Domain adaptation: Retraining NMT with translation memories

Maria Mäkinen
Master's thesis
Master's programme in translation and interpreting
Faculty of Arts
University of Helsinki
Supervisor: Prof. Lauri Carlson
May 2019

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Humanistinen tiedekunta	Koulutusohjelma – Utbildningsprogram – Degree Programme Kääntämisen ja tulkkauksen maisteriohjelma	
Opintosuunta – Studieinriktning – Study Track Käännösteknologia		
Tekijä – Författare – Author Maria Mäkinen		
Työn nimi – Arbetets titel – Title Domain adaptation: Retraining NMT with translation memories		
Työn laji – Arbetets art – Level Pro gradu -tutkielma	Aika – Datum – Month and year Toukokuu 2019	Sivumäärä– Sidoantal – Number of pages 43 s.; suomenkielinen lyhennelmä 5 s.

Tiivistelmä – Referat – Abstract

The topic of this thesis is domain adaptation of an NMT system by retraining it with translation memories. The translation memory used in the experiments is the EMEA corpus that consists of medical texts – mostly package leaflets. The NMT system used in the experiments is OpenNMT because it is completely free and easy to use.

The goal of this thesis is to find out how an NMT system can be adapted to a special domain, and if the translation quality improves after domain adaptation. The original plan was to continue training the pretrained model of OpenNMT with EMEA data, but this is not possible. Therefore, it is necessary to train a new baseline model with the same data as the pretrained model was trained with. After this two domain adaptation methods are tested: continuation training with EMEA data and continuation training with unknown terms.

In the manual evaluation, it turned out that domain adaptation with unknown terms worsens the translation quality drastically because all sentences are translated as single words. This method is only suitable for translating wordlists because it improved the translation of unknown terms. Domain adaptation with EMEA data, for the other hand, improves the translation quality significantly. The EMEA-retrained system translates long sentences and medical terms much better than the pretrained and the baseline models. Long and complicated terms are still difficult to translate but the EMEA-retrained model makes fewer errors than the other models.

The evaluation metrics used for automatic evaluation are BLEU and LeBLEU. BLEU is stricter than LeBLEU. The results are similar as in the manual evaluation: The EMEA-retrained model translates medical texts much better than the other models, and the translation quality of the UNK-retrained model is the worst of all.

It can be presumed that an NMT system needs contextual information so that it learns to translate terms and long sentences without transforming the text into a wordlist without sentences. In addition, it seems that long terms are translated in smaller pieces so that the NMT system possibly translates some pieces wrong, which results in that the whole term is wrong.

Avainsanat – Nyckelord – Keywords
NMT, neural machine translation, domain adaptation, translation memories, machine translation

Säilytyspaikka – Förvaringställe – Where deposited
Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgifter – Additional information

Table of contents

Figures

Tables

List of abbreviations

CNN	Convolutional neural network
FFN	Feed-forward neural network
LSTM	Long short-term memory
MT	Machine translation
NMT	Neural machine translation
RNN	Recurrent neural network
SMT	Statistical machine translation
TM	Translation memory
UNK	Unknown word or term

1 Introduction

Lately, artificial intelligence is being applied everywhere. All technology we use is becoming more and more intelligent.

The translation industry develops with the others: In the recent years, there have been many studies on neural machine translation which replaces the popular statistical machine translation (SMT). The researcher's goal is to generate machine translation systems which translate as well as people do. However, we are a long way from that because the translation quality has still room for improvement.

Especially in the medical domain, comprehensibility and a good quality of translations is important to ensure the safety of patients. Serious translation errors in package inserts, for instance, may involve the risk that patients take the wrong dose of medication or does not know what side effects may occur. The latter is the case when the text is too confusing. For these reasons it is important to improve the quality of machine translation (MT) output.

Because neural machine translation (NMT) as a research domain is quite new, it is necessary to explore it more. Therefore, I selected this topic for my master's thesis. More precisely, the aim was to find out how an NMT system can be adapted to a special domain by using translation memories, and if the translation quality of an adapted NMT system improves.

Domain adaptation for NMT systems has not been explored that much. In recent works (see e.g. Chinea-Ríos, Peris & Casacuberta 2017), the NMT system has been trained with machine translation output or back-translated data. My research differs from these insofar that I use translation memories to retrain a general purpose NMT system.

So far translation memories have only been used to train SMT systems (see e.g. Läubli, Fishel, Volk & Weibel 2013). They trained their SMT system with TMs on texts of the automobile industry in German-French and German-Italian. The results of the research are promising: the translation quality of their weighted SMT system, which pays special attention to the domain specific vocabulary, improved significantly at least when evaluated with BLEU. The weighted system beats the baseline system also in human evaluation tasks. (Läubli et al. 2013: pp. 333-337.) Therefore, it can be hypothesised that the translation quality improves also after retraining an NMT system with in-domain TMs.

This thesis is structured as follows: In the second chapter I give an overview of the theoretical background which is necessary for understanding, how an NMT system works. To do this, I first introduce SMT and after that I introduce different types of NMT systems and neural networks. In this case, the main focus is on pure NMT systems because the system, which will be used in my experiments, belongs to the same type. In this sub-chapter NMT systems are also compared to SMT systems. After that I will tell about automatic evaluation of machine translations and explain, why it is recommended to use several evaluation metrics rather than rely only on one single evaluation metric. Attention is especially turned to the BLEU and LeBLEU metrics. At the end of this chapter, I present studies on domain adaptation for NMT.

In the third chapter, I will introduce the NMT system, research method and translation memory. I introduce two different domain adaptation methods: continuation training with in-domain data and continuation training with unknown terms which are extracted of the in-domain corpus.

In chapter four I present the results of my own research. Firstly, explain how the NMT system can be adapted to the medical domain. I also compare the translations of the different NMT models with each other. Secondly, I present the BLEU and LeBLEU score for the different NMT models compare them with each other. The evaluation shows whether domain adaptation improves the translation quality or not. Thirdly, I present the most frequent translation errors and describe why the example translations are wrong.

Chapter five concludes my work: I summarise the most important pieces of the previous chapters, summarise the research results, and give some suggestions for future works.

2 Theoretical background

In this chapter, SMT systems and NMT systems are first introduced. The focus will be on NMT systems: their components, functionality, and different neural network types will be presented. After that both systems and their translation problems are compared. The second topic in this chapter is automatic evaluation of machine translation and especially the BLEU and LeBLEU metrics which are also used in the experiments. Finally, recent related work on domain adaptation for NMT is described.

2.1 Machine translation

2.1.1 Statistical machine translation

As the name suggests, the basis of statistical machine translation (SMT) systems' functionality are calculated probabilities for possible translation outputs from which the best one is selected (Vatsa, Joshi & Goswami 2010: p. 25).

The systems learn possible translations from aligned parallel corpora which are fed into it in a training phase. All SMT systems create phrase tables or translation tables which help them to find the right words for the words in the input sentences. These translation tables include not only input words and their translation but also the probabilities of the translations. The creation of translation tables is called word alignment. (Vatsa, Joshi & Goswami 2010: pp. 25-26.) In phrase-based SMT, word alignments are extended to word n-gram alignments. At translation time, the number of translation hypotheses can be restricted by using a beam search function which is part of the decoder (Koehn, Och & Marcu 2003: pp. 49-50).

In theory, SMT systems use the Bayes rule to calculate probabilities for possible translations (Vatsa, Joshi & Goswami 2010: p. 27):

$$\text{argmax } \Pr(e|g) = \frac{\Pr(e)\Pr(\frac{g}{e})}{\Pr(g)} = \Pr(e)\ \Pr(g|e)$$

Equation 1: *Bayes rule (Vatsa, Joshi & Goswami 2010: p. 27).*

The Bayes formula motivates the use of monolingual language models which are fed into the training phase – that means that SMT systems are not only trained with parallel corpora but also with language models (Och & Ney 2002: p. 295). The language model checks that the

translation is fluent target language. SMT systems which use translation models (see equation 1) as well as language models are also called noisy-channel models. (Koehn 2010: p. 7.)

Besides word and phrase-based SMT systems, there are also tree-based systems. Phrase-based SMT is the most common SMT system. The sentences to be translated are segmented whereby the segments are parts of the sentences (for example single terms). Word-based models, on the other hand, translate texts word by word, and tree-based systems create tree structures of the sentences with the help of learned grammatical information. (Koehn 2010: pp. 6-14.)

2.1.2 Neural machine translation and machine learning

Artificial intelligence has been a popular research topic for years. Recent advances in machine learning (in particular deep neural networks learning) has allowed researchers of translation technology to develop neural machine translation systems and reach better results by using such technologies. In addition, NMT systems require less memory than other MT systems. (Cho, van Merriënboer & Bahdanau 2014: p. 1.)

While the structure of an SMT system with multiple components and subcomponents is more complex, a pure NMT system is simpler with at best only one neural network, which consists of an encoder, which encrypts the source sentences into vectors, and a decoder, which builds the target sentences. Generally, an NMT system also contains at least one hidden layer – also known as softmax layer. (Zhang & Zong 2015: p. 23.)

As in SMT, the simplest main formula to find the best possible MT output using an NMT system which calculates the highest probability translation is the following equation:

$$\text{argmax}_y \, p(y|x)$$

Equation 2: *Equation for finding the best possible translation (Bahdanau, Cho & Bengio 2016: p. 2).*

In equation 1 the target sentence is y, x is the source sentence, and argmax_y represents the argument giving maximum probability, i.e. the best possible translation. As in SMT, the probability is calculated with help of parallel corpora which are fed into the system in the training phase. In practice, the NMT system maps the translations in the corpora into the neural network and learns optimal weights for source-target connections from them.

(Bahdanau, Cho & Bengio 2016: pp. 1-2.) Maximisation over token n-grams happens implicitly in the neural network training. The best translation of the whole sentence is chosen using beam search rather in the same way as in SMT. (Vaswani et al. 2017: p. 8.)

Many different network architectures for NMT have been proposed to date, and the field gets updated frequently. In the following, I sketch some of the popular architectures just at present.

The most commonly used network types in recent NMT systems are the recurrent neural network (RNN) (Zhang & Zong 2015: p. 23) and the transformer architecture (Vaswani et al. 2017).

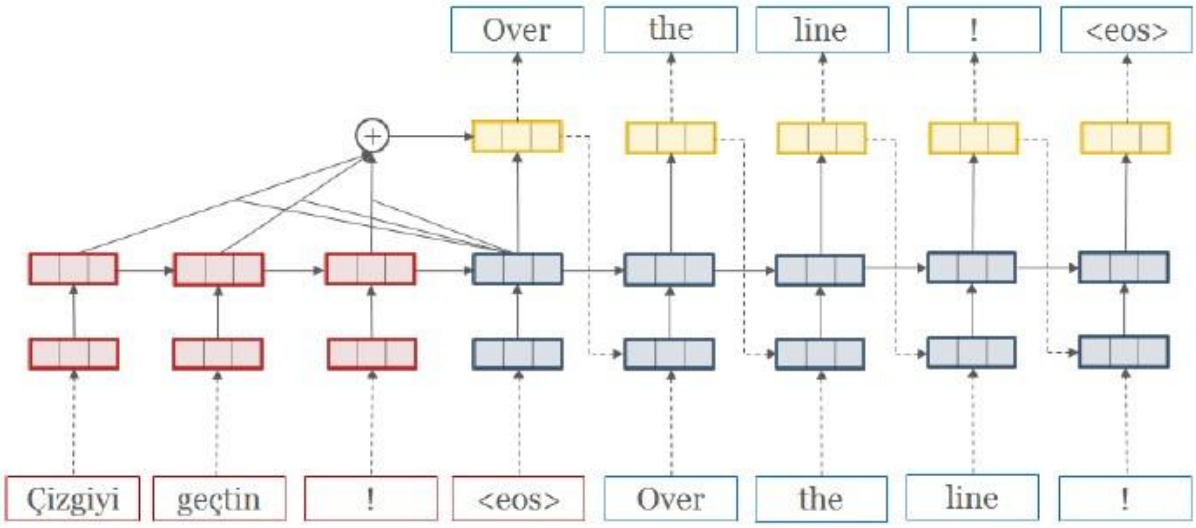

Figure 1: *The structure of an NMT system in which the red boxes are source RNNs, the blue boxes target RNNs and the yellow boxes hidden states (Klein, Kim, Deng, Nguyen, Senellart & Rush 2017: p. 67).*

The encoder and decoder deal with sequences of different lengths (Cho, van Merriënboer & Bahdanau 2014: p. 2). If the encoder were to put the source sentence into a single, fixed-length vector, the system has difficulties in translating long sentences. It is better to split the sentence into sequences and then convert these sequences into smaller vectors: $x = (x_1...x_{Tx})$. All sequence vectors of one sentence form a list, and the last vector of a sentence is marked with a special symbol (end-of-sentence, EOS). (Bahdanau, Cho & Bengio 2016: pp. 1-3.)

Between the encoder and decoder are layers representing hidden states which are updated at each time step during the training process. By updating the hidden states regularly, the NMT system can guess the next input sequence during the translation process. A context vector

helps the system to guess the next input sentences by using previously translated words $\{y_1...y_t'-1$). (Bahdanau, Cho & Bengio 2016: pp. 1-3.)

