university of groningen / faculty of science and engineering

Department of Artificial Intelligence

Master's thesis

Using Confidential Data for Domain Adaptation of Neural Machine Translation

Author

Sohyung Kim (S3475743)

Main Supervisor

Dr. Arianna Bisazza

Language Technology, University of Groningen

Internal Supervisor

Dr. Jennifer Spenader

Artificial Intelligence, University of Groningen

External Supervisor

Dr. Fatih Turkmen

Computer Science, University of Groningen

August 11, 2021

Abstract

Domain adaptation has led to remarkable achievements in Neural Machine Translation (NMT). Therefore, the availability of in-domain data remains essential to ensure the quality of NMT, especially in technical domains. However, obtaining such data is often challenging, and in many real-world scenarios this is further aggravated by data confidentiality or copyright concerns.

We study the problem of domain adaptation in NMT when domain-specific data cannot be shared due to confidentiality issues. We propose to fragment data into phrase pairs and use a shuffled and random sample to fine-tune a generic NMT model instead of using the full sentences. Despite the loss of long segments, we find that NMT quality can considerably benefit from this adaptation and that further gains can be obtained with a simple tagging technique.

Keywords — Confidential Data/ Domain Adaptation/ Neural Machine Translation/ Phrase Pairs/ Fine-tuning/ Transformer

Acknowledgements

Without all the support and help of many individuals, it would not have been possible to complete my Master's thesis. First and foremost, I would like to extend my deepest appreciation to my main advisor, Dr. Arianna Bisazza for unlimited supervision, support and patience throughout my graduate thesis work. She guided me by sharing her expertise and insights. Furthermore, due to our weekly online meeting, I could not get lost and could keep motivated, especially in the COVID-19 time. I am truly grateful to her for giving me such meaningful and valuable experience.

I also sincerely appreciate my other supervisors Dr. Jennifer Spenader and Dr. Fatih Turkmen who provided valuable comments and advice. With their guidance, my ideas could be more clear and concrete. I am thankful for them to devote their valuable time for my thesis.

Furthermore, I would like to express my gratitude towards my parents who support and encourage me to study abroad. Because of their dedication, I could focus on my work without any concern. Lastly, I would like to thank all my friends who have always supported me with warm comfort.

Nomenclature

AI	Artificial Intelligence
LSTM	Long Short-Term Memory
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
OOV	Out-Of-Vocabulary
SMT	Statistical Machine Translation

Contents

Chapter 1

Introduction

With the growth of the international market and the freedom to share information through the Internet, the need for translation is increasing exponentially. It is getting difficult to meet this explosive demand for translation by only human translators due to cost and time constraints. Accordingly, an approach to Machine Translation (MT), automatically translating one language into another language, has continuously received significant attention as a possible solution.

Therefore, many generations of MT models have evolved since the early 1950s, when the first practical MT models were suggested (Hutchins, 2007). For a while, Statistic Machine Translation (SMT) was the dominant framework for MT research and industry. However, more recently, Neural Machine Translation (NMT) (Bahdanau et al., 2014; Kalchbrenner & Blunsom, 2013), which uses neural network models to solve MT task, has replaced SMT with significant progress. One of the main reasons NMT outperforms other traditional MT models is that it is trained in an end-to-end fashion. While SMT consists of subsequent components that are trained separately, NMT instead combines all of the components into one big trainable encoder-decoder structured model. This allows NMT to be able to have better exploitation of context than SMT.

Although NMT yields state-of-art performance, it still shows several weaknesses. The most common problem is that NMT relies heavily on training data like other deep learning models because it has an architecture based on deep neural networks. This means that it is not only data-hungry but also very sensitive to the domain difference between training and test data. many research studies reported that NMT performs poorly for domain specific translation in low resource scenarios (Koehn & Knowles, 2017; Östling & Tiedemann, 2017; Sato et al., 2020; Zoph et al., 2016). However, in practice, large scale parallel data (several million sentences) is available only in limited language pairs and domains. To cope with this data scarcity, domain adaptation — using knowledge of the source domain to improve the performance of the target domain — is frequently used. In particular, as a conventional domain adaptation technique for NMT, fine-tuning (Chu et al., 2017; Freitag & Al-Onaizan, 2016) is employed in cases where large out-domain and relatively small in-domain parallel datasets are feasible.

The availability of high quality in-domain data, therefore, remains essential to ensure the quality of NMT, especially in technical domains (Koehn & Knowles, 2017). However, obtaining such data is still

challenging. In many real-world scenarios, this is further aggravated by data confidentiality or copyright concerns. For example, consider the following scenario: a translation company based on a pipeline of NMT and human post-editing, wants to access its clients' data and translations of that data to improve NMT quality. However, when the content of the data is highly sensitive, the owner of the data (the clients of the translation company) may limit or simply deny access to the data and its translations because of privacy concerns (Cancedda, 2012). Missing the opportunity to use such sensitive in-domain data can lead to considerably worse MT quality, higher post-editing efforts and subsequently higher translation costs for the data owners themselves. In this context, we begin to question the feasibility of exploiting the parallel datasets that cannot be used for reasons of confidentiality to improve NMT quality. If a NMT system can take advantage of any part of high-quality in-domain data, the data owner and the translation company could benefit together from reduced post-editing cost.

Our main observation is that, in natural language processing (NLP), when the complete data cannot be shared in its original form, releasing *fragmented* data can be considered as a compromise. The most well-known example of releasing fragmented data is *Google N-gram* (Michel et al., 2011). N-gram tables consisting of sequences of n words and their counts in a given corpus were routinely used to train count-based language models (Brants et al., 2007; Kneser & Ney, 1995) before the advent of neural methods. However, fragmented data like N-grams is not optimal for training state-of-the-art NMT models that are based on deep neural networks such as sequence-to-sequence LSTM (Sutskever et al., 2014) or Transformers (Vaswani et al., 2017). As mentioned above, one of the main strengths of these models is their ability to handle arbitrarily long contexts, which would be hindered by the use of fragmented data. In this thesis, we take a pragmatic approach and ask: If the data owner can *only* release fragmented data due to confidentiality issues, can this still benefit downstream NMT quality in any way? As a solution, we propose *phrase pairs* for a fragmented text format. Phrase pairs are one of the major components of phrase-based MT that keeps word alignments.

Motivated by the brittleness of NMT in out-of-domain settings (Koehn & Knowles, 2017) and the increasing availability of large pre-trained models (Ng et al., 2019), in this thesis, we focus on the task of adapting a strong-performing general-domain pre-trained NMT system to various technical domains. As a viable solution to exploit confidential data, we fine-tune phrase pairs of in-domain data as parallel sentences. Furthermore, to maximise the utility of phrase pairs for fine-tuning a NMT model while preserving the confidentiality of data, we devise various methods for presenting phrases to models.

1.1 Research Questions

This project aims to answer the following research question:

> *In the scenario where the original data is not shareable due to confidentiality issues and* ***only shuffled phrase pairs*** *can be released as a compromise, can this benefit downstream NMT quality in any way?*

We will address the main research question by answering the following sub-questions:

1. How much does the translation quality of out-of-domain models improve over the baseline models when fine-tuning on in-domain phrase pairs?
2. Does the use of shorter phrases (i.e. more fragmented data) lower translation quality?
3. Can the phrase adapted NMT model's translation quality be improved by applying tagging techniques to present phrase pairs to the NMT model?
4. When fine-tuning the NMT model on phrase pairs, are there any significant differences between different test domains?

1.2 Thesis Outline

This thesis is organised as follows. Chapter 2 provides some background on NMT, domain adaptation and regularization methods for fine-tuning NMT models. In Chapter 3, we propose our approach for answering the research questions with a motivated scenario. Chapter 4 provides all details of our experiments. We evaluate the results of all experiments and represent analysis of the results in chapter 5. Finally, we answer the research questions and conclude our findings with possible future work in Chapter 6.

Part of this thesis has been published as a workshop paper (Kim et al., 2021).

Chapter 2

Background

This chapter walks through the background required to understand our research. In Section 2.1, we explain what is machine translation (MT), and its evolution by covering several MT approaches in the past. We explain what is neural machine translation (NMT) and discuss on the state of the art architectures in Section 2.2. In Section 2.3, one of the common methods for improving NMT performance, domain adaptation will be covered. For training MT systems, assessment of system output is essential. Section 2.4 will describe automatic evaluation metric of MT and especially, BLEU. Training such NMT models are complex and can confront overfitting issues. In Section 2.5, we introduce multiple regularization methods to prevent overfitting. Lastly, we conduct a short research review related to using confidential data in NMT.

2.1 Machine Translation

Machine translation (MT) is a field of research in natural language processing (NLP), the task of converting an input consisting of a sequence of words in a source language into a sequence of words in a target language. This can be formulated as Equation 2.1.

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}}\, \Psi(\mathbf{s}, \mathbf{t}) \tag{2.1}$$

where $\mathbf{s}$ and $\mathbf{t}$ represent sentences of source and target languages, receptively. $\hat{\mathbf{t}}$ is generated translation and Ψ is a scoring function. In general, MT models consist of decoding $\mathbf{s}$ into $\mathbf{t}$ and learning algorithms for parameters of Ψ.

2.1.1 Quick MT history

MT began around the 1950s and has since developed numerous approaches. The early MT models were mostly about a direct word to word translation based on bilingual dictionaries, rule-based MT (RBMT). Linguistic experts created a large set of rules for each language pair. Based on the built-in rules, the MT system translates directly source words to corresponding target words. However, RBMT required too many linguistic rules that were manually created and adapting new rules was complicated (Lagarda

et al., 2009). Therefore, to reduce linguistic information and human involvement in building rules, the desire to teach MT systems through examples began to arise naturally.

In the 1990s, Brown et al., 1993 proposed corpus-based approaches as a beginning of modern statistical MT (SMT): a parallel corpus was needed with minimum linguistic information to train MT systems. Finally, Berger et al., 1994 introduced the first SMT model without any linguistic rules. The main idea of SMT is to learn a probabilistic model from samples of parallel corpora. Various models have been proposed for SMT, in particular the phrase-based SMT (PBSMT) model (Koehn et al., 2003a) has been widely used.

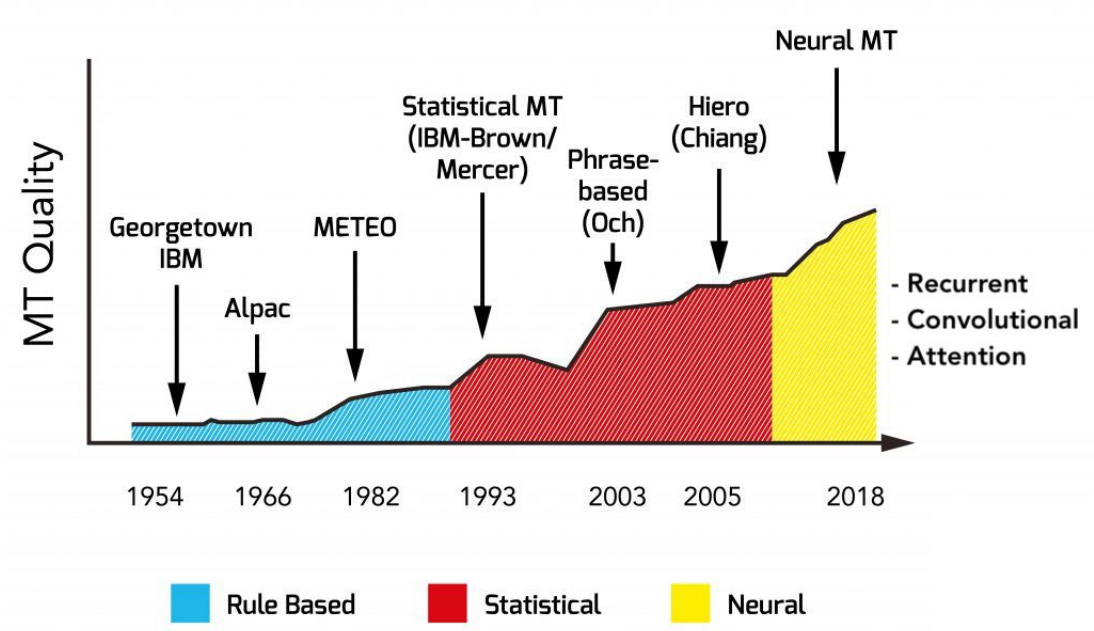


Figure 2.1: The improvement of MT quality: Since the 1950s, the MT systems have been evolved through multiple approaches. Currently, NMT dominates in MT. Figure taken from Iconic-Translation-Machines-Ltd, 2019.

2.2 Neural Machine Translation

More recently, advances in neural machine translation (NMT), an approach to MT that uses a neural network, have led to astonishing improvements in MT (Bahdanau et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017). As Figure 2.1 shows, NMT represents the current state-of-the-art in this task (Hassan et al., 2018). One of the strengths of NMT is unlike traditional SMT models, it can directly learn the mapping from the input sequence to the corresponding output sequence in an end-to-end manner. The encoder-decoder architecture enables NMT to exploit bigger context information than SMT models.

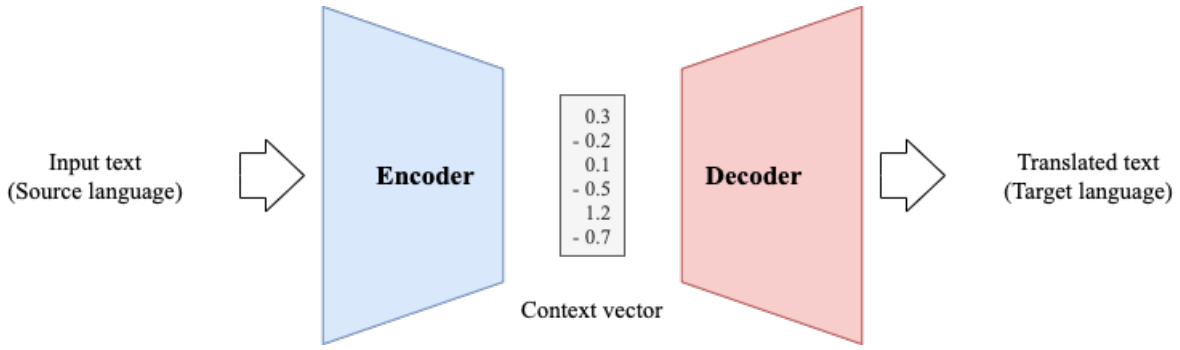


Figure 2.2: Encoder-Decoder architecture for NMT. The model builds a fixed size vector representation with an input sequence. The vector is generated into the output sequence by the decoder.

2.2.1 Encoder-Decoder architecture

NMT systems are based on the *encoder-decoder* architecture (Sutskever et al., 2014) which consists of the encoder and decoder. The encoder network receives an input sentence and converts it into a representation in a fixed-sized vector, so called context vector. This context vector contains all the valuable features and information of the input sequence. Then, the decoder network converts the vector into a sequence output in the target language. To do so, the decoder generates the output sentence word by word. The encoder and decoder networks are trained in end-to-end manner. This whole process of an encoder-decoder architecture for NMT is illustrated in Figure 2.2.

