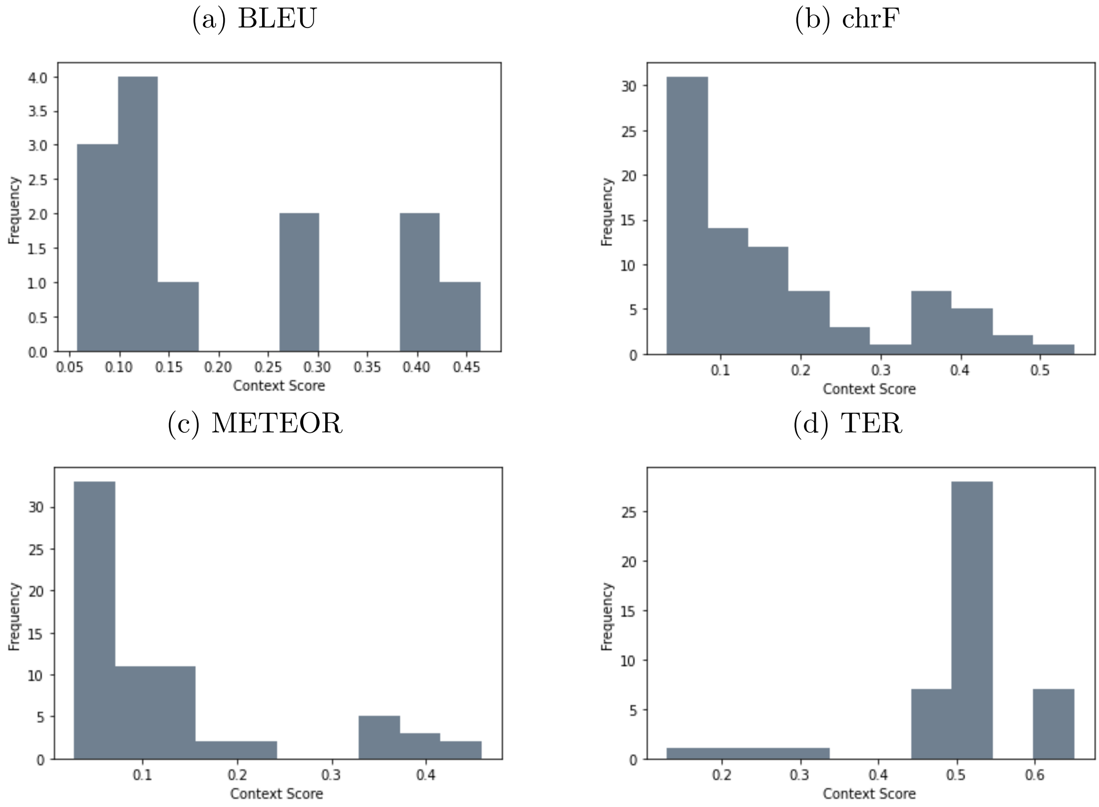
to that of the Hindi-to-English translation task (cf. Figure 4.2).

Chinese-to-English

As for Chinese-to-English, I similarly measured CSSs for the context-sensitive sentences when an improvement in translation quality is seen. The histogram distributions over the CSSs for BLEU, chrF, TER and METEOR are shown in Figure 4.5. Like Hindi-to-English and Spanish-to-English, I also wanted to see from where the context is coming when the translation quality of the context-sensitive sentences deteriorates. To this end, I measured CSSs for them when I see a drop in automatic evaluation scores. The distribution of the context-sensitive sentences over CSSs is shown in Figure 4.6. Similar to the Hind-to-English experiments, I conducted statistical significance tests and found that the differences in distributions are statistically significant.

When I compare these histogram plots in Figure 4.5 with those of the Hindi- and

Figure 4.5: Distribution of context-sensitive sentences over context span scores in the Chinese-to-English task for improvement in translation quality



(a) BLEU

(b) chrF

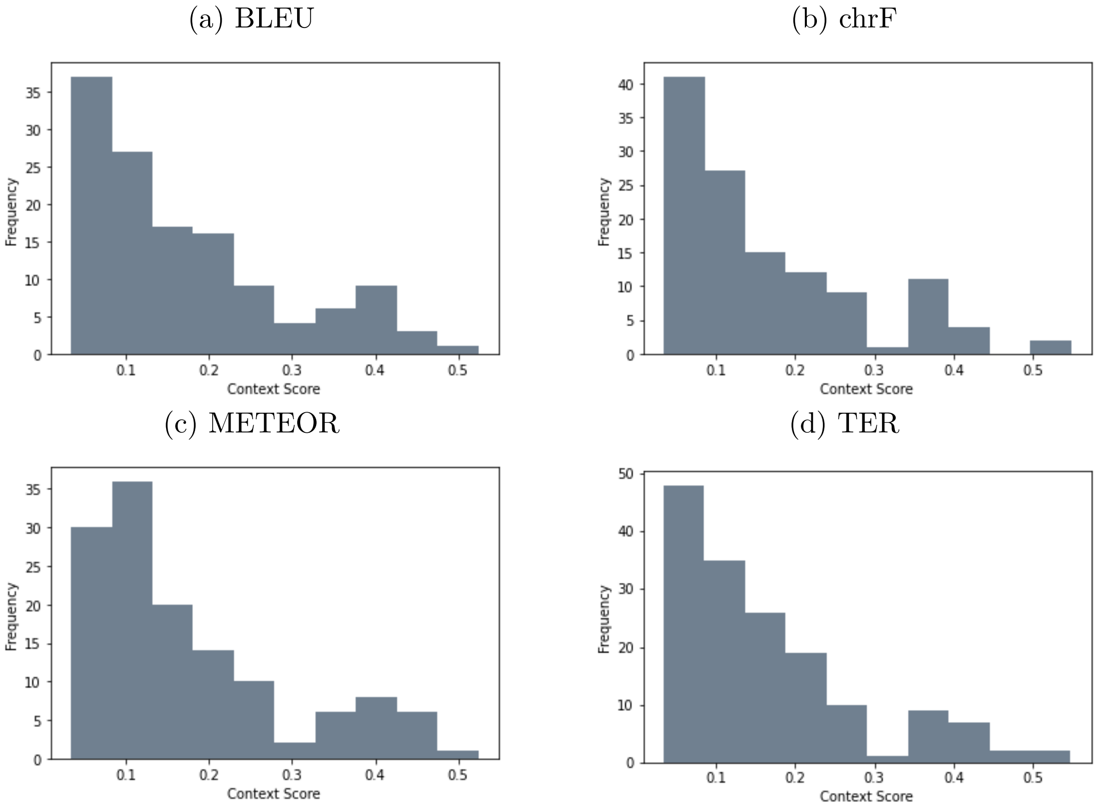
(c) METEOR

(d) TER

Spanish-to-English tasks, I see that they are quite similar to each other. In other words, translation quality tends to improve more when the context is farther away from the sentence being translated irrespective of the translation tasks. Interestingly, the TER metric shows the nearer context to be more useful, which is contrary to my findings on other metrics.

I see from Figure 4.6 that the characteristics of the histogram plots are quite similar to the plots of Figures 4.4 and 4.2. More specifically, histogram bars are high at the beginning (over low CSS bins) and low at the end (over high CSS bins). This indicates that contexts that are far away from the sentence being translated lead to deterioration in translation quality. Clearly, these findings are identical to those of the Hind-to-English and Spanish-to-English translation tasks (cf. Sections 4.2.2 and 4.2.2).

Figure 4.6: Distribution of context-sensitive sentences over context span scores in the Chinese-to-English task for deterioration in translation quality.



4.2.3 Distribution of Context in Document

In Section 4.2.2, I presented a number of graphs and discussed findings from the graphs. I came up with some key findings. For example, distant context can positively impact document-level MT. I also found that contexts that are far away from the sentence being translated may sometimes lead to deterioration in translation quality. To gain a better understanding of the nature of the broader context that helps improve or deteriorate the translation quality, I carried out further analysis. For this, I selected sentences from the context-sensitive class. I present the outcome of the analysis for the three translation tasks below.

I first selected a source Hindi sentence from the Hindi-to-English translation task and plotted the sentence's context scores in relation to its translation quality. This graph is shown in Figure 4.7. The y-axis of the graph represents the CSS of the Hindi sentence for the provided context. The x-axis of Figure 4.7 represents the index of sentences that are part of the context, with the index of the Hindi sentence being translated as 30. The graph's colour-coded points represent the quality of the HAN's English translations of the Hindi source sentence compared to that of the baseline MT system. Red points indicate a positive result (HAN > Baseline), where HAN outperforms the baseline. Green points indicate a negative result (HAN < Baseline), with HAN underperforming compared to the baseline system. Light green points signify a neutral outcome (HAN == Baseline), where HAN produces translations of equivalent quality to those generated by the baseline system.

Figure 4.7: Distribution of context scores for context-sensitive sentence for Hindi-to-English.

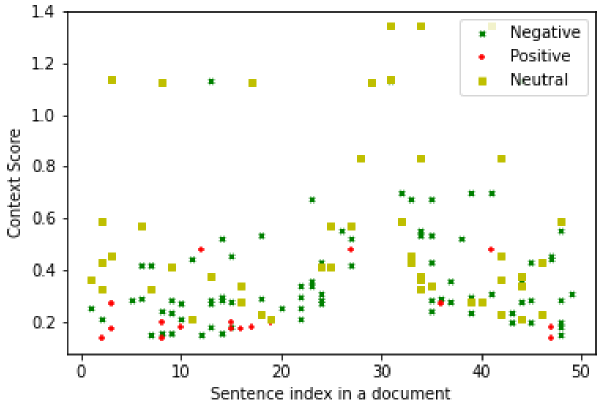
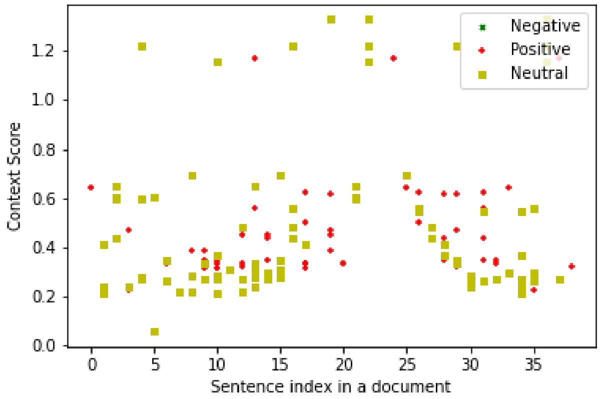
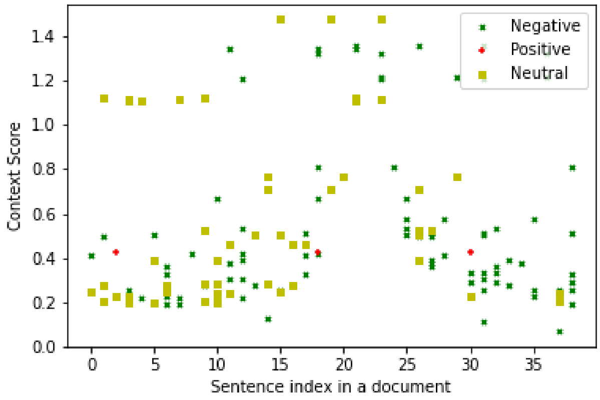


Figure 4.8: Distribution of context scores for context-sensitive sentence for Spanish-to-English.



