Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Evaluation of prefix constraints and their impact on related and unrelated language pairs

Bachelorthesis
at Research Group Knowledge Technology, WTM
M. Sc. Tayfun Alpay

Department of Informatics
MIN-Faculty
Universität Hamburg

submitted by
Vincent Dahmen
Course of study: Informatik
Matrikelnr.: 6689845
on
August 13, 2019

Examiners: M. Sc. Tayfun Alpay
Dr. Stefan Heinrich

Abstract

Translations need to be delivered in a cost-effective and time-efficient way. However, conventional translations require time and are rather expensive. Therefore machine translation was recently developed and became the preferred and widely used alternative. Neural Machine Translation is a subset of machine translation, which already gives remarkable results but can be further optimized. One of the currently discussed optimization strategies are domain control mechanisms such as the use of prefix constraints.

However, the impact of using prefix constraints on translation quality is not well explored.

This thesis aims at investigating the effects of prefix constraints on the performance of NMT across two language pairs of different relatedness.

For three domains (Law, Finance, Medical) one open corpus is available in German-English (related language pair) and Czech-English (unrelated language pair) respectively.

Data preprocessing included joining the 3 domain-specific corpora into a multidomain corpus and reducing corpus length via splitting into small(er) logical units. Then a multidomain corpus was built, resulting in four corpora (3 from different domains each and one from all three domains). The number of tokens was reduced, and prefix constraints were applied via byte pair encoding resulting in 4 multidomain corpora.

These datasets were used to train an Encoder-Decoder Recurrent Neuronal Network.

Three different scoring systems (BLEU for measuring translation precision, METEOR for assessing comprehensibility of translation, ROUGE for evaluating domain specialization) were used to evaluate the impact of prefix constraints on the performance of the NMT applied on the related and distant language pair.

Data reduction did not affect the characteristics of the reduced datasets compared to the original data sets as indicated by the word length (Characters/word) and sentence length (words/sentence).

Evaluation of translation performance revealed that using prefix constraints enhanced generalizability at the expense of specialized knowledge. Translation precision improved more than comprehensibility.

Language comparison revealed that the metrics of translation quality (indicated by the selected 3 diff scoring systems) were different in the related (GE-EN) compared to the unrelated language pair (CZ-EN). This might be due to language-related differences in the domains used for training, as demonstrated here upon using the three selected corpora.

Further research is needed to predict the potential benefit of applying the domain control mechanism prefix constraints on a given language pair. Special attention should be given to the impact of language-related differences in the domains.

Zusammenfassung

Übersetzungen müssen schnell und kostengünstig sein. Herkömliche Übersetzungen sind zeitintensiv und insgesamt eher teuer. Machischnelle Übersetzungssysteme wurden deshalb in letzer Zeit immer beliebeter und sind inzwischen allgemein akzetiert. Neuronale Maschinenübersetzungen bezeichnen eine Untergruppe der machinellen Übersetzung, die bereits heute bemerkenswert gute Ergebenisse erzielen und nach aktueller Erkenntis sogar noch Verbesserungspotential hat. Eine der am häufigsten diskutierten Optimisierungsstrategien beschreibt die Verwendung von "Prefix Constraints" als Domänen Adaptionmechanismus. Die Auswirkungen der Verwendung dieses Mechanismus auf die Übersetzungsqualität ist allerdigns nicht hinreichend untersucht.

Für drei Bereiche (Recht, Finanzen, Medizin) wurde jeweils ein offener Korpus mit Deutsch-Englisch als verwandtes und Tschechisch-Englisch als nicht verwandtes Sprachpaar verwendet. Die Datenvorverarbeitung beinhaltete die Zusammenführung der 3 domänenspezifischen Korpora zu einem Multi-Domain-Korpus, die Datenreduktion durch Aufteilung des Korpus in kleine(re) logische Einheiten und das Erstellen eines Multidomain-Korpus durch Rekombination dieser Einzelkorpora, die Kodierung in häufige Subsequenzen zur Reduzierung der Anzahl der Token und die Anwendung von des Domänadaptionsmechanismus "Prefix-Constraints", was zu 4 Multi-Domain-Korpora führte. Drei verschiedene Bewertungsmetriken (BLEU zur Messung der Übersetzungspräzision, METEOR zur Beurteilung der Verständlichkeit und ROUGE zur Bewertung der Domänenspezialisierung) wurden verwendet, um die Auswirkungen des Adaptionsmechansimus auf die Übersetzungsleistung des NMT unter der Berücksichting des Verwandheit der Sprachpaar untereinander zu bewerten. Die Auswertung ergab, dass die Verwendung des Mechansimus "Prefix Constraints" die Fähigkeit zur Erkennung und Reproduktion von Gemeinsamkeiten auf Kosten der Reprouktionsfähigkeit von Domänen-Jargons verbessert. Der Sprachvergleich ergab, dass sich die Übersetzungsqualität (indiziert durch die ausgewählten 3 Bewerungsmetriken) im verwandten Sprachpaar (DE-EN) im Vergleich zum nicht verwandten Sprachpaar (CZ-EN) unterschieden. Dies könnte auf sprachbedingte Unterschiede in den Domänen zurückzuführen sein. Weitere Forschung ist erforderlich, um den potenziellen Nutzen der Verwendung der "Prefx Constraints" ein bestimmtes Sprachpaar vorherzusagen. Besondere Aufmerksamkeit sollte der Ausprägung sprachbedingter Unterschiede zwischen Domänen geschenkt werden.

Contents

Chapter 1

Introduction

In times of progressing globalization the need for fast and reliable translation increases. Since human translations require skill and time not only to learn but as well as to make, they are often time-consuming and expensive. Thus lately, machine translations are on the rise, according to Castilho et al. (2017). The most used technique in machine translations are Neural Machine Translation (NMT) with Recurrent Neural Networks (RNN) since they produce the most promising results once they have been trained accordingly Wu et al. (2016)

1.1 Neural Networks (NN)

The term "Neural Networks" describes a class of machine learning algorithms, that are general function approximations. These algorithms (or networks) learn a specific mapping between an input and an output by calculating the prediction error and apply a corresponding change to the matrix (known as backpropagation) (Hecht-Nielsen, 1992). The algorithm that is used to apply a change to reduce the error is called the optimizer. Battou (2010) proposed to use the stochastic gradient descent to calculate and apply that change. Zeiler (2012) presented ADADELTA as an alternative to the stochastic gradient descent. The stochastic optimization method adam was introduced by Kingma & BA (2014), because of its ability to work on networks with large data sets due to its little memory requirements. All three methods are represented in current research. The amount of reduction that is applied to reduce the error is called the learning rate, which is set between 0 and 1 and describes how much the network should be changed in each step. A learning rate of 1 means, to take in the full error, whereas a rate of 0 means no fix the error at all (Bottou, 2010)

1.1.1 Recurrent Neural Networks (RNN)

So-called Recurrent Neural Networks (RNN) are a subclass of NN, which use an internal state to allow them to access their previous output in the next iteration (Zaremba et al., 2014). Thus RNNs can parse a sequence of inputs and produce out-

put sequences, e.g. speech recognition (Mikolov et al., 2010) or image transcription (Vinyals et al., 2015).

1.1.2 Attention

Plain RNNs have only limited memory to store their past output, limiting their usability when working with large or complex texts (Bahdanau, Cho, & Bengio, 2014). In comparison to "simple" NN, the attention-based networks address this issue by providing the NN with access to the input on top of the last output and implement a filter matrix to "blend in" certain features (Vaswani et al., 2017).

1.2 Neural Machine Translation

Neural machine translation (NMT) describes a branch of machine translations where RNN are used to translate text (Bahdanau et al., 2014). These are trained on input texts and their corresponding translations and allows the network to translate similar texts. NMTs are commonly used; for example, in Google Translate and can produce excellent results on short passages. With enough training data, they can even translate articles or more elaborate texts (Wu et al., 2016).

In NMT text collections from the same source are referred to as one corpus. A corpus may share the same context (domain) or is related to a particular event. Especially useful for NMT are corpora, which are available in multiple languages (aligned corpora) because they can be used to evaluate the performance of a model (Papineni et al., 2002). For this evaluation, the corpus is translated from one language into another with NMT, and the resulting translation is then compared to the available reference translation, to check the translation quality.

1.2.1 Byte-Pair-Encoding (BPE)

One of the main challenges in current NMTs is the choice of the available vocabulary. If too many words are included, the networks tend to forget rarely used words or become too general. On the other hand, if the vocabulary is too small, the networks cannot learn complex grammar forms. An approach to solve this is to split up long words into multiple symbols (Byte Pair Encoding)(Koehn & Knowles, 2017). Instead of using a list of actual words, the dictionary consists of numerous used character pairs or subwords. This allows the network to learn common base words as well as pre- and suffixes. Sennrich et al. (2015) have shown that this results in an overall improvement and allows the model to learn more complex grammar forms.

1.2.2 Content Domains

Another challenge in translations is context because each industry has its lingo. Contexts might be specialized through vocabularies like technical terms and stan-

dard shared definitions or even textual structures and tone. This often distinct set of rules, such as tone, vocabulary, or any other substantial similarity are defined as a domain (Koehn & Knowles, 2017).

In Neural Machine Translation, multiple mechanisms are known and used to translate across multiple domains; for example, prefix or side constraints (Chu et al., 2017).

1.2.3 Prefix/Side Constraints

Prefix or Side Constraints describe a class of domain control mechanisms, where additional tokens are provided for the neural network, to identify the domain of the text. These tokens can be direct modifications in the text, like the decoration of certain essential words or more general tags that are added in front of a sentence, word, or even corpus. The concept of side constraint was introduced by Sennrich & Birch (2016) and adapted for domain control by Kobus et al. (2016).