In the decoding process, the decoder network aims to choose the most likely output sequence from the target vocabulary. To search for the most proper tokens, various decoding algorithms have been proposed. Currently, beam search is widely used in state-of-the-art NMT systems. Instead of immediately selecting the word with the highest probability as the next word, beam search generates sentences by selecting words with high probability in the second and third candidate words (range is determined by beam size). It then calculates the sum of the probabilities of the candidate sentences, and generates the highest-scoring, 'best' output sentences.

2.2.2 Attention mechanism

In the previous section, the encoder compresses an input sequence into a single fixed-size vector representation called a context vector, and the decoder converts an output sequence from it. Although encoder-decoder architecture is effective, it has problems with long sequences. Since the encoder packs all the information and features into one fixed-size vector, the problem of information loss occurs. This leads to the decrease of translation quality in long input sentences.

As a solution, Bahdanau et al., 2014 introduced the attention mechanism. The main idea of attention is that at every time step when the decoder predicts the output word, the entire input sequence at the encoder is referred to once again. However, instead of referring to the whole input sentence with the same importance, some input words that are more related to the output word predicted at that time step are given more attention. In other words, with the attention mechanism, we can discover which input

word of the encoder is the most related to the decoder output at a specific time step.

2.2.3 Transformers

The current state-of-the-art NMT architecture is the transformer architecture (Vaswani et al., 2017). This architecture does not use recurrent layers commonly used in the encoder-decoder architectures for sequence-to-sequence models. The transformer uses multi-head self-attention to reduce sequential computation, making the training process more parallelizable while at the same time modelling more inter-word dependencies. This allows the model to reduce training time and alleviate long word distance dependency issues in a sentence. Figure 2.3 describes the architecture of the transformer.

Although the transformer does not use recurrent neural networks (RNN), it maintains an encoder-decoder structure. In the traditional encoder-decoder structure, the RNN in the encoder and in the decoder had t time-steps, but instead of this, in the transformer, the encoder and decoder consists of N units. In other words, a transformer processes a sentence at once rather than word by word. To provide the information of word order to the model, transformer applies positional encoding to the input before the self-attention layer. For example, in the sentence, "My cat loves to eat when I eat", the word "eat" has the same value in a general embedding but with positional encoding it has different embedding values depending on its position in the sentence. This way, transformer can learn the positional information of each word in a sentence without recursion.

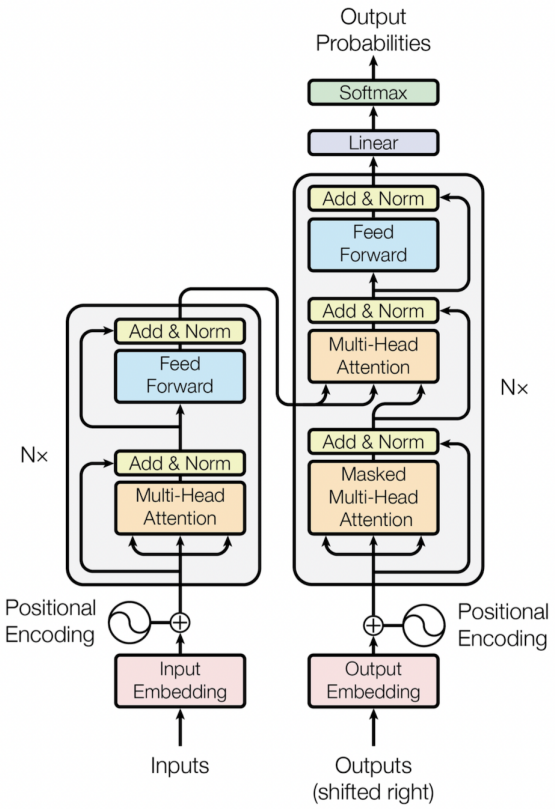


Figure 2.3: The Transformer architecture. Figure is taken from Vaswani et al., 2017.

2.3 Domain Adaptation for NMT

Deep learning approaches require huge amount of data but in practice, data scarcity issues are common. To tackle this problem, transfer learning is widely used. As Figure 2.4 describes, transfer learning exploits the knowledge obtained from previous related learning to the next task. In traditional learning, the model learns every task by training from the scratch, therefore it requires more time and more training datasets. However, transfer learning can *transfer* the knowledge to the target task with smaller target domain dataset. Domain adaptation is a related concept of transfer learning where the task remains the same but the domains are different. However, there seems to be some disagreement among researchers about the definition of transfer learning and domain adaptation. In many cases there are terminological inconsistencies which can cause confusion. In this thesis, we consider domain adaptation is a part of transfer learning.

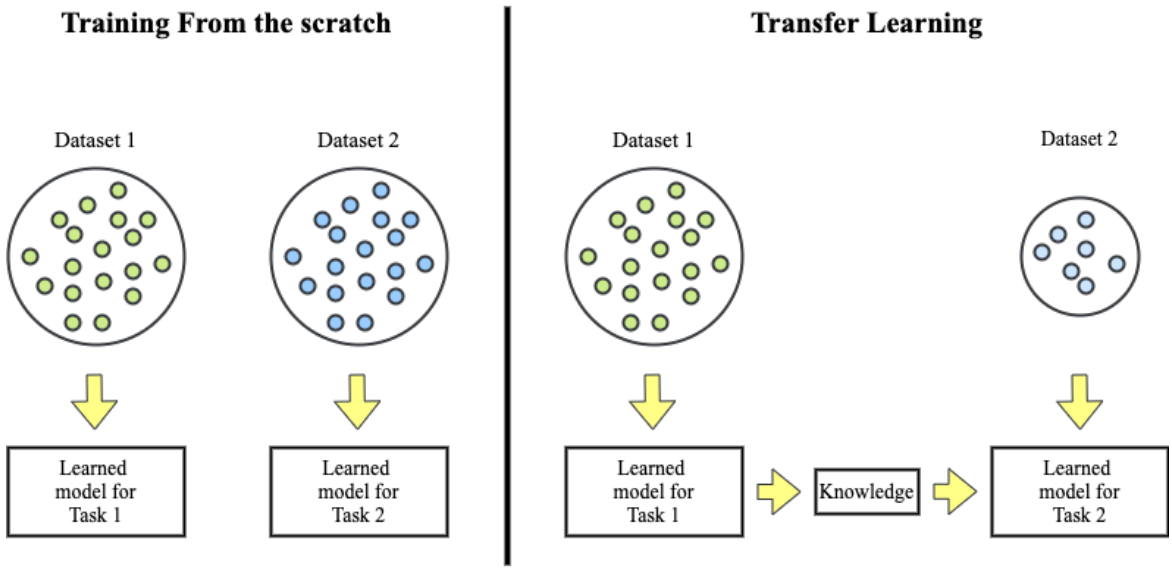


Figure 2.4: The comparison of transfer learning with traditional learning, where start training from scratch.

Domain adaptation is often applied in NMT systems where the target domain data is low resource. There are various methods for domain adaption for NMT, and a comprehensive review of the possible techniques are written by Chu and Wang, 2018. This can be categorised into two groups as Figure 2.5 represents: data and model centric.

Fine-tuning is the conventional way to apply domain adaption for NMT (Luong, Manning, et al., 2015; Sennrich et al., 2016b) in the scenario where a larger parallel out-of-domain data and a smaller parallel in-domain data are available. After pre-training a model on the out-of-domain data to pre-initialise weights of the model, fine-tuning is a way to slightly adjust a pre-trained model's weights by training further on target data.

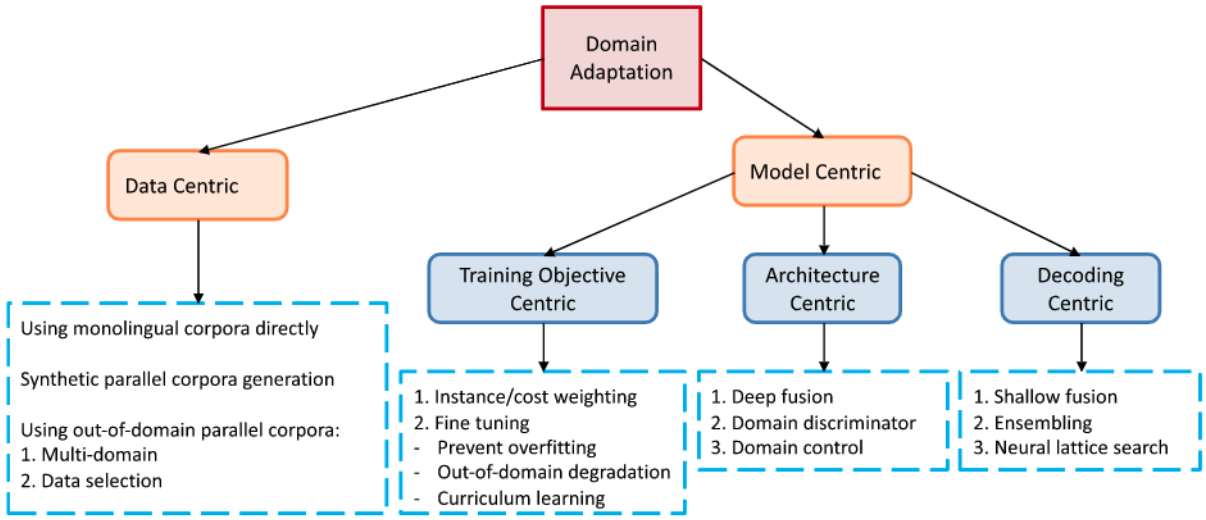


Figure 2.5: Overview of domain adaptation for NMT taken from Chu and Wang, 2018.

2.4 Automatic Evaluation Metric

Automated MT quality evaluation metrics are an indispensable part of the MT research area for verifying and analysing the efficiency of MT systems. Even if human translation evaluation is accurate and extensive, it is time-consuming, expensive, language-dependent and inherently subjective. Here, therefore, the need for automatic MT evaluation metrics arises. Automatic evaluation metrics provides a rapid, objective and consistent assessment of translation at a low price. The main idea of automatic evaluation metric is to measure how similar the model output (hypotheses text) is to a professional human translation (reference text) using statistical metrics. Multiple automatic evaluation methods of MT performance have been proposed.

2.4.1 BLEU

The **BiL**ingual **E**valuation **U**nderstudy (Papineni et al., 2002) (BLEU) method is the current standard for automatic evaluation metric in MT translations. The BLEU score of the generated translation is calculated by counting the number of N-grams[1] that overlap between the system translations and the references. BLEU is an N-gram precision metric that measures how close the system output is to a human translation and the precision score is defined as Equation 2.2.

$$p_n = \frac{\text{Number of overlapping N-grams in system output and reference}}{\text{Number of N-grams in system output}} \tag{2.2}$$

$$BP = \begin{cases} 1 & \text{if c} > \text{r} \\ e^{(1-r/c)} & \text{if c} \leq \text{r} \ , \end{cases} \tag{2.3}$$

[1]The size of N-gram can be from 1 to more than 4, and in general, size 4 is common.

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n log(p_n)\right) \tag{2.4}$$

However, precision-oriented metrics are biased to generate short outputs that only consist of high confident N-grams. To prevent this, a **brevity penalty** is applied to penalise the BLEU score if the output is shorter than the reference. Equation 2.3 shows the calculation of brevity penalty in BLEU: c and r represent the length of system output and reference, respectively. Finally, the BLEU score is defined in Equation 2.4 where commonly $N = 4$ and $w_n = \frac{1}{N}$. The BLEU scores range from 0 to 1 and we convert it to a percentage. The higher score represents the better translation: Figure 2.6 shows the interpretation of the BLEU scores by Google Translate.

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

Figure 2.6: Interpretation of BLEU score from Google Translate [2]

.

2.5 Regularization Techniques for fine-tuning

Neural network training often runs into the problem of overfitting, where the model performs exceptionally well on the training data but cannot predict the test data. In particular, the overfitting problem can easily arise in scenarios where large, high-performance pre-trained models are used for fine-tuning on typically small in-domain datasets (Geman et al., 1992). This is because neural networks cannot generalize to the unseen data, while decreasing the error on training data. To tackle this problem, various regularization methods are often applied that modify the learning algorithm for better generalisation of the model. In this section, we explain early stopping, dropout and weight decay.

2.5.1 Early stopping

When training a neural network with an iterative method, the training error decreases steadily, while the error on unseen examples (i.e. test or validation set) can worsen. To avoid the overfitting problem,

[2]This chart is from https://cloud.google.com/translate/automl/docs/evaluate

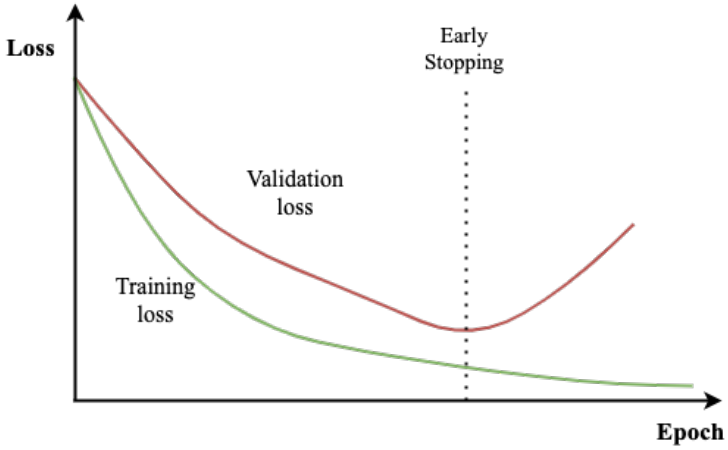


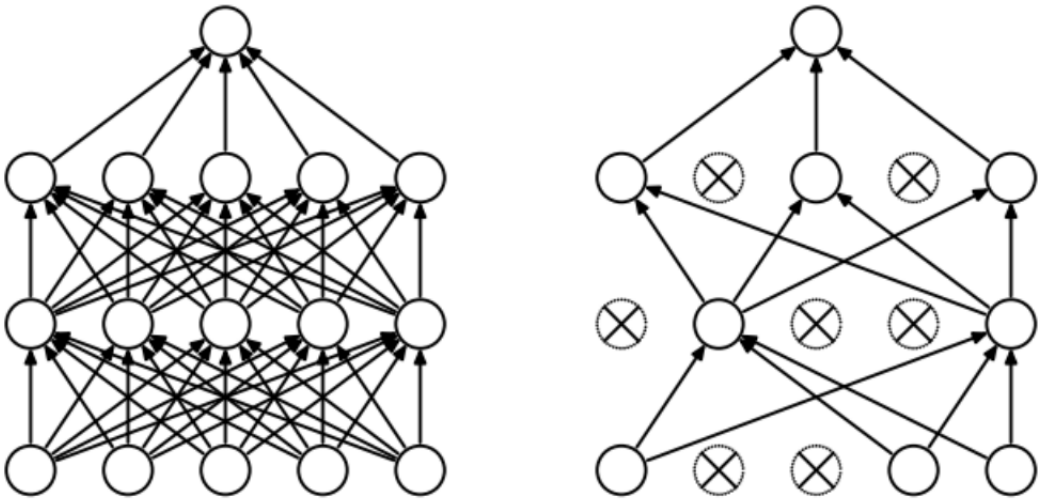
Figure 2.7: Early stopping: Learning curves describes how training and validation losses change over epochs. At a certain point, validation loss goes up whereas training loss keeps decreasing.

early stopping is the most common regularization method in deep learning because it is simple but very effective. As Figure 2.7 illustrates, early stopping aims to stop training the model when the validation set error reaches the lowest point. While training a model it is evaluated on the validation set after each epoch of training. The parameter settings are saved and training continues while this validation set error is progressively improved. Training terminates when the error on the validation set becomes worse or cannot progress compared to the previous validation error. Here, the error in the validation assumes the generalisation error.