Long short-term memory (LSTM) (see figure 2) is a helpful component in RNN based NMT systems. It consists of memory cells which filter information of the previous state of the neural network so as to pass long distance information through the network. (Hochreiter & Schmidhuber 1997: pp. 6-7.)

The LSTM is usually connected with the NMT system's hidden layer(s) which can also deliver information to the gate units (or the hidden layer itself contains gate units). If a memory cell uses the same gate units as another memory cell, it builds a memory block which can save more information than a single memory cell. (Hochreiter & Schmidhuber 1997: p. 8.) An NMT system can have multiple LSTMs but, if there are too many of them, the neural network works slower than it does with less LSTMs. In addition, training an NMT system is easier if the system is less complex. (Wu, Schuster, Chen, Le & Norouzi et al. 2016: pp. 4-5.)

Another type of neural network used for NMT systems is the convolutional neural network (CNN). Unlike RNNs, the size of CNNs is fixed. If it is necessary, the network can be enlarged by stacking multiple CNN layers one upon other. While RNNs create a new hidden state for the positions of every new sequence, CNNs do not need information about the previous time steps. Instead of hidden layers, there are convolution kernels in the encoder and decoder through which the MT input must go. (Gehring, Auli, Grangier, Yarats & Dauphin 2017: p. 1-3.)

One difference between NMT systems with RNNs and CNNs is that CNN use convolution instead of matrix multiplication to compute the best possible translation (Goodfellow, Bengio & Courville 2016: p. 326).

If an NMT system uses gated convolutional neural networks, learning the structure of a source sentence is very easy and fast by forming a tree structure of the sentence's pieces. The pieces are weighted after what the machine forms fixed-length vectors which represent the MT output. In general, an NMT system has four weight matrices and one hidden unit for doing this. (Cho, van Merriënboer & Bahdanau 2014: pp. 2-3.)

A popular improvement to NMT architecture is an attention mechanism which directs the NMT system to pay attention to specific (usually nearby) words or sequences in a source sentence (Bahdanau, Cho & Bengio 2016: p. 4.)

Another popular neural network type is the Transformer. Transformer is a model which consists of a feed-forward network (FFN) instead of an RNN. (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin 2017: p. 2.) FFNs are RNNs differ from each other in that RNNs have feedback connections which give feedback of the networks state to the output. FNNs do not have such connections. (Goodfellow et al. 2016: p. 164.)

Self-attention serves in this NMT system as a connection between input and output layers and it is related to the positions of input and output sequences. Thus, it is not necessary to compute a representation of these sequences. Both, encoder and decoder have six layers (Vaswani et al. 2017: p. 2.)

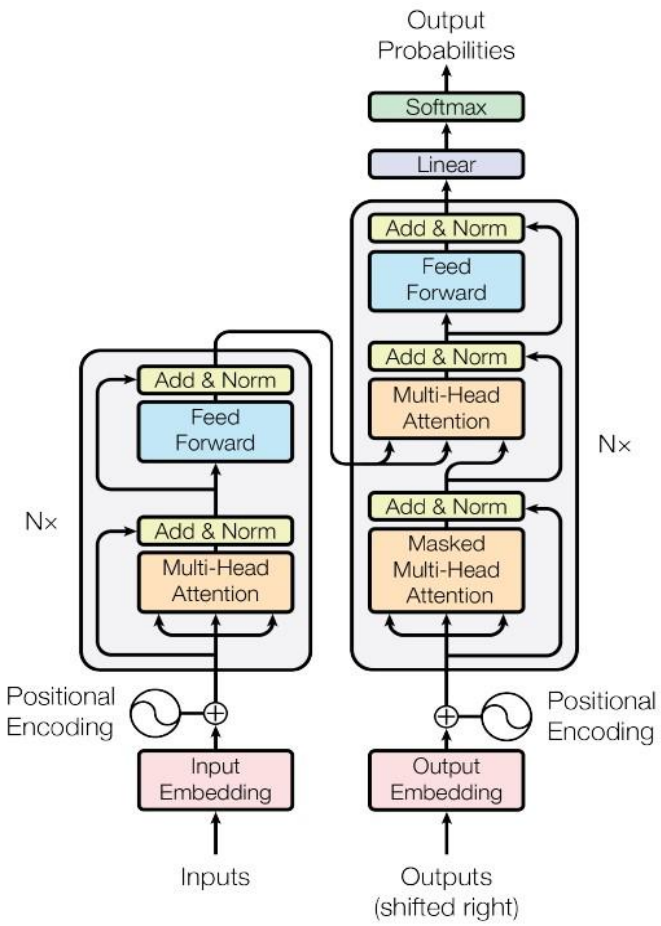

Figure 2: *The structure of the Transformer model.*

As can be seen in figure 2, the Transformer model includes three multi-head attention blocks. One of them is in the encoder (left) and two in the decoder (right). One of the decoder's multi-head attention block is connected to the encoder, so that the encoder's output can be fed into it. In addition, the encoder as well as the decoder has self-attention layers which take account of the positions of the input and output sequence. (Vaswani et al. 2017: p. 3-5.)

The Transformer's attention mechanism works a bit different than the attention mechanism of other systems. Vectors – i.e. queries and key-value pairs, in this case – are mapped to the layer's output which is a weighted sum of all mapped values (Vaswani et al. 2017: p. 4). The weights can be computed with a softmax function which can be used to calculate probabilities (Vaswani et al. 2017: p. 4; Goodfellow et al. 2016: p. 79).

There are three types of learning for NMT: supervised, semi-supervised and unsupervised learning. In supervised learning the machine does not get a separate model of the target language but rather it calculates the likelihood of translations and tries to maximize it to find the best translation for a certain source sentence.

In semi-supervised learning, however, the system uses a target-to-source translation model to build a translation (encoder) or it uses source-to-target model to build a reconstruction of the source sentence in the target language.

In unsupervised learning the NMT systems gets in the training phase a monolingual corpus which it uses to learn a certain language. The NMT system learns to translate with the help of autoencoders: the system builds a target-to-source translation model and uses it to construct a target sentence from the source sentence. (Cheng, Wu, He, Wu & Liu 2016: pp. 2-3.)

In summary, one difference between the mentioned learning types seems to be in the data which the NMT systems gets in the training phase: does it get language models for one or more languages, or does it train from monolingual corpora and build the language models itself.

NMT systems based on encoders and decoders can learn and align jointly, which means that the system can find the right positions (j) for words that it is currently translating, by soft-searching all the information that a certain word needs in a fluent target sentence. From the found positions the system builds context vectors which help to translate the word in the right way. The best translation of a word is in that case the word which is in position (i) and matches (i.e. has the highest probability) with the word in position j. (Bahdanau, Cho & Bengio 2016: pp. 1, 3.)

Because neural machine translation is a quite new field, there are problems which have not yet been solved. NMT systems have difficulties with long and complicated sentences as well as with small training vocabularies. (Cho, van Merriënboer & Bahdanau 2014: pp. 3-4.) If the NMT system's vocabulary does not contain a certain word, the system marks it as unknown (UNK) or copies it untranslated into the target text. Another possibility is to split unknown

word into pieces (“constituent characters” or “sub-word tokens”) and mark every piece with a location prefix (<B> for beginning, <M> for middle and <E> for end). (Wu, Schuster, Chen, Le & Norouzi et al. 2016: pp. 7-8.)

Reaching good quality is difficult especially for low-resource languages since there are only few parallel corpora and a limited number of domains, so the NMT system has not enough material for automatic learning in these languages. (Cheng, Wu, He, Wu, Sun & Liu 2016: pp. 1-2.)

2.1.3 Neural machine translation vs. statistical machine translation

If one takes a look at the publications on machine translation (see e.g. Koehn, Hoang, Birch, Callison-Burch et al. 2007 or Koehn, Och & Marcu 2003) in different databases, it can be seen that statistical machine translation was for a long time the most used and explored MT model (Vatsa, Joshi & Goswami 2010: p. 25). NMT is still in its infancy but the quality of the translation output is continuously improving. Koehn and Knowles (2017: pp. 28-29), for instance, have compared NMT output to SMT output. More precisely, they analysed the translations in respect of six error categories (which the researchers call challenges): 1) out-of-domain translations, 2) the effect of the amount of the training data, 3) rare words, 4) long sentences, 5) word alignment, and 6) beam search. The language pairs used in their experiments were English-Spanish and German-English.

According to Koehn and Knowles (2017: pp. 30-33), SMT systems perform much better in translating specialised texts – especially medical and legal terms are difficult to translate for NMT systems. People, who use translated specialised texts, “will be misled by hallucinated content in the NMT output” (Koehn & Knowles 2017: p. 30). Translating rare words does not really seem to be an irresolvable problem for NMT systems, because the results in the rare word challenge show that NMT systems are better with such words (mostly verbs, nouns, named entities and adjectives – i.e. inflected words – are difficult for both systems to translate). (Koehn & Knowles 2017: pp. 30-33.)

The results in the 2nd challenges are similar: the learning curve of NMT systems is much steeper than the learning curve of SMT systems but the quality of NMT’s translation output is much worse than of SMT’s output. If one trains an NMT system with 1/16 of the whole training data, the results are catastrophic; training with 1/1024 of the data, the output text has

nothing to do with the input text – one may think that they are two different texts. (Koehn & Knowles 2017: pp. 30-31.)

Long sentences are difficult for NMT systems, and the quality of the translations worsens remarkably when translating sentences with at least 60 words. A possible explanation for this is that NMT systems tend to shorten the output sentences. Augmenting the system's beam size can help in finding better translations but the output quality does not really improve. This means that SMT beats NMT also in these two challenges. (Koehn & Knowles 2017: pp. 33-37.)

Thanks to the attention mechanism, NMT systems can, for example, recognise the subjects and objects of verbs which improves the translation quality in an ideal case. Koehn and Knowles reached high results for both MT systems in their experiment, but they found that the results could have been better, because there was mismatch between the alignments, that the attention mechanism made, and the desired results. (Koehn & Knowles 2017: p.34-35.)

Bentivogli, Bisazza, Cettolo & Federico (2016: pp. 257-259) have also compared SMT and NMT output. They translated TEDtalks of the IWSLT 2015 corpus from English into German with an NMT system, a standard phrase-based SMT system, a hierarchical SMT system, and a hybrid SMT system (phrase-based and syntax-based).

The translated texts were analysed linguistically, and the researchers paid attention to the following categories: morphology, lexical errors, as well as word order because German is a language with inflected words and its word order is also different in main clauses and sub-clauses. (Bentivogli et al. 2016: pp. 257-258, 261).

Overall, NMT output was found to be qualitatively better than SMT output – no matter how long the input sentences are even though SMT systems cope better with sentences longer than 35 words. Lexical diversity in texts is easier to translate with an NMT system than with an SMT system. (Bentivogli et al. 2016: pp. 260-261.)

On morphological level (wrong word forms), the NMT system performed very well – it made even 19 % less errors than the SMT system. The NMT system also chose mostly the right words, and on closer examination, it turns out that its error rate on this level was 17 % less than the error rate of the SMT system. Word order is often difficult for machines but the NMT system used in Bentivogli et al,'s (2016: pp. 262-263) experiments found in most cases the

right places for the German words. Synonyms or morphological variants were not marked as word order errors in this case. (Bentivogli et al. 2016: pp. 262-263.)

Verbs, for instance, are often in wrong places in a sentence if a text has been translated with an SMT system – even the syntax-based systems have problems with this kind of words. An error category, where NMT performs worse, are nouns. Especially in deciding which noun is the object and which one the subject is difficult. Other problematic categories are prepositions, negation, and articles. Often the output is easy to read and understand but phrases with prepositions as one part (e.g. temporal phrases) can be in wrong positions. (Bentivogli et al. 2016: pp. 263-264.)

2.2 Automatic evaluation

2.2.1 BLEU

Automatic evaluation metrics were invented in order to reduce costs and save time – a human evaluation is more expensive and takes more time. Nevertheless, automatic evaluation does not work completely without people: Just like other evaluation metrics, BLEU compares machine translations to human translations; the calculated BLEU score shows how close an MT output is to a human translation. (Papineni, Roukos, Ward & Zhu 2002: p. 311.) A brevity penalty factor gives the most points to the translation which has the same length and same words in the same order as the reference sentence. (Papineni, Roukos, Ward & Zhu 2002: p. 315.)

The corpus level BLEU score is calculated as follows (Papineni et al. 2002: p. 315):

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

Equation 3: *Equation for calculating the BLEU score for an MT output. BP represents the brevity penalty. (Papineni et al. 2002: p. 315.)*

The formula calculates the geometric mean of sentence BLEU scores. BLEU sentence score measures the number of shared n-grams in source and target. The brevity penalty lowers the score if target is shorter than source. (Papineni et al. 2002: p. 315.)