1.2.4 Languages

Another challenge in NMT is language adaption and control. Johnson et al. (2017) have shown that certain languages can be translated without ever training on the actual language data (Zero-Shot Translation) and Luo et al. (2019) impressively demonstrated that an NMT was able to translate an extinct language. Hajic (2000) shows that a close affinity between languages simplifies certain translation actions. Mikolov et al. (2013) found that morphological features can improve the translation performance in certain situations. However, it is unknown to date which relationship languages and domain features have.

Since languages are grouped by language families and language branches, it can be suspected that similar languages may benefit from similar models.
A language family is a group of languages that descend from a common ancestor such as the Indo-European languages or the "Sino-Tibetan languages". A typical example for a language branch is the Germanic branch containing languages such as German, English, Dutch and Flemish in contrast to the Romance language branch with languages such as Latin, Italian, French and Spanish or the Slavonic language branch such as Czech and Russian (Wichmann et al., 2010). Languages within the same branch are closer related than outside of the branch and share some similarities in terms of vocabulary or certain structural similarities (Georgi et al., 2010).

It was shown by Kobus et al. (2016) that prefix constraints are beneficial in English-French. Takeno et al. (2017) adapted the method for English-Japanese translation. To date, no comparisons of the performance impact between the prefix constraints have been made. Thus the question arises if the use of prefix constrains mechanisms can be generalized and used with any language pair.

1.2.5 Aims

The problem of domain control is not yet solved and current research. In this thesis I want to evaluate the domain control mechanism prefix contraint and how it impacts the translation performance in different language pairs. Additionally I want to analyze if the relatedness of a language pair needs to be considered during multi domain corpus translation.

Chapter 2

Method

In the following section, I introduce and present the methods used to evaluate the performance impact of the domain control mechanism "prefix constraint" in NMT across different language pairs. I introduce the used data sets and describe my training setup for the neural networks.

The used domain control mechanism was adapted from Kobus et al. (2016), but the testing setup was changed to investigate the impact of prefix constraints in related and unrelated languages.

2.1 Data Selection and Preparation

To compare the performance impact of the domain control mechanism, I built a multi-domain corpus with two language pairs of different relatedness.

2.1.1 Language Pair Selection

The selected language pairs were Czech-English as a distant language pair and German-English as a related language pair. For both pairs, I used English as the source language.

2.1.2 Domain and Corpus Selection

I used the three domains FINANCE, LAW, and MEDICAL for my experiments. These domains share certain semantic qualities like a formal and precise language and a domain-specific jargon.

For the FINANCE domain, I used version 1 of the website and documentation of the European Central Bank (Tiedemann, 2012). This corpus will be referred to as ECB. For the LAW domain, I used version 7 of a parallel corpus extracted from the European Parliament website (Tiedemann, 2012). This corpus will be referred to as Europarl. For the last domain (MEDICAL) I used version 3 of a corpus made of PDF documents from the European Medicines Agency (Tiedemann, 2012), which will be referred to as EMEA.

All of the used corpora are accessible through the OPUS project (Tiedemann, 2012).

2.2 Data Preparation

A new corpus was created by reducing the domain corpora and combining them into one multi-domain corpus. The newly generated corpus was then prepared for neural network training. During the preprocessing, the words were split into tokens of high occurrence using BPE. Afterward, the domain control mechanism was applied to the data set.

2.2.1 Data Slicing

Since no document separation was available, I analyzed the corpora manually and split all corpora in smaller documents of logical units. Those units were then combined into a training set with roughly 70,000 example sentences and a validation set of 15,000 sentences. These new corpora were aligned in German-English and Czech-English. I calculated the distribution of word and sentence length for the original corpora, the validation, and the training data set.

2.2.2 BPE

Byte Pair Encoding was used to reduce the number of tokens to 32,000 as suggested by Kobus et al. (2016), by running the implementation of Sennrich et al. (2015). The unique sequence (@@) was used to mark word endings and pairs.

2.2.3 Prefix Constraints

Then the domain control mechanism 'prefix constraint' was applied as described by Kobus et al. (2016) For each domain, all English sentences were prefixed with a unique domain token. Since the BPE algorithm produced token with a specific format, the domain tokens needed only a different pre- and suffix to became unique.

The resulting corpora were: Not modified data in Czech-English (CZ-EN) and German-English (DE-EN), and Modified data in DE-EN and CS-EN. In the following, I will refer to them as Clean-de-en, Clean-cs-en, Tagged-de-en, and Tagged-cs-en.

2.3 Training and Optimization

To evaluate the performance impact of prefix constraints, I trained a neural network (Bahdanau et al., 2014) and performed one hyperparameter optimization round. The used parameter was adapted from multiple sources (Kobus et al., 2016; Rico Sennrich & Birch, 2016; Luong, Pham, & Manning, 2015) and will be explained subsequently.

2.3.1 Model

I used an Encode-Decoder Recurrent Neuronal Network with "Long Short-Term Memory" (Zaremba et al., 2014) gates, with a dropout probability of 0.3, two layers, and 1000 hidden states. On the source and target side, I used 500-word embeddings (Bahdanau et al., 2014). For the attention behavior (Vaswani et al., 2017) I used the "general" attention type (Luong et al., 2015) and the "softmax" function for the attention and the generator (Liu, Wen, Yu, & Yang, 2016).

2.3.2 Hyper Parameter

For the optimizer, I run multiple configurations. In all runs, a mini-batch size of 32 sentences for the training and 16 sentences for the validation was used. For the optimizer, I used stochastic gradient descents (Bottou, 2010), "Adam (Kingma & Ba, 2014)", and "ADADELTA (Zeiler, 2012)". I ran all of them with learning rates 1, 0.1 and 0.001 and started to decay the learning rate by 0.3 per epoch after 5 or 10 epochs and once per optimizer and learning rate without any decay. The gradient was set to be renormalized if the norm over the gradient vector exceeded 5.

2.3.3 Training

I used the framework of Klein et al. (2017) for the implementation of the model and the training procedure. I built an MQTT scheduler to coordinate the runs on a mixture of NVIDIA GTX 980, 1080 and 1080Ti GPUs (Light et al., 2017). Each model was trained for 18 epochs, which took between 2 and 3,5 hours depending on the GPU. English was used in all models as the source locale and Czech and German as the target locale. All models were trained multiple times to ensure the proper distribution of start vectors. During the training, I logged the training accuracy and calculated the validation accuracy after every epoch. However, the validation score was not used for early stopping.

2.4 Comparison

To compare the score impact and evaluate a possible connection with the relatedness of the languages, I picked the best models from the training and measured its performance with three metrics (BLEU(Papineni et al., 2002), METEOR(Banerjee & Lavie, 2005), ROUGE-L(Lin & Och, 2004)).

2.4.1 Metrics

I used three different metrics to measure the translation quality of the trained neural networks (BLEU, METEOR, and Rouge-L). An overall score was calculated by adding the rank of all scores. All data preprocessing steps were removed before calculating the scores. The scores were calculated with the implementations of

Sharma et al. (2017)(METEOR, ROUGE-L) and the OpenNMT-Project (Klein et al., 2017) (BLEU). Since the corpus was aligned, all scores were calculated in comparison to a conventional translation.

2.4.2 Model Selection

For each language pair, the best model was chosen according to overall score with and without prefix constraint resulting in four models: Clean-de-en, Clean-cs-en, Tagged-de-en, and Tagged-cs-en The validation data sets were translated using these models and metrics (BLEU(Papineni et al., 2002), METEOR(Banerjee & Lavie, 2005), ROUGE-L(Lin & Och, 2004)) were calculated for each resulting translation.

2.4.3 Prefix Constraint

Per domain, a data set of 1,200 randomly picked example sentences was prepared in the same way as the training data, except that the BPE merges were not relearned but reused to generate the same tokens. The data sets were then distributed into four sets, each containing 900 sentences either originating from the same domain (three sets) or combining all three domains into one set.

All sets were translated by the best models from the previous selection and scored with all three metrics.

2.4.4 Language Pairs

To compare the performance impact of the prefix constraints across the language pairs, I calculated the relative score change per metric and domain test set for both language pairs over the runs with and without prefix constraints.

Chapter 3

Results

In the following section, I present the results of the previously described procedures. First, I will show some statistics of my generated corpus and compare its characteristics to the source material. The next part comprises exemplary visualizations of the model training and detailed metrics of these examples. I will then rank the models and compare the impact of prefix constraints between the two language pairs, German-English and Czech-English.

3.1 Data Selection and Preparation

To ensure that the reduction of the data set did not change the defining characteristics of the respective corpora, I compared the number of words and sentence length on the reduced and original data set. The distributions are shown in figure 3.1 and 3.2.

The figures 3.1 and 3.2 both show a 3x4 matrix with three boxplots each. The first two plots per row show the distribution over the language pair DE-EN, and the 3rd and 4th over CS-EN. Each row shows the distribution of one domain.

All plots labeled with Q show the distribution of the original corpus. Plots labeled with T show the distribution of the training data, and V marks the validation data. Outliers are not shown in the plots. Instead, the whiskers are extended.

None of the plots show major differences between the reduces corpus to the original corpus. Therefore, the reduced corpora were accepted for further steps.

3.1.1 Number of words per Sentence

In most domains, sentences have between 0 and 50 words(interquartile range). In EMEA the sentence length was mostly between 0 and 25 words, and the median between 5 and 10 words. In all other domains the median was between 20 and 25 words per sentence.

