UPPSALA
UNIVERSITET

Cross-lingual Word Embeddings Beyond Zero-shot Machine Translation

Shifei Chen

Uppsala University
Department of Linguistics and Philology
Master Programme in Language Technology
Master's Thesis in Language Technology, 30 ECTS credits

September 13, 2020

Supervisors:
Ali Basirat, Uppsala University

Abstract

We explore the transferability of a multilingual neural machine translation model to unseen languages when the transfer is grounded solely on the cross-lingual word embeddings. Our experimental results show that the translation knowledge can transfer weakly to other languages and that the degree of transferability depends on the languages' relatedness. We also discuss the limiting aspects of the multilingual architectures that cause the weak translation transfer and suggest how to mitigate the limitations.

Contents

Preface

This thesis was finished under the supervision of Ali Basirat. I would like to thank him first for his guidance, inspiration and passion.

The Saga supercomputer [1] owned by UNINETT Sigma2 hosted all of the experiment in this thesis. Without it this thesis would not be possible.

Thank you Mr. Anders Wall and everyone in the Anders Wall Scholarship Foundation for sponsoring my Master's study. This opportunity led me to meet everyone in the Master Programme in Language Technology, from whom I have learned a lot during the 2-years journey.

Last but not least, I would like to say a thank you to my parents for their unconditional love and support; to all of my friends for the unique memories we have created; and to my girlfriend, who has always been next to me when the virus made everything unusual.

[1] https://www.sigma2.no/systems#saga

1. Introduction

The multilingual neural machine translation (NMT) aims at training a single translation model between multiple languages (Aharoni et al., 2019; M. Johnson et al., 2016). Among the appealing points of multilingual NMT models are their ability for zero-shot learning, to generalize and transfer a translation model to unseen language pairs (M. Johnson et al., 2016). In zero-shot learning, a multilingual model trained on a set of language pairs is tested on an unseen language pair whose elements are still in the set of individual training languages. The knowledge in this setting is transferred across the shared parameters of the model.

Kim et al., 2019 argues that one of the critical components responsible for the knowledge transfer in multilingual NMT is the embedding layers which train a cross-lingual vector space for all words of the languages. They show that multilingual translation models can be transferred to new languages if the model's cross-lingual vector space is aligned to the new language's vector space. However, the success of their approach is at the cost of aligning the word vectors and retraining the new language's translation model.

This research studies the importance of word representation in the multilingual NMT transfer model in a more controlled setting based on the pre-trained cross-lingual word embeddings (Ammar et al., 2016; Bojanowski et al., 2016; Joulin et al., 2018; Ruder et al., 2019). More specifically, we examine the transferability of a multilingual NMT when it is applied to a new test language. We use cross-lingual word embeddings as the source of transfer knowledge to the test languages and leave the translation model's shared parameters to model the interrelationships between the training languages. Our setting is different from the zero-shot setting of M. Johnson et al., 2016 in the way that one of the test languages does not belong to the set of individual training languages. It is also different from the Kim et al., 2019's setting, in that it does not retrain the embeddings and model parameters.

We hypothesize that a multilingual NMT model trained with pre-trained cross-lingual word embeddings should transfer reasonably to a new test language if the word representation has any role in the model's transferability. Our results on a set of test languages show that the translation model transfers only weakly to the unseen languages, and the amount of the transferability depends on the similarity of the test language to the training languages. Furthermore, when using aligned pre-trained word embeddings as the only transferable knowledge source, the performance will be negatively affected by the transformed output vector space, which need to be addressed in order to achieve meaningful translations.

The rest of the thesis is organized below:

Chapter 2 talks about the background and previous works when working in the transferability of multilingual word embeddings scope, including information about word embeddings, multilingual translation and related works from others.

Chapter 3 introduces our experiment method, whose result are discussed and analyzed in Chapter 4.

Finally, Chapter 5 gives out our conclusion. We also show samples of our multilingual NMT model in the Appendix A.

2. Background

2.1. Word Embeddings

2.1.1. Representing Words in Vectors

In Natural Language Processing, people need to convert the natural representation of words into forms that are more efficient for computers to process. The idea started with statistical language modeling introduced by Bengio et al., 2003. In 2013, Mikolov, Chen, et al., 2013 introduced Word2Vec, which encapsulates words and their latent information into vectors. Besides the benefit that it simplifies representation and storage of words for computers, it also enables the possibilities to calculate words and their semantic meanings just as vectors.

Take an example of vocabulary $V = \{\text{king}, \text{queen}, \text{man}, \text{woman}\}$, if we convert these words into vectors such as

$$\begin{aligned}\vec{k} &= \text{vec}(\text{king})\\ \vec{q} &= \text{vec}(\text{queen})\\ \vec{m} &= \text{vec}(\text{man})\\ \vec{w} &= \text{vec}(\text{woman})\end{aligned}$$

We could have an equation of

$$\vec{q} = \vec{k} - \vec{m} + \vec{w} \tag{2.1}$$

It is meaningful from both the mathematical perspective and the linguistic perspective. Figure 2.1 illustrates both perspectives in a vector space that contains these four vectors. Geometrically, the angle between $\vec{k}$ and $\vec{q}$ is small, together with the angle between $\vec{m}$ and $\vec{w}$. From the cosine similarity definition below

$$sim(x, y) = \cos(\theta) = \frac{x \cdot y}{||x||||y||}$$

It indicates that the cosine similarities of every two of them are high. Besides, Equation 2.1 is correct in this vector space too. We could get such an equation from basic vector calculation definitions.