Early stopping requires a validation set. To obtain it, the training data is split into a smaller training set and a validation set. In a scenario where cannot afford large in-domain data, some in-domain data will not be available for training. However, according to the results of Miceli Barone et al., 2017, early stopping is a powerful way to prevent overfitting, despite the disadvantage of reducing the training data.

2.5.2 Dropout

Dropout (Srivastava et al., 2014) has been widely used against overfitting problem in deep learning. As Figure 2.7 illustrates, the main idea of dropout is to randomly ”drop out” units and their corresponding connections while training the neural network. This alleviates over *co-adaptation* of units in the network. In the context of neural networks, co-adaptation refers to units that are highly correlated to each other. This may cause overfitting since these highly co-adapting units cannot detect features independently, and make it difficult to generalise to unseen data. Thus, to prevent co-adaptation, dropout makes units unreliable to others by killing random units. During training, the network is randomly sampled with a dropout probability p and only the sub-networks are trained. Afterwards, at test time, the neural network does not use dropout, and the units of the network have smaller weights than trained ones. In other words, each unit during testing is always used and the weights are multiplied by dropout probability p. This allows the network to easily estimate averaging the prediction of all sub-networks.



(a) Standard Neural Net (b) After applying dropout.

Figure 2.8: Dropout Neural Net Model (Srivastava et al., 2014): Left (a): A standard neural network. Right (b): This is one of the sub-networks. Dropout is applied on the standard network shown on the left and the random units are dropped.

2.5.3 Weight decay

When the training data is simple but the model is highly complex, overfitting occurs as the weights gradually increase during training. The larger the weights, the more it is affected by the training data, and the model fits too perfectly to the seen data. This is a phenomenon in which the model is affected by local noise and becomes fit to outliers. Weight decay is often used to avoid this problem by limiting the increase of weights to reduce the model complexity (Krogh & Hertz, 1992). Weight decay minimises a loss function by including a penalty on the L_2 Norm of the weights, and this is given by the following Equation 2.5:

$$L_{new}(w) = L_{original}(w) + \lambda w^T w, \tag{2.5}$$

where λ is a regularization parameter determining how to trade off the large weight penalty with the original loss L and $\lambda w^T w$ is a *regularizer*. Adding this regularizer to the loss function suppresses the weight increase.

2.6 Using confidential data in NMT

To our knowledge, the use of confidential data in MT has not received much attention recently. Cancedda, 2012 proposed an encryption-based method for phrase-based statistical machine translation (PB-SMT). However, PB-SMT is nowadays clearly outperformed by NMT (Bentivogli et al., 2016), which function completely differently compared to classical statistical models. Therefore, new solutions should be required to preserve data confidentiality for NMT systems.

In the broader context of NLP, secure multi-party computation (Feng et al., 2020) and homomorphic encryption (Al Badawi et al., 2020) have been used to provide strong privacy guarantees. Secure multiparty computation (MPC or SMPC) distributes a task across multiple parties while keeping each individual data private to other parties. Homomorphic encryption (HE) protects user's data when it is sent to outside of user's environment by encrypting. Operations or functions can be performed on encrypted data without decryption. However, these cryptographic methods incur high performance penalties such as slowing down training or dropping accuracy (see Riazi et al., 2019 for an overview of their performance in deep learning). More recent proposals have focused on the careful use of simpler cryptographic primitives while training a model over encrypted text due to confidentiality reasons. For instance, TextHide (Huang et al., 2020) allows to perform natural language understanding tasks while requiring the participants to complete an encryption step in a federated setting.

The aforementioned studies are mostly assuming accessibility of the original data, and focus on preventing explicit or implicit leakage of partial information while training the models on such data. However, in many real life cases, although the original data is essential to improve NLP models including NMT systems, we cannot have the right to access it. By contrast to previous studies, the novel approach taken in this thesis work is exploring the possibility of using fragmentation of confidential data for improving state-of-the-art NMT applications when no other in-domain data is available.

2.7 Using dictionary in NMT

Several studies have been done on using dictionaries or lexicons to improve the accuracy of rare words in NMT systems. Arthur et al., 2016 attempted to incorporate discrete probabilistic lexicons into the softmax layer of an NMT model to solve difficulties with the handling of low-frequency words. On the other hand, some studies suggest solving the issues by generating pseudo sentence pairs. Fadaee et al., 2017 proposed an effective translation data augmentation method for NMT systems in a low-resource scenario. Their approach consists of using a large monolingual corpus to train language models that generate new sentence pairs containing rare words in a new synthetic context.

More recently, Thompson et al., 2019 released Human Annotated Bilingual Lexicons (HABLex) that can easily be integrated into the training or decoding part of NMT systems. In this study, a pre-trained NMT system was fine-tuned on high quality bilingual lexicons created by bilingual experts. Furthermore, the authors applied an elastic weight consolidation (EWC) (Kirkpatrick et al., 2017) training method to prevent catastrophic forgetting. The results showed that combining HABLex and the EWC method improves NMT quality.

Unlike previous studies, our goal is not solving issues regarding rare words in low-resource NMT by generating new sentence pairs based on dictionaries or by using human-generated bilingual lexicons. In this thesis, we study whether a pre-trained NMT model can benefit from fine-tuning directly on bilingual short segments or dictionaries automatically extracted from the original in-domain text.

Chapter 3

Methodology

This chapter presents our approach for answering the research questions mentioned in Chapter 1. Before diving into the details of our methods, in Section 3.1, we first introduce a concrete scenario motivating the research questions: A translation company wants to improve the quality of NMT systems by only using fragmented data instead of full sentences as a compromise. Based on the scenario, we finally propose our solution for the given problem. Our approach consists of two parts: Extraction of fragment text and domain adaptation for NMT. In Section 3.2, we propose to use *phrase pairs* as a parallel corpus fragmentation method. In Section 3.3, we suggest domain adaptation of a NMT model with phrase pairs. Specifically, we choose fine-tuning, the most conventional way for domain adaptation in NMT. Finally, in Section 3.5, we present the pipeline to implement and evaluate the proposed method based on the motivated scenario.

3.1 Scenario

The translation industry has long been limited to human translation, but as the performance of MT models has improved in recent years, the use of MT is increasing to reduce costs and increase productivity. For instance, a law firm that operates across countries often needs to translate day-to-day manifold documents written in foreign languages. In this case, using only human translation is too expensive and time-consuming, therefore, MT based services are appropriate. In the early days, Statistical Machine Translation (SMT) dominated commercial MT systems, but by around 2016, many companies have started using deep learning approaches for production systems, such as Google Translate (Zhang et al., 2019), and DeepL. In this thesis, we focus on translation service based on the NMT pipeline, thus MT service denotes NMT.

Many MT-based translation companies apply various solutions to improve translation quality. The most widely used method is customising their models to the customer's domain because the general NMT performs poorly for domain-specific translations. In the aforementioned example, the law firm wants to translate the documents, which is often the legal domain, and using a generic NMT system can result in the loss of subtle nuances or terminology of the text. For instance, "stay" means "remain"

in general context but in legal domain, it often refers to the postponement or suspension of a judicial proceeding ("Stay. (n.d.) West's Encyclopedia of American Law, edition 2." 2008). When the NMT system is trained on data that is relevant to the organisation's domain, the ambiguous word problem can be lifted. As a result, many translation enterprises offer a custom MT system that uses supervised domain adaptation.

Although NMT systems show the state-of-art translation quality, it is still inferior to human translation quality (Läubli et al., 2018). To ensure official translation quality, MT companies provide NMT solutions with post-editing that indicates correcting grammar and spelling errors to improve the translation quality by human translators. If necessary, professional translators not only correct errors but also rewrite the text to meet the level of manual translation. Post-editing with NMT systems benefits cost savings, shorter lead times and reduced cognitive effort compared to manual translation, especially, domain-specific translation (Daems et al., 2017; Jia et al., 2019; Läubli et al., 2019; Toral et al., 2018).

3.1.1 Proposed scenario and solution

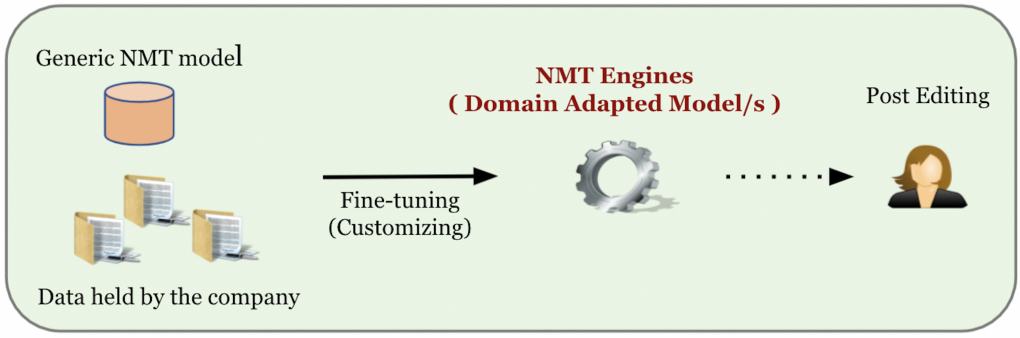


Figure 3.1: The overview of the pipeline translation company ***A***.

To clarify the motivation of our study, we consider a common case where a translation company called ***A*** provides professional translation services by using an NMT solution. As Figure 3.1 illustrates, its pipeline consists of NMT system and human post-editing. Company ***A*** customises the generic NMT model into multiple industry domains for each client to ensure the high NMT quality. To do this, each domain-specific data held by ***A*** is used for domain adaption. Then, specialised human translators refine the output of NMT engines as a post-editing step. Although ***A*** has many domain-specific datasets by itself, there may be limitations to using only them to reflect every clients' domains. It is likely difficult to apply more granular domains to an NMT system, or certain terminology frequently used by clients can not be handled properly, which increases post-editing costs and times.

As a solution, ***A*** wants to improve the quality of its NMT models by training or adapting them on the clients' previously translated documents. However, due to confidentiality and privacy concerns, the clients do not often share the original documents and its translation to ***A***. Instead of sharing the full text, as illustrated in Figure 3.2, we consider a scenario that clients of ***A*** provide their data and translation only in a fragmented form as a compromise for fine-tuning generic NMT system. If this kind of data can

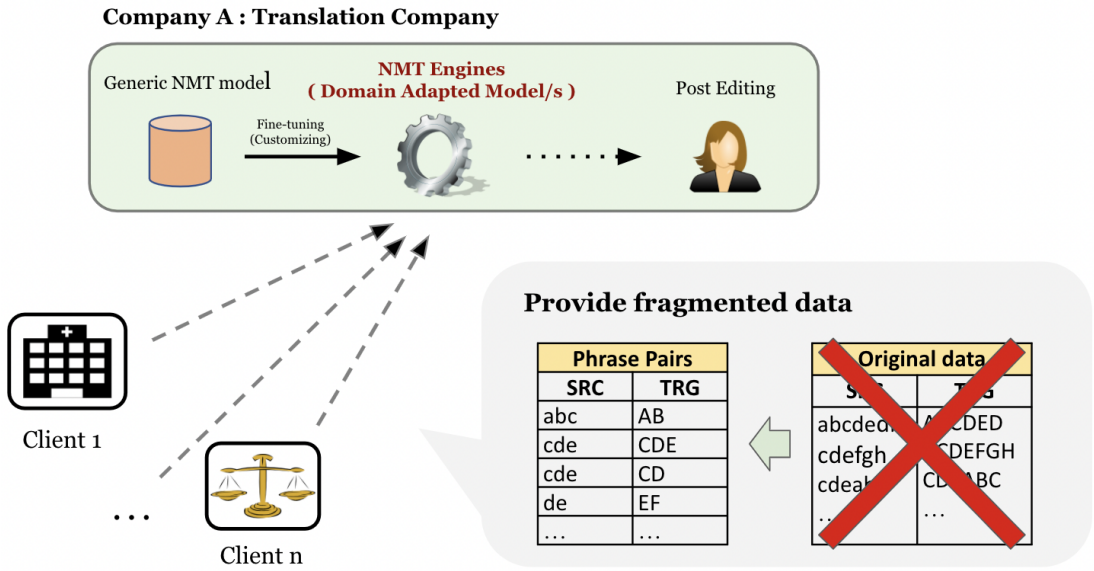


Figure 3.2: Proposal scenario : a Translation Company (***A***) uses fragmented confidential data from its clients to adapt a pre-trained generic NMT system to different industry (e.g. medical, legal) domains.

be used to improve the NMT model, both the clients and the company will benefit by abating human post-editing costs and turnaround times. Thus, we want to study the possibility of sharing fragmented data for improving utility while preserving the confidentiality of data.

We propose using *phrase pairs* as a text fragment method. In our proposed scenario, the full sentences of the original documents are not allowed for Company ***A*** and only extracted phrases are shareable. Therefore, we use the phrase pairs as parallel in-domain data for fine-tuning a baseline model. In addition, we assume that ***A*** already owns a high-performance generic NMT system, therefore, we choose a strong out-domain pre-trained NMT model.

3.1.2 Threat model

In this section, we detail the threat model that is used in this study to define properly the attacker's goal. We assume an *honest but curious* model in which the receiver of the partial data (e.g. the translation company) is not trusted or only partially trusted by the data owner. Honest but curious setting implies the adversary follows the rules but can later infer the information of the rules for malicious purposes (Goldreich, 2009). The main threat we focus on is the **full reconstruction** of the original text from a list of given n-grams of phrases rather than the protection of partial information (e.g. key phrases (Hard et al., 2018), names, social security numbers). Thus, the confidentiality consideration in our study is that 'secrets' of clients may leak when a NMT model is fine-tuned on confidential data that contains sensitive business information. The secrets stand for the core information from which the data owners obtain financial benefits. This setting is useful in various contexts where only partial data release is desired such as copyright protection. Examples of text where sensitive information is encoded in long

sequences (sentences or paragraphs) include patent applications, as well as not (yet) publicly available product analysis reports, drug reaction reports or technological processes of software. Our study is the first step in opening the possibility of using compromised fragmented data for NMT when the original data is unavailable.

3.2 Phrase Pairs

In the previous section, we presented a solution in which only fragments of the original data are used for domain adaptation of the NMT system. In NLP, releasing fragmented data in the form of N-grams has a long tradition such as Google N-grams (Michel et al., 2011). This corpus provides insight into annual trends through N-grams and statistics, but restricts access to specific subsets of documents. However, fixed-size N-gram extraction is not directly applicable to parallel data because it breaks translation equivalence with the target side. As a solution, we propose to use *phrase pairs* as a text fragmentation method.