In ECB, the maximum number of words was 70 for German and English. EMEA was found to contain sentences up to 40 words and Europal up to 60, but the Czech sentences tend to be about 10 words shorter than English and German. Only in the

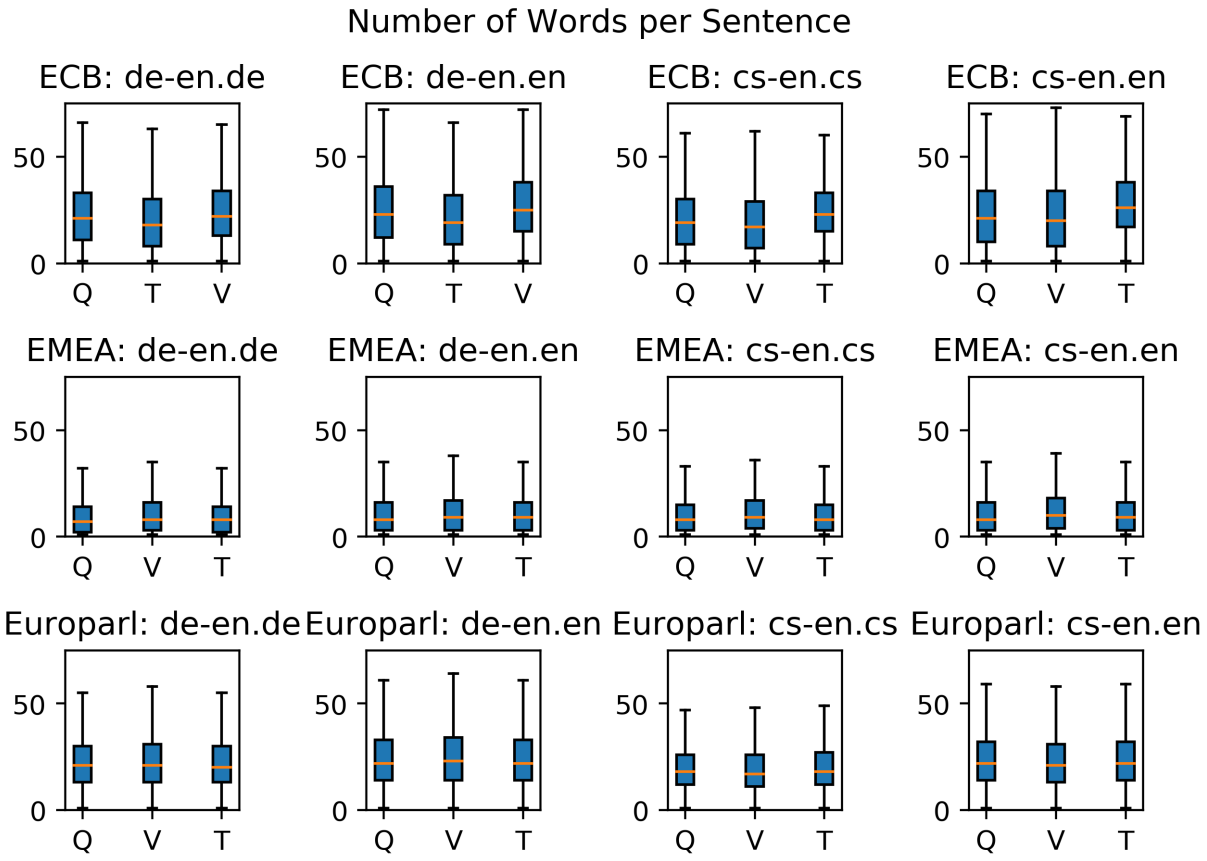

Figure 3.1: Distribution of number of words per sentence

English ECB data set, the interquartile range was slightly higher in the training set than in the original data. All other data sets showed no distinct differences.

3.1.2 Word Length

Words mostly had a length between 0 and 10 characters. The median ranges from 4 to 6 in all domains.

Median word lengths in all domains were between 1 and 15 characters in English, but 5 to 15 characters in Czech in all domains. The median word length in the German text was 7. The median word length for English words was found to be 7, Czech lies in between.

However, no distinct difference between the original and reduced data sets was found.

3.2 Training and Optimization

During the training of the models, the training accuracy and the validation accuracy were recorded. For a selection of models, the BLEU score was calculated

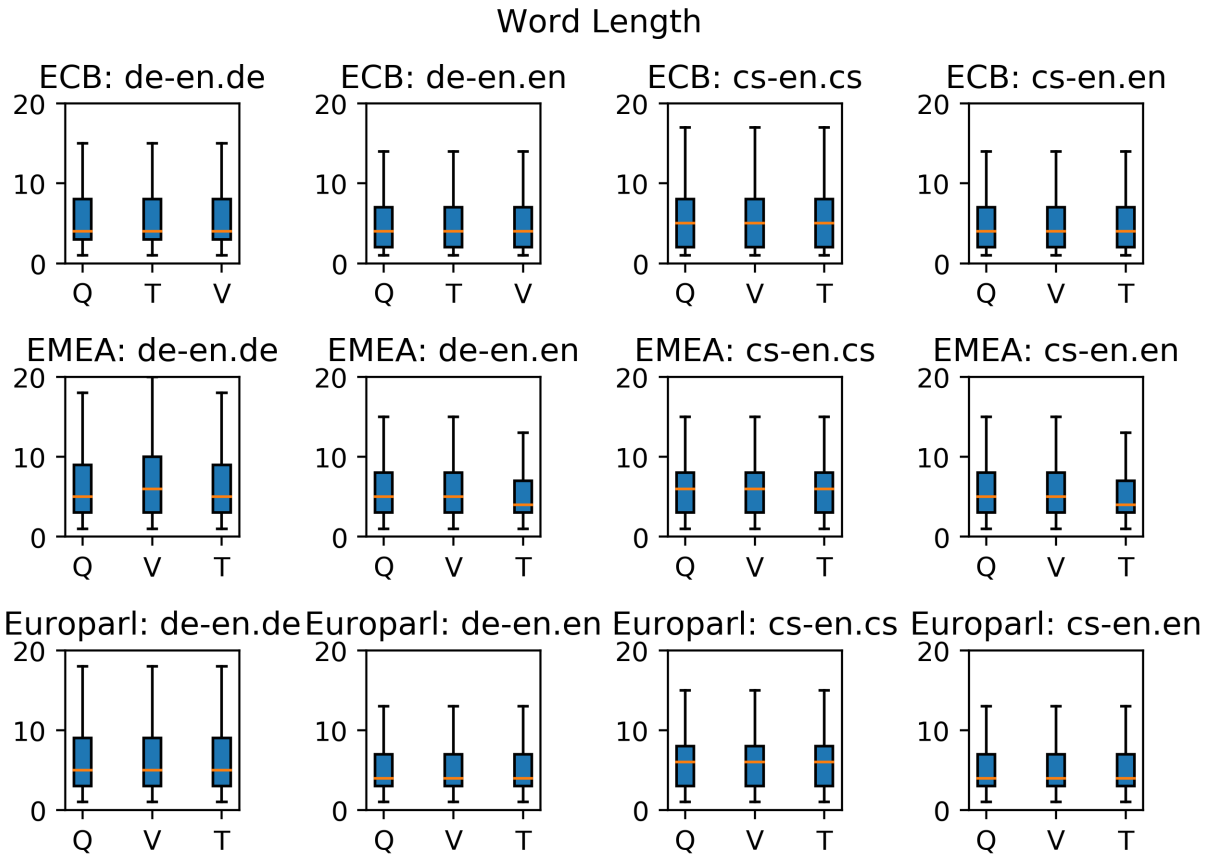

Figure 3.2: Distribuion of word lengths

to visualize the learning process. All models were used to translate the validation data sets after 6, 12, and 18 epochs.

3.2.1 Hyper Parameter

Table 3.5 shows the ranking of all training configurations for the corpus without prefix constraints in the DE-EN language pair. Models are characterized by the optimization method, learning rate, start of the decay, and METEOR score. Tables 3.1 3.2 3.3 3.4 show the best 3 models per corpus ranked by BLEU, METEOR, and ROUGE-L.

Tables with Scores and rankings

3.2.2 Training

Figures 3.3 and 3.4 shows the validation accuracy and the training accuracy for each training step. Both metrics increase in a logarithmic manner. The training accuracy meets the 20 % mark after 10,000 training steps and growth to 30% in the next 25,000 training steps. The validation accuracy hits the 15% mark after

Table 3.1: Top Configurations for English-German without Prefix Constraints

optim	learning rate	start decay steps after (Epoch)	BLEU
adam	0.001	5	23.9
adam	0.001	None	23.6
adadelta	1	5	23.6
			ROUGE-L
adam	0.001	5	47.5
adadelta	1	5	47.3
adam	0.001	None	46.6
			METEOR
adam	0.001	5	20.9
adadelta	1	5	20.9
adam	0.001	None	20.4
			Validation-Accuracy
adam	0.001	10	38.7
sgd	1	10	38.3
adam	0.001	None	38.2

Table 3.2: Top Configurations for English-German with Prefix Constraints

optim	learning rate	start decay steps after (Epoch)	BLEU
adadelta	1	5	22.670
adam	0.001	None	22.660
sgd	1	10	22.410
			ROUGE-L
adadelta	1	5	0.464
adam	0.001	None	0.456
sgd	1	None	0.455
			METEOR
adadelta	1	5	0.206
adam	0.001	None	0.200
sgd	1	None	0.194
			Validation-Accuracy
adam	0.001	5	38.517
adam	0.001	10	38.362
adam	0.001	None	38.114