Semantically, the word "king" and "queue" are closely related in the same category as well as the word "man" and "woman". It is meaningful to say "queue" is a "king" being replaced its "man" part by a female.

To turn words into vectors, one could use a simple one-hot encoding. Like in the example above we could make $\vec{k} = [1, 0, 0, 0]$. However, these one-hot vectors can merely capture any latent semantic information between different words. Recent vectorized word representations, or word embeddings, were learned through neural networks, such as Word2Vec, which learns word embeddings through a Skip-gram model or a Continuous Bag of Words model (Mikolov, Sutskever, et al., 2013). Both models are shown in Figure 2.1.1.

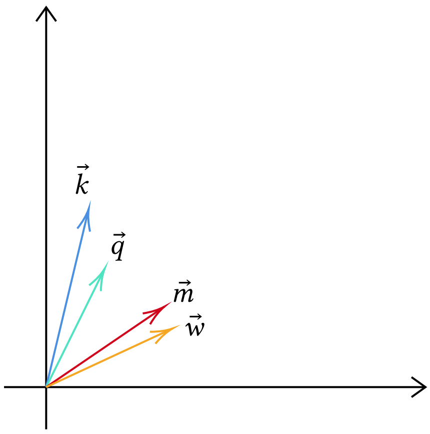

Figure 2.1.: Illustraion of a vector space where Equation 2.1 exists.

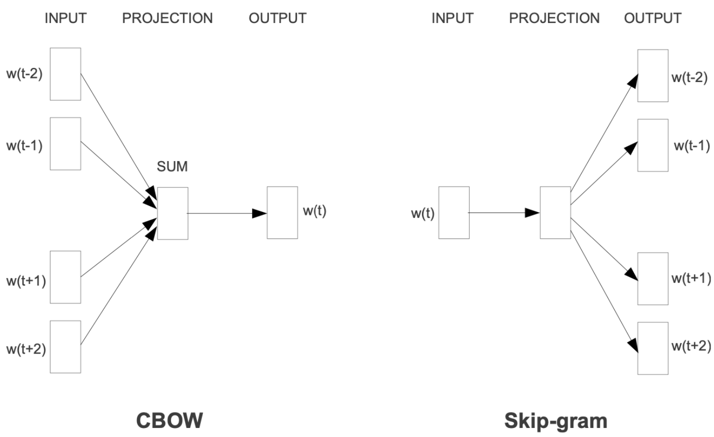

Figure 2.2.: The Skip-gram and the CBOW model. (Mikolov, Le, et al., 2013)

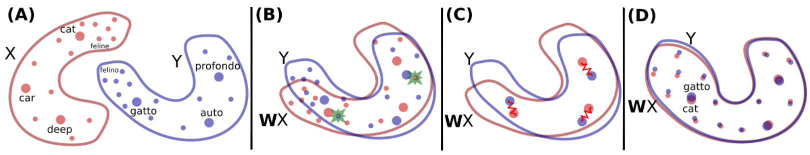

Figure 2.3.: Aligning bilingual vector spaces. (Conneau et al., 2017)

The Skip-gram model

When given a target word w, the model can produce vector representations that are good at predicting the words surrounding w within the context size of C. The probability of a context word w_k given a target word w is:

$$P(w_k|w) = \frac{\exp({v'_{w_k}}^{\top} v_w)}{\sum_{i=1}^{|V|} \exp({v'_i}^{\top} v_w)} \tag{2.2}$$

Here $|V|$ means the size of the whole vocabulary from the corpus. Both v' and v stand for the input vector representation, and the output vector representation of a word (Mikolov, Chen, et al., 2013). The aforementioned one-hot vectors are used in the Skip-gram model too. They can initialize the input representation v'.

The Continuous Bag of Words model (CBOW)

The other model, CBOW, works just like the other side of the coin. It predicts the target word w based on a bunch of context words $w_{-C}, w_{-C+1}..., w_{C-1}, w_C$ within the window size C, as the formula below:

$$P(w|w_{-C}, w_{-C+1}..., w_{C-1}, w_C) = \frac{\exp({v'_w}^{\top} \bar{v}_{w_k})}{\sum_{i=1}^{|V|} \exp({v'_{w_i}}^{\top} \bar{v}_{w_k})} \tag{2.3}$$

Here $\bar{v}_{w_k}$ means the sum of the context word w_k's vectorized representation, while v'_w means the input vector representations of word w as in the Skip-gram model.

The difference between these two models is that the CBOW model predicts the target word from multiple given context words, while the Skip-gram model predicts the context words from one given center word. Hence the Skip-gram model is better at predicting rare words because all of the words are treated equally in the *word AND context* relationship. While in the CBOW model, frequent words have advantages over rare words as they will have higher probabilities in a given context. The Skip-gram model is arguably the most popular method to learn word embeddings as it is both fast and robust, especially with less frequent words in the corpus. (Levy et al., 2015)

2.1.2. Multilingual Word Embeddings

Learned from approaches like the Skip-gram model or the CBOW model, vectorized word representations tend to cluster words with similar semantics (Mikolov, Le, et al., 2013). It then becomes attractive to see whether we could fit two or more languages into the same vector space. Word embeddings that consist of more than one language are called multilingual word embeddings.