Like N-grams, phrases are short sequences of consecutive words extracted from the input sentences. Unlike N-grams, phrases are always extracted in pairs from source-target sentence pairs in a way that is *consistent* with their word-level alignment. Formally, a phrase pair $(\bar{f}, \bar{e})$ is consistent with word alignment A if all source words $f_1, \cdots, f_n$ in $\bar{f}$ that have alignment points in A are connected with target words $e_1, \cdots, e_m$ in $\bar{e}$ and vice versa (Koehn, 2009; Koehn et al., 2003b). As Figure 3.3 illustrates, the words of the target language (German) are first automatically aligned (grey connecting lines) with the words of the source language (English) by a statistical alignment model. Then, phrase pairs of various lengths (denoted by boxes) are extracted. A user can limit the maximum length of the phrases. This example is extracted with maximum length 3 that denotes the source side phrases must consist of up to 3 words.

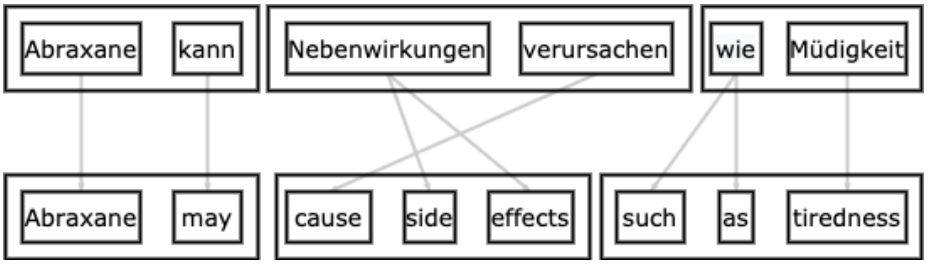


Figure 3.3: Sample of extracted phrases from EMEA training dataset.

Phrase pairs and their statistics constitute the main component of Phrase-based SMT (PB-SMT) systems, together with the target language model. In this work, however, we only use phrase extraction as a text fragmentation technique.

3.2.1 Preserving Confidentiality

We aim to ensure that the attacker cannot reconstruct the entire original document from the extracted phrases. After extraction, we shuffle the large set of phrase pairs extracted from the whole dataset and, finally, discard a random sample of phrase pairs (e.g. 50%) to preserve confidentiality. In the example of Figure 3.3, this would mean protecting the hypothetically sensitive connection between the drug name (*Abraxane*) and its reported side effect (*tiredness*).

3.3 Domain Adaptation

In our solution, phrase pairs are proposed as a compromise instead of using whole sentences to improve the NMT system. However, fragmented text such as phrases is not an appropriate form to train NMT systems. Phrase pairs are the main components for training most SMT models rather than whole sentences. SMT models obtain informative statistics from phrase pairs for translation. However, due to encoder-decoder architecture (Section 2.2), NMT models expect full sentences as input, which allows them to better exploit larger context. Indeed, this is one of the main strengths of NMT over the traditional SMT approach. As a result, training NMT on such fragmented data is likely to lead to very poor performance as a significant amount of contextual information is lost. Nonetheless, we postulate that phrase pairs may still contain very valuable information for the *adaptation* of a general-domain system to a specific target domain. In fact, much of domain adaptation has to do with learning new words or short phrases, as well as new senses for known words and phrases (Irvine et al., 2013).

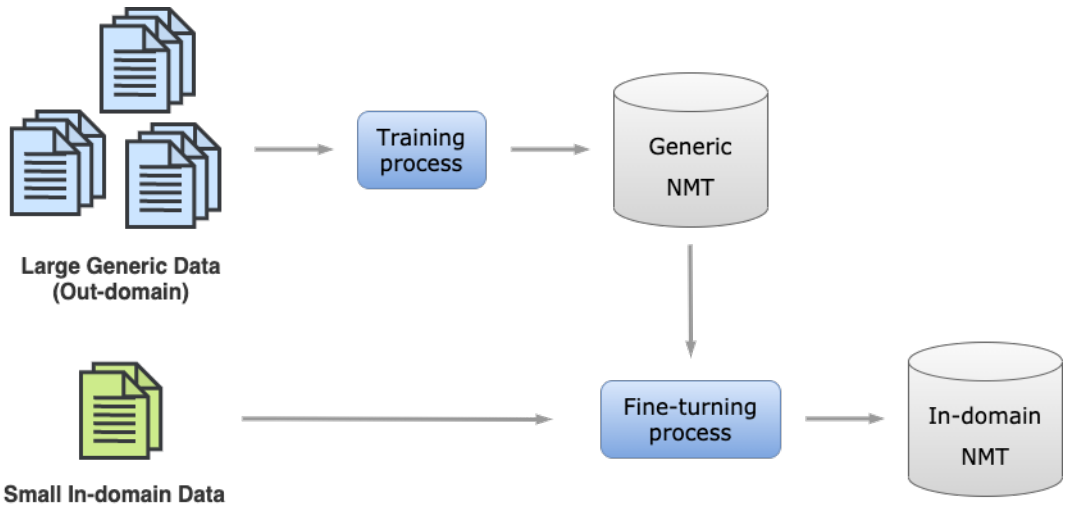


Figure 3.4: Fine-tuning for neural machine translation.

Our motivation for this study is to assess the possibility of using fragmented data without changing the architecture of a NMT system, thus we choose *fine-tuning* as the domain adaptation technique. Fine-tuning is a simple but effective strategy for domain adaptation of NMT and is currently considered the most conventional method (Luong, Manning, et al., 2015; Sennrich et al., 2016b). As Figure 3.4 describes,

in the fine-tuning process, an NMT model, which is pre-trained on large out-domain data, continues to train on small in-domain data until the convergence. We start by directly fine-tuning a general-domain NMT system on a random sample of phrase pairs (occurrences, not types) extracted from the in-domain dataset.

3.4 Tagging

In our solution, we present phrase pairs as input for the fine-tuning process of an NMT system. However, these samples are much shorter than the full sentences that were used for the pre-training process of the NMT model. This can cause phrase-adapted NMT systems to have a bias in producing relatively shorter translation outputs compared to fine-tuning on whole sentences. When the output is shorter than the reference, this is affected in lower evaluation. For instance, BLEU (Section 2.4) applies a brevity penalty to short translations compared to references. To prevent this bias, a fine-tuned NMT system on phrase pairs should be able to recognise the difference between phrases and whole sentences. This way, even when the model is fine-tuned on phrases, it may produce outputs of normal length when translating entire sentences.

To teach the NMT model the difference between phrases and whole sentences during fine-tuning, we use a simple tagging technique inspired by Sennrich et al., 2016a. The paper shows that adding artificial tags on source sentences of training data to encodes the use of politeness can control the politeness in target side at testing. Additional tags on either the source or target side help NMT systems to distinguish specific features of the data during training and finally improve translation quality at test time. This is a very effective and simple solution that does not require any changes to the architecture or other parameters. Recently, the tagging technique has also been used to differentiate various features such as multi-domains (Kobus et al., 2017), gender (Kuczmarski & Johnson, 2018) and different languages (Johnson et al., 2017). In our work, by appending tags to each phrase while fine-tuning, we provide the NMT system with information that phrases are a different type of input than normal sentences.

3.5 Pipeline

In this thesis work, experiments are constructed based on the scenario that was presented in Section 3.1. As Figure 3.5 illustrates, two pipelines are considered: The standard method that refers to a general method of fine-tuning using whole sentences and the proposed method using phrase pairs.

We use an pre-trained model, which already has high translation quality, as a *baseline model*. Using this pre-trained model gives us several benefits. This satisfies our scenario where the translation company ***A*** customises outsourced powerful NMT systems without changing the architecture, only through domain adaptation. In addition, it saves us time that can better be spent focusing on studying our approach.

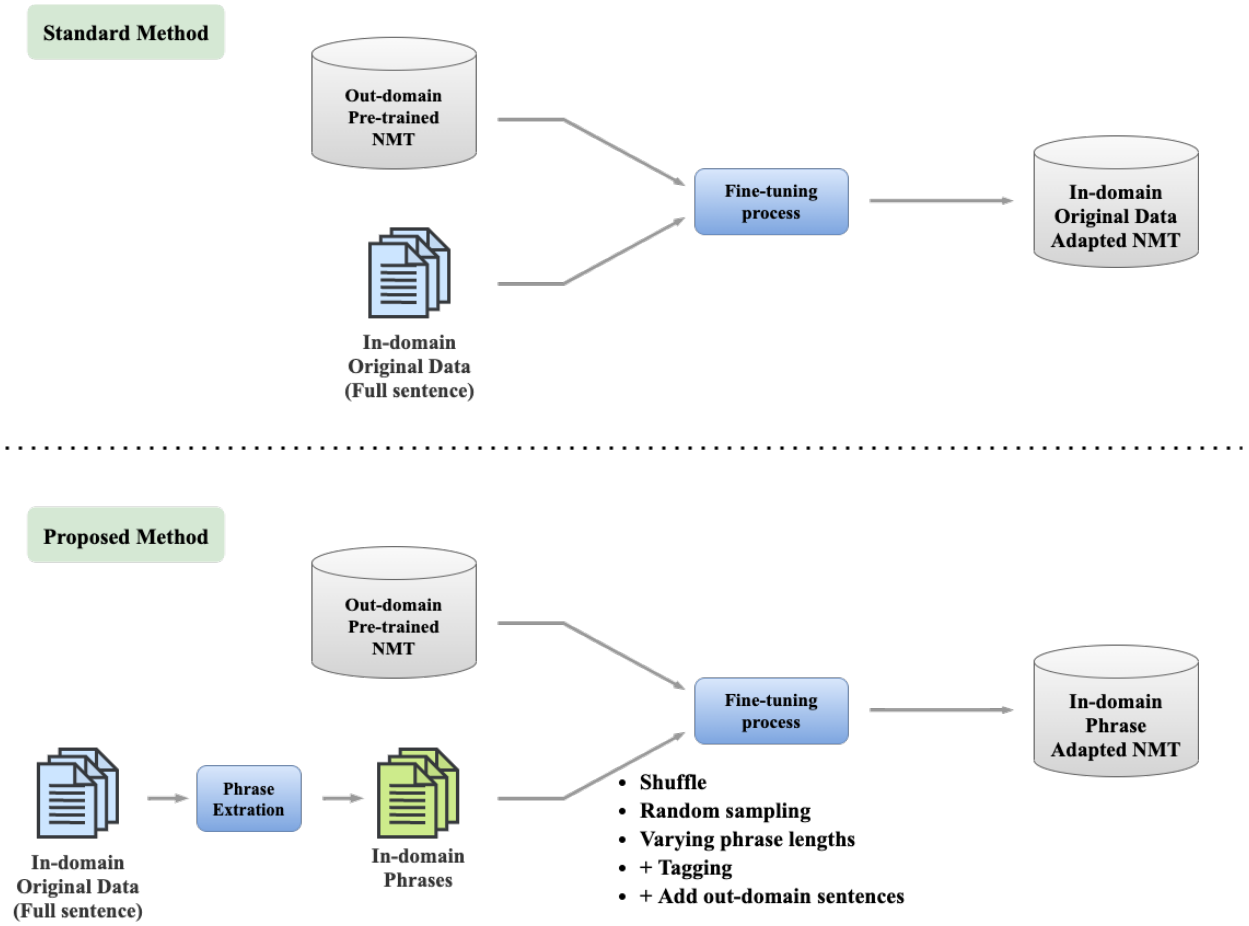


Figure 3.5: For the project, two pipelines consist of the standard method where the pre-trained NMT model is fine-tuned on in-domain original data and the proposed method where only phrase pairs are available.

3.5.1 Standard method

The standard method represents the common way of fine-tuning for NMT models (Fig 3.5, top). This means that an NMT system is fine-tuned on in-domain full sentences. However, in our scenario, whole sentences are not shareable due to confidentiality issues. We evaluate the standard method to establish the maximal potential gains by NMT domain adaptation that our method can reach in principle.

3.5.2 Proposed method

Our proposed method only uses phrases for domain adaptation rather than full sentences in the scenario where confidentiality is a concern and complete texts are not available. In the proposed method (Fig 3.5, bottom), we extract phrase pairs from the in-domain original data and feed them as a new in-domain parallel corpus to the baseline model. With this method, we investigate whether fine-tuning on phrases can also create a domain-adapted NMT system. If this fine-tuning does well, then we will have validated

a method to utilise fragmented text while retaining confidentiality.

As Section 3.1.2 mentioned, Our confidentiality concern focuses on the possibility of reconstructing the original documents from the phrases. To avoid this, after phrase extraction, we shuffle and randomly sample phrases at a certain rate (See Section 4.3 for details).

Using phrase pairs to train state-of-art NMT systems can cause poor performance because phrases contain less information compared to full sentences. Therefore, for maximising the utility of phrase pairs, we suggest several ways to present phrases to the baseline model. First, to test the effect of phrase length on translation quality, we experiment with the various maximum lengths of phrase pairs. We also expect that phrase-adopted NMT systems will have a bias in generating shorter sentences than an NMT model trained on full sentences. To alleviate this bias, we apply the tagging technique introduced in Section 3.4. We expect the tagging technique to help the model to recognising phrases as a different type of input rather than full sentences. Furthermore, inspired by Chu et al., 2017, we fine-tune the model on a mix of tagged in-domain phrase pairs and out-domain parallel sentences. The mixture of in-domain phrases and out-domain sentences is expected to help to generate longer sentences and to improve translation quality.

Chapter 4

Experimental Setup

In the previous chapter, we introduced our motivating scenario and the main ideas for answering the research questions. Based on this, the current chapter describes the setup of the experiments that are carried out in this thesis. Section 4.1 explains the baseline model that we use for all our experiments as an out-domain pre-trained model. Then, Section 4.2 provides the details of in-domain datasets and preprocessing steps applied to the datasets. Section 4.3 explains the process of extracting phrase pairs from the in-domain datasets. Finally, Section 4.4 covers the details of fine-tuning processes. In particular, we show the several methods for presenting phrase pairs to the NMT model during fine-tuning: varying maximum phrase lengths and a tagging system.

4.1 Baseline Model

We use the Transformer-based NMT system (Vaswani et al., 2017) pre-trained by Facebook for the WMT'19 news translation task (Ng et al., 2019). It is released as part of the Fairseq toolkit (Ott et al., 2019)[1]. This model basically follows a big Transformer architecture from Vaswani et al., 2017, except using the larger feed-forward network (FFN) sub-layers size (8192). It was trained on Paracrawl 27.7M sentences, and was fine-tuned on several previous years WMT news-test sets for an additional epoch. Fairseq team released an ensemble of these four models but for our study we only use 'model 1'.

Based on our scenario, to simulate a realistic production setup, a baseline model that has high performance for out-domain translation is required. This model satisfies this requirement: It was ranked first in the WMT'19 news competition (Barrault et al., 2019) with a BLEU score of 40.8 on a German-English news task.