We can see from Figure 4.7 that nearly all red dots are below 0.3. This emphasises that the sentences providing context that helps improve the translation quality mainly come from farther parts of the document. In contrast, the green dots

Figure 4.9: Distribution of context scores for context-sensitive sentence for Chinese-to-English.



appear to be fairly equally distributed on the x-y plane. This indicates that the sentences providing context that lead to deterioration in translation quality usually come from all parts of the document. I inspected the scatter plots that were prepared for several context-sensitive sentences of this translation task and observed that their characteristics are similar to that shown in Figure 4.7 in many instances. Accordingly, these broadly similar plots are excluded from the thesis as they do not provide much additional benefit to the discussion.

As in Hindi-to-English, I selected a Spanish context-sensitive sentence and a Chinese context-sensitive sentence and produced scatter plots for them in order to see the relatedness of the sentences' CSSs, translation quality and distance of the sentences appearing in the context from the sentences being translated. The plots are shown in Figures 4.8 and 4.9 for the Spanish-to-English and Chinese-to-English tasks, respectively. Note that the indices of the Spanish and Hindi sentences are 30 and 7, respectively. I see from the scatter plots that the CSSs of the red dots are usually lower than 0.6 and 0.4 for the Spanish-to-English and Chinese-to-English tasks, respectively. This again shows that the sentences of the context that causes improvement in translation quality mainly come from areas of the document that is far away from the sentence being translated.

4.2.4 Manual Evaluation

I manually looked at the sentences that I used for producing distribution of sentences of the contextual information (cf. Section 4.2.3), context itself to the sentences being translated by HAN and their translations. In Table 4.2, I show an example translation for Hindi-to-English. I see from Table 4.2 that a Hindi word प्रतिवाद appears in source sentence. The same word is seen in the second sentence of the context. I believe that this lexical overlap may lead to an improvement in translation quality. I also looked at the other translations when improvements in translation quality are seen. I observed that an improvement in translation quality is seen when there generally holds a relationship between the context-sensitive sentence and at least one of the sentences of the context. More specifically, I observed that whenever there is an improvement in translation quality, there is a high lexical overlap between the source sentence being translated and one of the sentences of the context used.

Table 4.2: Example for context-sensitive sentence showing lexical overlap.

		Doc index
Context1	इसलिए , यह माना गया कि अदालत में पहली बार पहचान पर भरोसा करना ठीक नहीं होगा .	8
Context2	याची ने प्रतिवाद किया कि याचियों द्वारा पाए गए वाउचर वर्ष 1988 के लिए रु . 1206 / – की राशि के लिए हैं और क्रमशः वर्ष 1989 के लिए रु . 924 / – की एक अतिरिक्त राशि और 1990 के लिए रु . 1672 / – का एक तीसरा वाउचर है .	47
Context3	इसके अतिरिक्त , जिस चिकित्सक ने शल्य चिकित्सा लेखों को तैयार किया था उससे एक गवाह के रूप में पूछ – ताछ नहीं की गई .	2
Source	प्रतिवादी , यह अभिकथन करते हैं कि वादी , सिनेमा पाराडिसो के नाम और अभिनाम कि संस्था का स्वत्वधारी / भागीदार है .	30
Reference	The Defendant , allege the Plaintiffs , is the Proprietor / partner of the concern by the name and style of Cinema Paradiso .	
MT	Defendant , the <unk> ensure that the plaintiff , the cinema is the partner / partner of the institute .	

As for Spanish-to-English, in Table 4.3, I present an example translation of a context-sensitive sentence that I considered for the scatter plot (see Figure 4.8; cf. Section 4.2.3). I can see from Table 4.3 that the Spanish word *inspira* appears in the second sentence of the context. I also see that the word *inspira* appears in the source Spanish sentence.

Table 4.3: Example for context-sensitive sentence showing lexical overlap

		Doc index
Context1	y entendí que él no sólo tenía un conocimiento enciclopédico de música sino que se identificaba con esta música a un nivel personal .	21
Context2	Durante la charla TED, el orador compartió una historia conmovedora que inspira a las personas a perseguir sus sueños a pesar de los desafíos que enfrenten.	23
Context3	hablaba de una forma muy jovial y sociable acerca de Yo @-@ Yo Ma y de Hillary Clinton y de cómo los Dodgers nunca llegarían a la Serie Mundial , todo debido a la traicionera ejecución del pasaje del primer violín en el último movimiento de la cuarta sinfonía de Beethoven .	7
Source	y la realidad de esa expresión nos alcanza a todos , nos mueve , nos inspira y nos une .	30
Reference	and the reality of that expression reaches all of us and moves us , inspires and unites us .	
MT	and the reality of that expression gives us all , moves us , inspire us and binds us .	

For the Chinese-to-English translation task, I similarly show an example translation of the context-sensitive Chinese sentence that I used to show the scatter plot in Section 4.2.3 (cf. Figure 4.9) in Table 4.4. I see from the table that the Chinese word **貝多芬** is part of the first sentence of the context. The word **貝多芬** also appears in the source Chinese sentence.

4.2.5 Relationship Between Context and Source Sentence

Since my core interest lies in identifying reasons why translation quality of context-sensitive sentences improves or deteriorates, I further tried to identify the relationship between the context-sensitive sentences and the context used that leads to improving or deteriorating translation quality. Considering the findings from my manual evaluation (cf. Section 4.2.4), I further measured the similarity between the sentences of the context used and the context-sensitive sentences. For this, I used sentence-Transformer (Reimers and Gurevych, 2019). sentence-Transformers

Table 4.4: Example for context-sensitive sentence showing lexical overlap

		Doc index
Context1	那些日子，管弦樂隊的．樂器包括小提琴和貝多芬	28
Context2	他的狂躁愤怒也转化成理解安静的好奇，和优雅．	18
Context3	结果呢，他的精神分裂症现在变得更容易发作最糟糕的表现是他发作后会消失几天在大街上流浪暴露着内心的恐惧，让心灵的煎熬在身上释放	10
Source	他说话很愉快，很合群提到了马友友和希拉里．克林顿道奇队不可能进入世界联赛而这都是因为最后一刻贝多芬第四交响乐中开始的那段变幻莫测的小提琴演奏起的作用	7
Reference	he was speaking in a very jovial and gregarious way about Yo @-@ Yo Ma and Hillary Clinton and how the Dodgers were never going to make the World Series , all because of the treacherous first violin passage work in the last movement of Beethoven &apos;s Fourth Symphony .	
MT	and he was talking very happy , very <unk> , and he talked about the <unk> and Hillary Clinton , the Dodgers , who were not going to go into the world , and that &apos;s because the last moment that Beethoven &apos;s <unk> started to play the role of	

are characterised by their ability to create contextually rich sentence embeddings. These models leverage the transformer architecture, renowned for its bidirectional context understanding, to process the entire sentence rather than focusing on individual words. These scores are produced for all context-sensitive sentences. My findings are reported in Table 4.5. Table 4.5 shows mean similarity scores for improvement and deterioration in translation quality for three language-pairs.

The mean similarity score is found to be 0.42 for Hindi-to-English when there is an improvement in translation quality. As for Spanish-to-English and Chinese-to-English, the mean similarity scores between sentences of the context used and the sentences being translated are found to be 0.23 and 0.29, respectively when there is an improvement in translation quality. I notice that across language pairs, similarity scores are lower when the translation improves. Conversely, similarity scores are comparatively higher when the translation quality deteriorates. This suggests that more similar context sentences might not always be helpful in enhancing translation.

	Translation Improvement	Translation Deterioration
Hindi-to-English	0.42	0.45
Spanish-to-English	0.23	0.26
Chinese-to-English	0.29	0.32

Table 4.5: Mean similarity scores for the context-sensitive sentences and the sentences of the context used during translation.

4.3 Discussion

In this study, I aimed at investigating the role of context in document-level NMT systems. As pointed out above, I made use of HAN for my investigation and considered three sentences as context. The sentences of the context are sampled randomly from the different positions of the document. The idea is to understand the role and origin of context in document-level NMT systems (in my case, HAN).

I proposed a metric that produces scores given the relative distance between sentences that form context span and a source sentence that is to be translated (cf. Section 4.2.1). This metric gives more weight to a context-sensitive sentence which is nearby the sentences of the context used and less weight to a context-sensitive sentence which is farther away the sentences of the context used.

My findings (cf. Section 4.2.2) suggest that incorporating document-level context into NMT models can lead to performance improvement, particularly when a broader range of contextual factors are considered. I further carried out sentence-level analysis (cf. Section 4.2.3) by picking up a specific context-sensitive sentence from each of the translation tasks. I obtained similar observations this time too. I found that specific sentences of the context that are far away from the sentence being translated generally help improve the translation.

I also conducted a thorough manual analysis by looking at context-sensitive sentences, context provided during translation and their target translations. I found that there holds a relationship or pattern between the sentence being translated and sentences of a context used (cf. Section 4.2.4). I further studied this relationship by computing the similarity between sentences used as a context and source sentences using sentence-Transformer. I found that less similarity between the context and the sentence being translated helps to improve translation.

Most document-level NMT systems that exist today often consider only a limited number of sentences as context, which is an inadequate approach to fully comprehending the text. By only focusing on n sentences, these systems inherently overlook the importance of broader context, which can be crucial for a more accurate understanding of the subject matter. My experiments demonstrate that expanding the scope to include farther context not only enhances the semantic understanding of the text but also provides a more holistic interpretation, which ultimately leads to better performance and more beneficial outcomes.