In the multilingual scenario, it is vital to align words in two different vector spaces to make word embeddings from different languages comparable. Figure 2.1.2 illustrated the alignment method from Conneau et al., 2017. Suppose there is a set of word pairs in their associated vectorized representation $\{x_i, y_i\}_{i \in \{1,...,n\}}$, the two vector spaces were aligned by a rotation matrix $W \in \mathbb{R}^{d \times d}$ as shown in process **(B)**, where we try to optimize the formula

$$\min_{W \in \mathbb{R}^{d \times d}} \frac{1}{n} \sum_{i=1}^{n} \ell(Wx_i, y_i)$$

Here ℓ is the loss function and it is usually the square loss function $\ell_2(x, y) = ||x-y||^2$. Then W is further refined in process **(C)**, where we choose frequent words as anchor points and minimize the distance between each correspondent anchor points by an energy function. After this, the refined W is then used to map all words in the dictionary during the inference process. We obtain the translation $t(i)$ of a given source word i in the formula

$$t(i) \in \underset{j \in \{1,\ldots,N\}}{\arg\min}\ \ell(Wx_i, y_j)$$

Again, the loss function ℓ is typically the square loss function. However, using the square loss function could make the model suffer from the "hubness problem". Conneau et al., 2017 counter reacted to the "hubness" problem by introducing the cross-domain similarity local scaling (CSLS), whose detail is beyond this thesis's scope.

Despite the theoretical feasibility proven by mathematic formulas, we need data to drive the alignment process. Take bilingual word embeddings as the first step, their initial alignment data for adversarial learning of the rotation matrix W could come from a bilingual dictionary (Mikolov, Le, et al., 2013). Also, there are other kinds of alignment using aligned data from sentence level, or even document level, even though word-level information is most common. By using word-level information, we can start with a pivot language (usually English) and map each other monolingual word embeddings by looking up translation dictionaries. This mapping process could also start with vectors only, where we choose a bilingual word embedding that shares a language (also typically English) with other bilingual embeddings and choose the other bilingual word embeddings by aligning their shared language subspace. Sentence-level parallel data is similar to the corpus in Machine Translation (MT), which contains sentence-aligned texts (Hermann and Blunsom, 2013). Document-level information is more common in the form of topic-aligned or class-aligned, such as Wikipedia data (Vulić and Moens, 2013). The alignment process of multilingual word embeddings is roughly the same as bilingual word embeddings, using parallel data from either word-level, sentence-level, or document-level (Ruder et al., 2019).

2.1.3. fastText

In this work, we have chosen fastText aligned word vectors[1] (Joulin et al., 2018) as my vectorized word representation. They are based on the pre-trained vectors computed from the Wikipedia corpus using fastText (Bojanowski et al., 2016). Not only because of the soon to be talked good performance, but fastText aligned word embeddings also has a large selection of languages available to use out of the box. It is ideal for implementing the cross-lingual experiments on aligned word embeddings.

fastText is an extension of the original Word2Vec methods, which uses sub-words to augment low-frequency and unseen words. Take word `low-key` as an example. As a whole word, its possibility in a given document would be much lower than its components, `low` and `key`. fastText learns its vectorized representation from a smaller n-gram sub-word level. It divides the whole word into sub-words units as below if we assume $n = 3$

[1] https://fasttext.cc/docs/en/aligned-vectors.html

<`lo`, `low`, `ow-`, `w-k`, `-ke`, `key`, `ey`>

Each sub-word has its own vectorized representation learned through a CBOW or Skip-gram model as in Word2Vec. The word vector for the whole word unit `<low-key>` is then the sum of all of its sub-word units' vectors. Hence its rareness would be compensated by two more frequent subwords `low` and `key`, even if it might not appear in the training document at all.

In terms of multilingual alignment, fastText improves the standard solution to the hubness problem by directly including the Relaxed CSLS (RCSLS) criterion into the model during both the learning and the inference phase. Before the work of Joulin et al., 2018, inverted softmax (ISF) Smith et al., 2017 or CSLS (Conneau et al., 2017) was only used in the inference time to address the hubness problem while square loss is still the loss function used in the training time. However, since both the ISF and the CSLS are not consistent with the square loss function in the training time, they will create a discrepancy between the learning of the translation model and the inference. According to the authors, fastText outperforms other alignment approaches by 3 to 4% on average.

2.2. Multilingual Neural Machine Translation (MNMT) Systems

2.2.1. Multilingual Neural Machine Translation

Neural Machine Translation (NMT) uses neural networks to learn the translation relationship between a source and a target language. Its power has gone over the traditional Statistical Machine Translation (SMT) and enabled things that were impossible in the past. One of that is Multilingual Neural Machine Translation, as shown in the Figure 2.2.1, it uses the same attentional encoder-decoder model as bilingual NMT but trains it on a multilingual corpus (M. Johnson et al., 2016). The benefit of such a multilingual system does not necessarily stop at higher translation performance between common languages like English, French, or Spanish; it also leverages additional information from high resource languages to low resource languages in the same semantic space (Ha et al., 2016). People identify this information leverage as a special form of transfer learning (Zoph et al., 2016), which could happen both in the horizontal or vertical direction (Lakew et al., 2019): In the horizontal direction, knowledge transfers from pre-trained data (such as word embeddings or language models) to the raw test data; in the vertical direction, knowledge transfers from languages to languages, which could either transfer from high resource to low resource languages or from seen languages to unseen languages. The latter is called zero-shot translation.