Thus, the human translations serve as reference translations with high quality because they are written by professional translators. In the evaluation process, BLEU weights the sentences and marks n-grams in the translations to be evaluated and calculates an average score for all matches by comparing the MT n-grams to the n-grams in the reference translation. The final BLEU score is then calculated by dividing the number of matching n-grams by the number of words in the MT. The more matches a sentence contain, the merrier the translation is. (Papineni et al. 2002: pp. 311-312.) All scores are between 0 and 1, where 0 indicates a wrong and 1 correct translation (Callison-Burch, Osborne & Koehn 2006: p. 2).

The calculating approach above is used to evaluate single sentences. Whole text blocks can be evaluated by using a modified n-gram precision: First, matches in single sentences are marked with n-grams, like in the normal precision measure. (Papineni et al. 2002: p. 313.) Recurring words are calculated as one single word. (Callison-Burch, Osborne & Koehn 2006: p. 2). Then, the n-grams in the particular sentences are count and added up together. Finally, the total number of MT n-grams is divided by the number of n-grams in the reference translation corpus. The result of this calculation is then the modified precision score for a text block. (Papineni et al. 2002: p. 313.)

Empirically, BLEU can distinguish good translations from bad ones quite well and the evaluation results correlate with the results of human evaluation (Papineni et al. 2002: p. 318). There are also critics who say that BLEU allows too much variation. This means that BLEU cannot well recognize synonyms or other lexical variants for which reason the BLEU score may increase while translation quality does not improve from a human perspective. (Callison-Burch, Osborne & Koehn 2006: pp. 3-5.)

Because BLEU has its weaknesses, it is better not to rely only on one single evaluation metric, the LeBLEU metric (introduced in the next section) and human evaluation to be sure that the translation quality really improved.

2.2.2 LeBLEU

Another evaluation metric is LeBLEU, an enhanced version of BLEU which performs better in evaluating morphologically rich languages while remaining language independent. Unlike BLEU, LeBLEU matches not only words or n-grams which are exactly the same as in the

reference translations, but also “fuzzy” n-grams by using a calculated Levenshtein letter edit distance. Thanks to the ability to match fuzzy n-grams, evaluating translations of inflected and compound words is easier. (Virpioja & Grönroos 2015: p. 412.)

This is specifically helpful when translating into languages such as French where the parts of a compound word are often flipped over: For example, *letter box* means *boîte aux lettres* in French. This example shows that the words letter and box have changed places in the translation (and there is an article between the word pieces).

LeBLEU does not give 0 points for the translation in the letter box example like BLEU would do. German is also a language with compounds which are difficult to translate and evaluate automatically and LeBLEU seems to be a better evaluation metric for this language than BLEU. (Virpioja & Grönroos 2015: pp. 414-415.)

That LeBLEU does not declare this example translation as incorrect results also from the fact, that its brevity penalty counts the number of characters instead of the number of tokens in the MT output and the reference translations (Virpioa & Grönroos 2015: pp. 412-413). LeBLEU calculates word and n-gram order arithmetically while BLEU calculates it geometrically (Virpioja & Grönroos 2015: p. 413).

The evaluation of MT output with LeBLEU consists of four phases: first it creates a collection of all hypothetic n-gram alternatives for a translation and their frequencies in the data. Then it compares the collected n-grams with the ones in the reference translations and calculates the letter edit score that is then collected and normalized after which these scores are added up separately for each possible n-gram order. Finally, the perfect n-gram orders must be calculated and multiplied by the brevity penalty of the metric. (Virpioja & Grönroos 2015: p. 413.)

In this sense, BLEU and LeBLEU calculate their scores quite equally – the difference is that BLEU does calculate letter edit scores for n-grams. Letter edit scores – or, in other words, Levenshtein distances – can be calculated with a particular extension module for the Python programming language – although, the extension is written in C. All n-grams are divided in groups or rather string lists: hypothetic translation n-grams and reference n-grams. Each n-gram of a list is compared with the n-grams of the second list and, if a match is found, the system calculates the length between the positions of the matched n-grams. The difference in

length between these two strings is the final Levenshtein score. (Virpioja & Grönroos 2015: p. 413.)

If all hypothetic n-grams are compared to all reference n-grams, calculating the LeBLEU score takes a lot of time. A more time-consuming resolution is to build subsets of hypothetic n-grams and calculate the Levenshtein score only for this subset. (Virpioja & Grönroos 2015: p. 413.)

2.3 Domain adaptation for NMT

So far domain adaptation has been researched more for SMT systems than NMT systems. That is probably because NMT is rather a new invention. In some papers, researchers introduce different ways to obtain synthetic data for domain adaptation which helps especially in rare languages where there is not enough parallel data available. (Chinea-Ríos, Peris & Casacuberta 2017: pp. 138-140; Niu, Denkowski & Carpuat 2018: pp. 1-2.)

Besides research papers on domain adaptation in which researchers present their domain adaptation methods for NMT, there is also a paper in which researchers give an overview of all domain adaptation methods known so far. Chu and Wang (2018: pp. 4-8) classify all methods into two main groups: data centric methods and model centric methods. Data centric means, in this context, that one generates or selects sentences or sequences of in-domain data using existing out-of-domain or synthetic monolingual corpora for this procedure. Model centric, on the contrary, means methods like fine-tuning existing corpora, domain controlling/discriminating, or joining two NMT models or in multi-domain corpora. The most existing domain adaptation methods have already been utilised in studies on domain adaptation in SMT and have then been applied on NMT systems, but only data centric methods are directly applicable on NMT systems. (Chu & Wang 2018: pp. 4-8.)

Chinea-Ríos, Peris and Casacuberta (2017) have achieved good results by developing a method for selecting suitable source sentences for domain adaptation from a large monolingual sentence pool. The selected sentences are then translated and used as synthetic data or for fine-tuning of existing parallel corpora in a specific language pair. They select the sentences by calculating and comparing continuous vector spaces of corpora – texts in the same domain have a similar vector space. In this vector space is a calculated centroid around which the adequate sentences are collected. (Chinea-Ríos et al. 2017.)

Chinea-Ríos et al. tested their method by training the Keras NMT system with news texts for the language pair English-Spanish. As special domains they chose technology (i.e. Xerox printer manuals), IT and e-commerce. In addition, they used also the open-source SMT system Moses[1] to compare NMT with SMT. It turned out that Keras' translation quality is worse than Moses' translation quality if one does not use synthetic data or fine-tune existing parallel corpora. In this respect the vector space method works well and using it for NMT improves translation quality remarkably. (Chinea-Ríos et al. 2013: pp. 139-143.)

Niu, Denkowski and Carpuat (2018: pp. 1-2), on the other hand, trained their NMT system with multiple languages to get better translations, adapted the system to a domain and reduce training costs. The languages pairs used in their experiments are German-English-German, Tagalog-English-Tagalog and Swahili-English-Swahili. Their training data consisted of news texts and weblog texts. The NMT system consisted of an attentional RNN encoder-decoder with only one LSTM layer. (Niu et al. 2018: pp. 1-3.)

Niu et al. (2018) found out in their experiment that back-translated synthetic parallel data could not improve the translation quality enough when translating with a uni-directional NMT system in low resource languages. The results were better if one used a bi-directional NMT system, which was trained with the same synthetic data as for training the uni-directional system, was used. In German-English the results were better if the synthetic data was added to the source side of a parallel corpus. Furthermore, using a bi-directional system reduced the training time by 15-30 %. (Niu et al. 2018: pp. 3-5.)

Chu, Dabre & Kurohashi (2017) state that the translation quality of vanilla NMT systems is worse than the translation quality of SMT systems. They tried to improve the quality by combining fine-tuned corpora with multi-domain corpora in the training phase. This procedure is called mixed fine tuning. First, they used an out-of-domain corpus for training their NMT system (KyotoNMT) and, second, they mixed the out-of-domain corpus (resource-rich) with several in-domain corpora (resource-poor) after which they continued training the system with this new mixed and fine-tuned corpus. The different domains were marked with artificial tags (<2domain>), so that the NMT system recognized them during the translation. Similar tags are also used to pay attention to politeness in the NMT output. (Chu et al. 2017: p. 1.)

[1] See Koehn et al. 2007 for more details.

In their experiments, Chu et al (2017) translated between the languages Chinese-English and Chinese-Japanese. Their poor-quality corpus consisted of spoken language, and it is fine-tuned with patent texts (NTCIR-CE, IWSLT-CE'15). For Japanese-Chinese, they used scientific texts to improve a Wikipedia corpus (ASPEC-CJ, WAT, Wiki-CJ). Chu et al. found out that NMT without domain adaptation is indeed worse than SMT (Moses) in translating resource-poor domains. If translating resource-rich domains, NMT is better than SMT. Training the system with mixed and fine-tuned corpora improves the quality in resource-poor domains so much that the NMT system is better than the SMT system. (Chu et al 2017.)

In all the researches previously mentioned in this chapter, the NMT systems were trained during the research process which takes a lot of time. A more time-saving method is to use a baseline model, i.e. an NMT system which is already pretrained with out-of-domain data. Freitag and Al-Onaizan (2016) used this method in their research. For domain adaptation they chose the TEDtalks corpus (IWSLT'15), and their baseline model was trained with the WMT'15 corpus as well as DARPA BOLT news corpus. The languages pairs in this experiment were German-English and Chinese-English. (Freitag & Al-Onaizan 2016.)

The translation quality of the baseline model was unsatisfactory, so their goal was to improve it by training the system with a small in-domain-data corpus. To avoid overfitting and quality worsening in general domains, Freitag and Al-Onaizan combined the further trained model with the original baseline model. They trained the model over 20 epochs until it overfitted. The researchers found out that one does not need so many epochs of training to achieve BLEU results: the translation quality improved after two epochs in German-English translations and after six epochs in Chinese-English translations. They achieved the best BLEU results with the combined NMT model. (Freitag & Al-Onaizan 2016: pp. 3-8.)

3 Material and research method

This chapter introduces the NMT system which is used in the experiments, as well as the training and translation material. The purpose is to use freely available and downloadable material and tools wherever applicable.

3.1 Neural machine translation system

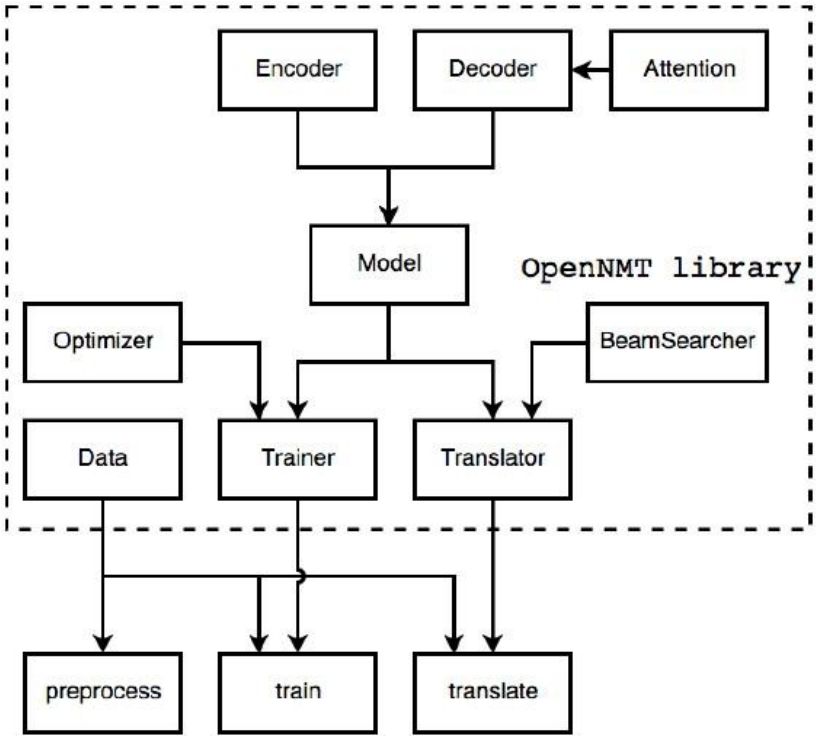

Figure 3: *Graphical representation of the OpenNMT-py library (Klein, Kim, Deng; Nguyen, Senellart & Rush 2018: p. 4).*

In this research I use the pretrained pyTorch version of the OpenNMT system with vanilla NMT models for English and German. This pure neural machine translation system was originally developed from a sequence-to-sequence MT system to which the developers added an attention mechanism as well as other functions like a speech recognizer and a parser. The pyTorch system, which Adam Lerer developed from the LuaTorch version, has been used in Facebook's AI Research project. (Klein et al. 2018: pp. 1-3; Klein et al. 2017: pp. 67-68.)

Before OpenNMT pyTorch had been developed, NMT researchers used a system called Nematus built by researchers of the University of Edinburgh, which was quite good, but it had also its weaknesses. The OpenNMT research team decided then to build a new system with

the strengths of Nematus and provide an open-source system for other users (e.g. the NLP community). (Klein et al. 2018: pp. 1-3; Klein et al. 2017: pp. 67-68.)

The three goals of developing OpenNMT-py were fast and efficient training, user-friendliness, that is good readable source code and modules which are maintained frequently), and the possibility to extend the system for research purposes. The system itself (as of 15 March 2018) consists, among other things, of training and deploying libraries, RNNs, beam search, and an LSTM which enables the MT system to remember what it translated and can translate sentences by means of the context. (Klein et al. 2018: pp. 1-3; Klein et al. 2017: p. 68.)