4.4 Conclusion

In this study, I investigated how context affects NMT. Specifically, I examined the context-sensitive class of sentences and the context span that this class of sentences could utilise. I conducted my experiments for Hindi-to-English, Spanish-to-English, and Chinese-to-English language pair. I observed that document-level MT systems benefit from incorporating a broader context. Similar findings were also seen in my sentence-level analysis. Furthermore, I found that having the context similarity to the source sentence helps improve the translation. My findings from this study offer a new perspective on how context integration into NMT may be approached. I also provide recommendations that can be considered when developing document-level NMT systems. In the future, I aim to conduct experiments with document-level systems that leverage large language models (LLMs) as their foundation. By harnessing the impressive capabilities of LLMs, I hope to advance my understanding of context and semantics within texts, thus improving the overall effectiveness of document-level analysis. This innovative approach has the potential to revolutionise the way we process and interpret vast amounts of information, leading to more accurate and efficient outcomes across various applications and industries.

Chapter 5

Terminology-Aware Mining for Improving Terminology Translation

5.1 Introduction

Terminology translation is the process of translating specialised terms from one language to another while preserving their specific meanings within a particular field or area of expertise. It involves a deep understanding of both the source and target languages and expertise in the relevant domain to ensure that the translated terms accurately convey the same concepts and ideas as in the original language.

This is an essential aspect of translation, particularly in specialised fields such as law, medicine, business, and technology, where accurate translation of specific jargon, abbreviations, and terms is essential. In many cases, these terms cannot be translated directly or literally, as they may have different meanings or may not even exist as lexical items *per se* in the target language. Terminology translation is often supported by tools like terminology databases or glossaries, which contain a list of source language terms and their specified translations in the target language. These resources help ensure consistency in translating specific terms across different texts and projects, but they are hard to obtain, making it challenging to consistently produce correct term translations when they are unavailable.

In Chapters 3 and 4, in order to address my research questions RQ1 and RQ2, I investigated the influence of context in document translation, and discovered the sensitivity of sentence to context. This led me to categorise the sentences in the test set accordingly. Interestingly as a secondary observation, I observed that contextual information impacts the translation of domain-terms (cf. Section 3.5.2, Table 3.8) within the context-sensitive set of sentences. Similar observations were made in Chapter 4 where I examined the span of context in document-level translation. During the manual evaluation (cf. Section 4.2.4), as a secondary observation, I noted that lexical overlap was influencing translation of domain-terms (cf. Tables 4.2, 4.3, and 4.4). This observation was consistent across three language pairs: Hindi-to-English, Spanish-to-English, and Chinese-to-English. The observations in Chapters 3 and 4 clearly indicate the influence of context on domain-term translation, which prompted me to investigate terminology translation using HAN, which I formalised as a new research question RQ3.

In recent times, LLMs (Devlin et al., 2019; Brown et al., 2020; Liu et al., 2020) have gained significant attention due to their remarkable performance in various NLP tasks. These models have proven effective in diverse applications, from information extraction to text generation. As a result, the NLP community is increasingly focused on harnessing their potential. This prompted me to compare the terminology translation capabilities of LLMs to that of HAN.

One of the key advantages of LLMs over document-level systems like HAN is that they often require smaller amounts of data for domain adaptation (by fine-tuning) compared to traditional machine learning models (built from scratch) (Devlin et al., 2019). Fine-tuning standard NMT models usually requires specialised domain data for translating domain text (Luong and Manning, 2015; Huang et al., 2023; Hung et al., 2023). By leveraging pre-trained knowledge, LLMs can be fine-tuned on specific domains with relatively limited amounts of data, making them a valuable resource for addressing domain-specific challenges in NLP. However, despite significant improvements in translation quality, NMT systems still struggle with translating terminology[1] (Alam et al., 2021). Even domain-adapted models are found to have difficulty with accurately translating domain-specific terminology (Sato et al., 2020).

This chapter discusses my experiments on terminology translation using LLMs. I conducted two experiments based on the core idea of "terminology-aware mining". In the first experiment, I applied an "on-the-fly adaptation" methodology, a dynamic approach that adapts the model instance-by-instance, incorporating domain-specific terms. My second experiment exploits the capabilities of LLMs to generate synthetic data based on domain terms. I then use this synthetic data to further adapt my model. My experimental findings suggest that terminology-aware mining effectively improves the model's understanding and translation proficiency of domain-specific terms.

I tested our approach on the French-to-English terminology translation task[2] for COVID-19 domain data. My findings show that the proposed approach helps improve terminology translation in COVID-19 domain data.

The rest of this chapter is organised as follows. Section 5.2 discusses work related to my study. Section 5.3 details the data I used in my experiments. I compare the performance of HAN with LLMs in Section 5.4. My methodologies are described in Section 5.5.1 and 5.7.1. My NMT models are explained in Section 5.6 and 5.7.2. The experiments and results are covered in Section 5.6.1 and 5.7.3. Finally, Section 5.8 summarises my work and discusses possible future research ideas.

5.2 Related Work

Although NMT models have shown significant improvement in many translation tasks, as pointed out above, translating terms of specific domains, such as medical

[1] SMT systems could do it better, as Moses has specific ways in which termbanks could be accessed and the translations contained therein enforced.

[2] https://www.statmt.org/wmt21/terminology-task.html

or technical (Ao and Acharya, 2021), still remains a challenge. This section highlights the foremost approaches in the literature that aim at improving terminology translation in terminology. These include:

- fine-tuning with domain-specific data: these help NMT models translate domain-specific terms more effectively (Nayak et al., 2020),
- data augmentation approaches, including generating synthetic data through back-translation or self-training: these methods expose the NMT model to a variety of examples, ultimately improving term recognition and translation (Fernando et al., 2020),
- incorporating external resources like glossaries, dictionaries, or terminology databases can assist NMT models in translating specialised terms more effectively (Scansani and Dugast, 2021),
- terminology injection during inference, using techniques like inline tags (Dinu et al., 2019), source-target alignments (Dougal and Lonsdale, 2020), or fixed source positions (Niehues, 2021) for reference terms, helps produce translations with accurate domain-specific terminology,
- introducing auxiliary objectives during training such as predicting masked source terms or generating domain-specific inflections (Michon et al., 2020) can better handle domain-specific terms during inference.

Standard NMT domain adaptation involves fine-tuning a generic NMT model using domain-specific data. Accordingly, it is essential to consider factors such as similarity or distinct domain features that characterise the specialised field to effectively select the appropriate data. In their study, Farajian et al. (2017) showed that fine-tuning an NMT general domain model using a sentence highly similar to the source-test sentence can improve the usage of domain-specific terminology after adaptation. Likewise, Li et al. (2018) conducted an experiment in which they fine-tuned an NMT general domain model on a small subset of bilingual training data acquired through a similarity search with the source test sentence. Their findings also indicated an improvement in translation performance. In their experiments, both Farajian et al. (2017) and Li et al. (2018) showed how only a small set of sentences based on similarity to that of the test sentence is sufficient to improve the quality of translation. However, it is crucial that the sentences used for fine-tuning exhibit considerable similarity to the sentences being translated; otherwise, this can lead to a deterioration in translation quality.

Unlike Farajian et al. (2017) and Li et al. (2018), who fine-tuned their models on fewer sentences for each test instance, Chen et al. (2020a) took a different approach by employing *n*-gram matching for the entire test set. Their study focused on matching and selecting *n*-grams from the training data which are most relevant to the entire test set rather than just individual sentences. By doing so, they were able to create a more comprehensive fine-tuning dataset, which in turn led to improved terminology translation.

Numerous studies have investigated ways to better incorporate technical terms into NMT systems during inference. For example, Dinu et al. (2019) added special tags to the source text sentence by identifying domain-specific terms. After translating, they found that these tags were correctly replaced with the appropriate terms in the target language. A similar approach was tried by Song et al. (2019), where they replaced specific phrases in the source text with pre-selected, domain-specific translations before translating. This made it easier for the system to use the correct domain-specific terms in the final translation. Michon et al. (2020) carried out a comparative analysis by experimenting with variations of inline terminology tags and discussed the optimal settings in the experiment that helped improve terminology translation. In their work, Dougal and Lonsdale (2020) added domain-specific terminology after the translation process as a post-processing step, replacing incorrect terms with approved ones using source-target alignments. This approach offers the benefit of not requiring the translation model to handle tags, so it could potentially be used to introduce terminology to MT system outputs. However, the effectiveness of this method relies on an effective alignment model. Let us recall the work of Chen et al. (2020a), where they developed constraint-aware training data by randomly choosing phrases from the reference translation to serve as constraints and subsequently merging them into the source sentence with the help of a separation symbol. Their method does not require alignments and solely depends on bilingual dictionaries during translation. They inserted the reference terminology at a fixed location in the source text, facilitating the model's learning of proper alignment. Similarly, Niehues (2021) also placed the reference terminology at a fixed point within the source text. However, his primary focus was on using the lemma of the term, which encouraged the model to learn the appropriate inflections for the given terminology. In their experiments, Lee et al. (2021) presented a technique that estimates the range of masked source terms during MT training, facilitating the integration of multi-word domain-specific terms in the translation process. They found that their models produced performance similar to that of Chen et al. (2020a) in terms of single-word accuracy, but improved performance when it came to translating multi-word terms.

Nayak et al. (2020) conducted an experiment in which they mined sentences from a large general domain corpus based on the presence of domain-specific terms in the test data. They then utilised the extracted data to fine-tune the model and observed improvements in terminology translation performance. Similar experiments were carried out by Haque et al. (2020), with their approach also demonstrating improvements in terminology translation.

In my experiments, I employ an approach similar to that used by Nayak et al. (2020) and Haque et al. (2020). The first experiment extends these methodologies by performing extraction and adaptation for each instance in the test data as in Farajian et al. (2017). This means that, instead of using a predetermined set of sentences containing domain-specific terms, we adapt my model (mBART) on a per-instance basis, allowing the model to better handle the domain-specific terminology in each test sentence. My proposed approach aims to provide a more tailored and flexible adaptation process, potentially resulting in more significant improvements in translation performance and domain-specific term management.