2.2.2. Zero-shot Machine Translation Systems

Zero-shot translation stands for translation between language pairs invisible to the multilingual NMT system during the training time. E.g., we build a multilingual NMT system with training language pairs of German-English and French-English while test its performance on a German-French scenario. In 2016, M. Johnson et al., 2016 first published their result on a zero-shot MT system. Their multilingual MT system, including an encoder, decoder, and attention module, requires no change to a standard NMT system. The only modification is in the training corpus, where they had introduced an artificial token at the beginning of each source sentence to denote the translation target language. Ha et al., 2016 also showed that their universal encoder and decoder model is capable of zero-shot MT. The concept of translation between

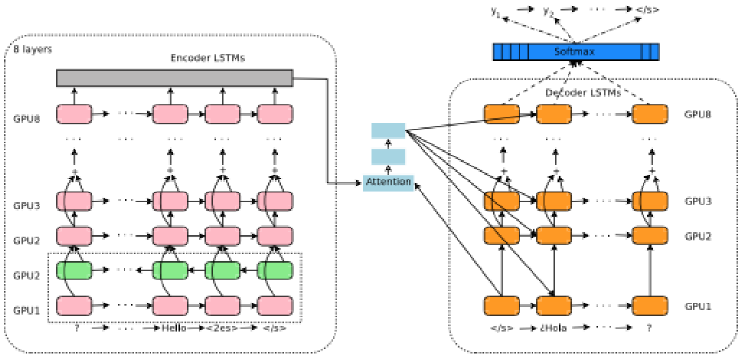

Figure 2.4.: Google's MNMT Architecturef from (M. Johnson et al., 2016)
.

unseen language pairs is attractive, especially for low-resource language pairs, though these two models both underperformed than a pivot based system.

There are two reasons that could explain the gap between a zero-shot system and a pivot based system, language bias (Arivazhagan et al., 2019; Ha et al., 2016, 2017) and poor generalization (Arivazhagan et al., 2019). Language bias means that during inference, the MT system tends to decode the target sentence into the wrong language, usually copying the source language or the bridging language Ha et al., 2016. It could be the consequence of always translating all source languages into the bridging language, hence make the model difficult to learn to translate the desired target language (Arivazhagan et al., 2019).

The other potential reason for the worse performance of a zero-shot system is poor generalization (Arivazhagan et al., 2019). When a zero-shot system is trained purely on the end-to-end translation objective, the model prefers to overfit the supervised translation direction features than learn more transferable language features.

To fix these two problems, there has been work on improving the preprocessing process (Lakew et al., 2018), parameter sharing (Blackwood et al., 2018; Firat et al., 2016), additional loss penalty functions (Arivazhagan et al., 2019) and pre-training modules using external information (Baziotis et al., 2020). In some cases, zero-shot system could achieve better performance than pivot based systems.

2.3. Previous Work

The transferability of multilingual neural machine translation models is vital from both the theoretical and practical perspectives. The theoretical importance of these models come to the way that they find the correspondence between language pairs (M. Johnson et al., 2016; Lu et al., 2018). The practical importance is due to their effective use for the translation of low-resource languages (Nguyen and Chiang, 2017; Zoph et al., 2016). M. Johnson et al., 2016 shows that multilingual NMT trained on a massive training set can generalize reasonably well in the zero-shot learning setting. This capability is further examined by Aharoni et al., 2019, demonstrating that multilingual NMT models transfer better when trained on a massive training set. Kim et al., 2019 shows that multilingual NMT models can be transferred to a new language when their embedding spaces are realigned to the embeddings of the new language.

In terms of the application of pre-trained aligned word embeddings in a multilingual NMT system, there are some successful applications, such as using the aligned embeddings as the embedding layer (Artetxe et al., 2017; Neishi et al., 2017), as the substitution of a supervised dictionary (Conneau et al., 2017), or as an external supplementary extension (Di Gangi and Federico, 2017). There are even cases where people successfully trained a MT system using very little or none parallel data (Conneau et al., 2017) and heavily rely on aligned word embeddings. Nevertheless, in most MT systems, using pre-trained word embeddings purely as the embedding layer will not outperform other models such as Transformers (Vaswani et al., 2017) and its other evolutions, largely because the training data for an MT system is usually several orders of magnitude larger than the monolingual pre-trained word embeddings. Typically pre-trained word embeddings are mainly introduced in MT systems dealing with low-resource languages.

For NMT systems focused on low resource language, Qi et al., 2018 looked into the question of when and why are pre-trained word embeddings useful. They found that pre-trained word embeddings are consistently useful for all languages. The gains would be more visible if the source and target language are similar, such as languages within the same family. Also, pre-trained word embeddings work well only on MT systems with moderate performance. Pre-trained word embeddings can not work when there is not enough data to train a basic MT system. Finally, aligned word embeddings are useful in a multilingual MT system. For bilingual MT systems, pre-trained word embeddings do not necessarily need to be aligned.

Moreover, aligned word embeddings do not work well for morphologically rich languages such as Russian and Belarusian. Qi et al., 2018 argues that this may be mainly due to the sparsity in the word embeddings files. Plus, most of the previous works target zero-shot language pairs, not on completely unseen languages. For language pairs $A \longrightarrow$ EN and EN $\longrightarrow B$, they are all interested in the unseen language pair $A \longrightarrow B$. For language pairs that include an unseen language C, whether it is on the source side, or the target side, it remains to see how universal word embeddings could help translate in this scenario.

It is then interesting to see how far can aligned word embeddings could go beyond known languages. As mentioned in Section 2.2.2, zero-shot translation analyzed this question by testing the multilingual NMT system on unseen language pairs — language in either source or target side of the translation is known to the system, but their paired combination remains unknown. In this work, we would like to take a step further to see how aligned word embeddings would work for zero-resource languages – Languages that are entirely unseen to the multilingual NMT setting.

3. Methodalogy

In this chapter, we will perform experiments in a universal word embedding based multilingual NMT system to see the transferability of the model on the test languages.