The pyTorch system has been pretrained with the WMT'15 dataset which contains 4.5 million sentence pairs (ca. 50.000 most frequent words for both languages). WMT'15 contains amongst others news texts (Newstest 2012, news commentary set) and texts from the European parliament. (OPENNMT; NLPL.)

In the current version (as of April 2019) (OPENNMT) of OpenNMT it is possible to use the Transformer model with self-attention and no LSTM (OPENNMT d) or convolution (Vaswani et al. 2017: p. 2).

There is also a component called positional encoding in the encoder and the decoder. This component remembers the positions of tokens in the sequence to be translated. In the Transformer, the token's positions are computed with sine and cosine functions. (Vaswani et al. 2017: p. 5-6.)

3.2 Settings

A neural network needs hyperparameters to be able to learn efficiently. Hyperparameters are parameters which the neural network user must set before the training process can begin. Tuning hyperparameters may affect the learning results. (Claesen & De Moor 2015: p. 1.) Using an NMT systems with wrong hyperparameter values may result in overfitting (Goodfellow et al. 2016: pp. 118-119).

In the training phase a neural network learns from the input data. If the neural network is used after training to, for example, translate texts it gets new input data with unseen information. The network should be able to link the new information to the old information so that it can translate also unknown words based on all what it has learned. If the system translates

unknown input sequences well the NMT system is generalized and the test error (or generalization error) is small. It is also possible that the system makes learning errors in the training phase and links between new and old information wrong. For this reason, a training error must be calculated on the test set. If the difference between the training error and the test error is too large, the neural network is overfitted. The opposite of overfitting is underfitting which means that the training error is too big. (Goodfellow et al. 2016: pp. 108-110.)

Overfitting can be prevented using a method called dropout. With this method randomly selected units can be eliminated from the network so that all its links to other units will also be disconnected. The units are selected with stochastic methods. After doing this, the neural network should be generalized, and it includes more noise than before dropout. (Srivastava, Hinton, Krizhevsky, Sutskever & Salakhutdinov 2014: pp. 1929-1933.)

Basic training	**continuation training**	**translation**
en-de pretrained	-	baseline
en-de pretrained	-	UNK terms
en-de baseline	-	baseline
en-de baseline	-	UNK terms
en-de baseline	-	test
EMEA-retrained	EMEA train	baseline
EMEA-retrained	EMEA train	UNK terms
EMEA-retrained	EMEA test	test
UNK-retrained	UNK terms	baseline
UNK-retrained	UNK terms	UNK terms

Table 1: *Summary of all NMT models, training data, and translations used in this thesis.*

I started out in my own experiments from the pretrained en-de translation model provided with the OpenNMT package. Some of the parameters and hyperparameters mentioned in its documentation are presented below:

For the en-de pretrained system, I used the standard settings of the older version OpenNMT pyTorch: Both the encoder and the decoder have two RNN layers (-enc_layer[2], -dec_layer[2]), and also the LSTM layer uses RNN gates. There are 500 hidden states at the encoder side as well as the decoder side ((-enc_rnn_size[500], -dec_rnn_size[500]). The decoder gets by every input step one context vector. The size of word embeddings for the encoder and decoder is 500 (-src_word_vec_size[500], -tgt_word_vec_size[500]). In order

that a machine can read and translate words, they must be converted into vector format. These word vectors are called word embeddings and they include the syntactic and semantic information which is associated to the words. (Almeida & Xexéo 2019: p. 3; Turian, Ratinov & Bengio 2010: p. 386.)

The Bahdanau (MLP) attention mechanism uses softmax to calculate which translation is the most probable one (-global_attention_function[softmax], -generator_function[softmax]). (OPENMT c.)

The pretrained model (as of 15 March 2018) had already been trained for over 200.000 steps. The system's learning rate during training and translating had been set to 1.0 (-learning_rate[1.0]). (OPENNMT c.)

The sizes of word embeddings for both the source and the target side (--src_word_vec_size, -src_word_vec_size and –tgt_word_vec_size, -tgt_word_vec_size) is also in the newest version set to 500 by default (OPENNMT c). The attention mechanism used in this model is the scaled-dot self-attention as described in Vaswani et al. (2017: p. 4). The number of hidden feed-forward network layers (--transformer_ff, -transformer_ff) is 2048 (OPENNMT c). The maximum length for source and target sequences (--src_seq_length, -src_seq_length and –tgt_seq_length, -tgt_seq_length) is 50 words (OPENNMT e).

In this thesis the tokenizer used is the sentencepiece model. The sentencepiece model works unsupervised (Domingo, García-Martínez, Helle, Casacuberta & Herranz 2018: p. 3). In general, this tokenizer is language independent, but it must be trained separately for all languages which are used for translation (Domingo et al. 2018: p. 3).

3.3 Translation memories

The purpose of translation memories is to facilitate and speed up human translation by offering complete translations of sentences or sequences which have previously been translated. For this reason, translation memories are most effective if one uses them for translating the same text types which contain the same sentences in form and content. (Koskenniemi 2013: p. 140.) Such translation databases and TM systems have been since their invention the most important translator's software which allows users to work more productive and efficient (Garcia 2009: pp. 199-201).

Most modern translation memories work also well if one translates texts which are in a different tense than previously translated texts. In general, translation memory systems cannot analyse languages. (Koskenniemi 2013: p. 140.) Instead of a language analysis, translation memory systems have a component which compares the contents of the new text with its database entries and searches for similar sequences. In practice, the system finds similarities between texts by detecting identical characters or tokens. Then the system proposes all the possible translations for a certain segment which the translator can use directly or alternatively modify one of the suggestions so that it fits better into the text to be translated. (Reinke 2018: pp. 55, 74; Stein 2018: p. 8.)

The EMEA (European medicines agencies) parallel corpus in the typical translation memory format (tmx) serves in this thesis as training and test data for OpenNMT. The corpus for English-German contains 364.005 sentence pairs and 9.88 million words. The texts contained in this corpus are for example package inserts and EU's decisions in medical law. (OPUS.) Medicine is a domain for this thesis, because such texts are easy to find and there are many texts which can be used for translating. Because of patient's health and secure it is also important that the translation quality of medical texts is good enough, so that everyone can use their medicaments in the right way.

As mentioned in chapter 2.1, MT systems translate texts segment by segment. Because the content in TMs is also segmented, it appears logical to use TMs for training an NMT system. Segments are easier to translate which results in better translation quality and less data loss.

3.4 Research method

The first step in my research was to install the OpenNMT pyTorch package (NLP-WIKI.) After this I installed the English-German version of OpenNMT-py.

The EMEA corpus was split into ten parts as in k-fold cross-validation (k = 10), so that different texts are used for training and translation. As per Bengio and Grandvalet (2004: pp. 1092-1093) k-fold cross-validation is a statistical function for dividing data into several parts or so-called subsets. All subsets have the same size. (Bengio & Grandvalet 2004: pp. 1092-1093.) The endpoint of a subset and the beginning point of the next subset are selected randomly (Cawley & Talbot 2010: p. 2082). A subset can be selected randomly by training the whole data k (for example k = 10) times on k-1 folds and, finally, the subset with the

smallest validation loss is chosen for further use (Ding, Tarokh & Yang 2018: p. 8-9). Due to time constraints, only the first two pieces were used for training and testing, respectively.

After this, I tested the setup by translating the EMEA data with the pretrained en-de system. The original idea was to continue training with the EMEA data, but this is not possible. Instead, a new baseline model was trained from scratch using the same training data which has also been used for training the pretrained model.

The new en-de baseline model was then trained using two different methods: in the first experiment the model was trained with EMEA data, and in the second experiment it was trained with unknown terms which were extracted of the EMEA corpus. The purpose of doing this is to adapt the system on a special domain like medicine in this case. The next step was then to translate medical texts with the adapted system.

The translated text was evaluated automatically with BLEU and LeBLEU. As mentioned in chapter 2.2 it is necessary to use additional ways to evaluate machine translation output, because BLEU is not reliable enough, so I evaluated the translations also with LeBLEU. In addition, I evaluated the translations myself. In this way it was possible to get more reliable results about the true translation quality: the content of translated sentences is as important as the fluency of the translation – if not even more important. In conclusion I compared my results with the results of the en-de baseline translation before domain adaptation.

According to Chu and Wang (2018: pp. 4-8; see chapter 2.3), my domain adaptation method belongs to the group with data centric methods because I trained a pretrained NMT system with an in-domain translation memory to adapt the system on a domain (in my case the domain is medicine). The translation memory can in my opinion be considered to be synthetic data because the tmx file consists of sequences which are gathered from other texts (for more information, see chapter 3.3).

4 Analysis

In this chapter I present my research results: Firstly, I describe the training and domain adaptation processes. Secondly, present the results of automatic evaluation. More detailed results with a presentation of some translation errors are presented after this.

4.1 Training

The original plan was to train the OpenNMT pretrained en-de model with my EMEA data. It turned out that the pretrained model did not permit continuation training at all (this was documented, but it came up on close reading of the OpenNMT community discussion threads).

In OpenNMT it is not generally workable to continue training a baseline model with new texts because each system uses a fixed vocabulary which cannot be extended during training. This means that the machine is not able to learn new vocabulary and it uses even after domain adaptation / training the old vocabulary which does not improve the quality of domain specific translations. See for example the following thread (see e.g. GITHUB) for more information about the vocabulary problem.

One possible solution to this problem is to retrain the whole system from scratch with the desired corpora like EMEA in this case. After continuous training (i.e. domain adaptation), the system learns the new vocabulary but also the old vocabulary if the system is trained with old and new corpora.

In this way, the translation after domain adaptation should be better than the first one because OpenNMT has learned new, domain-specific vocabulary. However, there is a risk that OpenNMT focuses in the learning process too much on the new vocabulary which worsens the translation quality if more general texts are translated.

A second way to the domain adaptation problem is to extract all term of the text to be translated which does not occur in the corpora of the pretrained model and to train the system with this wordlist. This method is also a bit risky because the system can overfit if it focuses too much on single terms without a context.

Another approach is to split words into sub-word-tokens. A small token vocabulary may cover a larger full form vocabulary. To test this, the sentencepiece tokenizer was on the

EMEA training set and compared the token vocabulary to that of the baseline system. All but one token of the EMEA set was included in the baseline sentencepiece vocabulary.

In this thesis, both domain adaptation methods are tested, and the results are then compared with each other.

4.2 Evaluation

Text	Model	BLEU	LeBLEU
EMEA test	EMEA retrained	0.0.51902	0.678062
EMEA train	UNK retrained	3.7E-05	0.000371
EMEA train	EMEA retrained	0.721412	0.791316
EMEA train	retrained	0.245369	0.572453
EMEA test	retrained	0.246455	0.573123
UNK	EMEA retrained	0.058277	0.152967
UNK	UNK retrained	0.055628	0.55526
UNK	retrained	0.041062	0.072695
UNK	pretrained	0.049933	0.089344
EMEA train	pretrained	0.274706	0.622514

Table 2: *Overview of the evaluation results on system level.*

Supposing that LeBLEU scores are between 0 and 1, the score for the en-de pretrained translation (see Table 2) is not too bad. The en-de baseline model (retrained), on turn, performed a bit worse than the en-de pretrained model, and the translation quality improved clearly after domain adaptation with EMEA texts (EMEA retrained).

The large drop in scores from the training to the test set may mean that the EMEA retrained model was overfitted. Nevertheless, domain adaptation with the EMEA text improved the translation quality.

As it turned out in more detailed analyses se chapter 4.3 and 4.4, the NMT systems has the most difficulties with long sentences and unknown words – the results are, thus, similar than those in Koehn's and Knowles' experiments (see chapter 2.1.3). Long sentences were sometimes (not always) shortened by leaving out parts of the content which results in incomplete sentences. Some out-of-domain terms, like *Beschlagnahmungen*, occurred in the translations before domain adaptation. Domain adaptation reduced the number of such terms

clearly. On the other hand, the trained models had surprisingly few problems with inflected verbs but a bit more problems with product names, like Abilify and Abseamed.

It seems that especially the unknown words were translated in smaller pieces which sometimes results in partly wrong translations. The beginning of the translation, for instance, can be correct but the ending is wrong. Extracting unknown terms of the EMEA corpus and training the system with them worsened the translation quality radically – this method works only if translating only the unknown terms. Table 3 establishes this hypothesis.