In my second experiment, we exploit the exceptional capabilities of LLMs to generate synthetic data. Initially, we generate sentences using domain-specific terms, followed by translating these sentences to form a synthetic corpus. This corpus is then used to fine-tune my model (GPT). My proposed approach holds significance beyond merely improving terminology handling by offering a potential solution for generating domain-specific data in zero-resource scenarios. This capability could be instrumental in improving the adaptability and performance of models operating in zero-resource or specialised domains, thereby helping improve translation quality in such scenarios.

I pointed out the rationale for investigating terminology translation in HAN at the beginning of the chapter. To the best of my knowledge, no one has investigated terminology translation in document-level NMT. I also stated the rationale for comparing HAN with LLMs on the terminology translation above. Section 5.4 elaborates and justifies the reason for investigating domain term translation using LLMs.

5.3 Dataset

In my experiment, we used French-to-English parallel data from WMT2021,[3] which includes sources such as Europarlv10, ParaCrawlv7.1, News Commentary v16, UN Parallel Corpus V1.0, CommonCrawl corpus, and 10^9French-English corpus. We combine these datasets, remove duplicates, and tokenise the text using Moses (Koehn et al., 2007)[4] tokeniser scripts. The resulting dataset consists of 44M unique sentence pairs. The terminologies for French-to-English translation were obtained from the TICO-19 project by Anastasopoulos et al. (2020),[5] focusing on the COVID-19 domain. The test set includes 2100 sentences containing 595 unique domain-specific

[3] https://www.statmt.org/wmt21/terminology-task.html

[4] https://github.com/moses-smt/mosesdecoder

[5] https://tico-19.github.io/

terms and a cumulative total of 2557 terms.

Despite not being a boundary-annotated, document-level dataset, I observed that it comprises segments from various articles, each precisely annotated for specific terms, making it suitable for document translation tasks.

I observed that not all sentences in the test set were annotated with terms. For my experiments in Section 5.5.1, these unannotated sentences were retained and translated by the baseline, whereas for the experiments in Section 5.7.1, we filtered out all sentences without term annotations, resulting in a reduced test set of 1,281 sentences, i.e. 819 sentences were filtered out.

5.4 HAN versus LLMs

5.4.1 NMT Model

The mBART (Multilingual BART) (Liu et al., 2020) model is a multilingual extension of the BART (Bidirectional and Auto-Regressive Transformers) (Lewis et al., 2020) model, a sequence-to-sequence pre-training framework for natural language understanding and generation tasks. mBART is trained on large-scale multilingual data, enabling it to perform well across various languages and tasks. The model follows the encoder-decoder architecture, where the encoder learns to capture the input's semantic information, and the decoder generates the output text based on the encoded representation. It is pre-trained using a combination of denoising auto-encoding and masked language modeling, which involves reconstructing corrupted text or predicting masked tokens. One key feature of mBART is its shared vocabulary across languages, making it easier to fine-tune the model for downstream tasks such as MT, summarisation, or sentiment analysis. By leveraging its pre-trained knowledge, mBART can achieve state-of-the-art performance on a wide range of NLP tasks and languages.

5.4.2 Experiments and Results

This section elaborates on my investigation on terminology translation using HAN and LLMs. The experiment compares the HAN model (discussed in detail in Chapter 3, Section 3.3.4) with LLMs, as referenced in Section 5.4.1, for the French-to-English language pair. My MT systems were evaluated using BLEU, COMET, and Term Count as evaluation metrics. Term Count (TC) measures the number of occurrences of accurately translated domain-specific terms by the MT system. I measured the performance of MT systems on the evaluation test set and the scores are shown in Table 5.1. From Table 5.1, it is observed that the mBART model significantly outperforms the HAN model in translating terminology, with a substantial improvement of 41.69%, clearly showcasing its superior performance. Furthermore, a statistical significance test revealed that the difference in TC is statistically significant. These findings clearly indicate that LLMs (mBART) outperform HAN in translating terminology, leading me to conduct further experiments on terminology translation using LLMs.

Table 5.1: Performance Comparison between HAN and mBART

	BLEU	Term Count	COMET
HAN	21.93	1535	0.654
mBART	27.63	2175	0.844

5.5 Instance-based adaptation

5.5.1 Methodology

This section presents the methodology for my instance-based adaptation approach using terminology-aware mining. Note that since baseline mBART statistically significantly outperformed HAN (differences in evaluation scores are massive; see Table 5.1), we chose mBART for these experiments.

Domain adaptation using terminology-aware mining

Terms or phrases appearing in domain-specific data may encode meanings or usages different to those when they appear in generic data. In order to obtain correct translations for terms or phrases of a domain text, Translation Service Providers normally use domain-specific terminology or glossaries, but obtaining such terminological resources is challenging as this process can be very expensive in terms of both cost and time. Automatically identifying and extracting domain-specific terminology from training data or external resources and integrating them into industrial translation workflows can partly alleviate this problem (Haque et al., 2018; Mouratidis et al., 2022). A notable obstacle to these approaches could be the training itself. Since the NMT training process is a highly time-consuming task, integrating terminology and training from scratch is not a feasible solution. In fact, this is unimaginable in a dynamic industrial setting where terminologies often need to be updated for translating newly arrived documents with particular styles. Alternatively, we could have certain situations where the training time may not be a concern, and the entire terminology is available at the training.

Adapting a generic NMT system to a specific domain and obtaining accurate translations for the domain-specific terms can be more challenging when one does not have domain-specific data. In this study, I investigate this specific scenario (i.e. unavailability of domain text) and systematically make use of large general-domain data in order to fine-tune my MT systems. First, I extract terms from the source sentence to be translated based on the term annotations provided in the test data. Then, I mine parallel sentences from the general domain parallel data based on the frequency of the extracted domain-specific terms in the parallel sentences. The extracted sentences are then used to fine-tune my NMT models. Note that the entire process (term extraction from the test sentence to be translated and mining parallel sentences from large generic data) is characterised as *on-the-fly* instance-based adaptation by Farajian et al. (2017).

In Algorithm 1, I present my approach for instance-based adaptation using

Algorithm 1 Algorithm for Instance-Based Adaptation Using Terminology-Aware Mining

```
for $src\_sent$ in $tst\_set$ do
    $D_{\text{Trm}}$ = Extract_trm($src\_sent$)
    $R_{\text{Sent}}$= Retrieve($max\_trm$(Data,$D_{\text{Trm}}$))
    $F_{\text{MT}}$=Finetune($\text{G}_{\text{MT}}$,$R_{\text{Sent}}$)
    Translate($F_{\text{MT}}$,$src\_sent$)
end for
```

terminology-aware mining. The algorithm leverages domain-specific terminology to adapt the NMT system by fine-tuning it on relevant instances from the general-domain parallel data.

The algorithm picks a source sentence (*src_sent*) from the test set (*tst_set*) and performs the following steps:

- **Extract** domain-specific terminology (D_{Trm}) from the source sentence to be translated using the **Extract_trm** function, which is designed to identify and extract terms unique to the given domain based on the annotations in the test data.
- **Retrieve** a sentence (R_{Sent}) from the general-domain parallel data based on the highest number of matching domain-specific terms in the test data. This is done using the **Retrieve** function in combination with the fuction *max_trm*, ensuring that the most relevant instances with maximum domain terms are selected for adaptation.
- **Fine-tune** the general-domain NMT system (G_{MT},R_{Sent}) using the retrieved sentence (R_{Sent}). The **Finetune** function updates the model parameters based on the domain-specific instance, resulting in a fine-tuned MT system (F_{MT}).
- **Translate** the source text (*src_sent*) using the fine-tuned NMT system (F_{MT}) to generate a domain-adapted translation.

5.6 Experimental Setup

In this study, I wanted to see how my proposed domain adaptation method of terminology-aware fine-tuning would work on mBART (cf. Section 5.4.1). I placed particular emphasis on terminology translation (cf. Section 5.5.1). My baseline model is a generic mBART-based NMT system. I apply the instance-based adaptation on mBART (see Algorithm 1). I expect that my terminology-aware mining techniques will be able to help adapt the baseline so that the model can correctly translate a larger number of domain-specific terms.

In order to thoroughly assess how my proposed terminology-aware adaptation process works on terminology translation, I carried out experiments with a different number of instances (one, three, and five) and epochs (one, three, and five) for fine-tuning. By examining the impact of varying numbers of sentence and epoch combinations on the model's performance and its handling of domain-specific terms, I aimed to gain a deeper understanding of the potential benefits and limitations of the proposed approach.

5.6.1 Experiments and Results

Table 5.2 shows the results that I obtained through my experiments. It displays BLEU, TC, and COMET scores for each of the test scenarios described in Section 5.6. I can see from the table that TC improves in two cases over the baseline. In both cases, the improvement occurs while fine-tuning using a single sentence only with three and five epochs. I conducted statistical significance tests for both cases using bootstrap resampling as described by Koehn (2004). My findings revealed that the differences in scores (baseline and adapted models (single sentence with three and five epochs)) and were statistically significant. Furthermore, the improvement in TC over the baseline MT system suggests that the proposed adaptation method effectively improves the generic NMT system's ability to handle domain-specific terminology. In order to further understand the results in Table 5.2, I visualise the results in Figures 5.1, 5.2 and 5.3.

Sentence	Epoch	BLEU	Term Count	COMET
Base		27.63	2175	0.844
1	1	26.60	2155	0.825
1	3	27.21	2191	0.826
1	5	27.68	2190	0.822
3	1	26.12	2115	0.829
3	3	26.24	2111	0.832
3	5	26.43	2094	0.832
5	1	25.39	2119	0.817
5	3	26.18	2109	0.833
5	5	26.30	2089	0.835

Table 5.2: Results of instance-based adaptation using terminology-aware mining (best setup: single sentence with three and five epochs).

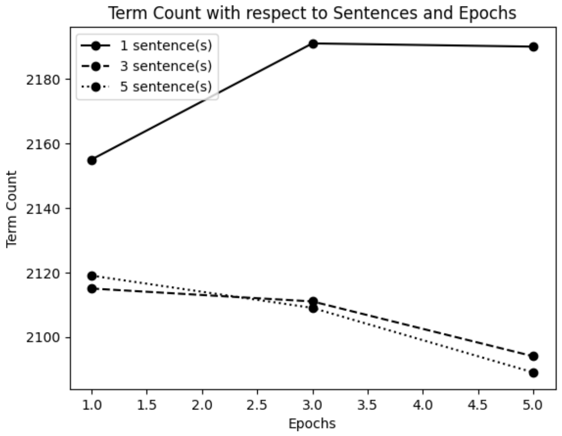


Figure 5.1: TC in relation to the number of sentences and epochs used in the adapted model.