Table 3.3: Top Configurations for English-Czech without Prefix Constraints

optim	learning rate	start decay steps after (Epoch)	BLEU
adam	0.01	5	23.5
sgd	1	5	14.0
adam	0.001	None	13.6
			ROUGE-L
adam	0.01	5	45.4
adam	0.001	5	38.5
sgd	1	5	38.4
			METEOR
adam	0.01	5	20.2
sgd	1	5	14.8
sgd	1	10	14.7
			Validation-Accuracy
sgd	1	None	50.6
adam	0.001	10	40.4
adam	0.001	None	37.8

Table 3.4: Top Configurations for English-Czech with Prefix Constraints

optim	learning rate	start decay steps after (Epoch)	BLEU
adam	0.001	10	23.840
adam	0.001	5	13.440
adam	0.001	None	12.910
			ROUGE-L
adam	0.001	10	0.450
adam	0.001	5	0.372
adam	0.001	None	0.367
			METEOR
adam	0.001	10	0.203
adam	0.001	5	0.146
adam	0.001	None	0.141
			Validation-Accuracy
sgd	1	None	33.654
adam	0.001	None	33.383
adam	0.001	10	33.203

Table 3.5: Top Configurations for English-Czech without Prefix Constraints

optim	learning rate	start decay steps after (Eoch)	METEOR	step
adam	0.001	5	20.9	23700
adadelta	1	5	20.9	35550
adam	0.001	None	20.4	23700
adadelta	1	10	20.0	35550
adadelta	1	None	19.6	35550
sgd	1	None	19.2	35550
adam	0.001	10	18.0	35550
sgd	1	10	16.9	35550
sgd	0.1	None	16.9	35550
sgd	0.1	10	15.7	35550
adadelta	0.1	None	12.6	35550
adadelta	0.1	10	12.5	35550
sgd	0.1	5	12.2	35550
adadelta	0.1	5	11.0	23700
sgd	0.01	10	05.6	35550
sgd	0.01	None	05.2	35550
adadelta	0.01	None	05.1	35550
adadelta	0.01	10	05.0	35550
adam	0.01	5	04.7	35550
sgd	0.01	5	04.6	10
adadelta	0.01	5	04.4	23700
adadelta	0.001	5	03.7	11850
adadelta	0.001	None	03.7	11850
adadelta	0.001	10	03.6	35550
sgd	0.001	5	03.4	23700
sgd	0.001	10	03.4	23700
sgd	0.001	None	03.4	23700
adam	0.01	None	02.2	11850
adam	1	5	00.0	11850
adam	1	None	00.0	35550
adam	1	10	00.0	35550
adam	0.1	5	00.0	35550
adam	0.1	None	00.0	35550
adam	0.1	10	00.0	35550
adam	0.01	10	000	None

5,000 training steps and growths over the next 30,000 training steps to a value of 20 %.

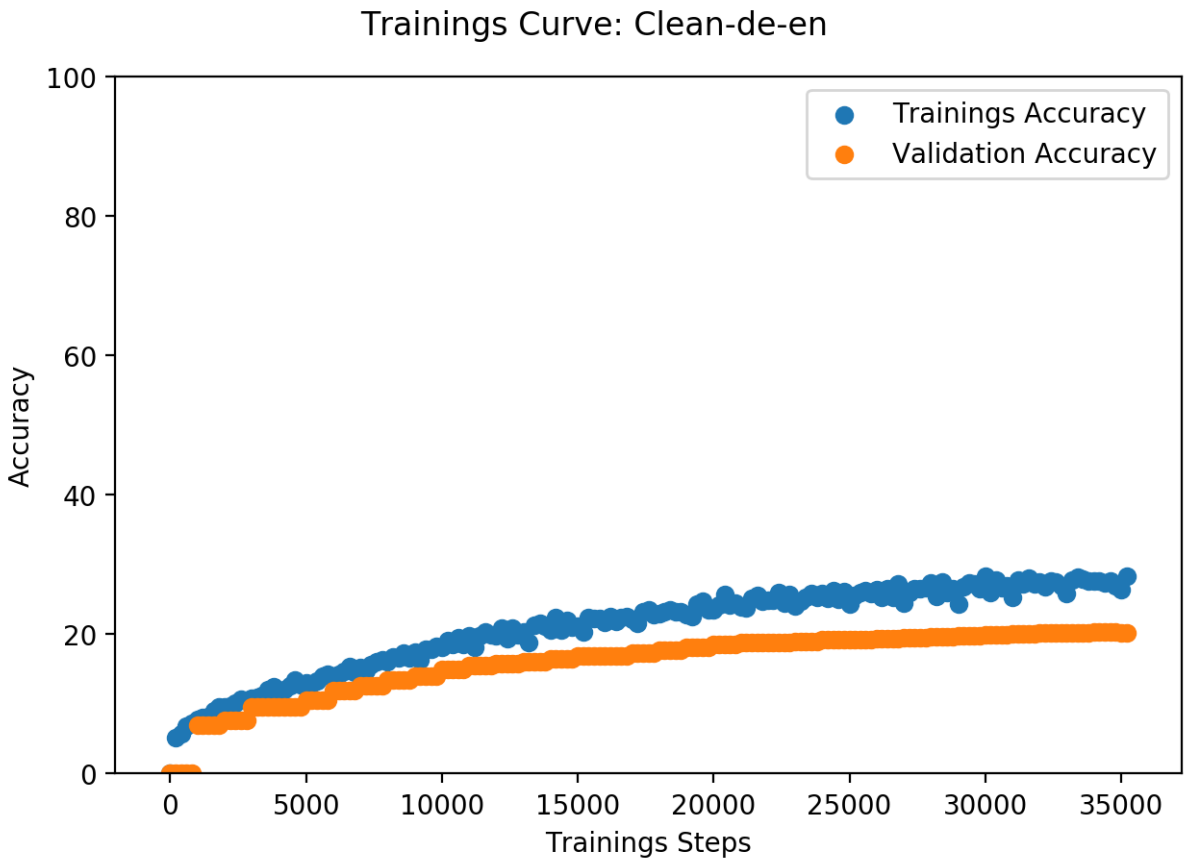

Figure 3.3: Example of a training curve of an unsuccsessfull training

The graph in 3.4 shows the same two metrics, which increase again ina logarithmic manner. The training accuracy reaches the 60 % mark after 10,000 training steps and gains another 20 % points over the remaining 25,000 training steps. After 15,000 steps, the curve begins to scatter in a sinus shape.

The validation accuracy reaches the 30 % mark after 5,000 training steps and gains another 10 % points over the next 10,000 training steps. Between step 15,000 and 35,000 the graph is constant around the 40 % mark.

The figure 3.5 shows the BLEU score and the validation accuracy in two graphs. For each metric, 5 different models are plotted, and the highest score is highlighted. In the first part of the graph, the plot scatters up to 10% points but stabilizes after 10,000 steps. The validation accuracy curve is logarithmic shaped and reaches the 30% mark after 5,000 training steps. It continues to grow until it reaches a plateau at the 40% mark after 10,000 steps.

The second graph shows the BLEU score for the same training. The first non-zero point is found after ca 3,000 training steps. The plots scatter for another 7,000 steps between the 10 % and 20 % mark with a linear growing tendency. Between

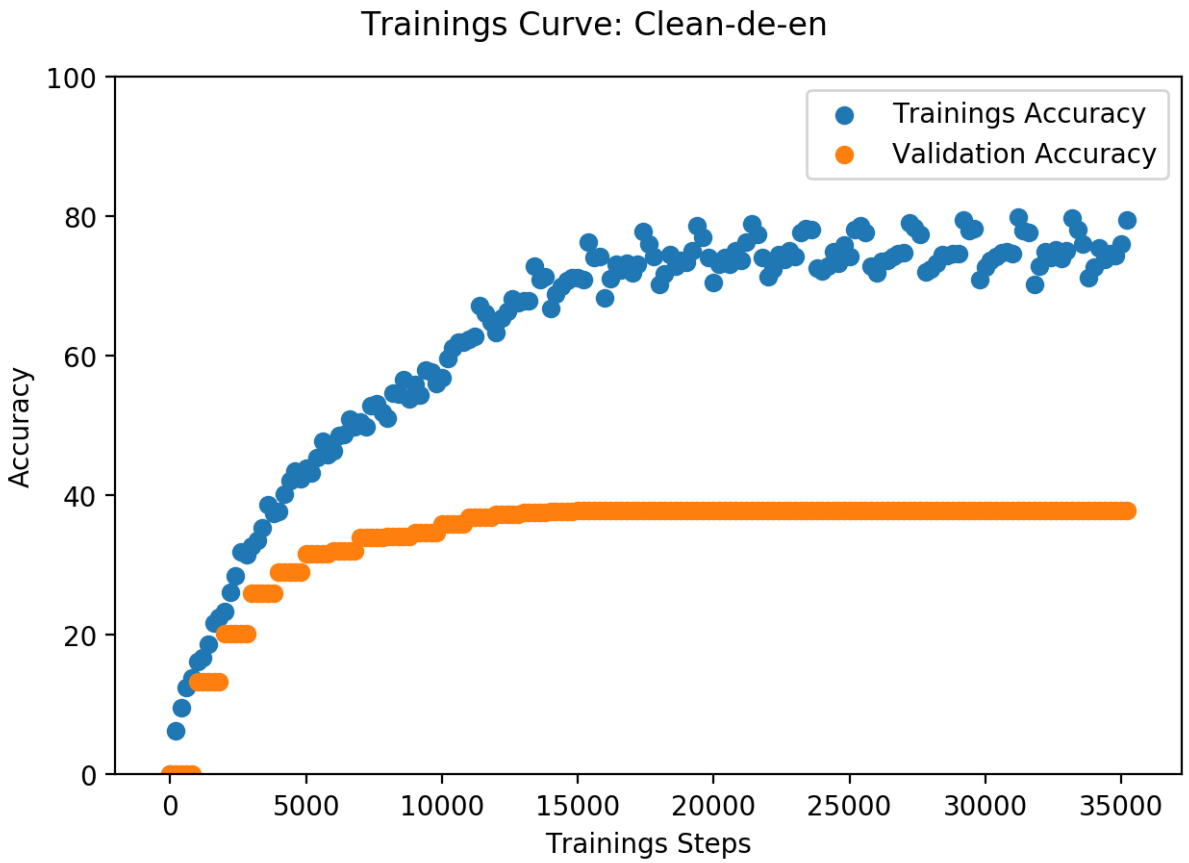

Figure 3.4: Example of a training curve of a succsessfull training

the 10,000th and the 20,000th training step, the points fluctuate between the 15 % and 20 % mark. During the last 15,000 steps curve reaches a plateau at the 20 % mark.