3.1. Theoratical Feasibility

As mentioned in Section 2.1.2, Mikolov, Le, et al., 2013 showed that there is a linear relationship between similar word embeddings in different languages. For each word pairs, assume their vector representations are $\{x_i, y_i\}_{i=1}^{n}$, we could calculate a transformation matrix W such that Wx_i approximates to y_i. In practice, people can learn W by optimizing the following target function.

$$\min_{W} \sum_{i=1}^{n} ||Wx_i - y_i||^2$$

Mikolov, Le, et al., 2013 also showed their result in the word/phrase translation task for the approximated word embedding mappings. For some subsets of words, around 70% of word embeddings match precisely with each other according to the P@5 score. If we relax the cosine similarity threshold to 0.6, the P@5 score would be as high as 90%.

To convert words into vectors to be calculated in the neural network, NMT systems should treat each word as a word embedding. The value of these word embeddings could be learned directly during translation, but then the initialization is a crucial step since poor initialization could lead to slow converge or worse local minima (Glorot and Bengio, 2010). The situation could be even more challenging when translating with very few parallel corpora since there is no data to help the embedding layer converge to its ideal state. Hence the word embedding mapping technique above becomes appealing.

Qi et al., 2018 explored how effective it is by using aligned pre-trained word embeddings in an NMT system. They found that regardless of languages, alignment is useful as long as it is applied in a multilingual setting. They believe that since both the source and the target side vector spaces are already aligned, the NMT system will learn how to transform a similar fashion from the source language to the target language.

Therefore, translating a completely unseen language is to test the transferability from known languages to an unknow language. It can be viewed as the question below — Given a vector space Z that consists of aligned word embeddings $\{a_i, b_i, c_i, ...\}$, how much does the NMT system knows about an unseen language A if it was only trained on the remaining languages? In theory, since the word embeddings are clustered by their semantic meanings in the same vector space Z, we should be able to build loose mappings between the semantic centers from both the source and the target sides. The generalization ability of the system is the key to answer this question. Hence we conducted some experiments below.

3.2. Experiment Settings

To get a basic multilingual MT system running, we chose English (EN), German (De), and French (FR) to be the training languages. Let C donate the final corpus, l donates the language-specific corpus fragment and Z is the set of corresponding candidate languages, the training language set is then $Z_{TRAIN} = l_{EN}, l_{DE}, l_{FR}$. For the test language set, we picked up Swedish (SV), Hungarian (HU), and Hebrew (HE) being his test languages. Therefore $Z_{TEST} = l_{SV}, l_{HU}, l_{HE}$.

For each experiment, the author trained a basic multilingual NMT system using a training corpus C_{TEST} with all three training languages, including all six directions from the cartesian product without duplicates as below.

$$C_{\text{TRAIN}} = \{x \times y \mid x, y \in Z_{\text{TRAIN}} \text{ and } x \neq y\} \tag{3.1}$$

The equation below means that the multilingual NMT system is tested on the test corpus with all three training languages and one of the test language. The test corpus consists of both translation directions of three different training language and that only test language.

$$C_{\text{TEST}} = \{(x, y) \cup (y, x) \mid x \in Z_{\text{TRAIN}} \text{ and } y \in Z_{\text{TEST}}\} \tag{3.2}$$

we designed the experiments and picked up the training and target languages based on Language Similarity. Qi et al., 2018 observed that pre-trained word embeddings are useful for languages from the same language family. The closer their relationship is, the higher the performance improvement is. Aligned word embeddings will also be beneficial if applied in a multilingual NMT system consisting of languages from the same language family.

3.2.1. Corpus and Preprocessing

we have used the TED talk subtitle corpus from Qi et al., 2018 [1] to train the multilingual NMT. The whole corpus has roughly 2.7×10^6 sentences split into three parts, train, dev, test at the ratio of $0.95 : 0.025 : 0.025$.

To build up the corpus for each experiment, the author has modified the original script from Qi et al., 2018 and added a few customized features. In short, the script will extract shared sentences from each part of the split corpus to form up a common intersection used in training, developing, and testing. Since the experiments consist of languages that are relatively common in the TED project, this fine-tuned corpus is not too different from the original corpus, hence after all the sizes for the train, dev, and test split were kept.

For preprocessing, since the original TED corpus is already tokenized by Moses. Then the system Neubig et al., 2018 turned all of the text into lower cases and applied a sentence length filter to remove any long sentences with more than 60 words. This sentence length filter prevents inferior performance in training. After that, when building the i2w and w2i index for the pre-trained embeddings, we have also removed any words that are less frequent than two times to stop the system from overfitting by low-frequency words. All of the preprocess functions are built upon the built-in XNMT preprocess features (Neubig et al., 2018).

3.2.2. Neural Network

The neural network is a modified version of the one from Qi et al., 2018, which is built with XNMT Neubig et al., 2018. The only change is doubling the encoding layer to a

[1] https://github.com/neulab/word-embeddings-for-nmt

2-layer-bidirectional LSTM network, thus having more parameters to accommodate the additional information in a multilingual scenario. Everything else is the same as the original experiment settings, including the encoder-decoder model with attention (Bahdanau et al., 2014) with a beam size of 5, trained using batches of size 32, dropout set to 0.1, the Adam optimizer (Kingma and Ba, 2014) and the evaluation metric BLEU score (Papineni et al., 2002). The initial learning starts at 0.0002 and decays by 0.5 when development BLEU score decreases (Denkowski and Neubig, 2017).

3.2.3. Embeddings

As metioned in Section 2.1.3, the embeddings used in the experiments are fastText aligned word embeddings[2]. They are based on the pre-trained vectors on Wikipedia[3] using fastText (Bojanowski et al., 2016). The alignment is performed using RCSLS as in Joulin et al., 2018.