Table 3 shows the difference between BLEU and LeBLEU scores if translating extracted unknown terms with the en-de pretrained model and comparing it to the reference translation. The quality of the reference translations (statistical machine translation) is often bad. The last column of the table shows separated BLEU and LeBLEU scores. Because LeBLEU allows more variation than BLEU (see chapters 2.2.1 and 2.2.2 for more details), the LeBLEU score is higher. It means that the translation is partially right.

source phrase	**reference (SMT)**	**pretrained**	**BLEU-LeBLEU**	**text**	**system**	**BLEU**	**LeBLEU**
Priapism	Priapismus	Primatismus	-0.636364	UNK-terms	en-de pretrained	0.000000	0.636364
newborn.	Säugling	neugeboren.	-0.636364	UNK-terms	en-de pretrained	0.000000	0.636364
macroscopic	makroskopischen	Makroskopie	-0.636364	UNK-term	en-de pretrained	0.000000	0.636364
medicine).	eingestellt	Medizin).	-0.626388	UNK-terms	en-de pretrained	0.000000	0.626388
medicine),	Arzneimittel	Medizin),	-0.626388	UNK-term	en-de pretrained	0.000000	0.626388
end-users.	Endbenutzer	End-User	-0.626388	UNK-terms	en-de pretrained	0.000000	0.626388
Undergoing	Dialysepflichtige	Untergang	-0.626388	UNK-terms	en-de pretrained	0.000000	0.626388
Povidone	Polyvidone	Populone	-0.625000	UNK-terms	en-de pretrained	0.000000	0.625000

leukemia	leukemia	Leukämie	-0.625000	UNK-terms	en-de pretrained	0.000000	0.625000
Erectile	Erektiler	Erektion	-0.625000	UNK-terms	en-de pretrained	0.000000	0.625000
Brytania	Brytania,	Bryanien	-0.625000	UNK-terms	en-de pretrained	0.000000	0.625000
albicans	albicans-Infektion	Albikane	-0.625000	UNK-terms	en-de pretrained	0.000000	0.625000

Table 3: *Difference between BLEU and LeBLEU scores on translated UNK terms.*

4.3 Baseline translations

4.3.1 En-de pretrained

The first text is translated with the pretrained version of OpenNMT-py. The settings are as described in chapter 3.2.

Basically, the translations are fluent, but the content does often not make sense. Some translations are, on the other hand, quite good:

(1) **en:** The needle should be kept under the skin for at least 6 seconds to make sure the entire dose is injected.

de: Die Nadel sollte mindestens 6 Sekunden unter der Haut gehalten werden, um sicherzustellen, dass die gesamte Dosis injiziert wird.

(2) **en:** The blood glucose lowering effect of insulin is due to the facilitated uptake of glucose following binding of insulin to receptors on muscle and fat cells and to the simultaneous inhibition of glucose output from the liver.

de: Die Blutzuckersenkende Wirkung des Insulins ist auf die vereinfachte Aufnahme von Glukose nach der Bindung von Insulin an Rezeptoren an Muskel- und Fettzellen und auf die gleichzeitige Hemmung der Glukoseproduktion aus der Leber zurückzuführen.

(3) **en:** The daily insulin requirement may be higher in patients with insulin resistance (e.g. during puberty or due to obesity) and lower in patients with residual, endogenous insulin production.

de: Bei Patienten mit Insulinresistenz (z.B. während der Pubertät oder aufgrund von Fettleibigkeit) kann der tägliche Insulinbedarf höher und bei Patienten mit Restinsulinresistenz niedriger sein.

Especially long sentences and sentences with many terms seem to be difficult to translate – many of such sentences are often unfinished like in example (4):

(4) **en:** It is available as 5 mg, 10 mg, 15 mg, and 30 mg tablets, as 10 mg, 15 mg and 30 mg orodispersible tablets (tablets that dissolve in the mouth), as an oral solution (1 mg/ ml) and as asolution for infection (7.5 mg/ ml).

de: Es ist als 5 mg, 10 mg, 15 mg und 30 mg Tabletten, als 10 mg, 15 mg und 30 mg orodispersible Tabletten (Tabletten, die sich im Mund auflösen), als orale Lösung (1 mg/ ml) und als Lösung für die Injektion (7,5 mg/ ml).

After the list of different types of the drug “Abilify”, a word like for example “erhältlich” (en. available) is missing.

(5) **en:** Hypersensitivity to the active substance or to any of the excipients.

de: Hypersensibilität gegenüber dem aktiven Stoff oder den Ausführenden.

(6) **en:** An additional finding was cholelithiasis as a consequence of precipitation of sulphate conjugates of hydroxy metabolites of aripiprazole in the bile of monkeys after repeated oral dosing at 25 to 125 mg/ kg/ day (1 to 3 times the mean steady-state AUC at the maximum recommended clinical dose or 16 to 81 times the maximum recommended human dose based on mg/ m2).

de: Ein weiteres Ergebnis war die Choleraelilyse als Folge der Niederschlagung von Sulfatkonjugaten von hydroxyen Metaboliten von aripiprazol in der Galle der Affen nach wiederholter oraler Dosierung bei 25 bis 125 mg/kg/ Tag (1 bis 3 Mal so hoch wie die durchschnittliche konstante AUC bei der maximal empfohlenen klinischen Dosis oder 16 bis 81 Mal so hoch wie die empfohlene menschliche

(7) **en:** Abseamed can be used to reduce exposure to allogeneic blood transfusions in adult non-iron deficient patients prior to major elective orthopaedic surgery, having a high perceived risk for transfusion complications.

de: Abseamed kann verwendet werden, um die Exposition gegenüber allogenen Bluttransfusionen bei nicht-eisernen Patienten, die vor einer wichtigen elektrisch-orthopädischen Chirurgie leiden, zu verringern, die ein hohes Risiko für Transfusionskomplikationen aufweist.

In example (5) it is not clearly identifiable which word is missing. Maybe the drug includes another active substance? The same can be said about example (6) which is also a too complex sentence and therefore difficult to read. Example (7), in turn, has some bizarre terms: *nicht-eiserne Patienten* (literal translation: patients who are not of iron, but in the source text it means people who do not suffer from iron deficient). A better translation would be […] bei Patienten, die nicht an Eisenmagel (or Eisenmangelanämie) leiden […]. The second bizarre term is *elektrisch-orthopädische Chirurgie* (literally translation: electric-orthopaedic surgery). *Orthopaedic surgery* (de. orthopädische Chirurgie) is indeed an existing term and it can be done with electrical tools but an *electric-orthopaedic surgery* does not seem to exist.

Another interesting aspect is that the same phrase has different translations:

(8) **en:** oral use

de: Mündliche Nutzung / Mündliche Verwendung.

Furthermore, the phrases in example (8) are titles or subtitles for which reason it should not end with a full stop like the second phrase.

The pretrained OpenNMT-py system copies sometimes the English word into the translation if it does not know how to translate it and it also cannot guess it:

(9) **en:** When seen, it is usually in those with significant risk factors such as history of diabetes, thyroid disorder or pituitary adenoma.

de: Wenn man sieht, dann in der Regel in solchen Fällen mit signifikanten Risikofaktoren wie der Geschichte von Diabetes, Schilddrüsenstörung oder pituitary adenoma.

In example (9) the problem is the term *pituitary adenoma* which means *Hypophysenadenom* (see e.g. WIKIPEDIA) in German. The question is of a disease.

Example (10) shows a sentence with a list of terms (diseases, in this case):

(10) **en:** Cardiac disorders Uncommon: tachycardia* Nervous System disorders Common: extrapyramidal disorder, akathisia, tremor, dizziness, somnolence, sedation, headache Eye disorders Common: blurred vision Gastrointestinal disorders Common: dyspepsia, vomiting, nausea, constipation, salivary hypersecretion

de: Kardinale Erkrankungen Ungewöhnliches: Herzinfarkt* Nervous System Störungen Common: extrapyramidal disorder, akathisia, tremor, dizziness, somnolence, sedation, Kysepsie, Erbrechen, Überlkeit, Verzückung, Hypersekreta

More common terms are translated well (or the system has guessed) but the rarer terms are again copied from English. In addition, the beginning of the sentence is confusing, and the comma is missing after *Herzinfarkt*. In addition, the translation does not include all pieces of the source text; symptomps like *eye disorders*, headache and the term *gastrointestinal disorders* are missing. The missing comma in example (10) is not a singular case:

(11) **en:** Lactose monohydrate Maize starch Microcrystalline cellulose Hydroxypropyl cellulose Magnesium stearate

de: Lactose Monohydrat Maize Stärke Microcrystalline Zellulose Hydroxypropyl Zellulose Magnesium Stearate

This (see example (11)) does not sound as a long name of a chemical compound – it is rather a list of ingredients of a medicinal product. The same problem occurs in the source sentence, so possibly the printed version includes a list with bullet points which are missing in the corpus.

(12) **en:** Therefore, aripiprazole should be used with caution in patients who have a history of seizure disorder or have conditions associated with seizures.

de: Daher sollte aripiprazole mit Vorsicht bei Patienten verwendet werden, die eine Geschichte von Beschlagnahmungsstörungen haben oder mit Anfällen in Zusammenhang stehen.

Example (12) sounds so that this drug is not suitable for “criminals” (*Beschlagnahmung* means that something has been confiscated and *Anfälle* means seizures in English).

Sometimes the NMT system guesses terms but it guesses mostly wrong:

(13) **en:** Aripiprazole was excreted in the milk of treated rats during lactation.

de: Aripiprazole wurde während der Laktation in der Milch behandelter Ratten ausgegraben.

In example (13) the word *ausgegraben* is wrong because the sentence means literally that aripiprazole has been grubbed out of the in milk medicated rats.

It seems that especially the unknown words were translated in smaller pieces which sometimes results in partly wrong translations. The beginning of the translation, for instance, can be correct but the ending is wrong.

This hypothesis can be proved by extracting all terms from the source text and translating them. If one part is translated wrong, the whole term is mostly wrong because the NMT system does not know the term. This can also be the cause why some terms have different translations in the target text:

(14) **en:** Abseamed

de: Abseamed / Abseamierte / Absaamierte / Absaamed

(15) **en:** Pharmacokinetics

de: Pharmakokinetik / Pharmakkinetik / pharmakikokinetik / Apothekerkinetik

(16) **en:** 1-(6-Chloro-3-pyridylmethylen-N-nitro-imidazolidin-2-ylideneamine

de: 1-(6-Chloro-3-pyridylmethylen-N-nitro-imidazolidin-2-ylideneamine

(17) **en:** gonadotropin-containing / gonadotropin-releasing

de: Gonadotropin-Besatzung / Gonadotrophin-Freigabe / Gonadotrophin-Release

(18) **en:** bumpy

de: Die Lage des Hotels ist sehr gut.

(19) **en:** drugs

de: Drogen

Sometimes the translations are out-of-domain, like in example (18), or they use obsolete vocabulary (see example (19). In example (19), the problem is the term *drugs* which has a different meaning in English and German. In German drugs are rather substances like cannabis or cocaine. The word *Drogen* in a medical context is obsolete – unless e.g. cannabis is used as medication (DUDEN). But if it is about other, "normal", medicines it is better to use the term *Medikament(e)* to avoid misunderstandings. In English it means on the one hand the same as in German and on the other hand it means medicaments. The term *drugs* in a medical context does not seem to be obsolete in English. (CAMBRIDGE DICTIONARY.)

Another problem which occurs in the translation is that words and phrases, which are written in capital letters are mostly translated wrong:

(20) **en:** NAME OF THE MEDICINAL PRODUCT

de: NAME des MEDICINAL PRODUKT

(21) **en:** DATE OF THE AUTHORISATION/ RENEWAL OF THE AUTHORISATION

de: ERSTLICHE AUTHORISATION/ RENEWAL DER AUTHORISATION

(22) **en:** GENERAL CLASSIFICATION FOR SUPPLY

de: GENERKLASSIFIC

Such words and phrases are often copied from English or translated partially.

A third problem is that the NMT system repeats sometimes words:

(23) **en:** General

de: General General General General General General General General

The reason may be that the system does not know the word and it cannot guess the right word. Another reason may be that the most sentences in the source text are long, so the systems thinks that it has to translate suddenly occurring short phrases and single words (e.g. in titles) longer. Because there are no other words in the phrase demonstrated in example (23) the system repeated the same word several times.

4.3.2 En-de baseline (retrained)

The following example sentences, terms, and phrases are translated with the retrained (from scratch) en-de model.

Overall, the translation contains mostly the same translation errors as the previous one. The errors occur in both long sentences and short sentences, but terminological errors seem to be more frequent in long sentences. Long sentences with mostly frequent words are often translated well.

(24) **en:** It is available as 5 mg, 10 mg, 15 mg and 30 mg tablets, as 10 mg, 15 mg and 30 mg orodispersible tablets (tablets that dissolve in the mouth), as an oral solution (1 mg/ ml) and as a solution for injection (7.5 mg/ ml).

de: Es ist als 5 mg, 10 mg, 15 mg und 30 g Tablett erhältlich, als orale Lösung (1 mg/ 30 mg/l.

Unlike example (4), the translation in example (24) contains the word *erhältlich*. The problem in this sentence is that it is much shorter than the source sentence. The NMT system cut all information after the first part of the sentence (“It is available as 5 mg, 10 mg, 15 mg and 30 mg tablets”), then it began the next part (“als orale Lösung”) and replaced the rest with “1 mg/ 30 mg/l”. The result is an incomplete and incomprehensible sentence. There are more sentences with “hallucinated” content and wrong term translations:

(25) **en:** 1 ml contains 150 mg of gamithromycin 1 mg of monothioglycerol

de: 1 ml enthält 150 mg Glücksspiel 1 mg monothioglycerol

In example (25), the source sentence contains the term *gamithromycin* which is a name of an active substance (or another chemical substance) of a medicine. The NMT system confuses *gamithromycin* with gambling (de: *Glücksspiel*).