3.3 Comparison and Evaluation

3.3.1 Candidate Selection

For the evaluation I picked the following four combinations, which are highlighted in the tables 3.1 3.2 3.3 3.4.

The configurations were as follows:

3.3.2 Prefix Constraints

The figures 3.7 3.8 3.9 show the absolute performance per domain for the three scores BLEU, ROUGE-L and METEOR. All figures show two graphs with four groups of two bar diagrams each. The top graph shows the absolute score for the language pair German-English and the bottom graph for Czech-English. The first

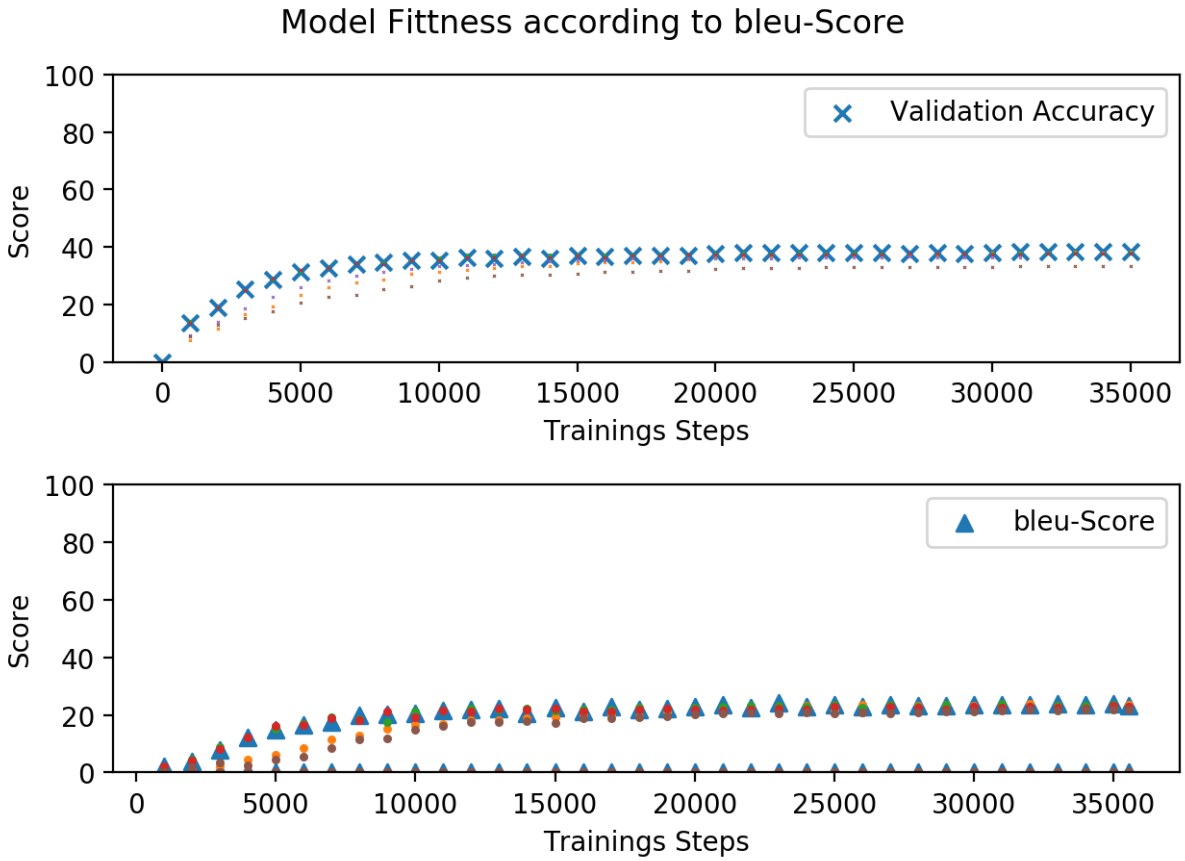

Figure 3.5: BLEU score and the validation accuracy for the best five models

bar in each group represents the performance without prefix constraints, and the second bar with prefix constraints.

BLEU

In both diagrams (Fig 3.7) the domain data sets show similar performance scores within the groups, but the performance per domain varied between 8% and 31% points for then German-English pair and between 15% and 41% for the Czech-English pair. The median scores for the German-English pair were 31% for ECB, 20% for EMEA, 8% for Europarl, and 18.5% for the mixed data set. In the Czech-English pair, the model scored a median 40% over ECB, 20.5% for EMEA, 15% for 23.5% the mixed set.

Rouge

The scores (Fig 3.8) are similar for all domains except EMEA. They rank for German-English between 25% and 54%, and between 34% and 65% in Czech-English. The median score for ECB in German-English is 53%, for Czech-English 63.5%, and Europarl 25% for German-English and 33.5% for Czech-English. In

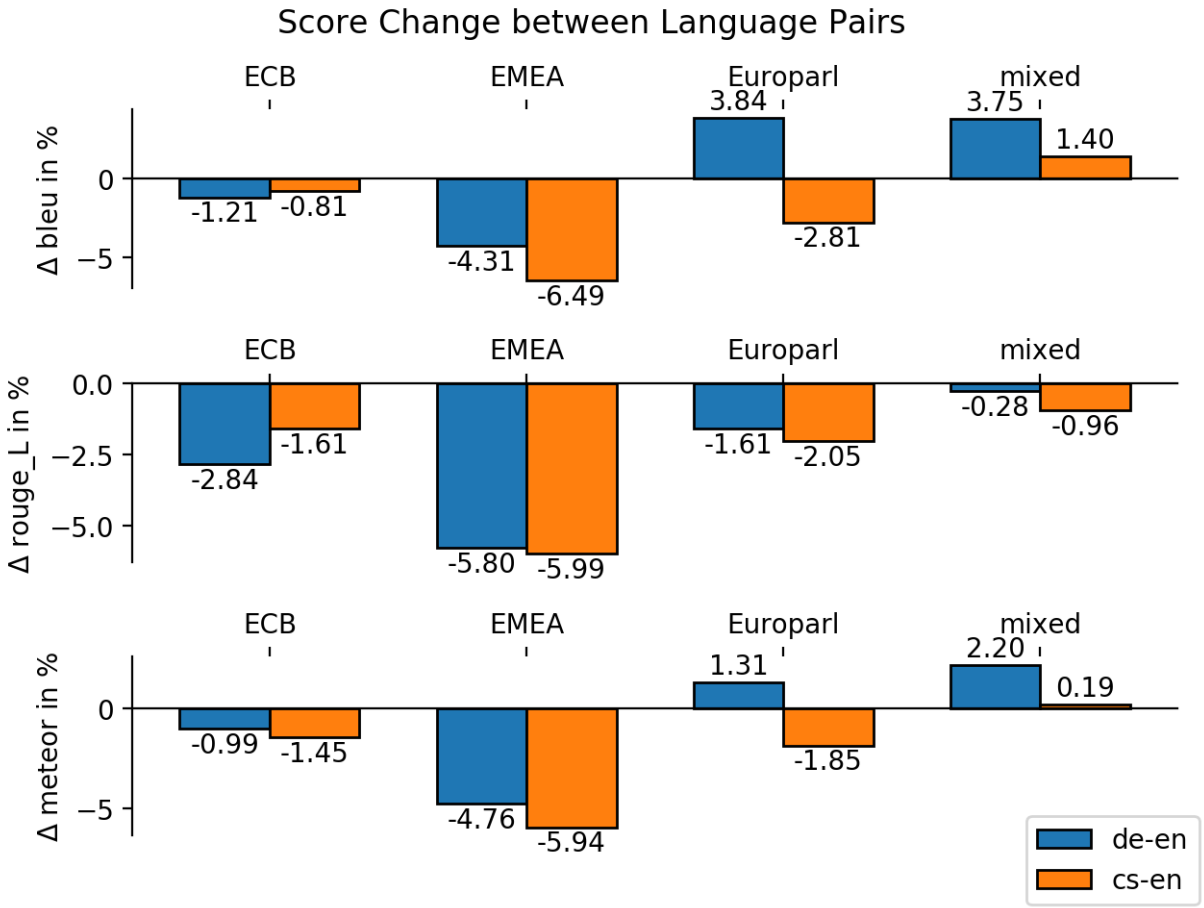

Figure 3.6: Score changes between language pairs for each domain

the EMEA domain, the German-English with prefix constrains scored 42.62% and without 45.25%. In Czech-English the model with prefix-constraints scored 41.07% and the model without 43.69%. The mixed data set had a score of 38% in German-English and 45% in Czech-English.