4.2 Data

We simulate the scenario of confidential data by using publicly available datasets in several three domains: medicine descriptions, software manual and EU legislation. We consider these domains because they

[1] https://github.com/pytorch/fairseq

generally contain sensitive information that an adversary may abuse for profit or other reasons.

We evaluate our approach on a German to English translation task. For different technical domains, we choose EMEA (medical), GNOME (software) and JRC-Acquis (legal) (Steinberger et al., 2006). EMEA is a parallel text of medical guidelines from European Medicines Agency. GNOME is a collection of the text from GNOME desktop environment and software platform. JRC is a collection of legislative text from the European Union. All the public corpora are from the OPUS project (Tiedemann, 2012)[2].

4.2.1 Preprocessing

For a realistic simulation of a professional translation scenario, we split the datasets by documents. This allows keeping track of the documents from which the sentences in the datasets were extracted. In future work, this could help the quantification of reconstructing the original documents from the extracted phrases.

All datasets are tokenized by Moses Tokenizer (Koehn et al., 2007). To remove some defects from the parallel corpora that can affect the quality of NMT models, we filter out sentence pairs where either the source or target sentence is empty. We also remove duplicate sentences per documents.

Type	Domain	Sentences	Tokens (DE)	Tokens (EN)
Train	EMEA	10k	199k	209k
	GNOME		179k	194k
	JRC		279k	396k
Validation	EMEA	150	3k	3k
	GNOME		3k	3k
	JRC		4k	5k
Test	EMEA	2k	38k	42k
	GNOME		29k	30k
	JRC		53k	82k

Table 4.1: Details of datasets used in our fine-tuning experiments. The number of Tokens are round down to the nearest thousand.

To segment our data, we use the same sub-word algorithms and split rules applied on the baseline model (Section 4.1) in the Fairseq pipeline. NMT uses a fixed vocabulary size that can cause out-of-vocabulary (OOV) word issues. To mitigate this, Fairseq uses Byte Pair Encoding (BPE) (Sennrich et al., 2016c) for its subword segmentation algorithm. Especially, the baseline model used the FastBPE implementation[3] and we apply it to encode our data into sub-tokens. In addition, the baseline NMT model was pre-trained on a separate corpus, and the dictionary was built based on it. When fine-tuning on additional data, it is crucial to ensure that the new data gets consistent indices with the dictionary

[2] https://opus.nlpl.eu/
[3] https://github.com/glample/fastBPE

that it was originally trained on. Thus, we use the original dictionary from FAIRSEQ for embedding our datasets.

We assure that there is a limit to the possible amount of in-domain dataset. The original text of phrases that may be given to the translation company would still consist of a relatively small number of sentences. Based on this, we reserve 10K sentence pairs and 2K sentence pairs for training and test set respectively. For early stopping (Section 2.5.1), we reserve only 150 sentence pairs as a small validation set. The details of data statistics are shown in Table 4.1. We release the benchmarks at https://github.com/Sohyo/Using-Confidential-Data-for-NMT.

4.3 Phrase Extraction

In Chapter 3, we suggest *phrase extraction* as a text fragmentation technique in our experiments. It is a major element in the pipeline of Phrase-based SMT (PBSMT) model (Koehn et al., 2003a). PBSMT uses phrase pairs obtained by using a word-aligned parallel corpus as base units in the translation model sub-component. For our experiments, only the phrase extraction step is used.

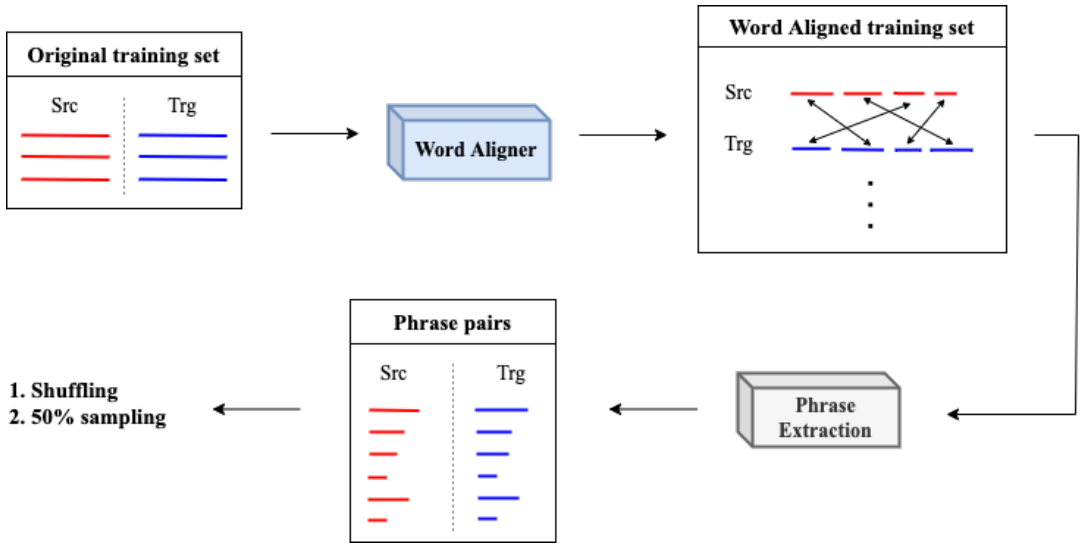


Figure 4.1: Process of phrase extraction. A word aligner establishes word alignments of parallel training datasets. Based on the bi-directional word alignments, phrase pairs are extracted. The lengths of phrases are less than or equal to a selected maximum phrase length. After phrase extraction, we shuffle the phrase pairs and sample 50% of them.

Figure 4.1 describes the steps of phrase extraction. We first word-align the in-domain datasets using FASTALIGN (Dyer et al., 2013)[4], an unsupervised word aligner. In this step, the word aligner finds the best word alignments in two separate directions (source-to-target and target-to-source) and then combines them based on the symmetrization heuristics to obtain the alignment A (Koehn et al., 2003a; Och et al., 1999). In our experiments, we take the *union* of the two alignments for the symmetrization

[4] https://github.com/clab/fast_align

A. Then we use the phrase extraction utility from the Moses phrase-based SMT toolkit Koehn et al., 2007[5] to extract all phrases consistent with A. After the phrase extraction step, our dataset has been fragmented into a list of aligned phrases of various lengths. A maximum source-side phrase length should be specified. We experimented with this number by setting the maximum phrase length to 1, 4 or 7. An example of extracted phrase pairs is shown in Table 4.2.

To prevent the full reconstruction of the original data by using the phrases, we randomly discard 50% of the all extracted phrases. Note that, because phrase pairs are considered as new parallel inputs for fine-tuning the NMT system, we extract phrases only from the training set.

	German	**English**
Original sentence	Bearbeiten der Spalten in der Nachrichtenliste	Editing the columns displayed in the list of messages
Phrase	Bearbeiten der Spalten	Editing the columns displayed
	Bearbeiten	Editing
	der Spalten	the columns displayed
	in der Nachrichtenliste	in the list of messages

Table 4.2: Extracted phrase pairs example from GNOME. The phrases are extracted from the given original German-English sentences with maximum length 4.

4.4 Fine-tuning

In Section 3.5, we presented the two pipelines of our experiments consisting of the standard and proposed method, respectively. Both methods use fine-tuning, which indicates continuing to train the pre-trained model on an in-domain training set.

To establish an upper bound of the translation quality of domain adaptation, we first conduct the standard method, a common fine-tuning technique for domain adaptation. To do so, we fine-tune the baseline model on full sentences of the original in-domain datasets. On the other hand, during fine-tuning for the proposed method, we provide phrase pairs to the models as if they were sentence pairs. After the phrase extraction step, the in-domain phrase datasets has many duplicates. Finally, to study the utility of the phrase pairs, we experiment with various setups for the fine-tuning process.

4.4.1 Varying maximum phrase lengths

To explore the difference in the efficiency of phrases of different length, we fine-tune by using various maximum phrase lengths. Although the sentence structure or context information is all broken, we conjecture the NMT model can learn new terminology from fine-tuning on maximum length 1, equivalent

[5] http://www.statmt.org/moses

to a dictionary of the in-domain dataset. The longer the phrase (max lengths 4 and 7), the richer the amount of in-domain information for domain adaptation the phrase is expected to hold.

4.4.2 Tagging

As explained in Section 3.4, we hypothesise that a tagging technique on phrases can prevent the short output bias in phrase-adapted NMT models. We experiment with a simple tagging technique by adding <PT> and </PT> at the front and end of each phrase respectively, in both source and target side. We expect this tagging system allows an NMT system can recognise the difference between the original sentences and phrases. Therefore, tags are only added to the training phrase datasets. During testing, full sentences with no tags are given to the model. Table 4.3 shows an example a pair of phrases with tags from the EMEA dataset.

Sentence	De	In den meisten Fällen wird der Serumferritinwert simultan zum Anstieg des Hämatokritwertes abfallen .
	En	In most cases , the ferritin values in the serum fall simultaneously with the rise in packed cell volume .
Phrase + tags	De	<PT> Hämatokritwertes abfallen </PT>
	En	<PT> packed cell volume </PT>

Table 4.3: A sample of a pair of full sentences and a phrase pair with tags. The phrase pair is one of the extracted phrases from the given parallel sentence.

4.4.3 Fine-tuning Hyperparameters

We apply the hyper-parameters described by Ng et al., 2019 with only a few adjustments inspired from previous work on fine-tuning regularization (Miceli Barone et al., 2017). Specifically, we divide the original learning rate by 4 to 0.000175 following the suggestion of Miceli Barone et al., 2017. Since using a smaller learning rate is preferred for fine-tuning a pre-trained model. The pre-trained model is already relatively good, therefore we do not want to distort the weights of the pre-trained model by changing them hastily or recklessly. For early stopping, we use a small (full-sentence) validation set in each domain (150 sentences, see Table 4.1) because we want to avoid reducing the amount of in-domain training data available for fine-tuning. We set the weight decay rate to 0.0001 and dropout probability to 0.2 after varying experiments. All details of other parameters of the fine-tuning are described in Table 4.4. We run all the experiments on a node with a NVIDIA V100 GPU in the Peregrine high-performance computing (HPC) cluster of the University of Groningen.[6]

[6] https://portal.hpc.rug.nl/public/start.html

Hyperparameter	**Value**
Maximum number of tokens in a batch	4096 tokens
Optimizer	Adam
Learning rate	0.00017
Epoch	20
Best checkpoint metric	BLEU
Beam search size	5
Drop out	0.2
Weight decay	0.0001
Label smoothing	0.1

Table 4.4: Hyperparameters for all fine-tuning Transformer based baseline (Section 4.1) experiments.

Chapter 5

Results and Analysis

In this chapter, we present all experimental results and their analysis. Section 5.1 shows the main results of the tagging technique and different maximal phrase lengths experiments on the three domains. In Section 5.2, to provide further explanations from various perspectives on the findings we performed analyses on the domain differences and samples of system translations. To examine the way of improving our main results, we conducted additional experiments: 1) fine-tuning on a mix of out-of-domain sentences and in-domain phrases 2) fine-tuning on the maximum length of 7 phrases where the phrases shorter than 5 words were removed. Section 5.3 reports the results of the additional experiments. Furthermore, after finishing all of the experiments and analyses, we found the JRC dataset to contain an extra noise. In Section 5.4, we re-conducted the experiments on the cleaned the JRC data and compare it with the original results.

5.1 Main Results

We evaluate the quality of NMT models by BLEU (Section 2.4) computed with SacreBLEU (Post, 2018). The phrase-adapted models are compared to the non-adapted baseline model (Ng et al., 2019), and to fine-tuned model on the original (non fragmented) dataset. This demonstrates the lower and upper bounds of the translation quality improvement by domain adaptation, and objectively evaluates our proposed method.

Main results are reported in Table 5.1. In all domains, the NMT models fine-tuned on phrases outperform the BLEU scores of the baseline model. The effect of maximum phrase length and tagging technique on the translation quality of the phrase-adapted models yields varying results across domains. For EMEA and GNOME domains, the BLEU scores of fine-tuning on phrases nearly reached the scores of fine-tuning on complete sentences. On the other hand, fine-tuning on phrases of the JRC domain achieves a relatively small improvement compared to other domains with a score increase of only up to +1.4 BLEU.

Besides BLEU, we report the average lengths of the translations from all the NMT models in Figure 5.1: the baseline, fine-tuning on maximum phrase length 4 with and without tags and fine-tuning

	Baseline *(No fine-tune)*	Fine-tuning						
		Phrase pairs						Original data *(Sentences)*
		Max length 1		Max length 4		Max length 7		
		No Tag	Tag	No Tag	Tag	No Tag	Tag	
EMEA	35.5	36.4	38.0	39.1	40.5	41.5	37.2	45.2
GNOME	29.8	29.7	33.5	36.0	37.0	35.8	36.8	38.9
JRC	29.0	28.8	30.4	29.4	30.0	29.2	29.7	54.7

Table 5.1: Main results : BLEU scores of German-English NMT in three different domains: medical (EMEA), software (GNOME), and legal (JRC). The baseline is the pre-trained Fairseq WMT19 news system (Ng et al., 2019) based on Transformer (Vaswani et al., 2017) and ranked first in the WMT19 competition. For fine-tuning on phrase pairs, multiple maximum phrase lengths (1 (dictionary), 4 and 7) are used with and without tags.

on the original full sentences. We compare them with reference sentences to examine brevity issues that cause the lower BLEU score. Sentence length is measured by the total amount of space-separated tokens. As we expected, in general, the non-tagged phrase-adapted models generate shorter translations than the reference and the baseline and the full-sentence adapted models. Adding tags on phrases improves the bias of short sentences to some extent.

Our main finding is that phrase pairs can indeed be used to fine-tune a NMT model without any changes to the architecture or the need of specific fine-tuning algorithms. It is relevant for our scenario because even translation companies without significant in-house NMT expertise could easily apply our solution to their workflow. Our approach is also applicable in cases where translation company ***A*** uses NMT as an outsourced (cloud-based) service, by sending the provider phrase pairs instead of full sentences for model adaptation. In the following, we describe in depth the effects of the tagging technique, phrase length and domains on the main results.

5.1.1 Effect of phrase tagging

We expected that the tagging technique would help an NMT system differentiate between the phrase and the original dataset and eventually could improve translation quality. In most cases, adding tags to phrases increase the BLEU scores, gaining up to +3.8 BLEU (fine-tuning on the GNOME domain with maximum length 1). In addition, the tagging technique seems to be more effective, especially for short phrases. When NMT is fine-tuned on dictionaries of in-domain datasets (i.e., maximum phrase length of 1), without tags, the model cannot take advantage of domain adaptation in the GNOME and JRC domains, but with tags, the increase in BLEU score stands out compared to other longer phrase lengths.