Each of the fastText word embedding file is language-specific and contains word embeddings in 300 dimensions. We concatenated different language files to build up multilingual word embedding files for the multilingual NMT system. If there is a shared word w with two different vector values $\vec{v_a}$ and $\vec{v_b}$ in different embedding files, the average value of both vectors $v_{\vec{mean}}$ will be the new vector.

$$v_{\vec{mean}} = (\vec{v_a} + \vec{v_b})/2 \tag{3.3}$$

In this way, there are possibilities that both of the unique semantic values in the two words w_a and w_b could be lost, as there are cases that word with distant meaning share the same spelling in different languages. However, people could also argue that many words with the same spelling do have a similar meaning. For example, the word `café` means the same thing in English and French, as English borrowed that word from French. Later in the experiment, there will also be a different attempt where the system treats each word as a unique word even though they might share the same spelling. Both of the results will be available below.

[2]https://fasttext.cc/docs/en/aligned-vectors.html
[3]https://www.wikipedia.org/

4. Results and Analysis

As anticipated, Swedish will get the best result among all three test languages. Its performance could even be on the same level as the baseline multilingual NMT consists of only the training languages — EN, DE, and FR. The other two test languages' performance will not be close to the Swedish one, and they could be less than 10 BLEU scores.

In the results shown in Table 4.1, Swedish, Hungarian and Hebrew all got unpredicted low BLEU scores. The expected high perfromance from Swedish did not appear in the experiment result. All of the three languages only achieved around 1 BLEU score. Also, since the system hardly translates any of the languages, it is hard to tell the relationship between language similarity and the model's performance. Nevertheless, the relatively low results on the test languages compared with the training languages indicate that cross-lingual embeddings are not rich enough for the model transfer in machine translation. However, when it comes to a random setting with no pre-trained embeddings, we see that the translation model trained with cross-lingual embeddings performs substantially better (Avg BLEU=1.2) than a model trained with random embeddings (BLEU=0.1).

In XNMT, one can also see the individual precision scores form 1 to 4 grams in the translation text. By looking at that details, all three languages had a significantly better unigram precision score than their bigram, trigram, and quadgram precision score. The bigram precision score in Swedish was about half of its unigram scores. The precison score on trigrams and quadgrams are close to 0 on all languages, which again is a sign showing the multilingual NMT system has little transferability from known training languages to an unknown test language.

To increase the transferability from known languages to unknown languages, we have tried various techniques, such as increasing dropout rate. We have observed small improvements (average 0.5 BLEU score increase), but since this technique improves the zero-shot performance at the cost of supervised translation directions Arivazhagan et al., 2019, we decided to explore other approaches in below.

4.1. The Effect of Source/Target Language Annotation

Previous initial results opened up some follow up experiments to see what could be improved. The first improvement is to alter the way the target language annotation in the source sentences, inspired by Blackwood et al., 2018. In the original corpus building script, it will add a custom `__{lang_id}__` token at the front of each source

Language	BLEU	1gram	2gram	3gram	4gram
EN+DE+FR	29.22	0.57	0.34	0.24	0.16
SV	1.12	0.16	0.02	0.00	0.00
HU	1.12	0.18	0.02	0.00	0.00
HE	1.02	0.16	0.02	0.00	0.00

Table 4.1.: Initial results for SV, HU and HE on the baseline system (Target language annotation only, dropout=0.3, trained on mixed language branch corpus.)

Languages	TGT	SRC	Full
EN+DE+FR	29.22	17.59	28.73
SV	1.12	0.00	1.16
HU	1.12	0.00	1.12
HE	1.02	0.00	1.02

Table 4.2.: BLEU scores for different language annotations (Target only, source only and full annotation)

sentence, as suggested by M. Johnson et al., 2016. A sentence in the annotated source text whose target language is German would look like

```
__de__ and we struggle with how to deal with them .
```

Later two other tokens — a single source token and a source token together with a target token, were added into the experiments. Hence a sentence in English would look like this.

```
__en__ and we struggle with how to deal with them .
```

When it needs to be translated into German, the annotation would then become

```
__en__ __de__ and we struggle with how to deal with them .
```

The results are in Table 4.2. Adding a source token together with the target token did not change the overall result much, but removing the target token had a negative impact on the final BLEU score. This discovery is in line with previous claims and results (Blackwood et al., 2018; M. Johnson et al., 2016), as the multilingual NMT system requires a target token at the beginning of each sentence to help identify the target language. The difference between different langauge annotations indicated that a word embedding based multilingual NMT system would also automatically learn the source language during training. People should only annotate the target language into the source text.

4.2. The Effect of Language Similarity

Previous hypothesis believes that Swedish will perform better the Hungarian and Hebrew for its closer relationship to the training languages. To deeper undetstand the question, we have further designed experiments to see if language similarity will improve the results.

The additional experiments will still use Swedish as the test langauge while remove French as the training language to homogenize the training language set more towards Swedish. In the previous training language set, French is the only training language that is Romanic. By replacing French with Danish, all of the training languages are now Germanic, as well as the test language Swedish. We have also included two more Germanic languages as the training language, Dutch (NL) and Norwegian (NO). We began with experiment trained on English, German and Danish, and added the additional training languages one by one in the next two expriments. Everything else is the same. The results are shown in Table 4.3.