METEOR

The scores (Fig 3.9) are similar across all domains and language pairs, except ECB. In Czech-English the scores ranged between 18.5% (EMEA), 17% (Europarl), 20% (mixed) and 30% in ECB The ECB in German-English achieved 25%, where both EMEA and the mixed corpus scored 18%. The Europarl corpus reached 14.5%

3.3.3 Language Pairs

The figure 3.10 shows three graphs with four groups of two bars. Each graph represents the relative improvement of one metric measured over a model trained without prefix constraints to the model trained with prefix constraints. The grouped

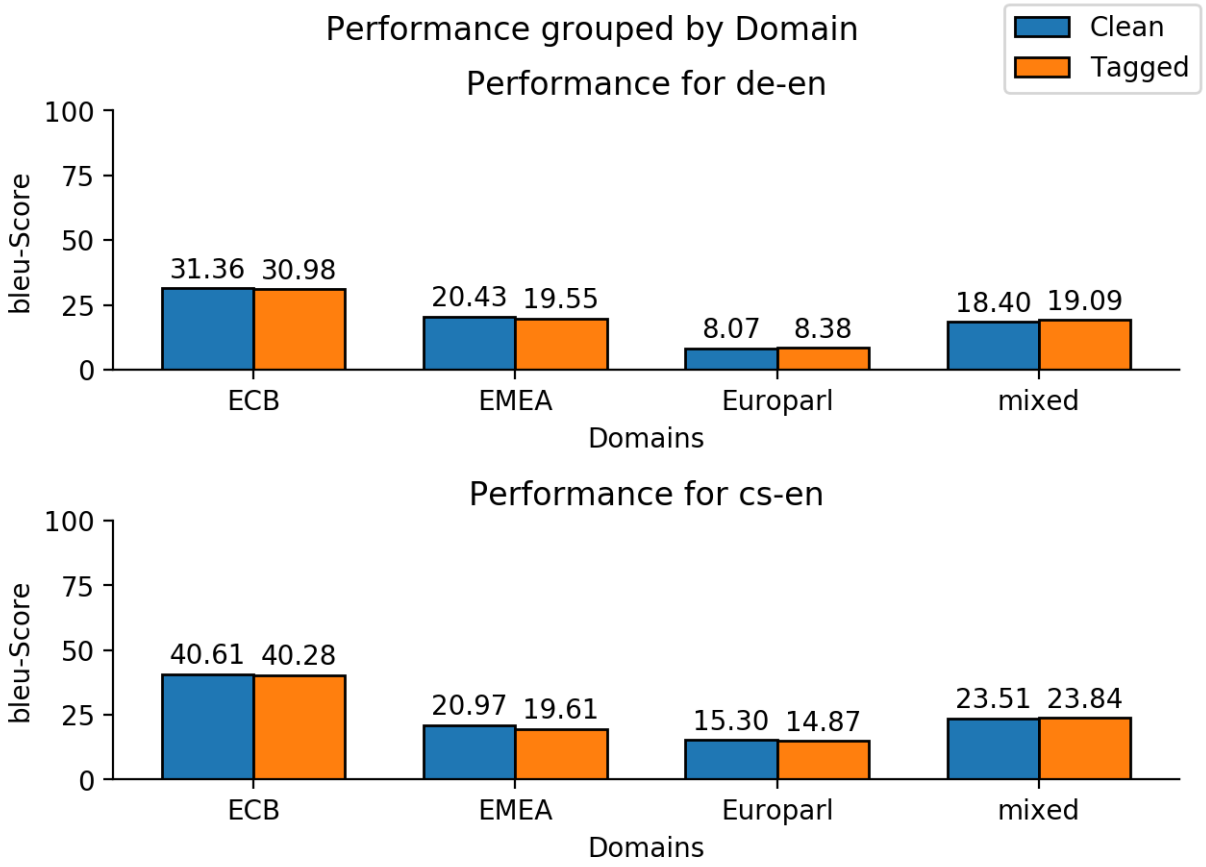

Figure 3.7: Comparison of BLEU score for each domain for with and without prefix constraints

bars represent the language pairs, where the first bar is German-English and the second pair represents Czech-English.

In the first row, the score change for BLEU is displayed. The score decreased for the domains ECB and EMEA and increased in the mixed data set. For the German-English pair, the score improved by 3.81% and decreased by 2.81% for the Czech-English pair.

The ROUGE-L metric shown in the second row decreased in all language pairs and domains: In the mixed data set by less than 1%, in ECB and Europarl between 1.6% and 2.8% and in EMEA by nearly 6%.

The last row shows the performance change for the METEOR score. It looks similar to the BLEU score change. For ECB and EMEA the performance decreased, and on the mixed data set the score improved over both language pairs. However, the improvement was smaller and in the Czech-English pair only at 0.19%. In the Europarl test set the METEOR score improved by only 1.31% in German-English where the Czech-English pair's score decreased by 1.85%

All scores are relative differences and not actual score points.

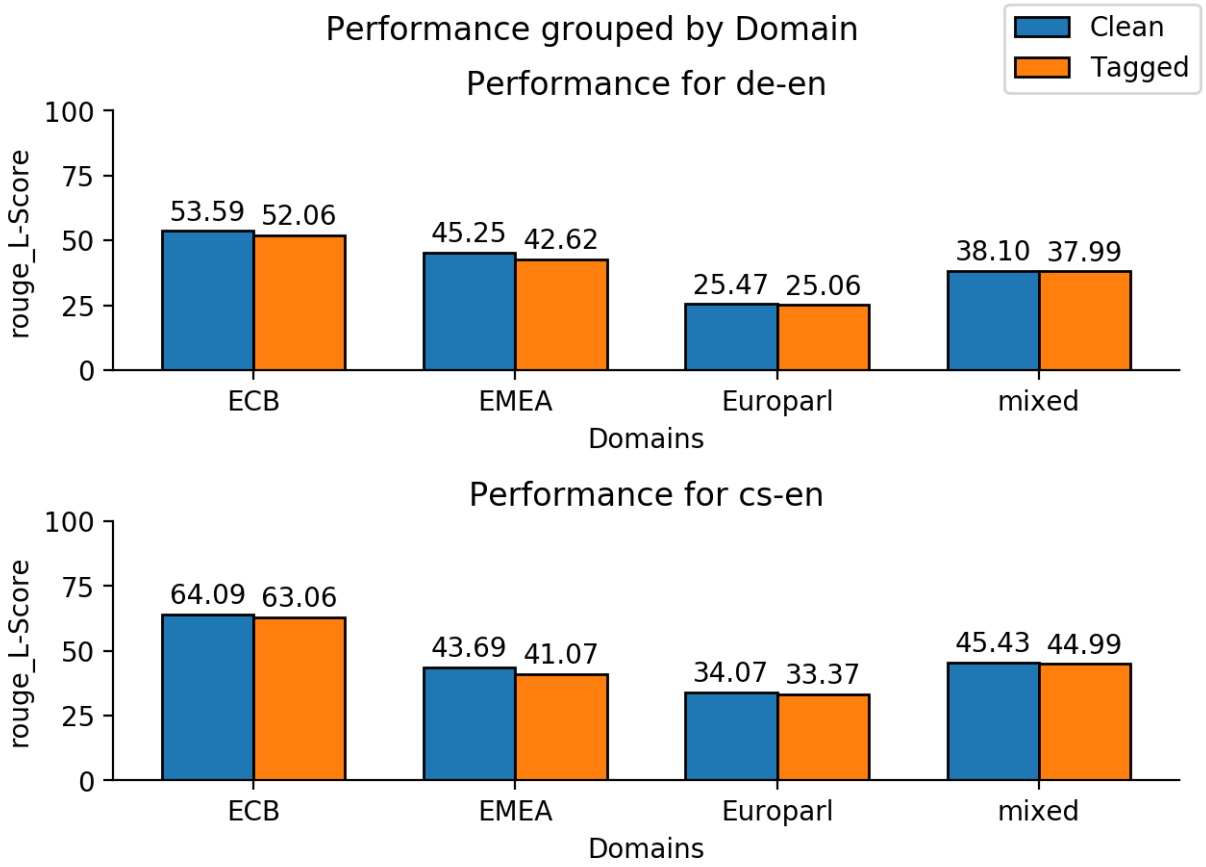

Figure 3.8: Comparison of ROUGE-L score for each domain for with and without prefix constraints