Figure 5.1 shows that tagging yields to slightly longer system outputs, suggesting the model indeed learned to associate the <PT>, </PT> tags with shorter training samples. While differences look small, they have a large impact on BLEU because of the Brevity Penalty (Papineni et al., 2002). As a notable

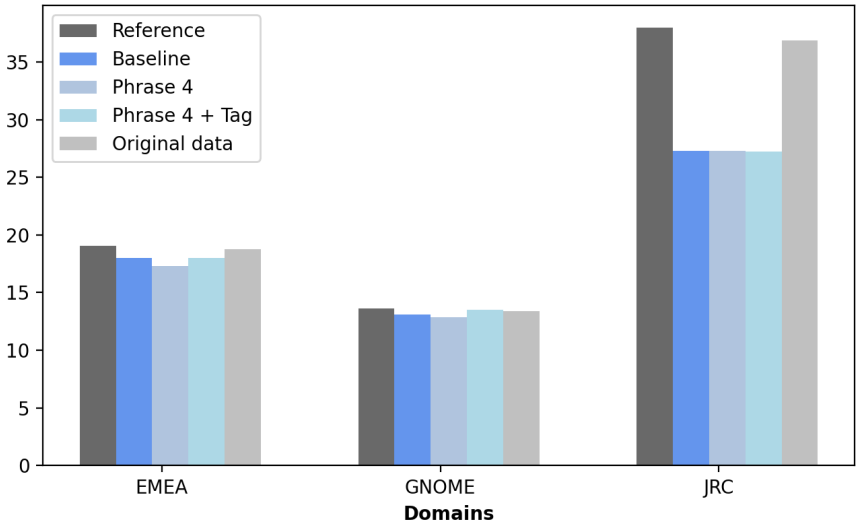


Figure 5.1: Average length (in words) of reference translations and outputs of different NMT systems.

exception to this positive trend, BLEU score decreases with tagging on EMEA with maximum length 7.

5.1.2 Effect of phrase length

We expected longer phrases to be considerably more useful for fine-tuning NMT models than short ones. This is because longer phrases may have more contextual information, at the expense of less confidentiality protection. However, our experimental results for three different maximum phrase lengths (1, 4 and 7) show that the increase in phrase length is not proportional to the increase in BLEU. The NMT models fine-tuned on phrases with a maximum length of 1 have a higher BLEU score (+1.6) than the baseline model only in the EMEA domain and decrease slightly in other domains. Fine-tuning on a maximum length of 4 phrases results in higher BLEU scores than dictionary-adapted models (except for tagged JRC domains). By contrast, increasing the maximum length from 4 to 7 does not have a positive effect on BLEU but actually lowers it in the GNOME and JRC domains. This counter-intuitive result may be due to the fact that increasing the maximum length leads to a much larger number of extracted phrases that are redundant and overlapping. Previous work on lexicon-augmented NMT also reported negative results when fine-tuning on very large numbers of segments (Thompson et al., 2019).

5.1.3 Domain differences

The benefits of fine-tuning on phrases appear to vary strongly across domains: on EMEA we obtain large gains but there is still space for improvement, on GNOME our approach nears the maximum gain achievable with original data fine-tuning, whereas on JRC gains are small and scores remain very far from the ceiling. To explain these results, we inspected our in-domain datasets and specifically looked for peculiarities of the JRC dataset. We find that JRC is rather different in terms of sentence length distribution, with much longer sentences on average. As shown in Figure 5.1, only fine-tuning on

original data leads to reasonably long outputs, whereas baseline and phrase-adapted systems all generate sentences that are, on average, about 10 words shorter then they should be. This suggests that our tagging technique is not sufficient to address the shorter-output bias in a robust way.

5.2 Analysis

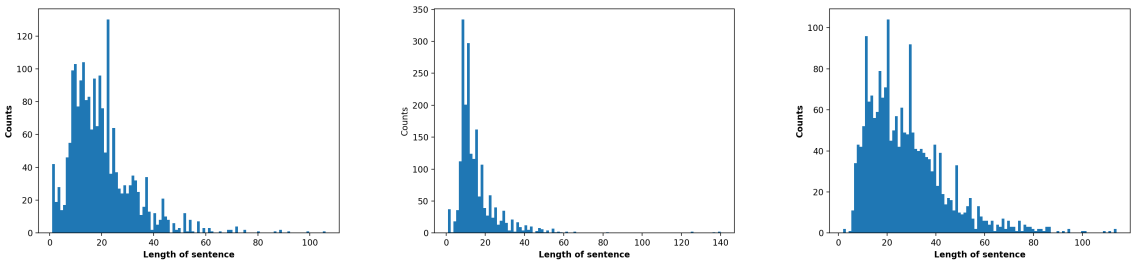
In the previous section, we presented the main experimental results. It clearly indicates that using phrases with tags for fine-tuning NMT models can achieve translation quality improvements. In this section, we analyse the difference observed among datasets and all the fine-tuned models (on tagged phrases and full sentences) in more depth to gain some insights from the results.

5.2.1 Domain analysis

In this thesis, we applied our proposed method to three different technical domains. Every domain has its characteristics, and an NMT model learns them while fine-tuning on the target domain sentences. However, our approach uses fragmented sentences — phrase pairs — instead of whole sentences for fine-tuning. According to the main results, phrase-adapted NMT models were unable to reach the BLEU scores of the fine-tuned NMT models on full sentences, and how close these models reached their upper bound differed significantly across domains. We explore how domain specificity affects domain adaptation in phrases through several analyses.

Sentence length differences in domains

As we mentioned in Section 5.1.3, we discovered that the average sentence length is significantly different across every domain. To take a closer look, we report the distribution of German sentence lengths for all domains in Figure 5.2. It is clear that the JRC domain data consists of longer sentences than other domains, while most sentences of the GNOME domain are less than 20 words. In the main results, the JRC domain has the biggest room for potential BLEU gain of domain adaptation but using phrases of JRC for fine-tuning improved only +1.4 BLEU. On the other hand, fine-tuning on phrases of the GNOME domain reached close to the maximum potential gain of the BLEU score. We may explain this



(a) EMEA test set (b) GNOME test set (c) JRC test set

Figure 5.2: German sentence length distributions for each domain's test set.

with the following assumption: Training NMT models on short phrases may give good translations for short sentences and less good for long sentences. Thus, when evaluating phrase-adapted models, domains with long sentences may have a disadvantage in scoring BLEU compared to other domains, as they may translate less good for long test sentences.

To verify this assumption, we investigate whether the phrase-adapted models in each domain show translation quality differences for different length sentences. We re-evaluate all NMT models with test sets of different sentence lengths: baseline model, the models fine-tuned on tagged phrases in maximum length 4 and the models fine-tuned on the original full sentences. We split the test set for each domain into three different sentence lengths: short ($1 \sim 9$ words), middle ($10 \sim 19$ words) and long (more than 20 words). The numbers of short, medium and long sentences of each domain's test set are described in Table 5.2.

	Sentence length of testset		
	Short ($1 \sim 9$)	**Middle** ($10 \sim 19$)	**Long** ($20 \sim$)
EMEA	320	859	822
GNOME	538	1064	399
JRC	133	610	1258

Table 5.2: Number of sentences in the test set divided by short, medium and long sentence length.

	Test set	**Baseline** (No fine-tuning)	**Fine-Tuning**	
			Tagged phrases (Max length 4)	**Original data**
EMEA	Short	16.8	23.0	33.8
	Middle	35.8	43.2	48.2
	Long	37.8	41.1	45.3
	All	35.5	40.5	45.2
GNOME	Short	21.3	28.3	35.5
	Middel	26.5	34.9	37.6
	Long	36.8	41.8	41.8
	All	29.8	37.0	38.9
JRC	Short	32.6	37.8	48.8
	Middle	31.4	33.4	52.4
	Long	28.5	29.0	55.3
	All	29.0	30.0	54.7

Table 5.3: BLEU scores by different length test sets for the baseline, the tagged-phrase-adapted models (maximum length 4) and the model fine-tuned on original full sentences.

In Table 5.3, the BLEU scores of the baseline, the tagged phrase-adapted NMT models and the sentence-adapted NMT models in different lengths test set are reported. We expected that NMT fine-tuned on phrases should perform well for short sentence translations and perform relatively less good for long sentence translations. Contrary to our expectations, phrase-adapted NMT models generally perform poorly for short sentence translation except for fine-tuning on the phrases of JRC. In fact, in EMEA and GNOME, the phrase adapted models achieved the highest BLEU for middle and long sentence translations, respectively.

However, this only compares the translation performance itself in the sentence length of the phrase-adapted model in each domain. To determine why the effect of fine-tuning on phrases differs among domains, we also consider comparing the difference in translation improvement from the baseline in each domain. We observe that the BLEU scores of the baseline for different sentence lengths already are varying across domains. For example, the baseline works better for long sentences than for short sentences in the GNOME domain, and it is opposite in the JRC domain. Therefore, we compare the improvements in translation quality across domains for different sentence lengths from non-adapted model to adapted models.

We draw line graphs of the change in BLEU scores from the lower bound (the baseline model) through the phrase-adapted model to the upper bound (fine-tuning on the full sentences) in Figure 5.3. The common feature found in all domains is that fine-tuning on phrases can not only improve the translation quality for short sentences but also long sentences. We also observe that the BLEU gains from fine-tuning on phrases for short translation are the biggest in all domains. However, the aspects of BLEU gain at different translation lengths are very different for each domain. Interestingly, in the EMEA and GNOME domains, the possible gain of domain adaptation is the biggest for short sentences, but in the JRC domains, for long sentences. On the EMEA and GNOME domains, all NMT models have lower performance for translating short sentences than other sentence lengths. On the contrary, in the JRC domain, the baseline and the phrase-adapted model have the highest BLEU score for short sentences, whereas the sentence-adapted model has the highest BLEU score for long sentences. But still, this does not provide a sufficient explanation as to why fine-tuning on the JRC phrases results in small improvements over other domains, and why on GNOME can hit close to the ceiling of BLEU gain.

Vocabulary differences in domains

We hypothesise when a domain has more new vocabularies compared to the pre-training data, it can benefit relatively more by fine-tuning on phrases. Even if the amount of useful information is reduced due to fragmentation of sentences, NMT still can exploit these unseen words information through domain adaptation. Therefore, we study whether a domain with the largest amount of new words can improve BLEU scores over other domains.

In our experiments, our model uses a byte pair encoding (BPE) to handle out of vocabulary (OOV) words. After applying BPE, the vocabulary size of a training dataset can be smaller since not all characters or subwords can remain without being merged into larger subword units. In other words, if a dictionary of domain data contains more new words from the pre-training data than other domains, the

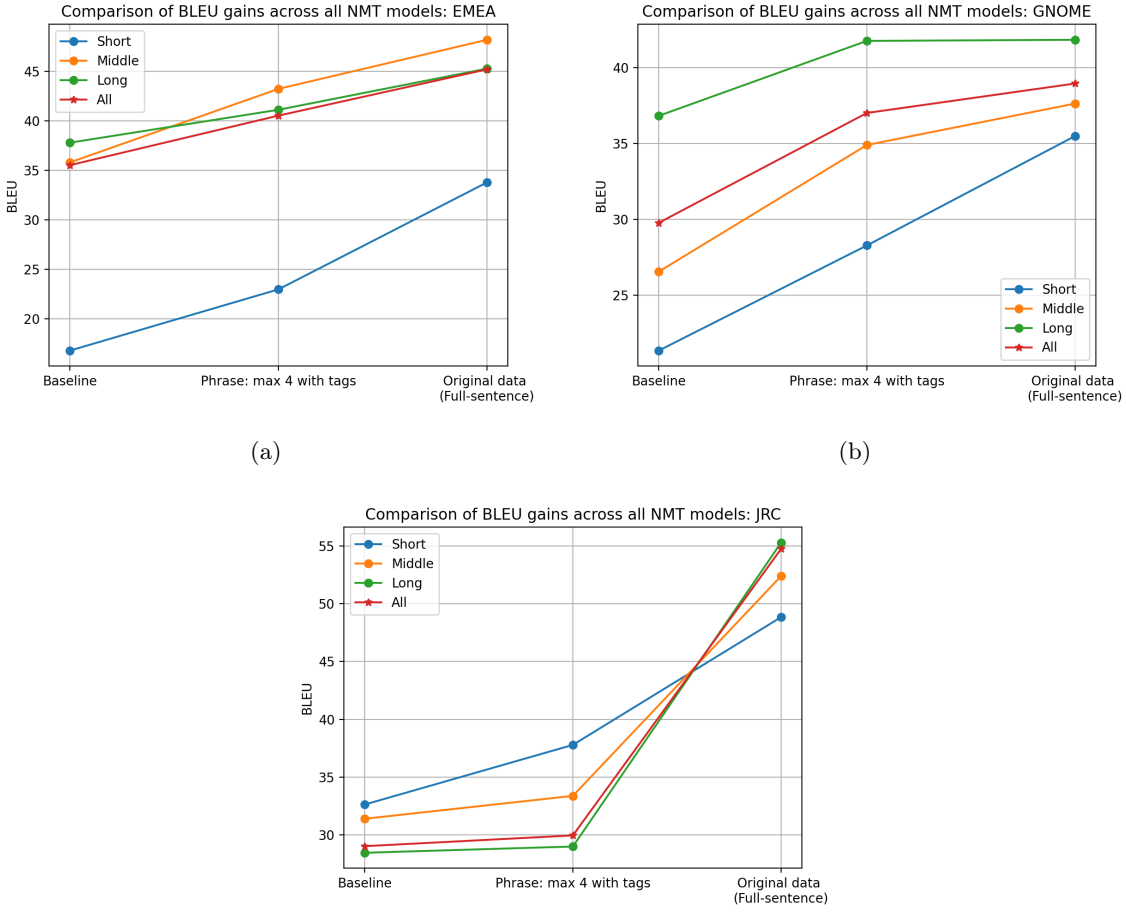


Figure 5.3: Comparison of BLEU gains of all NMT models for various test sets: Baseline, fine-tuning on tagged phrases (max length 4) and fine-tuning on original data(full-sentences). The test set consists of three length groups: short, middle and long (See Table 5.2 for details).

reduction range of the dictionary size after BPE is smaller than the others. For that reason, to determine which domain contains the largest number of new words compared to the pre-training data, we need to examine the change rate of the subword unit numbers after applying BPE to the training data.

In Table 5.4, we report the number of unique tokens before and after applying BPE in each domain's training set and the change rates of the unique tokens for easy comparison among different sized datasets. Although the EMEA domain has the smallest vocabulary among the domains, it contains the largest amount of new words than others. On the other hand, the JRC domain has the largest vocabulary but the smallest amount of new words. This may explain the results (Table 5.1) where fine-tuning on the EMEA dictionary (maximum length of phrase 1) improves the translation quality, but not on other domains' dictionaries.

		Before BPE	**After BPE**	**Rate (After / Before)**
EMEA	De	7.9K	6.2K	0.79
	En	6.5K	5.5K	0.85
GNOME	De	10K	7.2K	0.72
	En	6.8K	5.5K	0.81
JRC	De	15.8K	9.5K	0.60
	En	16.8K	12.1K	0.72

Table 5.4: The number of unique words (before applying BPE) and unique subword units (after applying BPE) and their change rate in each domain. The bigger rate indicates having more new words compared to the pre-training vocabulary.