In Figure 5.1, I show the performance of my adapted MT systems for the French-to-English translation task using TC scores. The graph presents the results for different combinations of sentences (one, three, and five) and epochs (one, three, and five) in the fine-tuning process. The x-axis represents the number of epochs, and the y-axis represents TC. The lines with varying markers correspond to the different epoch combinations. In Figure 5.1, I observe that increasing the number of sentences used for fine-tuning does not contribute significantly to the improvement of terminology translation performance. Rather, I find that increasing the number of epochs for a single sentence is more beneficial. This finding suggests that the model may benefit from more focused training, concentrating its learning efforts on a smaller number of sentences for a longer period of time (i.e., more epochs).

By doing so, the model can potentially gain a deeper understanding of the specific domain terminology, which in turn can lead to better translation performance with respect to the domain-specific terms.

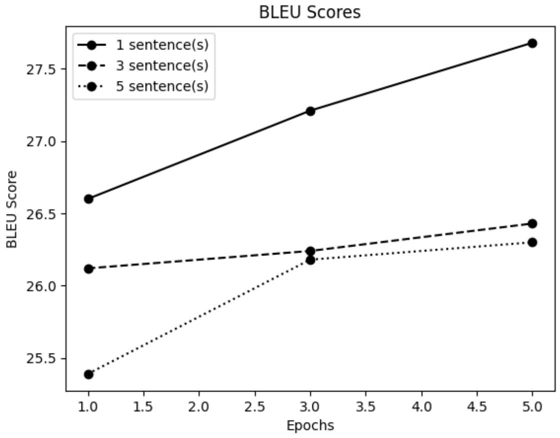


Figure 5.2: BLEU scores in relation to the number of sentences and epochs used in the adapted model.

In Figure 5.2, I have plotted the performance of my adapted MT systems using BLEU scores to analyse the relationship between the number of sentences, the number of epochs, and the translation quality. The x-axis represents the number of epochs, and the y-axis represents the BLEU scores. The lines with varying markers correspond to different sentence combinations. I observe that increasing the number of sentences does not consistently improve translation quality, showing mixed results. A part of this finding resembles the findings in terms of TC (as in Figure 5.1), where adding more sentences offered no improvement. This suggests that adding more sentences to the fine-tuning data may not guarantee better translation outcomes.

While the graphs for TC and BLEU display a similar trend, it is crucial to understand that an increase in the BLEU score does not necessarily indicate an improvement in terminology. In fact, alterations made to the adapted model might have led to improvement in the meta-language (i.e. the words in the rest of the sentence, not the terms themsleves) without directly translating to substantial improvements in the translation of domain-specific terms.

In Figure 5.3, I plotted my MT systems' performance based on COMET scores

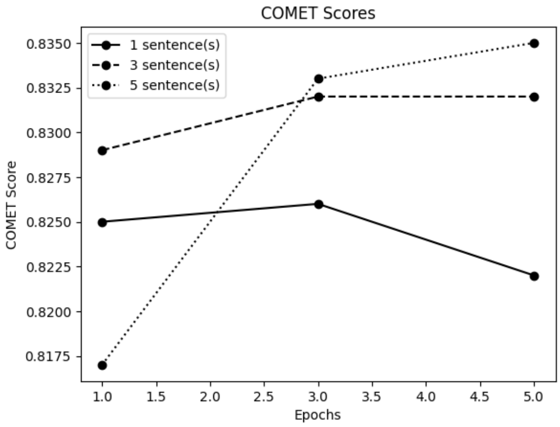


Figure 5.3: COMET scores in relation to the number of sentences and epochs used in the adapted model.

to analyse the relationship between the number of instances used for training and epochs. The x-axis represents the number of epochs, and the y-axis represents the COMET scores. The lines with varying markers correspond to different sentence combinations. I see that the COMET scores exhibit a different trend. When the number of sentences is increased, the translation quality measured by COMET scores appears to improve. This contrasts with the trends that were observed in terms of TC and BLEU. I also see that increasing the number of sentences did not consistently lead to better translation quality.

The discrepancy between the trends observed for three metrics (COMET, BLEU and TC) could be attributed to the differences in the evaluation metrics. While TC and BLEU scores focus on specific aspects of translation quality, such as the handling of domain-specific terminology and *n*-gram overlaps between the reference and the translation, the COMET metric is designed to provide a more holistic assessment of translation quality by considering factors such as fluency, adequacy, semanticity and style. I plan to investigate this in more detail in future work, by conducting a thorough manual analysis encompassing all these issues.

5.6.2 Analysis of Terminology Improvements

Table 5.2 presents the results of my experiments aimed at improving terminology translation using instance-based adaptation. I discovered that the TC scores for the

Table 5.3: Comparison of terminology translation counts: baseline vs. domain-adapted MT system.

Category	**Count**
Unique to Base	101
Unique to Adapted	117
Common to Both	2074

adapted MT system are found to be high in two cases (i.e. setup: a single sentence using three and five epochs). As for analysing translations produced by the MT systems, I choose the best-performing adapted MT system (i.e., one sentence and three epochs).

In order to further understand the terminology translation results presented in Table 5.2, I provide a detailed analysis of these results, along with a comparison between the baseline and the best domain-adapted MT systems in Table 5.3. The row labeled "Base" corresponds to the baseline MT system and has 101 unique terms, indicating the exclusive number of terminology translations by the model. The row titled "Adapted" refers to the best domain-adapted model, which has 117 unique domain-specific terms. Additionally, a third row highlights the shared terminology, listing 2074 terms.

Table 5.4: Example: adapted MT system correctly translates terminology.

Source	dans environ 14 % des cas , la COVID-19 entraîne une atteinte plus sévère nécessitant une hospitalisation , tandis que les 6% de cas restants développent une forme grave de la **maladie** nécessitant des soins intensifs .
Reference	in ca 14% cases , covid-19 develops into a more severe disease requiring hospitalisation while the remaining 6% cases experience critical **illness** requiring intensive care .
Baseline MT	in about 14% of cases, covid-19 causes a more severe condition requiring hospitalization, while the remaining 6% develop a serious form of the **disease** requiring intensive care.
Adapted MT	in about 14% of the cases, covid-19 leads to more severe illness requiring hospitalization, while 6% of the remaining cases develop a serious form of serious **illness** requiring intensive care

To further understand how the two models differ when it comes to the quality of terminology translation, I select an example sentence from the test set. In Table 5.4, I present translations of the sentences by the baseline and adapted MT systems. I can see from the table that the adapted MT system demonstrates an improvement over the baseline MT system, where the domain term “maladie” in the source sentence is accurately translated as “illness” by the adapted MT system. In contrast, the baseline system incorrectly translates it as “disease”. However, it is worth noting that the baseline system still provides a decent translation. While it may not capture the exact terminology, the overall semantic content of the sentence is preserved, demonstrating the robustness of the baseline system.

Table 5.5: Comparison of multi-word terminology counts: baseline vs. domain-adapted MT system.

Category	**Count**
Unique to Base	25
Unique to Adapted	44
Common to Both	251

I observed that the adapted MT system better handles the translation of multi-word terms. Table 5.5 details terminology translations: the baseline MT system with 25 unique terms, the domain-adapted system with 44 unique terms, and 251 terms shared by both models.

Table 5.6: Example: adapted MT system correctly translates multi-word terminology.

Source	la ventilation mécanique devient plus complexe avec le développement du **syndrome de détresse respiratoire aiguë** (SDRA) au cours de la COVID-19 et l’ oxygénation devient plus difficile .
Reference	mechanical ventilation becomes more complex as **acute respiratory distress syndrome** (ards) develops in covid-19 and oxygenation becomes increasingly difficult .
Baseline MT	mechanical ventilation becomes more complex with the development of **acute respiratory disorder syndrome** (sdra) during covid-19 and oxygenation becomes more difficult.
Adapted MT	mechanical ventilation becomes increasingly complex as **acute respiratory distress syndrome** (ards) develops in covid-19 and oxygenation becomes increasingly difficult.

In Table 5.6, I show another example translation. This time, I chose a source sentence that contains a multi-word term. I see from the table that the adapted MT system shows improvement over the baseline MT system where the multi-word term “syndrome de détresse respiratoire aiguë” in the source sentence is accurately translated as “acute respiratory distress syndrome” by the adapted MT system. In contrast, the baseline system incorrectly translates it as “acute respiratory disorder syndrome”.

5.7 Terminology-aware Synthetic Data Generation for Domain Adaptation

5.7.1 Methodology

My second approach involves extracting terms from the source test sentence based on the term annotations in the test data. I then leverage the capabilities of LLMs to generate new source sentences using these extracted terms. These source sentences are subsequently translated to create a synthetic parallel corpus. Finally, I utilise this synthetic parallel corpus to fine-tune my NMT model. The methodology in Algorithm 2 draws inspiration from my experiments with Algorithm 1. The primary distinctions are that Algorithm 2 adapts models for the entire test set instead of individual instances, and uses synthetic data[6] instead of general domain data for sentence mining.

Algorithm 2 Algorithm for Generating Terminology-aware Synthetic Data for Domain Adaptation

```
for $src\_sent$ in $tst\_set$ do
    $D_{\text{Trm}}$ = Extract_trm($src\_sent$)
    $G_{\text{Sent}}$= Generate_sent($D_{\text{Trm}}$)
    $S_{\text{data}}$= Store_bitext($G_{\text{Sent}}$,Translate($G_{\text{Sent}}$))
end for
$F_{\text{MT}}$=Fine_tune($G_{\text{MT}}$,$S_{\text{data}}$)
Translate($F_{\text{MT}}$,$tst\_set$)
```

[6] double the size of the test set

I present my methodology for domain adaptation using synthetic data in Algorithm 2 which employs domain-specific terminology to adapt the NMT system by fine-tuning it using synthetic data.

The algorithm picks a source sentence (*src_sent*) from the test set (*tst_set*) and performs the following steps:

- **Extract** domain-specific terminology (D_{Trm}) from the source sentence that is being translated using the **Extract_trm** function, which is designed to identify and extract terms unique to the given domain based on the annotations in the test data.