The en-de baseline model tends also to repeat words and phrases more often than the en-de pretrained model.

(26) **en:** Cardiac disorders Uncommon: tachycardia* Nervous System disorders Common: extrapyramidal disorder, akathisia, tremor, dizziness, somnolence, sedation, headache Eye disorders Common: blurred vision Gastrointestinal disorders Common: dyspepsia, vomiting, nausea, constipation, salivary hypersecretion

de: Herzkrankheiten: Störungen des Körpers, Störungen des Körpers, Störungen des Organismus, Störungen des Organismus, Störungen des Organismus und der Organe.

In example (26) the system was not at all able to translate the list of cardiac disorders, nervous system disorders, eye disorders and gastrointestinal disorders. This shows that terms are difficult to translate. The same problem occurs also in the translation of unknown terms:

(27) **en:** keeps

de: Für die Einhaltung der Einhaltung der Einhaltungs- und Einhaltungsvorschriften sorgt die Einhaltung der Einhaltung der Einhaltungs- und Einhaltungsforschriften.

(28) **en:** Antipsychotics/Neuroleptics

de: Psychiater/Neurolepismus

Antipsychotic is a medicine but “Psychiater” is psychiatrist.

(29) **en:** pharmacokinetic/pharmacodynamic

de: Pharmakovigilanz

(30) **en:** 21-gauge

de: 21-gaudi

En-de baseline copies some terms directly from Englisch but often it tries to translate them, like in example (28). Another example for translation tries is example (31) which occurs in Abilify's package leaflet:

(31) **en:** ABILIFY

de: Abilifie / Abilify / ABILIFolez / Absurdität

Sometimes also frequent words are copied from English. In example (32) the same word is translated wrong at the beginning and right at the end of the sentence:

(32) **en:** Pregnancy and breast-feeding You should not take ABILIFY if you are pregnant unless you have discussed this with your doctor.

de: Pregnancy und Stillen Sie sollten ABILIFY nicht nehmen, wenn sie schwanger sind, es seidenn, Sie haben mit Ihrem Arzt darüber diskutiert.

In addition, example (33) shows another translation problem: words and phrases which are written with capital letters. The same problem occurred also in the previous translation (en-de pretrained).

(33) **en:** CLINICAL PARTICULARS

de: CLINICAL PARTICULAR

The translation of the en-de baseline model contains also errors which the translation of the en-de pretrained model does not contain. The en-de baseline model tries to be polite when it translates – it uses, for example, formal language:

(34) **en:** Store in the original package in order to protect from moisture.

de: Lagern Sie im Originalpaket, um vor Feuchtigkeit zu schützen.

But sometimes it is all but not polite:

(35) **en:** In patients with severe hepatic impairment, the data available are insufficient to establish recommendations.

de: Bei Patienten mit schwerer hässlicher Behinderung reichen die verfügbaren Daten nicht aus, um Empfehlungen zu geben.

The problem in this sentence is the phrase *patients with severe hepatic impairment* which is translated as *Patienten mit schwerer hässlicher Behinderung* (literally in English: *patients with severe ugly (mental) handicap*). The word can indeed be translated with the word *beeinträchtigen* but in this context the translation is wrong because *hepatic impairment* is a liver disease and has nothing to do with handicaps. It is not right to insult other people (with or without (mental) handicap), so such translations should never be published without proofreading.

4.4 Translation after domain adaptation

4.4.1 EMEA retrained

This text is translated with the domain adapted EMEA retrained model. The translations are more fluent than the translations before domain adaptation (en-de pretrained and en-de baseline) and, all in all, the EMEA retrained system seems to translate terms better than the out-of-domain systems.

(36) **en:** Enhanced efficacy at doses higher than a daily dose of 15 mg has not been demonstrated although individual patients may benefit from a higher dose.

de: Eine gesteigerte Wirksamkeit bei Dosierungen über einer Tagesdosis von 15 mg wurde nicht nachgewiesen, obwohl einzelne Patienten von einer höreren Dosis profitieren können.

(37) **en:** The maximum daily dose should not exceed 30 mg.

de: Die maximale Tagesdosis sollte 30 mg nicht überschreiten.

(38) **en:** Store in the original package in order to protect from moisture.

de: In der Oroginalverpackung aufbewahren, um den Inhalt vor Feuchtigkeit zu schützen.

Even though the translations are better than before domain adaptation, not all problems are solved. The shortening problem, for example, is not completely solved:

(39) **en:** Aripiprazole exhibited high binding affinity in vitro for dopamine D2 and D3, serotonin 5HT1a and 5HT2a receptors and moderate affinity for dopamine D4, serotonin HT2c and 5HT7, alpha-1 adrenergic and histamine H1 receptors.

de: Aripiprazol zeigte in vitro eine hohe Bindungs Affinität zu Dopamin D2 und D3, Serotonin 5HT1a und 5HT2a-Rezeptoren und moderater

(40) **en:** As an increased incidence of thrombotic vascular events (TVEs) has been observed in cancer patients receiving erythropoiesis-stimulating agents (see section 4.8), this risk should be carefully weighed against the benefit to be derived from treatment (with epoetin alfa) particularly in cancer patients with an increased risk of thrombotic vascular events, such as obesity and patients with a prior history of TVEs (e.g. deep vein thrombosis or pulmonary embolism).

de: Da es in der Vorgeschichte zu einer erhöhten Inzidenz thrombovaskulärer Ereignisse (TVEs) bei Tumorpatienten unter Erythropoese-stimulierenden Ar

(41) **en:** Precise risk estimates for hyperglycaemia-related adverse events in patients treated with ABILIFY and with other atypical antipsychotic agents are not available to allow direct comparisons.

de: Patienten, die mit antipsychotischen Wirkstoffen einschließlich ABILIFY behandelt werden, sollten auf auf Anzeichen und Symptome einer Hyperglykämie (wie z.B.

The translation in example (41) is not only uncomplete but also the content is wrong. As per the source sentence, there are no enough risk estimates to be able to say how frequent hyperglycaemia-related adverse events are if patients use Abilify. But in the translation is said that patients should pay attention to the symptoms of hyperglycaemia.

(42) **en:** It is available as 5 mg, 10 mg, 15 mg and 30 mg tablets, as 10 mg, 15 mg and 30 mg orodispersible tablets (tablets that dissolve in the mouth), as an oral solution (1 mg/ ml) and as a solution for injection (7.5 mg/ ml).

de: Es ist als 5 mg, 10 mg, 15 mg und 30 mg Tabletten in Form von 10 mg, 15 mg und 30 mg Schmelztabletten (Tabletten, die sich in der Mund-Lösung auflösen) erhältlich.

The terms *Tabletten* (en: tablets) and *Schmelztabletten* (en: orodispersible tablets) are correct but the rest of the sentence ("as an oral solution […]") is missing. The phrase *Tabletten, die sich in der Mund-Lösung auflösen* is unnecessary complex. It is not necessary to mention that the tablet dissolves in the spittle (*Spucke* would be the correct translation and not *Mund-Lösung*) because everybody knows it. It would be better the phrase in question as follows: *Tabletten, die sich him Mund auflösen*. Another error in this translation is the use of "in Form von": as per the source sentence, Abilify is available as normal tablets, orodispersible tablets, oral solution and solution for injection. As per the translation, Abilify is only available as orodispersible tablets.

Sometimes the EMEA retrained model still has trouble translating medical terms:

(43) **en:** Actual scores in rating scales used as secondary endpoints, including PANSS and the Montgomery-Asberg Depression Rating Scale schowed a significant improvement over haloperidol.

de: Actuelle Scores der Bewertungsskala, die als sekundäre verwendet wurden, einschließlich PANSS und der Pangomerie-Asberg-Konzentration Scale, zeigten eine signifikante Verbesserung gegenüber Haloperidol.

Frequent words like *aktuell* (in the translation *Actuelle*) and *Skala* (in the translation, once, *Scale*) are translated incorrect and *Montgomery* has transformed into *Pangomerie*.

(44) **en:** Abseamed can be used to reduce exposure to allogenic blood transfusions in adult non-iron deficient patients prior to major elective orthopaedic surgery, having a high perceived risk for transfusion complications.

de: Abseamed kann zur Reduktion der Exposition gegenüber allogenen Bluttransfusionen bei erwachsenen Patienten mit nicht-ironmangel vor einem elektiven orthopädischen Eingriff mit einem hohen Risiko für Transfusionen verwendet werden.

The term *non-iron deficient* seems still to make difficulties (anyhow, it is now clear that *Eisenmangel* would be a good translation for this term), but the translation of *elective orthopaedic surgery* is better than before: "[…] vor einem elektiven (not *elektrischen*) orthopädischen Eingriff […]".

The translation of UNK terms indicates also that (long) medical and chemical terms are mostly difficult to translate:

(45) **en:** 1-(6Chloro-3-pyridylmethyl)-N-nitro-imidazolidin-2-ylideneamine

de: 1-

(46) **en:** 1,1,3,3,3-pentafluoro-2(fluoromethoxy)propene

de: 1,1,3,3,3-pentafluoro-2(fluoromethoxy)propene

(47) **en:** haloperidol-controlled

de: Haloperidol

(48) **en:** *Based

de: * Hauptprodukt

Long names of chemical compounds, for example, are copied from English but most of the terms are translated into German.

Phrases which are written in capital letters are often translated better than before (i.e. there are less phrases which are copied directly from English):

(49) **en:** CLINICAL PARTICULARS

de: KLINISCHE ANGABEN

(50) **en:** NAME OF THE MEDICINAL PRODUCT

de: BEZEICHNUNG DES ARZNEIMITTELS

Apparently, the system does not normalise case. When the training data has capitalised terms, they get translated, otherwise not.

4.4.2 UNK retrained

The UNK retrained model translates sentences with single words, so this method is not suitable for domain adaptation. After training with a term list, the result of translating a text is rather a word list than a whole text.

(51) **en:** Tardive Dyskinesia: in clinical trials of one year or less duration, there were uncommon reports of treatment emergent dyskinesia during treatment with aripiprazole.

de: Tardive Dyskinesie

(52) **en:** After treatment period of 7-10 days, the condition of the dog should be re-evaluated in order to establish the need for continuation of treatment.

de: nach

The reason for such short translations is possibly that the system learns words and terms without context which results in that it is no longer able to translate whole sentences.

The UNK retrained model seems to translate some of the terms, like tardive dyskinesia, right. The translation of the extracted unknown terms shows that the model tends to translate many terms into German:

(53) **en:** complementary-determining

de: Komplementarität-bestimmende / komplementaritätsbestimmende

(54) **en:** Pharmacokinetic

de: Pharmakokinetik

(55) **en:** 1-(6-Chloro-3-pyridylmethyl)-N-nitro-imidazolidin-2-ylideneamine

de: 1-(6-Chloro-3-pyridylmethyl)-N-nitro-imidazolidin-2-ylidenamin

(56) **en:** aminotransferase/aspartate

de: Aminotransferase-Anstieg

Sometimes the system copies terms partially from English, like in example (55). The partially copied and modified terms are mostly long and difficult to translate.

5 Conclusion and discussion

The aim of this master's thesis was to find out how an NMT system can be adapted to a special domain, and if the domain adaptation improves the quality of in-domain translations. In the theoretical part, the function of SMT systems and NMT systems was explained. The function of both systems bases upon calculated probabilities of possible translation hypotheses. The translation output is always the translation with the highest probability.

The focus was on NMT systems because such a system was used in the own experiments. There are many different types of neural networks – the most relevant network types for this thesis are the recurrent neural network and the Transformer model because OpenNMT, which is the NMT system used in the experiments, uses these network types.

A comparison of NMT and SMT systems showed that SMT systems can translate specialised texts better than NMT systems without domain adaptation; NMT systems, on turn, are better on morphological and grammatical level. Furthermore, long sentences are difficult to translate for machine translation systems.

So far, domain adaptation by using translation memories has only been researched for SMT systems. NMT system have been adapted into special domains using other methods like generating synthetic parallel corpora with source sentences and machine translated target sentences. In all previous researches, the translation quality improved after domain adaptation, so I hypothesised that the quality of my own translations also improves.

The original plan was to use the pretrained model of OpenNMT and continue to train it with in-domain data. Because this was not possible, the NMT system had to be retrained from scratch. After that the system could be adapted into a special domain by continuing the training with in-domain data. Another method was to extract unknown terms of the EMEA data and train the NMT system with them. The translation memory used for training was the EMEA corpus for English and German. The BLEU and LeBLEU metrics were used to evaluate the translations automatically.

The manual and automatic evaluation showed that the translation quality on in-domain texts of the EMEA retrained system was much better than the translation quality of the en-de pretrained model. The LeBLEU scores were mostly higher than the BLEU scores because LeBLEU is less strict than BLEU in evaluating words, phrases, and sentences which are partially right.