Similarity between training and test set

When the training and test sets are more similar, NMT systems may achieve higher BLEU scores because training data contains more useful information to predict well at testing. We measure the similarity between training and test sets from each domain. Firstly, we need to define what we mean by the similarity between two datasets. In this thesis, we measure the similarity between two datasets by how many phrases overlap. For example, when extracting phrases of the same length from two data sets, the more overlapping phrases, the closer the datasets are.

The rate of overlapping phrases in the training and test sets for all domains are reported in the Figure 5.4. The similarity rates of EMEA and GNOME domains are lower than that of JRC at from 1 to 4 words. For longer phrases consisting of over 5 words, there is no difference in the proportion of overlapping phrases among domains. Therefore, the JRC domain has the most similarity between training and test datasets. However, since fine-tuning on phrases of the JRC domain has rather low BLEU than other domains, we expected that the similarity rate of the JRC domain is lower than others. Thus, this also cannot provide sufficient explanation why fine-tuning on phrases of the JRC domain gains

less than 2 points of BLEU while other domains achieved larger BLEU points.

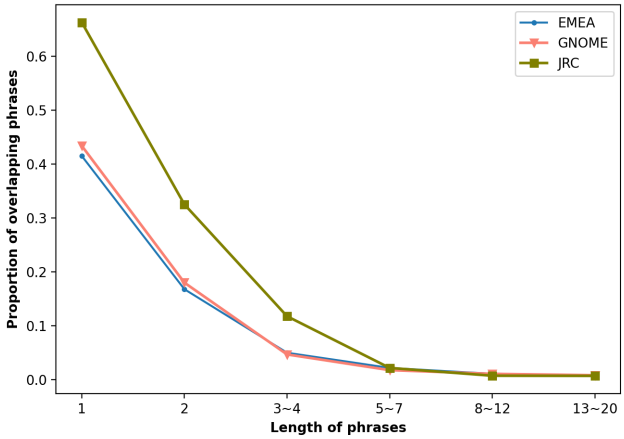


Figure 5.4: Overlapping phrases in every domain on their corresponding train and test sets.

5.2.2 Qualitative analysis: Translation examples

To give an overview of what the generated translations look like, we provide some examples of outputs and reference as a qualitative analysis. We want to utilise this to find interesting patterns that might have been missed when the experimental results are analysed only through statistical analysis.

Source	Injektion alle 8-24 Stunden (6-12 Stunden bei Patienten unter 6 Jahren) wiederholen, bis die Gefahr für den Patienten vorüber ist.
Reference	Repeat injections every 8 to 24 hours (6 to 12 hours for patients under the age of 6) until threat is resolved.
Baseline	Repeat the injection every 8-24 hours (6-12 hours in patients under 6 years of age) until the danger to the patient is over.
No tag, max 4	injection every 8-24 hours (6-12 hours in patients under 6 years of age) until the risk to the patient is over.
Tagged max 4	Repeat the injection every 8-24 hours (6-12 hours in patients under 6 years of age) until the risk to the patient is over.
Original data	Repeat the injection every 8-24 hours (6-12 hours in patients under 6 years of age) until the risk to the patient is over.

Table 5.5: Translation samples: EMEA

Table 5.5 shows reference and samples of the domain EMEA. We often observe in the output from phrase-adapted models that the first word of the sentences are missing or there are no capital letters for the first words. Different vocabularies are used for generating output from the NMT models. In this example, ”threat” in the reference sentence is translated into ”danger” in the baseline ”risk” for all

other NMT models. In addition, the translation lengths of the phrase-adapted models are slightly shorter than non-adapted and sentence-adapted models. Another example of translation in the EMEA domain is reported in Table 5.6 and again we observe that most phrase-adapted models use more technical word (paediatric) than baseline and sentence-adapted model (children). The use of the word "pediatric" in this sentence is incorrect, but from this example, we can infer that in practice fine-tuning on phrases can focus more on domain-specific vocabulary compared to baseline and sentence-adapted models.

Source	Diese Wirkungen sind bei Kindern, älteren Patienten oder im Falle einer Überdosierung wahrscheinlicher
Reference	These effects may be more likely to occur in children, elderly patients, or in cases of overdose
Baseline	These effects are more likely in children, elderly patients or in the event of an overdose
No tag, max 4	These effects are more likely in paediatric, elderly patients, or in case of overdose
Tagged max 4	These effects are more likely in paediatric, elderly patients or in case of overdose
No tag, max 7	These effects are more likely in paediatric, elderly patients or in the case of overdose
Tagged max 7	These effects are more likely in children, the elderly or in the case of overdose
Original data	These effects are more likely in children, elderly patients or in the event of an overdose

Table 5.6: Translation samples: EMEA

Source	Spiele könnten nur teilweise übersetzt sein, so dass es schwieriger ist, diese zu spielen.
Reference	You may be exposed to partially translated games making it more difficult to play.
Baseline	Games may only be partially translated, so it is more difficult to play them.
No tag, max 4	games might only be partially translated, so it is more difficult to play them.
Tagged max 4	Games might only be partially compiled, so it is more difficult to play them.
No tag, max 7	games might only be partially compiled, so it is more difficult to play them.
Tagged max 7	Games might only be partially compiled, so it is more difficult to play them.
Original data	Games might be only partially translated, making them more difficult to play.

Table 5.7: Translation samples: GNOME

In Table 5.7 we report translation samples of the GNOME domain. As mentioned earlier, this sample also has the lowercase error of the first word in the translation of the phrase-adapted models without tags. Some of the translation of the models fine-tuned on phrases contain domain-specific words compared to other models. This pattern can also be seen in the sample, where the word "translated" is translated into a software-related word "compiled" (GNOME is software domain). This example also shows even if the usage of "compiled" is incorrect but it is interesting that it is a very obvious software vocabulary than

”translated”.

5.3 Additional Experiments

Our main results showed that using phrases with the tagging technique for domain adaptation can improve translation quality of NMT models. Based on this, we conducted several additional experiments to further improve our approach: 1) Fine-tuning on a mix of out-domain sentences and target domain phrases to mitigate brevity issues. 2) For maximum phrase length 7 data, we set minimum phrase length to prevent very large numbers of segments. The results are reported in Table 5.8 and 5.9.

5.3.1 Mixed data: in-domain phrases and out-domain sentences

Besides the tagging technique, we also consider fine-tuning on a mix of in-domain phrases and out-domain full sentences to avoid the shorter output bias. For the out-domain sentences, because the baseline model was trained on previous years WMT newstest sets, including newstest2012, newstest2013, newstest2015 and newstest2017, we reserve 5K sentences from them. The out-domain sentences are mixed with the tagged phrases to generate a mixed dataset and we fine-tune the NMT model with this.

The results are reported in Table 5.8. Contrary to what we expected, fine-tuning on mix datasets cannot improve the BLEU scores over the tagging technique except the JRC domain with max length 7. Furthermore, in most cases, BLEU was decreased or only similar compared to that fine-tuning only on phrases.

	Baseline	**Fine-Tuning**						
		Max length 4			**Max length 7**			**Original data**
		No tag	**Tag**	**Mix**	**No tag**	**Tag**	**Mix**	
EMEA	35.5	39.1	40.5	39.6	41.5	37.2	35.4	45.2
GNOME	29.8	36.0	37.0	33.7	35.8	36.8	34.8	38.9
JRC	29.0	29.4	30.0	29.5	29.2	29.7	30.5	54.7

Table 5.8: The BLEU scores of the additional experimental results are compared with the BLEU scores of the main results. The fine-tuning results for mix of out-domain full sentences and in-domain phrases are reported.

5.3.2 Setting minimum length for maximum length 7 phrases

We hypothesised that the long phrases may contain more information than short phrases but we could not observe that longer phrases are more informative than shorter ones for fine-tuning. However, as we mentioned in Section 5.1.2, we suspect this is due to the interference of the bigger number of duplicate and overlapping phrases in the phrases of the bigger number of maximum length. After extracting phrases from the original full sentences, due to the way of extraction process, increasing the maximum length

leads to a much larger number of extracted phrases that are redundant and overlapping. In fact, when extracting phrase pairs with a maximum length of 7, the lengths of 1 to 7 phrases are included, and more phrases of each length are extracted than when the maximum length is 4. To reduce the effect of the overlapping phrases, we experiment with a minimum phrase length when we use a maximum length of 7 phrases for fine-tuning. For fine-tuning, we remove phrases that are shorter than 5 from the extracted phrases with a maximum length of 7. Furthermore, we conduct this experiment with and without tags.

The results are reported in Table 5.9. In the EMEA and JRC domains, fine-tuning on phrase length 5∼7 has the higher BLEU scores than on phrases length 4. On the GNOME domain, the BLUE score of fine-tuning on phrases length 5∼7 is comparable to that of maximum length 7. On the other hand, when using the tagging technique, the effect of reducing short phrases is +1.6 BLEU on EMEA, -3.1 BLEU on GNOME and maintained on JRC. To sum up, setting the minimum length of phrases does not have a positive effect on the BLUE score.

	Baseline	**Fine-Tuning**						
		Max length 4		**Max length 7**				**Original**
		No tag	**Tag**	**No tag**	**Min 5 no tag**	**Tag**	**Min 5 tag**	**data**
EMEA	35.5	39.1	40.5	41.5	40.2	37.2	41.8	45.2
GNOME	29.8	36.0	37.0	35.8	35.8	36.8	32.7	38.9
JRC	29.0	29.4	30.0	29.2	30.2	29.7	30.2	54.7

Table 5.9: The BLEU scores of the additional experimental results are compared with the BLEU scores of the main results. The fine-tuning results for phrases that consist of 5 up to 7 words are reported.

5.4 The JRC domain

We conducted several analyses to answer why the BLEU gains of fine-tuning for JRC phrases is insignificant, but we could not find a good explanation. However, we found that the average length of English sentences of JRC datasets is considerably longer than German, more than 10 words, as Table 5.10 shows. Only after completing most experiments and analyses in this thesis, we discovered a noise in the JRC dataset that some source sentences are copied in the target side. In other words, where the source sentence is **S** and its translation is **T**, the noised target reference consist of **S** and **T** together. An example of this noise is reported in Table 5.11 where the reference has the copy of source sentence. We suspect that the phrase-adapted model could not learn the noise, while fine-tuning on original data could. Furthermore, when we extract the phrases with this noise, the quality of phrases may drop. This would explain why fine-tuning on full sentences was so successful in JRC (the model could learn to copy the input), and conversely fine-tuning on phrases did not work very well.

		Train	Test
EMEA	De	20.9	19.8
	En	21.9	21.8
GNOME	De	18.9	15.3
	En	20.4	15.8
JRC	De	28.9	27.7
	En	40.6	41.9

Table 5.10: The average length of sentences from each dataset

Source	Der Missionsleiter der EUMM wird vom Rat der Europäischen Union ernannt.
Reference	Der Missionsleiter der EUMM wird vom Rat der Europäischen Union ernannt. The Head of Mission of the EUMM shall be appointed by the Council of the European Union.
Baseline	The Head of Mission of the EUMM shall be appointed by the Council of the European Union.
No tag, max 4	The Head of Mission of the EUMM shall be appointed by the Council of the European Union.
Tagged max 4	The Head of Mission of the EUMM shall be appointed by the Council of the European Union.
No tag, max 7	The Head of Mission of the EUMM shall be appointed by the Council of the European Union.
Tagged max 7	The Head of Mission of the EUMM shall be appointed by the Council of the European Union.
Original data	Der Missionsleiter der EUMM wird vom Rat der Europäischen Union ernannt. The Head of Mission of the EUMM shall be appointed by the Council of the European Union.

Table 5.11: Translation sample of JRC. This sample shows the noise that the target reference consists of copy of source sentence and its translation. Fine-tuning on whole original sentences can learn this noise but fine-tuning on phrases cannot.

We investigated how this noise affects the translation quality in fine-tuning on the JRC domain. We remove this noise from the JRC datasets to create new JRC dataset and also extract new phrases from the new training set. Then, we fine-tune the NMT model on the new JRC datasets. The results is reported in Table 5.12. The baseline improves +7.6 BLEU but fine-tuning on full sentences deteriorates -16 BLEU. In other words, the potential gain of domain adaptation significantly decreases. The BLEU score of the phrase-adapted model (with tags, max 4) increases +0.7 BLEU and close to the ceiling. This shows that removing the noise has a significant effect on the BLEU scores of the JRC domain and leads

to results that are more consistent with the other two domains (see Table 5.1).

	Baseline (No fine-tuning)	**Fine-tuning**	
		Tagged max 4 phrase	**Original data** (Full sentences)
Cleaned JRC	37.6	38.3	38.7
Original JRC	29.0	30.0	54.7

Table 5.12: BLEU scores of the baseline, the phrase-adapted model (with tag, maximum length 4) , and the sentence-adapted model fine-tuned on the old JRC and new JRC datasets.

Chapter 6

Discussion and Conclusion

In this chapter, we review and summarise our findings. First, we discuss the experimental results by answering research questions in Section 6.1. Afterwards, Section 6.2 suggests possible directions for future work regarding our approach and limitation of this thesis work. Finally, we conclude our work in Section 6.3.

6.1 Answer to Research Questions

Based on the results that we obtained by fine-tuning NMT systems on phrases with tagging in various target domains, we can answer the research questions we addressed in Section 1.1. We will answer the sub-questions and then the main research question.

1. How much does the translation quality of out-of-domain models improve over the baseline models when fine-tuning on in-domain phrase pairs?

We evaluated the translation quality of the NMT systems by BLEU. For an accurate assessment of the performance of our approach, we established the lower and upper bounds of BLEU gain using the baseline and sentence adapted models, respectively. The baseline's BLEU scores are reached 29 BLEU on JRC, 29.8 BLEU on GNOME and 35.5 BLEU on EMEA in target domain test sets and fine-tuning on in-domain sentences improved them +25.7 BLEU on JRC, +9.1 BLEU on GNOME and +9.7 BLEU on EMEA. By the results of fine-tuning on in-domain phrases, we ascertained that the phrase-adapted models boosted their translation quality over the baseline in all domains. The BLEU gains over the non-adapted baseline vary between +7.0 on EMEA and +1.4 on JRC depending on the maximum length of the phrases or the tagging technique. The details of the difference across the domain will be discussed in the following research questions.

The results of fine-tuning on the cleaned JRC datasets appeared quite different from fine-tuning on the original datasets. The gap between the lower and upper bounds became very small with a score of 1.1 BLEU. The phrase-adapted model (with the tagging technique and the maximum length of 4 phrases) increased +0.7 BLEU from the baseline model.

To sum up, in every domain, although the sentence-adapted models gained higher BLEU scores than the phrase-adapted models, fine-tuning on in-domain phrase pairs indeed improved the translation quality of out-of-domain NMT models over the baseline model.

2. Does the use of shorter phrases (i.e. more fragmented data) lower translation quality?

We hypothesised that shorter phrases would be significantly less useful for fine-tuning, but would better preserve confidentiality. Therefore, we expected that fine-tuning on longer phrases would result in a higher BLEU score. However, by experimenting with different maximum phrase lengths (1, 4 and 7), we observed that longer phrase datasets do not consistently guarantee a better BLEU score. Increasing the maximum length of phrases from 1 to 4 improved the performance of NMT systems in all domains, however from 4 to 7 worsens it in the GNOME and JRC domains.