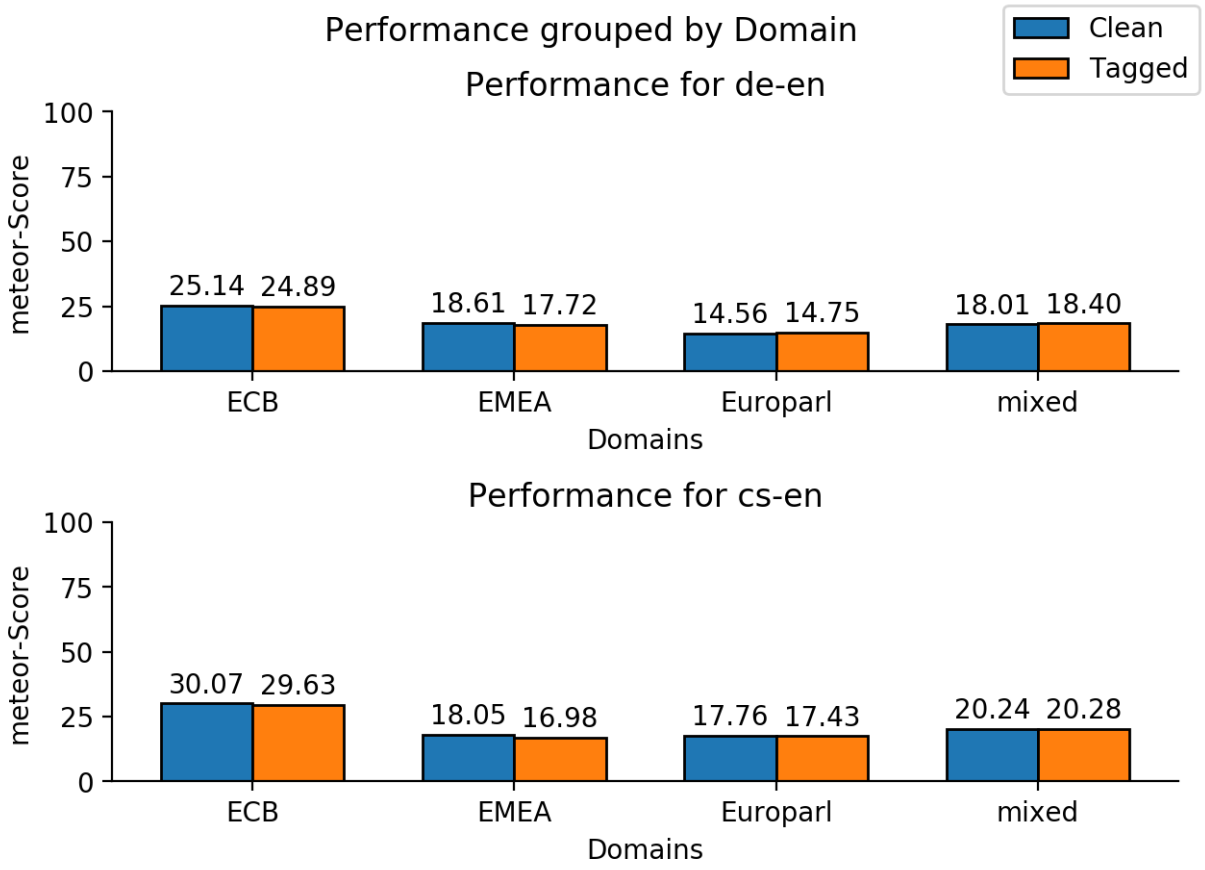

Figure 3.9: Comparison of METEOR score for each domain with and without prefix constraints

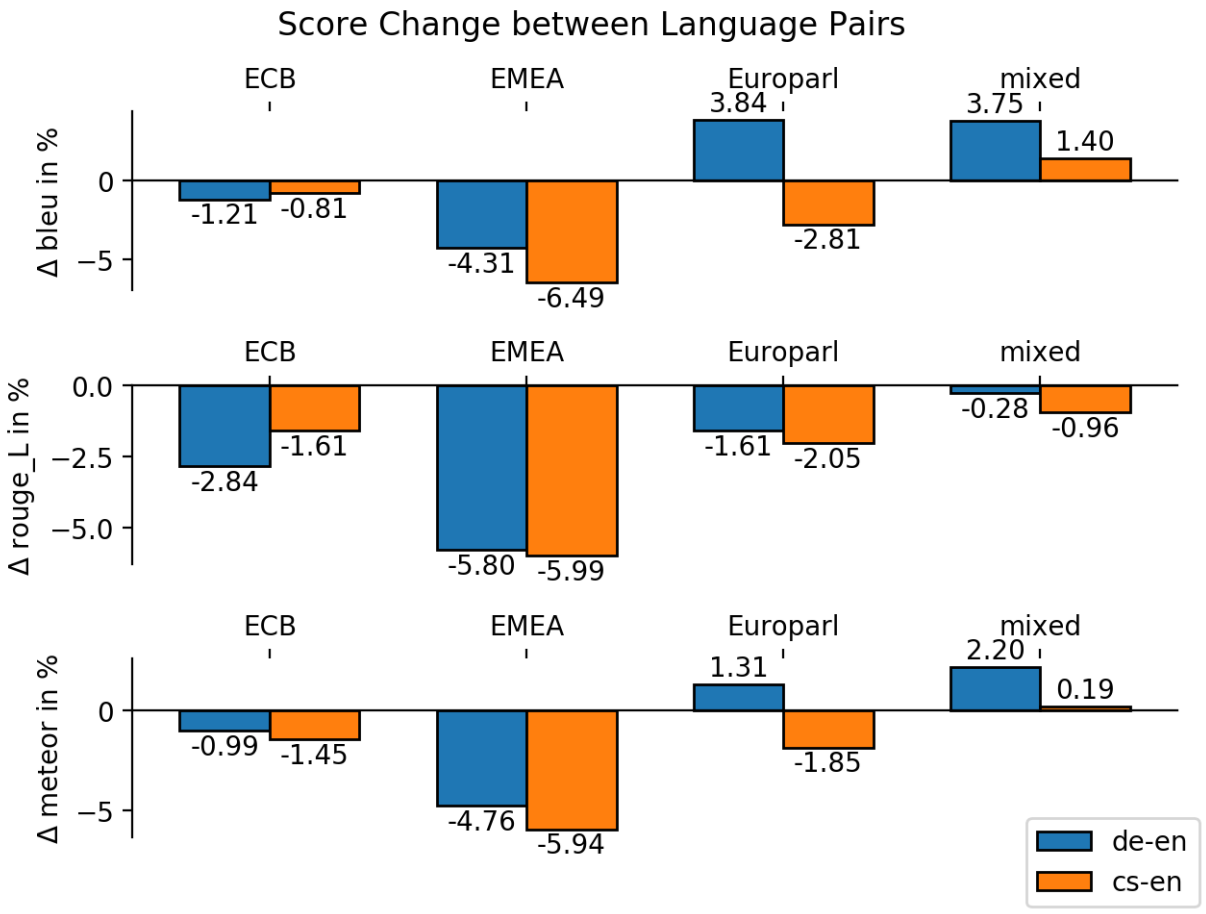

Figure 3.10: Score change (BLEU, ROUGE-L, METEOR)between language pairs

Chapter 4

Discussion

In the following section, I review my results and compare them to the current state of the art and discuss a possible connection between the relatedness of language pairs and the performance impact of the addition of prefix constraints.

4.1 Data Selection

The distributions of the number of words and the word length show that the domains had some structural differences. The EMEA corpus consisted in general of shorter sentences with longer words. However, while the median word and sentence length differed a little, the ECB and Europarl corpus seemed to be similarly structured, especially compared to the EMEA corpus. Since both ECB and Europarl are transcripts of meetings while EMEA consists mostly of patients' information, the similarities seem to be reasonable.

4.2 Data Preparation

During the data preparation step, I reduced the data sets by a large amount. I separated the corpus into smaller logical units by analyzing the text manually, since no index was available for my data. I choose the number of words and word length as two simple metrics to evaluate the original and reduced corpora. As seen in 2.1, the characteristics of the source corpora seem to be preserved in the reduced data sets. In the EMEA corpus, the word length distribution differs obviously from the original corpus. Since this corpus contains many brand names, I assumed that the variation in the small logical units was rather high.

Table 4.1: Guidline for the Interpretation of BLEU scores

BLEU-Score			Interpretation
0	<	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	<	100	Quality often bettern than humans

4.3 Training and Optimization

4.3.1 Training

The overall performance of the best models, according to BLEU, indicates successful training. According to LAvie (2010) the achieved score of 20 % points in BLEU can be interpreted roughly as an understandable but bad translation as seen in table 4.1 The top 5 models from each corpus reached that mark after 10,000 training steps, which is equivalent to 5 epochs.

The plots 3.3 and 3.4 show typical training curves as described by Lipton et al.(2015)

The validation accuracy in figure 3.4 indicates no overfitting, but the difference to the training accuracy shows additional potential with a larger data set.

4.4 Evaluation

4.4.1 Metric Interpretation

I used three different metrics to measure the quality of the translations of the trained neural networks in comparison to a conventional translation.

BLEU

The BLEU score is computed by measuring the difference between word groups and is calculated over the whole text. The metric was developed by Papineni et al. (2002) to measure translation quality. In this thesis, the BLEU score is used to measure the overall translation precision. The score can be interpreted according to Lavie (2010). As shown in figure 3.7, the best model achieved a score of 24 %, which indicates an understandable translation.

METEOR

The METEOR metric was developed to improve the correlation of human judgment with the translation score. It is calculated on sentence levels (Banerjee & Lavie, 2005). I used this score to represent the comprehensibility of the translation. The best model scored 20 % in the mixed data set. However, the METEOR score can only be used for comparisons and not as scale.

ROUGE-L

The ROGUE-L score is calculated by measuring the similarity of longer subsequences. This score was developed to compare summaries (Lin & Och, 2004). Since the score benefits from structural and long word matches, I used this score to measure the domain specialization. The interpretation of the ROUGE score is only evaluated on summaries, so the absolute score can not be interpreted here. However, its change is often used in model comparison (Sharma et al., 2017)

Overall

The overall score summarizes all three scores and is calculated through the addition of the ranks minus the total number of models times the number of metrics for the models. It represents the fit of the model in comparison to all models, that were trained on the same data.

4.4.2 Model Selection

During the model selection, the ranking using the overall score was similar to the ranking by individual scores. Since all scores rate the similarities to a reference text, an overall correlation is expected. This tendency is already known and described by other authors, for example Pryzbocki et al. (2009).

4.4.3 Prefix constraints

In 3.2.2 I described how the models performed according to METEOR in the different domains and how the performance changed in relation to the absolute score. All selected models performed best on the ECB and benefits most from prefix constraints in the Europarl corpus. The performance impact in the mixed test set was similar to the described impact by Kobus et al.(2016)

The overall performance showed a slight improvement when prefix constraints were provided but decreased in most corpora containing only one domain.

BLEU and METEOR both show that the general translation precision and comprehensibility of the mixed data set increased. The prefix constraint aimed to improve translations of mixed data sets by highlighting one essential meta information. Therefore, if prefix constraints are available, the network can focus more on the similarities between domains instead of learning specific features from it.