The most translation errors occurred in long sentences which were often shortened or incomplete translated. Medical and chemical terms were especially difficult to translate wherefore many of them were copied directly from English into the translation. The domain adapted model (EMEA retrained) tended to translate also the difficult terms, and the results were often good. The translation of unknown terms shows that the EMEA retrained model translates terms better than the other models.

In addition, it turned out that domain adaptation with unknown terms worsens the translation quality dramatically because the UNK retrained model translates all with single words. Consequently, domain adaptation with UNK terms is not possible. The reason is probably that the NMT system needs always a context to be able to translate whole sentences. If it is trained with words without context, the system has difficulties to translate whole sentences. It is better to train NMT systems with whole texts.

An idea for future work could be to find out how many training steps are enough to improve translation quality without overfitting the system. Another idea is to find ways to improve the translation quality even more.

References

Almeida, Felipe & Xexéo, Geraldo 2019: Word Embeddings: A Survey. *arXiv:1901.09060v1 [cs.CL]*.

Bahdanau, Dzimitry; Cho, KyungHyun & Bengio, Yoshua 2016: Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473v7 [cs.CL]*.

Bengio, Yoshua & Grandvalet, Yves 2004: No Unbiased Estimator of the Variance of K-Fold Cross-Validation. In: Journal of Machine Learning Research, vol. 5, pp. 1089-1105.

Bentivogli, Luisa; Bisazza, Arianna; Cettolo, Mauro & Federico, Marcello 2016: Neural versus Phrase-Based Machine Translation Quality: A Case Study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 257-267.

Callison-Burch, Chris; Osborne, Miles & Koehn, Philipp 2006: Re-evaluating the Role of Bleu in Machine Translation Research. School of Informatics, University of Edinburgh.

Cawley, Gavin C. & Talbot, Nicola L. C. 2010: On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. In: *Journal of Machine Learning Research*, vol. 11, pp. 2079-2107.

CAMBRIDGE DICTIONARY: “drug”. <https://dictionary.cambridge.org/dictionary/english/drug>. (accessed 11 March 2019.)

Chen, Yong; Xu, Wei; He, Zhongjun; He, Wei; Wu, Hua; Sun, Maosong; Liu, Yang 2016: Semi-Supervised Learning for Neural Machine Translation. a*rXiv:1606.04596v3 [cs.CL]*.

Chinea-Ríos, Mara; Peris, Álvaro & Casacuberta, Francisco 2017: Adapting Neural Machine Translation with Parallel Synthetic Data. In: *Proceedings of the Conference on Machine Translation (WMT),* vol 1, pp. 138-147. Association for Computational Linguistics.

Cho, KyungHyun; Merriënboer, Bart von; Bahdanau, Dzimitry & Bengio, Yoshua 2014: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv:1409.1259v2 [cs.CL]*.

Chu, Chenhui & Wang, Rui 2018: A Survey of Domain Adaptation for Neural Machine Translation. a*rXiv:1806.00258v1 [cs.CL]*.

Chu, Chenhui; Dabre, Raj & Kurohashi, Sadao 2017: An Empirical Comparison of Simple Domain Adaptation Methods for Neural Machine Translation. *arXiv:1701.03214v2 [cs.CL]*.

Claesen, Marc & Moor, Bart De 2015: Hyperparameter Search in Machine Learning. In: The XI Metaheuristics International Conference. *arXiv:1502.02127v2 [cs.LG]*.

Costa-Jussà, Marta R.; Allauzen, Alexandre; Barrault, Loïc; Cho, Kyunghun & Schwenk, Holger 2017: Introduction to the special issue on deep learning approaches for machine translation. In: *Computer Speech & Language, vol 46*, pp. 367-373.

Crego, Josep; Kim, Jungi; Klein, Guillaume; Rebollo, Anabel; Yang, Kathy; Senellart, Jean; Akhanov, Egor; Brunelle, Patrice; Coquard, Aurélien; Deng, Yongchao; Enoue, Satoshi; Geiss, Chiyo; Johanson, Joshua; Khalsa, Ardas; Khiari, Raoum; Ko, Byeongil; Kobus, Catherine; Lorieux, Jean; Martins, Leidiana; Nguyen, Dang-Chuan; Priori, Alexandra; Riccardi, Thomas; Segal, Natalia; Servan, Christophe; Tiquet, Cyril; Wang, Bo; Yang, Jin; Zhang, Dakun; Zhou, Jing & Zoldan, Peter 2016: SYSTRAN's Pure Neural Machine Translation Systems. *arXiv:1610.05540v1 [cs.CL]*.

Ding, Jie; Tarokh, Vahid & Yuhong, Yang 2018: Model Selection Techniques – An Overview. In*: IEEE Signal Processing Magazine*. *arXiv:1810.09583v1 [stat.ML]*.

Domingo, Miguel; García-Martínez, Mercedes; Helle, Alexandre; Casacuberta, Francisco & Herranz, Manuel 2018: How Much Does Tokenization Affect Neural Machine Translation? *arXiv:1812.08621v3 [cs.CL]*.

DUDEN: "Droge, die". <https://www.duden.de/rechtschreibung/Droge>. (accessed 11 March 2019.)

Freitag, Markus & Al-Onaizan, Yaser 2016: Fast Domain Adaptation for Neural Machine Translation. *arXiv:1612.06897v1 [cs.CL]*.

Garcia, Ignacio 2009: Beyond Translation Memory: Computers and the Professional Translator. In: *The Journal of Specialised Translation*. < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.523.867&rep=rep1&type=pdf>.

Gehring, Jonas; Auli, Michael; Grangier, David; Yarats, Denis & Dauphin, Yann N. 2017: Convolutional Sequence to Sequence Learning. *arXiv:1705.03122v3 [cs.CL]*.

GITHUB: "How do I continue training with new dataset on a pretrained model like transfer learning (Domain adaptation)". < https://github.com/OpenNMT/OpenNMT-py/issues/728>. (accessed 7 May 2019.)

Goodfellow, Ian; Bengio, Yoshua & Courville, Aaron 2016: "Deep Learning". MIT Press. Available at: < http://www.deeplearningbook.org/>. (accessed 9 April 2019.)

Hochreiter, Sepp & Schmidhuber, Jürgen 1997: Long Short-Term Memory. In: *Neural Computation* 9(8), pp. 1735-1780.

Jung, Alexander 2018: Machine Learning: Basic Principles. *arXiv:1805.05052v8 [cs.CL]*.

Klein, Guillaume; Kim, Yoon; Deng, Yuntian; Nguyen, Vincent; Senellart, Jean & Rush, Alexander M. 2018: OpenNMT: Neural Machine Translation Toolkit 2018. *arXiv:1805.11462v1 [cs.CL]*.

Klein, Guillaume; Kim, Yoon; Deng, Yuntian; Nguyen, Vincent; Senellart, Jean & Rush, Alexander M. 2017: OpenNMT: Open-Source Toolkit for Neural Machine Translation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*. pp. 67-72. Vancouver, Canada.

Koehn, Philipp & Knowles, Rebecca 2017: Six Challenges for Neural Machine Translation. In: *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28-39.

Koehn, Philipp 2010: Statistical Machine Translation. Cambridge University Press. New York.

Koeh, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondřej; Constantin, Alexandra & Herbst, Evan 2007: Moses: Open Source Toolkit for Statistical Machine Translation. In: *Proceedings of the ACL 2007 Demo and Poster Sessions, Prague*, pp. 177-180.

Koehn, Philipp; Och, Franz Josef & Marcu, Daniel 2003: Statistical Phrase-Based Translation. In: *Proceedings of HLT-NAACL, Edmonton*, pp. 48-54.

Koskenniemi, Kimmo 2013: Johdatus kieliteknologiaan, sen merkitykseen ja sovelluksiin. Nykykielten laitoksen oppimateriaalia 1. Helsingin yliopisto. < https://helda.helsinki.fi/bitstream/handle/10138/38503/kt-johd.pdf?sequence=1&isAllowed=y>. (accessed 8 January 2019.)

Kubat, Miroslav 2017: An Introduction to Machine Learning. Springer International Publishing Switzerland.

Läubli, Samuel; Fishel, Mark; Volk, Martin & Weibel, Manuela 2013: Combining Statistical Machine Translation and Translation Memories with Domain Adaptation. In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), Linköping Electronic Conference Proceedings #85*, pp. 331-474.

Niu, Xing; Denkowski, Michael & Carpuat, Marine 2018: Bi-Directional Neural Machine Translation with Synthetic Parallel Data. *arxiv:1805.11213v2 [cs.CL]*.

NLPL-WIKI: “Neural Machine Translation”. <https://nlp.stanford.edu/projects/nmt/>. Stanford NLP Group. (accessed 12 December 2018.)

NLP-WIKI: Translation/opennmt-py. <http://wiki.nlpl.eu/index.php/Translation/opennmt-py>. (accessed 21 January 2019.)

Och, Franz Joseph & Ney, Hermann 2002: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In: *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 295-302.

OPENNMT: "PyTorch Models". <http://opennmt.net/Models-py/>. Harvardnlp. (accessed 9 April 2018.)

OPENNMT b: "Example: Translation". <http://opennmt.net/OpenNMT-py/extended.html>. (accessed 12 December 2018.)

OPENNMT c: "Train". <http://opennmt.net/OpenNMT-py/options/train.html>. (accessed 4 february 2019.)

OPENNMT d: "Modules". <http://opennmt.net/OpenNMT-py/onmt.modules.html>. (accessed 9 April 2019.)

OPENNMT e: "Preprocess". < http://opennmt.net/OpenNMT-py/options/preprocess.html>. (accessed 11 April 2019.)

OPUS: "EMEA". <http://opus.nlpl.eu/EMEA.php>. (accessed 12 December 2018.)

Papineni, Kishore; Roukos, Salim; Ward, Todd & Zhu, Wei-Jing 2002: BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318.*

Park, Jaehong; Song, Jongyoon & Yoon, Sungroh 2017: Building a Neural Machine Translation System Using Only Synthetic Parallel Data. *arXiv:1704.00253v4 [cs.CL]*.

Paulsen Christensen, Tina & Schjoldager, Anne 2011: The Impact of Translation Memory ™ Technology on Cognitive Processes: Student-Translators' Retrospective Comments in an Online Questionnaire. In: *Proceedings of the 8th International NLPCS Workshop. Special Theme: Human-Machine Interaction in Translation.* Copenhagen Business School.

Reinke, Uwe 2018: State of the art in Translation Memory Technology. In: *Language technologies for a multilingual Europe.* Freie Universität Berlin. < https://books.google.fi/books?hl=fi&lr=&id=hC9tDwAAQBAJ&oi=fnd&pg=PA55&dq=translation+memory+technology&ots=nM0WyAwmA-&sig=G5XyzleceEkUugRNLqPUPf0KzwY&redir_esc=y#v=onepage&q&f=false>.

Rumelhart, David E.; Hinton, Geoffrey E. & Williams, Ronald J. 1986: Learning representations by back-propagating errors. In: *Nature, vol. 323(9),* pp. 533-536.

Sennrich, Rico; Haddow, Barry & Birch, Alexandra 2016: Improving Neural Machine Translation Models with Monolingual Data. *arXiv:1511.06709v4 [cs.CL]*.

Shen, Shiqi; Cheng, Yong; He, Zhongjun; He, Wei; Wu, Hua; Sun, Maosong & Liu, Yang 2016: Minimum Risk Training for Neural Machine Translation. *arXiv:1512.02433v3 [cs.CL]*.

Srivastava, Nitish; Hinton, Geoffrey; Krizhevsky, Alex; Sutskever, Ilya & Salakhutdinov, Ruslan 2014: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In: *Journal of Machine Learning Research*, vol. 15, pp. 1929-1958.

Stein, Daniel 2018: Machine translation: Past, present and future. In: *Language technologies for a multilingual Europe.* Freie Universität Berlin. < https://books.google.fi/books?hl=fi&lr=&id=hC9tDwAAQBAJ&oi=fnd&pg=PA55&dq=translation+memory+technology&ots=nM0WyAwmA-&sig=G5XyzleceEkUugRNLqPUPf0KzwY&redir_esc=y#v=onepage&q&f=false>.

Turian, Joseph; Ratinov, Lev & Bengio, Yoshua 2010: Word representations: A simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 384-394. Uppsala, Sweden.

Virpioja, Sami & Grönroos, Stig-Arne 2015: LeBLEU: N-gram-based Translation Evaluation Score for Morphologically Complex Languages. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 411-416.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukas & Polosukhin, Illia 2017: Attention Is All You Need. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, Canada, USA. arXiv:1706.03762v5 [cs.CL].

Vatsa, Mukesh; Joshi, Nikita & Goswami, Sumit 2010: Statistical Machine Translation. In: DESIDOC. Journal of Library & Information Technology, vol. 30(4), pp. 25-32.

Wikipedia: “Hypophysenadenom”. <https://de.wikipedia.org/wiki/Hypophysenadenom>. (accessed 7 May 2019.)