- **Generate** a synthetic sentence (G_{Sent}) by leveraging the capabilities of the LLM, using domain-specific terms (D_{Trm}) derived from the test data, with the following prompt being used " Generate a French sentence using the term *trm*"

- **Translate** the synthetically generated sentence (G_{Sent}) using the baseline NMT system. The synthetic sentence (G_{Sent}) and its translation **Translate**(G_{Sent}) are then stored to form the bitext for the synthetic data (S_{data}), with the following prompt being used " Translate the following French sentence into English: *sent*".

- Once I have generated synthetic data for all the terms in the test set and formed a synthetic corpus, I fine-tune the general-domain MT system (G_{MT}, S_{data}). The **fine-tune** function updates the model parameters based on the domain-specific synthetic data, resulting in an adapted MT system (F_{MT})

- **Translate** the source test data (*tst_set*) using the adapted MT system (F_{MT}) to generate a domain-specific translations.

5.7.2 Experimental setup

NMT Model

The GPT (Brown et al., 2020), developed by OpenAI,[7] represents a significant breakthrough in NLP. GPT is based on the Transformer architecture, which is explicitly designed for generating data sequences, such as text. These models rely on an attention mechanism that assigns different weights to words in an input sequence, recognising their relevance for generating the output sequence. Interestingly, unlike traditional bidirectional Transformer models, GPT models operate unidirectionally, considering only the context to the left of a target word during training. GPT is pre-trained on a large amount of text data and then fine-tuned for tasks such as question answering, translation, and summarisation. This model series has undergone multiple iterations, each improving capacity and performance. The most recent, GPT-4, possesses billions of parameters and generates human-like text, blurring the line to some extent between human and AI-generated content.

In this research, I assess the effectiveness of my proposed methodology for domain adaptation. This method involves generating synthetic data using the GPT model based on domain-specific terms for fine-tuning (cf. Section 5.7.1). I utilised the davinci model (part of OpenAI's GPT-3 series) as my baseline and applied terminology-based adaptation to it (cf. Algorithm 2). I show that my proposed approach to generate synthetic data based on terminology-aware mining using LLMs helps improve the adaptation of the baseline model, thereby improving translation accuracy for more domain-specific terms.

Table 5.7: Results of domain adaptation using terminology-aware using synthetic data.

Models	BLEU	Term Count	COMET
davinci(base)	24.13	1827	0.8021
text-davinci-002	26.72	2135	0.8463
Adapt	27.72	2250	0.8518

[7] https://openai.com/

5.7.3 Experiments and results

Similar to my experimental setups in Section 5.6.1, I evaluated my MT systems using BLEU, COMET, and TC as my evaluation metrics. My results are shown in Table 5.7, where I observe that my domain-adapted model improves over the performance of the baseline davinci system by 23.15% in translating domain-specific terms. Interestingly, my domain-adapted model also outperforms the text-davinci-002 (part of the GPT-3.5 series) by 5.38% in translating domain-specific terms. This overall improvement shows the effectiveness of my approach in managing domain-specific terminology, thereby leading to substantial improvements in MT accuracy (the accuracy improved by 12.7% relative to the BLEU score and by 6.19% compared to the COMET score over the baseline). As above, I conducted statistical significance tests for both cases using bootstrap resampling. My findings reveal that the differences in scores were statistically significant. Furthermore, improvements by TC over the baseline MT system and the advanced text-davinci-002 model suggest that the proposed method effectively improves the generic NMT system's ability to translate domain-specific terminology.

5.7.4 Analysis of Terminology Improvements

Table 5.7 presents the results of my experiments, which aimed at improving terminology translation by generating synthetic data. I found that the adapted MT system outperforms the baseline and the advanced text-davinci-002 model.

Table 5.8: Comparison of terminology counts: baseline (davinci) vs. domain-adapted MT system.

Category	**Count**
Unique to Base (davinci)	128
Unique to Adapted	551
Common to Both	1699

In Table 5.8, I compare terminology translations between the baseline (davinci) and domain-adapted MT systems. This table displays the unique and shared terminology counts for each model. The row labeled “Base” represents the baseline

MT system with 128 unique terms, highlighting the exclusive translations from this model. The "Adapted" row corresponds to the domain-adapted model, featuring 551 unique domain-specific terms. The shared terminology of 1699 terms translated by both models is shown in the final row of the table.

Table 5.9: Example: adapted MT system correctly translates terminology.

Source	les complications associées à la COVID-19 incluent la **septicémie** , les troubles de la coagulation et les lésions cardiaques , rénales et hépatiques .
Reference	complications associated with covid-19 include **sepsis**, abnormal clotting and damage to the heart , kidneys and liver .
Baseline MT	the complications associated with covid-19 include **septicemia**, bleeding disorders and heart, kidney and liver damage
Adapted MT	the complications associated with covid-19 include **sepsis**, bleeding disorders and cardiac, renal and liver damage.

I select an example sentence from the test set to understand further how the two models differ regarding the quality of terminology translation. In Table 5.9, I present translations of the sentence by the baseline and adapted MT systems. I can see from the table that the adapted MT system demonstrates improvement over the baseline MT system, where the domain term "septicémie" in the source sentence is correctly translated as "sepsis" by the adapted MT system. In contrast, the baseline system incorrectly translates it as "septicemia". However, the baseline system still produces a decent translation. While it may not capture the exact terminology, the overall semantic content of the sentence is preserved, demonstrating the robustness of the baseline system.

Table 5.10: Comparison of multi-word terminology counts: baseline (davinci) vs. domain-adapted MT system.

Category	**Count**
Unique to Base (davinci)	20
Unique to Adapted	84
Common to Both	227

I observed that the adapted MT system better handles the translation of multi-word terms, as evidenced in Table 5.10. Table 5.10 presents the terminology translations: the baseline MT system with 20 unique multi-word terms, the domain-adapted system with 84 unique multi-word terms, and 227 terms that are shared by both models.

Table 5.11: Example: adapted MT system correctly translates multi-word terminology.

Source	le sous-dénombrement des cas modérés peut entraîner une surévaluation des **taux de mortalité** .
Reference	the under-counting of mild cases can cause the **mortality rate** to be overestimated .
Baseline MT	the undercounting of moderate cases can lead to an overvaluation of **the rates of mortality.**
Adapted MT	the undercounting of moderate cases can lead to an overestimation of **mortality rate**.

In Table 5.11, I show another example translation. This time, I chose a source sentence that contains a multi-word term. I observe that the multi-word term “taux de mortalité” in the source sentence is accurately translated as “mortality rate ” by the adapted MT system. In contrast, the baseline system incorrectly translates it as “the rates of mortality”.

Table 5.12: Comparison of terminology counts: text-davinci-002 vs. domain-adapted MT system.

Category	**Count**
Unique to davinci-002	108
Unique to Adapted	223
Common to Both	2027

Similar to the analysis of the baseline model and my adapted model, I now perform an analysis comparing the adapted system with the more advanced text-davinci-002 model. In Table in 5.12, I compare text-davinci-002 and the domain-adapted MT systems on their terminology translation capabilities. An example translation is shown in Table 5.13. Multi-word terminology translation is shown in Table 5.14, with a relevant example in Table 5.15. Once again, my observations indicate that the methodology I have proposed effectively improves terminology translation. Furthermore, I also find that my domain-adapted model improves performance

Table 5.13: Example: adapted MT system correctly translates terminology.

Source	le CDC recommande également aux individus de se laver les mains souvent au savon et à l' eau pendant au moins 20 secondes , particulièrement après avoir été aux toilettes ou quand les mains sont visiblement sales ; avant de manger ; et après s' être mouché , avoir toussé , ou **éternué** .
Reference	the cdc also recommends that individuals wash hands often with soap and water for at least 20 seconds , especially after going to the toilet or when hands are visibly dirty , before eating and after blowing one 's nose , coughing or **sneezing** . .
Baseline MT	the cdc also recommends that individuals wash their hands often with soap and water for at least 20 seconds, particularly after going to the bathroom or when their hands are visibly dirty; before eating; and after blowing their nose, coughing, or **sneez**
Adapted MT	the cdc also recommends that individuals wash their hands often with soap and water for at least 20 seconds, particularly after using the toilet or when hands are visibly dirty; before eating; and after sneezing, coughing, or **sneezing.**

over the advanced text-davinci-002 model thereby reaffirming the effectiveness of my approach.

Table 5.14: Comparison of multi-word terminology counts: text-davinci-002 vs. domain-adapted MT system.

Category	**Count**
Unique to davinci-002	32
Unique to Adapted	43
Common to Both	268

Table 5.15: Example: adapted MT system correctly translates multi-word terminology.

Source	le CDC recommande que les personnes suspectées d' être porteuses du virus portent un simple **masque facial** .
Reference	the cdc recommends that those who suspect they carry the virus wear a simple **face mask** .
Baseline MT	the cdc recommends that people suspected of being carriers of the virus wear a simple **facial mask**.
Adapted MT	the cdc recommends that people suspected of being carriers of the virus wear a simple **face mask.**

5.8 Conclusion and Future Work

In this chapter, I discussed my experiments based on terminology-aware mining. I carried out my experiments for French-to-English translation. I began experiments on translating terms using HAN. However, with the emergence of LLMs in NLP, I decided to compare them with HAN in order to ensure that my research was as up-to-date as possible. Our results showed that LLMs performed better than HAN in term translation. Therefore, I deciced to conduct further experiments on terminology translation using LLMs.

In the first experiment, I investigated an approach called instance-based adaptation. My results demonstrated that the proposed approach was successful in improving terminology translation. Furthermore, I discover that increasing the number of sentences used for fine-tuning does not significantly impact the improvement of terminology translation performance. Instead, a more efficient strategy appears to be one that considers a high number of epochs for a single sentence. This observation suggests that the model may benefit from more focused training, concentrating its learning efforts on a single sentence over an extended period (i.e., more epochs). I evaluated my MT systems using BLEU and COMET evaluation metrics. I observe that the BLEU metric correlates with the TC, while the COMET metric shows improvements for the adapted model with an increased number of sentences. I also found that the adapted model outperformed the baseline when translating multi-word terms. My current proposed approach fine-tunes all instances, irrespective of whether a test instance requires fine-tuning or not, which may lead to the deterioration in translation quality for some sentences. In the future, I plan to identify those sentences that require fine-tuning and adapt only to them.