As the results shows, the system gained most improvements when Danish and Norwegian were added. Desipte Dutch and Swedish are both Germanic languages, Dutch does not help the multilingual NMT system to learn how to translate from

Language	BLEU	1gram	2gram	3gram	4gram
EN+DE+FR	1.12	0.16	0.02	0.00	0.00
EN+DE+DA	4.1	0.28	0.07	0.02	0.01
+NL	3.3	0.23	0.05	0.02	0.00
+NO	4.7	0.31	0.07	0.03	0.01

Table 4.3.: Results for langauge similarity tested on the Swedish language. Three other Germanic languages DA, NL and NO were added one by one into the training corpus.

Swedish or into Swedish a lot. This confirms that close languages would benefit each other more than distant languages in a multilingual NMT systemn using pre-trained word embeddings (Qi et al., 2018).

Swedish, Danish and Norwegian have deep historical relations to each other. As a result, these languages share many vocabularies as well as grammar and syntax ruls. To study how much of such kind of benefit were brought by shared vocabluaries or their similar syntax, each word in the training corpus was tagged by its source language token to distinguish its origins. Punctuations are not distinguished among languages, which means they don't receive a language-specific token. The word embeddings also tagged to point it to the correct source word. A Swedish sentence that needs to be translated into German is then

```
__de__ <<sv>>och <<sv>>vi <<sv>>kämpar <<sv>>med <<sv>>dem .
```

The assumption behind the word origin token is that, if the result suffers when each word differs by its origin language, the multilingual NMT system would primarily translate by shared vocabularies between language; if its results still holds after the modificaiton, it would primarily learn translation from shared syntax information instead.

The system has obtained 1.7 BLEU score on the EN+DE+FR to SV experiment. It showed that if each word is no longer allowed to be shared between languages the models's performance would dramatically decrease, hence most of the improvements were brought by the fact that Swedish, Danish and Norwegian have a large amount of common vocabluaries. On the other side, it also indicates that the system didn't learn too much syntactic information during training. Even though these languages have similar grammar structures, the system didn't catch it very well, otherwise we would see smaller BLEU score gap between the results as the close grammar relationship will be preserved in the embedding layer. The multilingual NMT system here primarily learns lexicon translation.

4.3. The Effect of the Transformed Vector Space

In addition to the low language similarity between the training languages and the test languages (except in the case of Swedish), we also hypothesis that the poor transferability is due to transformed vector space in the translation model. We believe that after a series of linear opeartion from the neural network onto the word embeddings, the output vector space is no longer aligned with the input vector space.

The whole experiment is based on the hypothesis that our NMT system have already learned the genrally mapping between words in the source vector space and the ones in the target vector space, even though the correct word in the target word space hasn't been seen by the system during training. However since every aligned word embeddings are grouped by their semantics, the correct target word should also be around the wrong output word.

Language	BLEU	1gram	2gram	3gram	4gram
EN+DE+DA → SV	0.65	0.14	0.01	0.00	0.00
SV → EN+DE+DA	6.00	0.33	0.08	0.03	0.01

Table 4.4.: Results for individual translation direction between EN+DE+DA and SV.

By looking closely from the predicted translation in Table 4.1, we have observed the contrary — Almost non of the words in the output text got translated to the correct word in the desired languages. They were either translated into one of the training languages, or were entirely copied directly from the source text, see Appendix A. The BLEU score gains were from punctuations and a small collection of words shared between languages (e.g. property nouns).

Taking a step further, when analyze both translation directions there are other traces to support our transformed vector space hypothesis. We have conducted comparisions on both directions on the Swedish language experiment, as shown in Table 4.4. When tranlsating from SV to the combined EN+DE+DA text, we could achieve almost 6 BLEU scores, which is much better than the nearly zero score when translating from the other direction. Also compared to the combined precision scores on the same experiment from Table 4.3, the results of the translation direction EN+DE+DA to SV contributed almost nothing to the combined translation performance.

Thus, it is highly possible to suspect the output vectors from the model's decoder have been altered and are no longer in the same vector space as the input word embeddings. In this case, the transformed vector space has also made less sense to search for the correct word vector neighbours close to the predicted ouput vector in the output vector space. However, there opens a new possible to translate from completely unknown languages to known languages. We could perform a lexicon replacement based on the Euclidean distance between words in the unknown language and one of the known languages, then feed the processed text into a translation model which has already been trained on known languages. During the whole process, the unknown langauge remains untouched by the translation model hence it still qualifies as zero-resource translation. We will discuss the lexicon replacement process below.

4.3.1. Lexicon Replacement By Euclidean Distance

Suppose we now have a vector space S that contains aligned word embeddings in the unknown language and the known languages, respectively. We donate them as W_x and W_k For each $w_x \in W_x$ there exists at least one mapping to a target word in the known languages $w_k \in W_k$. We are looking for that specific w_k that are with in a specific radius of the original w_s. The distance should still be relatively small so that w_s and w_k are both considered to be a effective translation of each other.

In theory to determine the nearyby neighbour w_k we can use different kinds of metrics. Here we have chosen to use the Euclidean distance where determines the distance between w_s and w_k as

$$d(w_s, w_k) = \sqrt{\sum_{i=1}^{n} (w_{s_i} - w_{k_i})^2} \tag{4.1}$$

The distance d is a variable here and its value needs to be determined as well. Hence there are experiments to test the distance argument d by different experiments, ranging from $d = 0.25$ to $d = 5$.