The difference between improvement in the BLEU (indicating precision) and METEOR (indicating comprehensibility) score, indicates that the prefix constraints helped more with the structural features than the content features of the corpora.

The ROUGE-L indicates that the addition of prefix constraints decrease the domain specialization in all domains, but its effects on the mixed dataset are rather small. Since the ECB and Europarl corpus are similar, the generalization does not impact the performance as much as on the EMEA test set.

Overall, generalizability seems to improve at the expense of specialized knowledge. The translation precision improves more than the comprehensibility.

4.4.4 Language Comparison

The comparison of the score change between the language pairs in figure 3.10 shows two major differences.

The first difference is the change in translation quality (measured by BLEU and METEOR) in the Europarl test set. In the English-German pair, the translation quality improved with the addition of prefix constraints. In contrast, in English-Czech, translation quality decreased. This indicates that the addition of prefix constraints improved the domain-specific translation in German-English.

The second difference is the quantity of the performance change The performance loss in the precision (BLEU score) and comprehensibility (METEOR score) on the domain sets in the Czech-English pair are higher, and the improvement in the mixed data set smaller.

Zoph & Knight (2016) used multiple input languages for a neural machine translation and showed that combining languages can improve translation quality. They found that the improvement is greater when distant languages are used. This indicates that domains or at least ambiguities are encoded differently in languages. While the classification of domain membership might be of similar difficulty over all domains, the pairwise differentiation may be different.

This can also be seen in figure 3.7, which indicates that the domain differences between FINANCE (ECB) and LAW (Euroarl) are in the selected data more obvious in Czech-English than in German-English.

Overall the comparison shows that prefix constraints improve the translation comprehensibility and precision whenever the domain membership is not expressed clearly through structural differences.

Since the difficulty of domain differentiation and domain classification may differ distinctly in languages, the impact of prefix constraints impact is not always beneficial. My results indicate that the domain classes are more clearly defined in Czech-English than in German-English. However, additional research is needed to prove this for other language pairs and in general.

4.5 Limitations

During my training, I ran into issues, that might have had an impact on my results and interpretation.

4.5.1 The relatedness of English-Czech and English-German and the generalization for the distance between languages

In my testing, I used Czech and German as languages that originated from different language families. However, they are still related and are closer to English as for example, Japanese or Chinese. The results need to be verified with more data from these language pairs. Also, only two language pairs were tested, and the indication described in 4.4.4 needs to be evaluated on more language pairs in the future.

4.5.2 The domain Selection and the Corpus Metrics

I used only three domains and compared very simple metrics on the corpora, and my indication (4.4.4) and observation (2.2.3) based upon these differences between the domains. To increase the confidence in my results, I suggest evaluating the same metrics on more domains. More complex metrics could be used to review my data and prove the described relationship.

4.5.3 Hyper Parameter Selection and Optimization

I ran only one round of hyper parameter optimization and compared only a few parameters and values. My training curves showed typical aspects of successful training, but better configurations may exist. This can be examined by rerunning the optimization with more configurations.

4.5.4 Better Performance of the distant Language Pair

In my results, the absolute score in all metrics was higher for Czech-English. This does not affect my comparison since my discussion interpreted only the relative change in the scores. On top of that, as Pryzbocki et al. (2009) explained, the scores do not represent the overall translation capabilities and are only comparable within the same language and the same corpus. In my training, the Czech-English models ran for 10 % more steps, since the corpora splits were slightly different then in the German-English corpora. However, as presented in 2.1, the overall key metrics were similar, and the models were trained on the same number of epochs. As mentioned in the limitations (4.5.3), the parameter selection may not have included the best model. I suggest using an exact alignment for all three languages and the training with additional parameter configurations to evaluate the absolute score differences.

4.6 Perspective

Based on my results it would be interesting to evaluate the following aspects:

4.6.1 More Domain and Language Comparison

The indication, that the performance impact of prefix constraints is connected or even correlates with the domain distinction between the language pairs should be evaluated with more data. The evaluation with more and more diverse domains may reveal additional insights on the impact of prefix constraints and the connection with the domain distinction differences. Additional languages for the training with English source texts, as well as new and different language pairs should be tested.

4.6.2 Evaluation with full Corpora

I ran my models on a reduced data set. The training without the reduction can improve the general ability of my findings.

4.6.3 Training with additional Attention Types

In the thesis, I used only a very simple attention type for the models. However, since the prefix constraints add additional meta information, a more complex attention type might benefit immensely from this mechanism.

4.7 Future work

Based on my findings, I suggest the following future projects:

4.7.1 Reviewing of Domain Control Mechanism on different Language Pairs

Most domain control and adaption mechanism are only tested for one or two language pairs. However, my findings indicate that language selection can have a huge impact on the performance of the domain control mechanism. Therefore, I suggest reevaluating the findings for the domain control mechanism for additional language pairs. Especially since Tkeno et al. (2017)rely on the work of Kobus et al. (2016) but use largely unrelated languages.

4.7.2 Creation of a relatedness Score

To measure the relatedness of language pairs and calculate a possible correlation with the performance on translation systems, I suggest the creation of a new and simple score. The Foreign Service Institute uses the necessary study time to rank

the training difficulty of language pairs. I suggest the creation and evaluation of a score based on the median training time. This information can be easily collected for humans and tested for neural machine translation and can then be to represent a translation disadvantage.

4.7.3 Extend the OPUS Project

The OPUS project hosts a variety of corpus data. However, the data is only available in an unstructured format. I invite all researchers to submit their training data and preparation scripts to generate a structured data bank. Projects like OpenNMT can benefit hugely from a shared collection of preparation scripts because unity on the data sets is key to further research in the whole field.

Chapter 5

Conclusion

In conclusion, the problem of domain control in neural machine translation is very challenging and not solved in any translation-related area. The impact of the domain control mechanism prefix constraints can not always be predicted and depends on the similarity of domain distinction in the languages. The domain and language selection should always be considered when choosing any domain adaption mechanism. The connection between the impact of prefix constraints and the relatedness needs to be researched further.

References

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

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of compstat'2010* (pp. 177–186). Springer.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 109–120.

Chu, C., Dabre, R., & Kurohashi, S. (2017). An empirical comparison of simple domain adaptation methods for neural machine translation. *arXiv preprint arXiv:1701.03214*.

Georgi, R., Xia, F., & Lewis, W. (2010). Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 385–393).

Hajic, J. (2000). Machine translation of very close languages. In *Sixth applied natural language processing conference* (pp. 7–12).

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65–93). Elsevier.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... others (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proc. acl*. Retrieved from `https://doi.org/10.18653/v1/P17-4012` doi: 10.18653/v1/P17-4012

Kobus, C., Crego, J. M., & Senellart, J. (2016). Domain control for neural machine translation. *CoRR*, *abs/1612.06140*. Retrieved from `http://arxiv.org/abs/1612.06140`

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation.

arXiv preprint arXiv:1706.03872.

Lavie, A. (2010). Evaluating the output of machine translation systems. *AMTA Tutorial*, 86.

Light, R. A., et al. (2017). Mosquitto: server and client implementation of the mqtt protocol. *J. Open Source Software*, *2*(13), 265.

Lin, C.-Y., & Och, F. (2004). Looking for a few good metrics: Rouge and its evaluation. In *Ntcir workshop.*

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2015). Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

Liu, W., Wen, Y., Yu, Z., & Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. In *Icml* (Vol. 2, p. 7).

Luo, J., Cao, Y., & Barzilay, R. (2019). *Neural decipherment via minimum-cost flow: from ugaritic to linear b.*

Luong, M.-T., Pham, H., & Manning, C. D. (2015). *Effective approaches to attention-based neural machine translation.*

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association.*

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `https://doi.org/10.3115/1073083.1073135` doi: 10.3115/1073083.1073135

Przybocki, M., Peterson, K., Bronsart, S., & Sanders, G. (2009). The nist 2008 metrics for machine translation challenge—overview, methodology, metrics, and results. *Machine Translation*, *23*(2-3), 71–103.

Rico Sennrich, B. H., & Birch, A. (2016). Controlling politeness in neural machine translation via side constraints. In *Naacl.*

Sennrich, R., Haddow, B., & Birch, A. (2015). *Neural machine translation of rare words with subword units.*

Sharma, S., El Asri, L., Schulz, H., & Zumer, J. (2017). Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, *abs/1706.09799*. Retrieved from `http://arxiv.org/abs/1706.09799`

Takeno, S., Nagata, M., & Yamamoto, K. (2017). Controlling target features in neural machine translation via prefix constraints. In *Proceedings of the 4th workshop on asian translation (wat2017)* (pp. 55–63).

Tiedemann, J. (2012, may). Parallel data, tools and interfaces in opus. In N. C. C. Chair) et al. (Eds.), *Proceedings of the eight international conference on language resources and evaluation (lrec'12).* Istanbul, Turkey: European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 5998–6008). Curran Associates, Inc. Retrieved from `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3156–3164).

Wichmann, S., Müller, A., & Velupillai, V. (2010). Homelands of the world's language families: A quantitative approach. *Diachronica*, *27*(2), 247–276.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation.*

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zoph, B., & Knight, K. (2016). Multi-source neural translation. *CoRR*, *abs/1601.00710*. Retrieved from `http://arxiv.org/abs/1601.00710`

Erklärung der Urheberschaft

Hiermit versichere ich an Eides statt, dass ich die vorliegende Bachelorthesis im Studiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel - insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Ort, Datum Unterschrift

Erklärung zur Veröffentlichung

Ich stimme der Einstellung der Bachelorthesis in die Bibliothek des Fachbereichs Informatik zu.

Ort, Datum Unterschrift