In my second experiment, I exploited the capabilities of LLMs to generate synthetic data based on domain-specific terms. My results demonstrate that the proposed approach helps improve terminology translation. I evaluated my MT systems using BLEU and COMET evaluation metrics. My approach improved the baseline significantly, and this has been observed across all metrics. Interestingly, my adapted model also improved over the next-generation LLM models, indicating the effectiveness of my approach. My analysis also indicated that the adapted model outperformed the baseline when translating multi-word terms. In the future, I aim to increase the quantity of my synthetic data and evaluate its performance. I would also like to experiment with this approach on different language pairs.

The experiments discussed above clearly indicate the effectiveness of my approach, which is based on terminology-aware mining, in improving terminology translation. It is important to note that both of these experiments were conducted in zero-resource scenarios (in the first experiment, data was mined from a general domain corpus and in the second experiment, synthetic data was generated because no COVID-19-related domain data was available. The test set used for the experiments belonged to the COVID-19 domain) This suggests that my proposed approaches are not only beneficial for improving terminology translation but also valuable in overcoming the challenge of domain data scarcity (zero-resource scenarios). In the future, it would be interesting to try my approach in different zero-resource situations to see how well it works.

Chapter 6

Conclusion and Future Work

MT has undergone many changes, from rule-based systems (Hutchins, 1986) to the more advanced neural models (Vaswani et al., 2017) we see today. However, the progress of MT has its challenges. While significant progress has been made in translating individual sentences, a complete approach (document-level MT) that looks at the whole document and understands its full meaning has remained hard to achieve.

This study focused mainly on understanding document-level MT systems. First, I looked at how the surrounding context of a sentence affects its translation. This involved examining how a sentence and its context relate and how this relationship influences the translation process. Next, I focused on the "context span" in document-level MT systems. Here, "context span" refers to the amount of context the MT system uses when translating a sentence. Understanding this is important as it helps us see how MT systems use available information in a document to provide accurate and relevant translations. Finally, I examined the new generation of document-level systems based on LLMs. I were interested in understanding the effectiveness of these LLMs in translating domain-specific terminology. I also proposed and tested ways to improve these new systems, making them more reliable for translating domain-specific terms.

6.1 Research Questions

I formulated three research questions in Chapter 1 of this thesis. This section briefly explains how I addressed each of them and summarises my findings.

- **RQ1: How important is contextual information for improving translation in a document, and are there specific categories of sentences that demand contextual understanding more than others?**

  In Chapter 2 of my thesis, the addressed RQ explores the reasons behind the improved performance of context-incorporated MT models like HAN, compared to those not using context. I sought to understand how a document's context influences sentence translation. I used the HAN model for my experiments, an MT system designed explicitly for document-level translation, incorporating prior sentences as context. As anticipated, the HAN model proved to be sensitive to context. My results showed that context-aware NMT systems significantly outperform context-agnostic ones, evidenced by higher BLEU, chrF, TER, and METEOR scores. The experiments included translations for three morphologically different language pairs, Hindi-to-English, Spanish-to-English, and Chinese-to-English, with the context-aware models consistently performing better across all pairs.

  I found that contextual information is crucial for improving translation, a finding consistent across my investigations involving all language pairs. My experiments identified specific sentence categories requiring more contextual understanding than others. These experiments not only revealed the significant impact of understanding sentence natures and identifying the correct context on the quality of document-level MT, but also led us to classify test set sentences into three categories: (i) context-sensitive sentences, (ii) normal sentences, and (iii) context-insensitive sentences. While the translation quality of context-sensitive sentences is heavily influenced by the presence or absence of appropriate contextual information, context-insensitive sentences show no sensitivity to contextual variations in terms of translation quality.

- **RQ2: What is the ideal context span that can be incorporated into document-level translation systems to improve translation?**

  I addressed this question in Chapter 3 of my thesis. The goal was to identify the optimal amount of context that should be considered during translation. Similar to my experiments on the previous RQ, I used HAN for my investigation and considered three sentences as context. The context sentences were sampled randomly from the different positions of the document. The idea was to understand the role and origin of context in document-level NMT systems. I conducted experiments with three morphologically distinct language pairs: Hindi-to-English, Spanish-to-English, and Chinese-to-English.

  I proposed a metric that produces CSS scores given the relative distance between sentences that form context span and a source sentence to be translated. This metric gives more weight to a context-sensitive sentence near the sentences of the context used and less to a context-sensitive sentence farther away from the sentences used. My findings showed that incorporating document-level context into NMT models can lead to performance improvements, primarily when a broader range of contextual factors are considered. I further carried out sentence-level analysis by selecting a specific context-sensitive sentence from each translation task. I found that specific sentences of the context far away from the sentence being translated generally help improve the translation.

  My investigations showed that document-level MT systems benefit significantly from incorporating a broader context. I specifically examined the context-sensitive class of sentences, determining the context span these sentences could effectively utilise. I also conducted a thorough manual analysis by looking at context-sensitive sentences, the context provided during translation and their target translations. I found a relationship or pattern between the sentence being translated and sentences of a context used. I computed the similarity between sentences used as a context and source sentences using sentence transformers. I found a specific pattern that differentiates the sentences of the context and improves translation from those that do not.

- **RQ3: How effective are document-level translation systems and LLMs at translating domain-specific terminology, and to what extent can approaches based on terminology-aware mining improve the accuracy of domain-term translation in these systems ?**

  I addressed this question in Chapter 4 of my thesis. In the experiments conducted for the English-to-French language pair, I explored two approaches to translating domain-specific terminology through LLMs: the first approach is an instance-based adaptation based on terminology-aware mining. In this experiment, I found that increasing the number of sentences used for fine-tuning does not substantially improve the performance of terminology translation; instead, a more effective approach involves using more epochs for a single sentence. This implies that models might perform better with intensive training on one sentence for longer, focusing their learning on that sentence. The second approach involved leveraging the impressive capability of LLMs to generate synthetic data based on terminology-aware mining. I found that my adapted model outperformed the baseline and the next-generation series of models (GPT-3.5), indicating the effectiveness of the proposed approach. Both experiments were evaluated using BLEU and COMET metrics, indicating that document-level translation systems effectively translate domain-specific terminology. The results suggest that domain adaptation methods based on terminology-aware mining can improve the accuracy of domain-term translation.

6.2 Future Work

6.2.1 Further Experiments on Research Question

In this section, I discuss additional experiments that can be conducted to further explore the research question already addressed.

- **RQ1: How important is contextual information for improving translation in a document, and are there specific categories of sentences that demand contextual understanding more than others?**

  This RQ was addressed in Chapter 3 of my thesis. Future experiments will involve examining the characteristics of context-sensitive class of sentences with a specific focus on understanding the reason for their sensitivity to context. I also plan to examine how the presence of domain-specific terminology either in the context or in the source sentence influences the quality of translation.

- **RQ2: What is the ideal context span that can be incorporated into document-level translation systems to improve translation?**

  This RQ was addressed in Chapter 4 of my thesis. As new document-level systems, particularly those based on LLMs, continue to emerge, my research plans include investigating their context span. Furthermore, I intend to broaden my inquiry to include various languages and genres of text. This approach will enable me to understand how the context window behaves under different linguistic scenarios and textual styles, providing a detailed understanding of their working and effectiveness.

- **RQ3: How effective are document-level translation systems and LLMs at translating domain-specific terminology, and to what extent can approaches based on terminology-aware mining improve the accuracy of domain-term translation in these systems ?**

  This RQ was addressed in Chapter 5 of my thesis. My first experiment in Chapter 5 on instance-based adaptation fine-tunes all instances, regardless of whether they require fine-tuning, which is computationally expensive and potentially degrades the quality of some translations. This method could be further improved by identifying and adapting particular sentences that require fine-tuning. In my second experiment, I plan to expand the synthetic data further to observe its impact on terminology translation. Additionally, I plan to experiment with hyperparameter tuning of LLMs and understand its impact on terminology translation. Furthermore, I would like to expand my experiments for different language pairs to understand the effectiveness of the proposed approaches.

Broadening the Scope: Other Areas to Explore

The future of document-level MT is closely connected to the rise of LLMs. These models are known for their impressive ability to understand and create text that feels human-like, and they are changing how I approach document-level MT. Traditional models often have difficulty understanding the context and how different document parts connect, but LLMs can keep the flow across longer texts. This ability helps make translations more accurate and maintains the particular meanings of the original text. As LLMs continue to improve, using even more advanced methods and learning from a broader range of sources, their potential to transform document-level MT becomes clear. Combining LLMs with document-level MT is an exciting development in MT. Which could lead to new ways of translating whole documents with greater precision and accuracy than ever before.

While document-level MT has advanced with the support of end-to-end learning frameworks from neural models, considerable work remains to be done. Improvements are needed in context modeling and developing context-dependent evaluation strategies. In the following, I will explore some possible future research directions.

- **Analysing the Context Window of LLMs**: I plan to examine how LLMs use a context window when translating documents. I will investigate the right size for this window and examine if changed is needed based on the document or language. By studying these questions, I aim to make LLMs better at translating, making the results more accurate and closer to a human translation. This research could be essential in improving document-level MT systems using LLMs.

- **Leveraging Context to Resolve Specific Linguistic Challenges**: As part of my future work, I plan to explore how LLMs utilise context to handle challenging language aspects in translating documents. This includes understanding how context helps the model decode slang, idioms, or technical terms that might be confusing. I also plan to investigate how LLMs use context to maintain consistent meaning across different document parts and how they interpret phrases not found in dictionaries.

- **Evaluation metrics**: Most commonly used automatic evaluation metrics like BLEU and METEOR do not consider the text's underlying discourse structure. Even though these metrics have been standard for evaluating MT outputs for almost two decades, they have flaws. Using a single reference translation for evaluation is ineffective (Way, 2018). There must be a balance between automatic and manual evaluation methods for document-level MT. This balanced approach could make manual evaluation more cost-effective and provide better insight into discourse phenomena than current automatic metrics do. However, evaluation test sets, primarily designed for specific language pairs, solve only part of the problem. Developing an evaluation metric suitable specifically document-level systems would be a valuable contribution to the field.

- **Developing Language Resources for Document-Level MT**: Most current datasets for MT consist of aligned sentence pairs; however, there is a need for datasets with entire documents translated and aligned into different languages. The existing datasets miss some language features, such as discourse connectors and term annotations, making it hard to improve document-level MT. The problem is bigger when translating dialogues. This is because datasets from sources like movie subtitles often fail to show who is speaking each line. Moreover, the MT community needs more datasets for languages with complicated grammar and vocabulary. These would help test and improve document-level MT methods.