Wu, Yonghui; Schuster, Mike; Chen, Zhifeng; Le, Quoc V.; Norouzi, Mohammad; Macherey, Wolfgang; Krikun, Maxim; Cao, Yuan; Gao, Qin; Macherey, Klaus; Klinger, Jeff; Shah, Apurva; Johnson, Melvin; Liu, Xiaobing; Kaiser, Łukasz; Gouws, Stephan; Kato, Yoshiko; Kudo, Taku; Kazawa, Hideto; Stevens, Keith; Kurian, George; Patil, Nishant; Wang, Wei; Young, Cliff; Smith, Jason; Riesa, Jason; Rudnick, Alex; Vinyals, Oriol; Corrado, Greg; Hughes, Macduff & Dean, Jeffrey 2016: Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144v2 [cs.CL]*.

Zhang, Jiajun & Zong, Chengqing 2015: Deep Neural Networks in Machine Translation: An Overview. In: *IEEE Intelligent Systems, vol. 30(5)*, pp. 16-25.

Annexe

Lyhennelmä suomeksi

Helsingin yliopisto

Humanistinen tiedekunta

Kääntämisen ja tulkkauksen maisteriohjelma

Maria Mäkinen: Domain adaptation: Retraining NMT with translation memories

Pro gradu -tutkielma: 41 s.; suomenkielinen lyhennelmä 6 s.

Toukokuu 2019

Johdanto

Tekoälyä käytetään nykyään joka paikassa, ja kaikesta käyttämästämme teknologiasta on tulossa aina vain älykkäämpää.

Käännösalakin kehittyy muiden alojen mukana: neuraalikonekääntäminen (neural machine translation) on korvaamassa tähän asti paljon käytetyn tilastollisen konekääntämisen. Alan tutkijoiden tavoitteena on rakentaa konekäännin, joka kääntää yhtä hyvin kuin ihminen, mutta siitä tavoitteesta ollaan tällä hetkellä vielä kaukana, koska konekäännösten laadussa on vielä parantamisen varaa.

Erityisesti lääketieteessä käännösten laatu on tärkeää, jotta potilaat saavat oikeanlaista tietoa käyttämistään lääkkeistä. Jos esimerkiksi pakkausselosteen käännöksessä on vakavia virheitä, potilas voi ottaa hänelle määrättyä lääkettä väärin tai hän ei tiedä, mitä sivuvaikutuksia lääke saattaa aiheuttaa.

Neuraalikonekääntimen adaptoimista erikoisalalle ei ole tähän mennessä vielä kovin paljon tutkittu, mikä oli syynä sille, miksi valitsin sen pro gradu -tutkielmani aiheeksi. Tarkemmin sanottuna tutkin, miten neuraalikonekääntimen voi adaptoida erikoisalalle kouluttumalla konekäännintä käännösmuisteilla, ja paraneeko käännöksen laatu adaptoinnin jälkeen.

Vastaavan laista tutkimusta, jossa erikoisalalle adaptointiin käytetään käännösmuisteja, on tehty tähän mennessä tilastollista konekäännintä käyttäen. Neuraalikonekääntimien

adaptoinnissa erikoisaloille on tähän mennessä käytetty lähinnä muita tekniikoita kuten synteettisen korpuksen luomista konekääntämällä yksikielisiä korpuksia tai takaisinkääntämällä jo käännettyjä tekstejä. Tutkimuksen tulokset ovat kuitenkin lupaavia, minkä takia voidaan olettaa, että tässä pro gradu -tutkielmassa esitellyssä tutkimuksessa erikoisalalle adaptointi parantaisi myös käännöslaatua.

Tutkielma etenee siten, että ensin esitellään tilastollisen ja neuraalikonekääntämisen toimintaa sekä vertaillaan niitä toisiinsa. Tämän jälkeen tutustutaan konekäännösten automaattiseen evaluointiin – tarkemmin sanoen kahteen evaluointimittaan, BLEU:hun ja LeBLEU:hun. Tämän jälkeen kerrotaan, miten neuraalikonekäännin on aiemmissa tutkimuksissa adaptoitu erikoisalalle.

Teorettisen taustan jälkeen vuorossa on metodin ja aineiston esittely. Käännösmuistina käytetään EMEA-korpusta, joka sisältää lähinnä lääkkeiden pakkausselosteita englanniksi ja saksaksi. Konekääntimenä käytetään OpenNMT:n esikoulutettua konekäännintä. Luvussa esitellään kaksi erilaista adaptointitekniikkaa: konekääntimen jatkokoulutusta EMEA-teksteillä ja jatkokoulutusta konekääntimelle tuntemattomilla lääketieteellisillä termeillä.

Analyysiosassa annetaan ensin esimerkkejä siitä, millaisia erikoisalakäännöksiä esikoulutetulla ja jatkokoulutetuilla OpenNMT:llä tulee. Luvussa esitellään käännösesimerkein erilaisia käännösvirheitä. Evaluointiluvussa puolestaan esitellään automaattisen evaluoinnin tulokset ja verrataan niitä aiempaan tutkimukseen.

Teoreettinen tausta

Tilastollinen ja neuraalikonekääntäminen

Tilastollisessa konekääntämisessä kone laskee eri käännösvaihtoehdoille tilastollisen todennäköisyyden. Näistä käännösvaihtoehdoista valitaan lopulliseksi se käännös, joka on saanut parhaan tuloksen eli on todennäköisimmin oikea. Millaisia käännösvaihtoehtoja voi millekin lauseille tai sanoille tulla, opitaan paralleelikorpuksista, jotka on syötetty konekääntimeen koulutusvaiheessa. Korpuksen avulla tilastollinen konekäännin pystyy

muodostamaan fraasi- tai käännöstaulukoita, joihin on merkitty käännösvaihtoehdot todennäköisyyksineen sekä lähtökieliset lauseet.

Myös neuraalikonekääntäminen perustuu laskettuihin todennäköisyyksiin, mutta – toisin kuin tilastollisessa konekääntämisessä – konekääntimelle ei tarvitse antaa koulutusvaiheessa valmiita kieli- tai käännösmalleja, koska se oppii tekoälyn ansiosta itsenäisesti kielten rakenteita koulutusvaiheessa syötettyjä paralleelikorpuksia apuna käyttäen, ja pystyy huomioimaan myös sanojen ja fraasien kontekstin kääntäessään. Neuraalikonekäännin ei tee käännöksistä käännöstaulukoita, vaan jakaa käännettävän tekstin sekvensseihin. Näin ollen tekstiä käännetään sekvenssi kerrallaan.

Neuraalikonekääntimissä yleisimmin käytetty neuroverkkotyyppi on rekurrentti neuroverkko (recurrent neural network). Muita neuroverkkotyyppejä ovat muun muassa konvoluutioverkko (convolutional neural network), ”myötäkytkentäverkko” (feedforward neural network) sekä Transformer. OpenNMT käyttää RNN:ia ja Transformeria.

Erityisesti pitkät lauseet ja harvinaiset tai taivutetut sanat aiheuttavat sekä tilastollisille että neuraalikonekääntimille käännösongelmia, joihin on yritetty löytää ratkaisuja. Vertailevista tutkimuksista käy ilmi, että tilastollinen konekäännin selviytyy näistä ongelmista paremmin, mutta jos neuraalikonekääntimiä kehitetään paremmiksi, ne voivat selviytyä kyseisistä ongelmista jopa tilastollisia konekääntimiä paremmin. Yleisellä tasolla neuraalikonekääntimien käännöslaatua pidetään kuitenkin parempana ja sujuvampana kuin tilastollisten konekäänninten käännöslaatua, koska se oppii kieliopin paremmin.

Automaattinen evaluointi

Automaattiset evaluointimitat, kuten BLEU ja LeBLEU on keksitty, jotta käännösten tarkistamisesta tulisi nopeampaa ja halvempaa. Käännösten laadun arviointi onnistuu siten, että konekäännöstä verrataan ihmiskäännöksiin. Kone- ja ihmiskäännöksistä etsitään yhtäläisyyksiä (samanlaisia n-grammeja), joiden perusteella konekäännöksille lasketaan pisteet. Sekä BLEU:ssä että LeBLEU:ssä käännökset pisteytetään nollasta yhteen – nolla tarkoittaa tällöin erittäin huonoa käännöstä ja yksi täydellistä käännöstä. BLEU on arvosteluissa tiukempi, koska se ei salli kielellistä vaihtelua (synonyymit, eri sanajärjestykset)

yhtä hyvin kuin LeBLEU, minkä takia BLEU:tä onkin kritisoitu. LeBLEU on BLEU:n paranneltu versio. Mahdollisten arviointivirheiden takia onkin suositeltavaa käyttää käännösten evaluoinnissa useampaa kuin yhtä mittaa sekä mahdollisesti manuaalista, ihmisten tekemää evaluointia.

Aineisto ja metodi

Tutkimuksessa käytettiin OpenNMT:n esikoulutettua versiota. Konekääntimen asetuksia ei muutettu tutkimuksen aikana, vaan tekstit käännettiin perusasetuksia käyttäen. Käännösmuistina käytettiin EMEA-korpusta, joka sisältää lähinnä lääkkeiden pakkausselosteita. Korpus jaettiin k-fold -menetelmällä kymmeneen osaan; yhtä osaa käytettiin koulutukseen ja toista osaa kääntämiseen. Näin koulutus- ja testiaineistona ei käytetty samaa tekstiä. Tekstiä käännettiin englannista saksaan. Ensimmäinen EMEA-käännös tehtiin esikoulutettua konekäännintä käyttämällä, minkä jälkeen konekäännintä olisi pitänyt jatkokouluttaa EMEA-aineistolla.

Kävi kuitenkin ilmi, ettei jatkokoulutus onnistu. Ongelma ratkaistiin siten, että koulutettiin täysin uusi konekäännin esikoulutetun konekääntimen pohjalta – koulutuksessa käytettiin siis samoja korpuksia, joita esikoulutetunkin version kouluttamiseen oli käytetty. Tätä uutta konekäännintä pystyi jatkokouluttamaan ongelmitta. Jatkokoulutusta eli erikoisalalle adaptointia testattiin kahdella eri tavalla: kouluttamalla konekäännintä EMEA-tekstillä sekä kouluttamalla sitä EMEA-korpuksesta eristetyillä ns. tuntemattomilla termeillä. Tuntemattomat termit ovat sellaisia termejä, jotka esiintyvät EMEA teksteissä, mutta eivät esikoulutusaineistossa.

Analyysi

Analyysivaiheessa OpenNMT-konekääntimen eri versioiden (esikoulutettu, uudelleenkoulutettu, teksteillä jatkokoulutettu ja termeillä jatkokoulutettu) käännöslaatua verrattiin keskenään.

Tutkimuksessa selvisi, että esikoulutetun konekääntimen käännöslaatu erikoisalatekstiä käännettäessä ei ollut huonoimmasta päästä. Lisäksi uudelleen koulutettu konekäännin oli ennen erikoisalalle adaptointia hieman huonompi kuin esikoulutettu konekäännin. Jatkokouluttamalla uudelleen koulutettua konekäännintä EMEA-aineistolla käännöksen laatu parani huomattavasti. Laadun parannus näkyi selkeästi paitsi koko tekstin myös termilistan käännöksissä.

Yleisimmät ongelmat olivat pitkien lauseiden ja vaikeiden termien kääntäminen ja kokonaan isoin kirjaimin kirjoitetut lauseet ja fraasit. Pitkien lauseiden käännökset olivat usein keskeneräisiä tai hieman epäsujuvia. Usein tällaisissa lauseissa oli myös paljon termejä, jotka oli käännetty väärin. Termien kääntäminen väärin saattaa johtua siitä, että pitkät ja vaikeat termit käännetään pienemmissä paloissa; jos yksi pala menee väärin, todennäköisesti koko termi käännetään väärin. Samat käännösongelmat näkyivät kaikkien käytettyjen konekääntimien käännöksissä, mutta ongelmia oli vähiten EMEA-tekstillä koulutetun konekääntimen käännöksissä.

Tuntemattomilla termeillä koulutetun konekääntimen käännöslaatu oli kaikista huonoin, sillä se käänsi kokonaiset lauseet yksittäisinä sanoina, jolloin teksti näytti pikemminkin sanalistalta. Kun tällä konekääntimellä käänsi pelkkää termilistaa, käännöksen laatu parani hieman.

Lopuksi

Tämän tutkimuksen perusteella voi sanoa, että neuraalikonekääntimen adaptoiminen erikoisalalle onnistuu kouluttamalla oman versionsa OpenNMT:stä ja jatkokouluttamalla sitä erikoisalateksteillä. Näin käännöksen laatu paranee huomattavasti. Sen sijaan pelkillä erikoisalatermeillä kouluttaminen ei toimi, koska kone oppii kääntämään pelkkiä termejä eikä ollenkaan kokonaisia lauseita. Molemmilla metodeilla on mahdollista parantaa pelkkien termien käännöslaatua. Lauseiden kääntämiseen konekäännin tarvitsee ilmeisesti kontekstin, jotta se oppii kääntämään termit oikein kulloisessakin kontekstissa.

Mahdollisissa jatkotutkimuksissa voisi yrittää selvittää, miten käännöslaatua voisi parantaa vielä enemmän ja onko ylipäätään mahdollista kääntää tekstejä koneella niin hyvin, että sen laatu vastaa ihmisen tekemän käännöksen laatua.