We suspected that as one of the reasons for this result, the number of overlapping phrases increases as the maximum length increases when extracting phrases. To investigate the effect of reducing overlapping phrases in maximum length 7 datasets, we experimented with a minimum phrase length (5 words). The experimental results showed various aspects depending on the domain. Fine-tuning on the length $5 \sim 7$ phrases in the EMEA domain still results in a higher BLEU than a maximum length 4, but in a decrease of the BLEU than a maximum length 7. In the GNOME domain, setting a minimum length of phrases did not have a positive effect on BLEU. However, in the JRC domain, our approach improved BLEU and eventually had a higher BLEU score than fine-tuning on maximum length 4. To sum up, using shorter phrases does not always reduce translation quality, and the effect is domain dependent. The use of maximum length 4 appears to work best across the board.

3. Can the phrase adapted NMT model's translation quality be improved by applying tagging techniques to present phrase pairs to the NMT model?

When fine-tuning NMT systems on phrases, we are concerned with a bias: the model generates shorter translations compared to fine-tuning on original sentences. To alleviate this bias, we used the tagging technique on phrases. In general, adding tags on phrases raised the BLEU scores. Especially, the positive effect of tagging was notable when fine-tuning on dictionary, +1.6 BLEU on JRC, +1.6 BLEU on EMEA and +3.8 BLEU on GNOME. Only in the EMEA domain with the maximum length of 7, the BLEU score dropped with tagging technique.

As an additional experiment, we reduced the total number of phrases in the maximum length 7 datasets by removing phrases shorter than 5 words. Here, on EMEA benefit the model's performance by tagging but interestingly, on other domains deteriorated or stayed the same.

4. When fine-tuning the NMT model on phrase pairs, are there any significant differences between different test domains?

The experimental results showed that the benefits of domain adaptation on phrases depended on the domain. We obtained the biggest improvement on the EMEA domain, but still almost 4 BLEU away

from the upper bound. On the GNOME domain, we improved the model performance close to the ceiling. However, the results of the JRC domain presented a significant deviation from our expectations. Despite the biggest benefit from domain adaptation, fine-tuning on phrases in the JRC domain achieved only minor improvements. We conducted analyses to explain this and discovered that some of target language sentences consist of copies of the corresponding source and its translation. According to Ott et al., 2018, the "copies" of source sentences can cause a significant effect on the translation quality. To investigate whether this noise affects the model's output, we removed the copies from the JRC datasets and fine-tuned again on them. This led to completely different results from the previous JRC data set. The maximum gain of the domain adaptation shrank significantly and fine-tuning on phrases nearly reached the ceiling.

In addition, we observed that the results of the experiments showed all domains have different patterns. Every domain scored the highest BLEU in a different combination of maximum length and tagging technique. On EMEA increase of the maximum length of the phrase was beneficial but on GNOME and JRC this was not the case. The effect of tagging technique differed across domains, especially in maximum length 7. Applying tagging on maximum length 7 phrases benefits on the GNOME and JRC domains but deteriorates on EMEA.

• **Main research question: In the scenario where the original data is not shareable due to confidentiality issues and *only shuffled phrase pairs* can be released as a compromise, can this benefit downstream NMT quality in any way?**

In this thesis, we considered the scenario where the translation company ***A*** wants to improve its NMT system by using the fragmented confidential data of the clients. We proposed phrase pairs as a fragmented format of the parallel corpus. After extracting phrases from the original data, we shuffled and randomly sampled them to preserve confidentiality. Then, we fine-tuned an NMT system on the phrase pairs to assess whether the model can take any advantage of it. Our experimental results show that NMT can benefit from fine-tuning on shuffled phrases when full sentences are not available. Still, the improvements by fine-tuning on phrases are lower compared to fine-tuning on full sentences which is a promising finding in practice because our approach does not need any change in architecture or other fine-tuning algorithms.

6.2 Limitations and Future Work

Based on our findings and limitations of this thesis work, we can contemplate several further research.

Quantifying the preservation of confidentiality

We proposed an initial solution using confidential data for domain adaptation of NMT models by fragmented parallel corpus, phrase pairs. When releasing sensitive data in a fragmented format, two aspects should be satisfied: confidentiality and usefulness. In this thesis work, we mainly focused on the latter

to examine whether there can be any benefit to using this type of data. On the other hand, we proposed relatively simple methods for preserving confidentiality but did not provide any formal guarantee. In particular, our threat model is that an adversary may reconstruct the original documents, and this can eventually leak core information. Therefore, the next step would be to quantify the degree of possible reconstruction of the original document from the phrases, which so far has only been done in the context of (monolingual) N-grams with fixed N (Gallé & Tealdi, 2015).

Experiments in more realistic circumstances

In this work, we simulated confidential data by using publicly available datasets. Therefore, future work can assess our approach by experimenting with actual confidential data in a more realistic manner. In addition, even if our work is more concerned with the reconstruction of the original text, the actual sensitive data may contain sensitive partial information as well. De-identification techniques are often applied to protect partial information like in clinical documents (Meystre et al., 2010). This may combined with our approach to have more reliable protection of confidentiality. It would be intriguing to examine how our method can still exploit fine-tuning on phrases from the de-identified dataset. For instance when certain key segments are deleted or substituted. Note that de-identification methods have to be applied while maintaining the alignments between source and target sentences.

Experiments in different setups

Our approach was only tested with a Transformer based model, which has a state-of-art performance in NMT and in German-English, which is a high-resource and related language pair. To investigate our method in different setups, we would like to assess the same experiments with different NMT models, such as CNN or LSTM based models, and with other language pairs including low-resource and more difficult ones. We also found that the improvements of translation quality of phrase-adapted models are varied across different domains. To explore the difference in more domains, we can experiment with other domains that we did not choose in this paper.

Monolingual fragmented data: N-grams

Our results show that a fragmented dataset still contains valuable information for fine-tuning an NMT system. We used phrase pairs as fragmented text because they can maintain the alignments of the parallel corpus. However, for some language pairs, the available parallel datasets are limited. As a solution for this, many studies utilising monolingual data for NMT systems have been conducted, and they have shown good results (Domhan & Hieber, 2017; Gulcehre et al., 2015; Lample et al., 2017). Therefore, extending our approach, we may consider N-grams as a monolingual dataset. For instance, using 'Back-Translation'(Sennrich et al., 2016b) we can generate a new pseudo parallel dataset and use this to fine-tune an NMT model. This way, we may easily apply the methods from Gallé and Tealdi, 2015.

6.3 Conclusion

We have studied the problem of domain adaptation of NMT models when domain-specific data cannot be shared due to confidentiality or copyright concerns. Inspired by a common NLP practice of sharing confidential data in the form of N-grams (Michel et al., 2011), we propose to use phrase extraction (Koehn et al., 2003b), shuffling and sub-sampling as a data fragmentation technique for translation data.

Our experiments on three different domains show that this type of data can be used to fine-tune NMT models leading to considerable improvements on top of a strong baseline and further gains when using a simple phrase tagging technique. We also find that the magnitude of these gains varies across domains. Our analysis and additional experiments show that fine-tuning on short segments can improve translation of short sentences as well as long sentences. Furthermore fine-tuning on phrases with a maximum length of 1 (dictionary) is shown to be beneficial when there are many new words in the training data compared to the pre-training data.

In conclusion, this thesis delivered good proof for the feasibility of using only in-domain phrases to fine-tune NMT systems without changing the architecture of the model when the original dataset is not available due to confidentiality.

Bibliography

Al Badawi, A., Hoang, L., Mun, C. F., Laine, K., & Aung, K. M. M. (2020). Privft: Private and fast text classification with homomorphic encryption. *IEEE Access*, *8*, 226544–226556.

Arthur, P., Neubig, G., & Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1557–1567. https://doi.org/10.18653/v1/D16-1162

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., & Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 1–61. https://doi.org/10.18653/v1/W19-5301

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: A case study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 257–267. https://doi.org/10.18653/v1/D16-1025

Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., Mercer, R. L., Printz, H., & Ures, L. (1994). The Candide system for machine translation. *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. https://aclanthology.org/H94-1028

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). Large language models in machine translation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 858–867. https://www.aclweb.org/anthology/D07-1090

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311. https://aclanthology.org/J93-2003

Cancedda, N. (2012). Private access to phrase tables for statistical machine translation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 23–27. https://www.aclweb.org/anthology/P12-2005

Chu, C., Dabre, R., & Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 385–391. https://doi.org/10.18653/v1/P17-2061

Chu, C., & Wang, R. (2018). A survey of domain adaptation for neural machine translation. *Proceedings of the 27th International Conference on Computational Linguistics*, 1304–1319. https://www.aclweb.org/anthology/C18-1111

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2017). Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta: Journal des traducteurs/Meta: Translators' Journal*, *62*(2), 245–270.

Domhan, T., & Hieber, F. (2017). Using target-side monolingual data for neural machine translation through multi-task learning. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1500–1505.

Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648. https://www.aclweb.org/anthology/N13-1073

Fadaee, M., Bisazza, A., & Monz, C. (2017). Data augmentation for low-resource neural machine translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 567–573. https://doi.org/10.18653/v1/P17-2090

Feng, Q., He, D., Liu, Z., Wang, H., & Choo, K.-K. R. (2020). Securenlp: A system for multi-party privacy-preserving natural language processing. *IEEE Transactions on Information Forensics and Security*, *15*, 3709–3721.

Freitag, M., & Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Gallé, M., & Tealdi, M. (2015). Reconstructing textual documents from n-grams. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 329–338.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, *4*(1), 1–58.

Goldreich, O. (2009). *Foundations of cryptography: Volume 2, basic applications*. Cambridge university press.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., & Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., & Ramage, D. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., et al. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Huang, Y., Song, Z., Chen, D., Li, K., & Arora, S. (2020). TextHide: Tackling data privacy in language understanding tasks. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1368–1382. https://doi.org/10.18653/v1/2020.findings-emnlp.123

Hutchins, J. (2007). Machine translation: A concise history. *Computer aided translation: Theory and practice*, *13*(29-70), 11.

Iconic-Translation-Machines-Ltd. (2019). *Machine translation quality*. https://iconictranslation.com/what-we-do/neural-machine-translation/

Irvine, A., Morgan, J., Carpuat, M., Daumé III, H., & Munteanu, D. (2013). Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, *1*, 429–440. https://doi.org/10.1162/tacl_a_00239

Jia, Y., Carl, M., & Wang, X. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? a product and process study. *The Journal of Specialised Translation*, *31*, 60–86.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351. https://doi.org/10.1162/tacl_a_00065

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709. https://www.aclweb.org/anthology/D13-1176

Kim, S., Bisazza, A., & Turkmen, F. (2021). Using confidential data for domain adaptation of neural machine translation. *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, 46–52. https://doi.org/10.18653/v1/2021.privatenlp-1.6

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing*, *1*, 181–184 vol.1. https://doi.org/10.1109/ICASSP.1995.479394

Kobus, C., Crego, J., & Senellart, J. (2017). Domain control for neural machine translation. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 372–378. https://doi.org/10.26615/978-954-452-049-6_049

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

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., & Herbst, E. (2007). Moses: Open

source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–180. https://www.aclweb.org/anthology/P07-2045

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28–39. https://doi.org/10.18653/v1/W17-3204

Koehn, P., Och, F. J., & Marcu, D. (2003a). Statistical phrase-based translation. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127–133. https://www.aclweb.org/anthology/N03-1017

Koehn, P., Och, F. J., & Marcu, D. (2003b). Statistical phrase-based translation. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127–133. https://www.aclweb.org/anthology/N03-1017

Krogh, A., & Hertz, J. A. (1992). A simple weight decay can improve generalization. *Advances in neural information processing systems*, 950–957.

Kuczmarski, J., & Johnson, M. (2018). Gender-aware natural language translation.

Lagarda, A.-L., Alabau, V., Casacuberta, F., Silva, R., & Dıaz-de-Liaño, E. (2009). Statistical post-editing of a rule-based machine translation system. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 217–220. https://aclanthology.org/N09-2055

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Läubli, S., Amrhein, C., Düggelin, P., Gonzalez, B., Zwahlen, A., & Volk, M. (2019). Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, 267–272. https://www.aclweb.org/anthology/W19-6626

Läubli, S., Sennrich, R., & Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4791–4796. https://doi.org/10.18653/v1/D18-1512

Luong, M.-T., Manning, C. D. et al. (2015). Stanford neural machine translation systems for spoken language domains. *Proceedings of the international workshop on spoken language translation*, 76–79.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC medical research methodology*, *10*(1), 1–16.

Miceli Barone, A. V., Haddow, B., Germann, U., & Sennrich, R. (2017). Regularization techniques for fine-tuning in neural machine translation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1489–1494. https://doi.org/10.18653/v1/D17-1156

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176–182.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR's WMT19 news translation task submission. *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 314–319. https://doi.org/10.18653/v1/W19-5333

Och, F. J., Tillmann, C., & Ney, H. (1999). Improved alignment models for statistical machine translation. *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. https://www.aclweb.org/anthology/W99-0604

Östling, R., & Tiedemann, J. (2017). Neural machine translation for low-resource languages. *arXiv preprint arXiv:1708.05729*.

Ott, M., Auli, M., Grangier, D., & Ranzato, M. (2018). Analyzing uncertainty in neural machine translation. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (pp. 3956–3965). PMLR. http://proceedings.mlr.press/v80/ott18a.html

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 48–53. https://doi.org/10.18653/v1/N19-4009

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Post, M. (2018). A call for clarity in reporting BLEU scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. https://www.aclweb.org/anthology/W18-6319

Riazi, M. S., Rouani, B. D., & Koushanfar, F. (2019). Deep learning on private data. *IEEE Security & Privacy*, *17*(6), 54–63.

Sato, S., Sakuma, J., Yoshinaga, N., Toyoda, M., & Kitsuregawa, M. (2020). Vocabulary adaptation for domain adaptation in neural machine translation. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4269–4279. https://doi.org/10.18653/v1/2020.findings-emnlp.381

Sennrich, R., Haddow, B., & Birch, A. (2016a). Controlling politeness in neural machine translation via side constraints. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 35–40.

Sennrich, R., Haddow, B., & Birch, A. (2016b). Improving neural machine translation models with monolingual data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96. https://doi.org/10.18653/v1/P16-1009

Sennrich, R., Haddow, B., & Birch, A. (2016c). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stay. (n.d.) west's encyclopedia of american law, edition 2. (2008). https://legal-dictionary.thefreedictionary.com/Stay

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 3104–3112.

Thompson, B., Knowles, R., Zhang, X., Khayrallah, H., Duh, K., & Koehn, P. (2019). HABLex: Human annotated bilingual lexicons for experiments in machine translation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1382–1387. https://doi.org/10.18653/v1/D19-1142

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Toral, A., Wieling, M., & Way, A. (2018). Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, *5*, 9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998–6008.

Zhang, W., Feng, Y., Meng, F., You, D., & Liu, Q. (2019). Bridging the gap between training and inference for neural machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4334–4343. https://doi.org/10.18653/v1/P19-1426

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.