- **Explicit Discourse-level Linguistic Annotation**: Automating the discourse annotation process can help develop and evaluate document-level MT systems. For example, acquiring annotations of discourse entities can directly influence their translation improving lexical cohesion. I believe annotating discourse phenomena like coreference and discourse markers is essential for advancing the field.

Bibliography

Agrawal, R., Turchi, M., and Negri, M. (2018). Contextual handling in neural machine translation: Look behind, ahead and on both sides. In Pérez-Ortiz, J. A. et al., editors, *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 11–20, Alacant, Spain. Universitat d'Alacant.

Alam, M. M. I., Kvapilíková, I., Anastasopoulos, A., Besacier, L., Dinu, G., Federico, M., Gallé, M., Jung, K., Koehn, P., and Nikoulina, V. (2021). Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P., Lazar, R., Lewis, W., Neubig, G., Niu, M., Öktem, A., Paquin, E., Tang, G., and Tur, S. (2020). TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ao, S. and Acharya, X. (2021). Learning ULMFiT and self-distillation with calibration for medical dialogue system. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 196–203, Trento, Italy. Association for Computational Linguistics.

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

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Bao, G., Teng, Z., and Zhang, Y. (2023). Target-side augmentation for document-level machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10725–10742, Toronto, Canada. Association for Computational Linguistics.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Castilho, S., Cavalheiro Camargo, J. L., Menezes, M., and Way, A. (2021). DELA corpus - a document-level corpus annotated with context-related issues. In *Proceedings of the Sixth Conference on Machine Translation*, pages 566–577, Online. Association for Computational Linguistics.

Castilho, S., Popović, M., and Way, A. (2020). On context span needed for machine translation evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

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

Cettolo, M., Niehues, J., Stuker, S., Bentivogli, L., Cattoni, R., and Federico, M. (2015). The iwslt 2015 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation.*, Trento, Italy.

Chen, G., Chen, Y., Wang, Y., and Li, V. O. (2020a). Lexical-constraint-aware neural machine translation via data augmentation. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, (IJCAI 2020a)*, pages 3587–3593. ijcai.org.

Chen, J., Li, X., Zhang, J., Zhou, C., Cui, J., Wang, B., and Su, J. (2020b). Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Dougal, D. K. and Lonsdale, D. (2020). Improving NMT quality using terminology injection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4820–4827, Marseille, France. European Language Resources Association.

Farajian, M. A., Turchi, M., Negri, M., and Federico, M. (2017). Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Fernando, A., Ranathunga, S., and Dias, G. (2020). Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. *ArXiv*, abs/2011.02821.

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces*, 6(2):291–309.

Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., and Zhang, Z. (2019). Star-transformer. In *Proceedings of the 2019 Conference of the North American Chapter of the Asso-*

ciation for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1315–1325, Minneapolis, Minnesota. Association for Computational Linguistics.

Haque, R., Hasanuzzaman, M., and Way, A. (2019). Investigating terminology translation in statistical and neural machine translation: A case study on english-to-hindi and hindi-to-english. In *Proceedings of RANLP 2019: Recent Advances in Natural Language Processing*, pages 437–446, Varna, Bulgaria.

Haque, R., Moslem, Y., and Way, A. (2020). Terminology-aware sentence mining for NMT domain adaptation: ADAPT's submission to the adap-MT 2020 English-to-Hindi AI translation shared task. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23, Patna, India. NLP Association of India (NLPAI).

Haque, R., Penkale, S., and Way, A. (2018). Termfinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation*, 52:365–400.

Hearne, M. and Way, A. (2011). Statistical machine translation: A guide for linguists and translators. *Language and Linguistics Compass*, 5(5):205–226.

Herold, C. and Ney, H. (2023a). Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

Herold, C. and Ney, H. (2023b). On search strategies for document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12827–12836, Toronto, Canada. Association for Computational Linguistics.

Huang, H., Wu, S., Liang, X., Zhou, Z., Yang, M., and Zhao, T. (2023). Iterative nearest neighbour machine translation for unsupervised domain adaptation. In

Findings of the Association for Computational Linguistics: ACL 2023, pages 13294–13301, Toronto, Canada. Association for Computational Linguistics.

Hung, C.-C., Lange, L., and Strötgen, J. (2023). TADA: Efficient task-agnostic domain adaptation for transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 487–503, Toronto, Canada. Association for Computational Linguistics.

Hutchins, W. J. (1986). *Machine Translation: Past, Present, Future.* John Wiley & Sons, Inc., USA.

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

Jiang, S., Wang, R., Li, Z., Utiyama, M., Chen, K., Sumita, E., Zhao, H., and Lu, B.-l. (2019). Document-level neural machine translation with associated memory network.

Jiang, S., Wang, R., Li, Z., Utiyama, M., Chen, K., Sumita, E., Zhao, H., and Lu, B.-l. (2021). Document-level neural machine translation with associated memory network. *IEICE TRANSACTIONS on Information and Systems*, E104-D(10):1712–1723.

Junczys-Dowmunt, M., Grundkiewicz, R., Guha, S., and Heafield, K. (2018). Approaching neural grammatical error correction as a low-resource machine translation task. 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 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

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

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

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., 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. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Kuang, S. and Xiong, D. (2018). Fusing recency into neural machine translation with an inter-sentence gate model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 607–617, Santa Fe, New Mexico, USA. 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.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lee, G., Yang, S., and Choi, E. (2021). Improving lexically constrained neural machine translation with source-conditioned masked span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 743–753, Online. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Li, X., Zhang, J., and Zong, C. (2018). One sentence one model for neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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

Luong, M.-T. and Manning, C. (2015). Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

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.

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

Macé, V. and Servan, C. (2019). Using whole document context in neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

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

Maruf, S., Martins, A. F. T., and Haffari, G. (2019). Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Maruf, S., Saleh, F., and Haffari, G. (2021). A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Michon, E., Crego, J., and Senellart, J. (2020). Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on 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.

Mouratidis, D., Mathe, E., Voutos, Y., Stamou, K., Kermanidis, K. L., Mylonas, P., and Kanavos, A. (2022). Domain-specific term extraction: A case study on greek maritime legal texts. In *Proceedings of the 12th Hellenic Conference on*

Artificial Intelligence, SETN '22, New York, NY, USA. Association for Computing Machinery.

Nayak, P., Haque, R., and Way, A. (2020). The ADAPT's submissions to the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 841–848, Online. Association for Computational Linguistics.

Niehues, J. (2021). Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Poncelas, A., Shterionov, D., Way, A., Maillette de Buy Wenniger, G., and Passban, P. (2018). Investigating backtranslation in neural machine translation. In Pérez-Ortiz, J. A., Sánchez-Martínez, F., Esplà-Gomis, M., Popović, M., Rico, C., Martins, A., Van den Bogaert, J., and Forcada, M. L., editors, *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 269–278, Alicante, Spain.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of* *the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rarrick, S., Quirk, C., and Lewis, W. (2011). MT detection in web-scraped parallel corpora. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Rehm, G. and Way, A. (2023). European language equality: Introduction. In Rehm, G. and Way, A., editors, *European Language Equality - A Strategic Agenda for Digital Language Equality*, Cognitive Technologies, pages 1–10. Springer.

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

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Sato, S., Sakuma, J., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M. (2020). Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4269–4279, Online. Association for Computational Linguistics.

Scansani, R. and Dugast, L. (2021). Glossary functionality in commercial machine translation: does it help? a first step to identify best practices for a language service provider. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Schamper, J., Rosendahl, J., Bahar, P., Kim, Y., Nix, A., and Ney, H. (2018). The RWTH Aachen University supervised machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 496–503, Belgium, Brussels. Association for Computational Linguistics.

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

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K., and Zhang, M. (2019). Code-switching for enhancing NMT with pre-specified 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 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Stojanovski, D. and Fraser, A. (2021). Addressing zero-resource domains using document-level context in neural machine translation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 80–93, Kyiv, Ukraine. Association for Computational Linguistics.

Sukhbaatar, S., Grave, E., Bojanowski, P., and Joulin, A. (2019). Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 331–335, Florence, Italy. Association for Computational Linguistics.

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

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Tan, X., Zhang, L., Xiong, D., and Zhou, G. (2019). Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.

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.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Tu, Z., Liu, Y., Shi, S., and Zhang, T. (2018). Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008, Long Beach, CA, USA. 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, X., Weston, J., Auli, M., and Jernite, Y. (2019). Improving conditioning in context-aware sequence to sequence models. *ArXiv*, abs/1911.09728.

Way, A. (2018). *Quality Expectations of Machine Translation*, pages 159–178. Springer International Publishing, Cham.

Way, A. and Hearne, M. (2011). On the role of translations in state-of-the-art statistical machine translation. *Language and Linguistics Compass*, 5(5):227–248.

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

Xu, H., Xiong, D., van Genabith, J., and Liu, Q. (2020a). Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3933–3940. International Joint Conferences on Artificial Intelligence Organization. Main track.

Xu, H., Xiong, D., van Genabith, J., and Liu, Q. (2020b). Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *International Joint Conference on Artificial Intelligence*, pages 3933–3940, Yokohama, Japan.

Yamagishi, H. and Komachi, M. (2020). Improving context-aware neural machine translation with target-side context. In Nguyen, L.-M., Phan, X.-H., Hasida, K., and Tojo, S., editors, *Computational Linguistics*, pages 112–122, Singapore. Springer Singapore.

Yang, Z., Zhang, J., Meng, F., Gu, S., Feng, Y., and Zhou, J. (2019). Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1527–1537, Hong Kong, China. Association for Computational Linguistics.

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

Yun, H., Hwang, Y., and Jung, K. (2020). Improving context-aware neural machine translation using self-attentive sentence embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9498–9506, Newyork,USA.

Zhang, B., Bapna, A., Johnson, M., Dabirmoghaddam, A., Arivazhagan, N., and Firat, O. (2022). Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192, Dublin, Ireland. 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.

Zhang, L., Zhang, T., Zhang, H., Yang, B., Ye, W., and Zhang, S. (2021). Multi-hop transformer for document-level machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3953–3963, Online. Association for Computational Linguistics.

Zhang, Z., Li, J., Tao, S., and Yang, H. (2023). Lexical translation inconsistency-aware document-level translation repair. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12492–12505, Toronto, Canada. Association for Computational Linguistics.

Zheng, Z., Yue, X., Huang, S., Chen, J., and Birch, A. (2021). Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, page 551, Yokohama, Yokohama, Japan. IJCAI.