The algorithm is described in Algorithm 1

Input: hypothesis H, source language embeddings E_s, target language embeddings E_t, distance threshold D
Result: Updated hypothesis H' with words being replaced by their neighbors in the desired language
Build kd-tree T from E_s **for** $l \in H$ **do** each line l in the source hypothesis H
    for $w \in l$ **do** each word w in line l
        if w *is a punctuation* **then**
            skip w;
        else if w *is an unknown word* **then**
            skip w;
        else
            query distance $d(w, w')$ for w in T;
            if $d < D$ **then**
                replace w with the correspoding w'
            end
        end
    end
end

Algorithm 1: Pesudo code for output hypothesis word subsitution. Each word in the NMT output hypothesis that are not in the desired language will be replaced by its cloeset neighbour in that language.

Performing a distance query on a vector space that has more than 3×10^6 vectors is slow, especially when all these vectors are considered to be high dimensional vectors. The code was implemented with SciPy (Virtanen et al., 2019). There are algothrims like KD-tree (Maneewongvatana and Mount, n.d.) that could reduce the calcualtion time for low-dimensional vectors, but for vectors that are higher than 20 dimentions it is not necessarily faster than brutal force.[1] On the other hand, based on the Johnson–Lindenstrauss theorem (W. B. Johnson and Lindenstrauss, 1984), a vector space should have at least more than 300 dimensions to distinguish 1×10^6 vectors in it. As the aligned vector space in fastText contains more than 3×10^6 words, the dimensions could not be compressed any more or you are at risk of not be able to distinguish each word. All in all, the script is slow at subsitute every word in the output hypothesis into the correspoding one in the desired language.

We have performed the lexicon replacement experiment on the SV ↔ EN+DE+DA text from the same TED text corpus (Qi et al., 2018), feeded into an already trained tranlsation model based on the NO ↔ EN+DE+DA corpus, using NO as the pivot language. When applying the previously mention Algorithm 1 on the SV ↔ EN+DE+DA text, we replace all of the source text started without the target langauge token `__SV__`. In other words, we translate all of the source sentence in Swedish to Norwegian first, leaving all of the other source sentences (sentences in EN, DE or DA) untouched. The target translation reference is also remained as is.

The BLEU scores are shown in Table 4.5. In order to demostate the difference of applying the same algorithm on the input and the output vector space, we have also selected result from $d = 0$ to $d = 4$ and compared them with the results when Algorithm 1 was performed on the output vector space. The comparision is in Figure 4.1.

From both Table 4.5 and Figure 4.1, we can see a noticeable improvement over the baseline result as the BLEU score doubled. Even compared with the SV ↔ EN+DE+DA

[1] As descibed on the API document, "High-dimensional nearest-neighbor queries are a substantial open problem in computer science.", https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.html

d value	BLEU	1gram	2gram	3gram	4gram
SV ↔ EN+DE+DA	4.1	0.28	0.07	0.02	0.01
No replacement	2.99	0.27	0.05	0.02	0.00
0.25	2.99	0.27	0.05	0.02	0.00
0.5	2.99	0.27	0.05	0.02	0.00
1	6.18	0.34	0.10	0.04	0.01
2	6.17	0.34	0.10	0.04	0.01
3	6.17	0.34	0.10	0.04	0.01
4	6.17	0.34	0.10	0.04	0.01
0	6.00	0.33	0.08	0.03	0.01

Table 4.5.: Results for the lexicon replacement experiments with different d thresholds. Tested on SV text using NO as the pivot language on the NO ↔ EN+DE+DA translation model. $d = 0$ stands for no threshold control (replace every word).

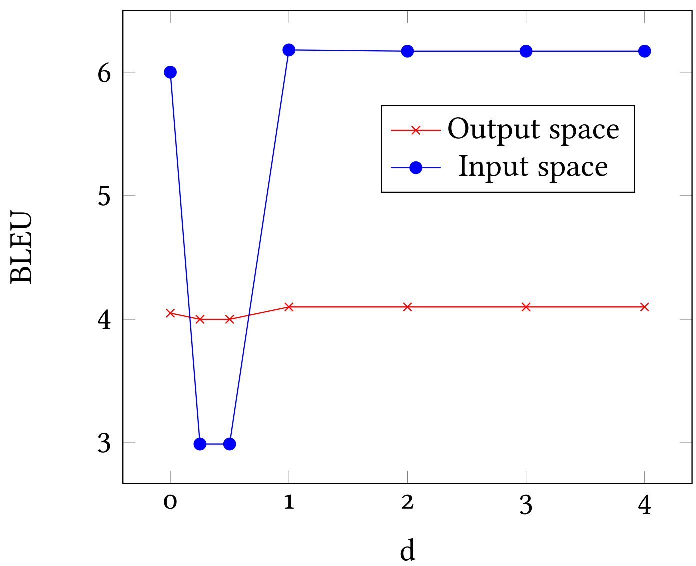

Figure 4.1.: BLEU scores for the lexicon replacement algorithm applied on both the input and the output vector space.

model, the result from the lexicon replacement experiment is still leading ahead for about 2 BLEU scores. Thus it demostrates that our lexicon replacement hypothesis is effective within the input vector space.

Moreover, the value d will affect the translation model. As we increase d from 0.5 to 1 we see a big tranlation quality improvement. On the other side, increasing d after $d = 1$ will have no positive impact on the model's performance. When we set $d = 0$ to remove the threshold control its performance even dropped around 0.2 points, mostly due to the lowered bigram precison score. We conclude that there is a sweet spot for the d value, though its accurate value needs to be fine tuned to adapt different source and pivot language combinations.

Finally, when comparing the results of lexicon replacement on the input vector space and the results on the output vector space, we confirmed the output vector space did changed during the translation tranining process by the model's neural network. Unlike the input vector space, excuting lexicon replacement on the output space does not have a leap on the BLEU score no matter how the distance threshold d changes. Hence it is not worth to perform operations like lexicon replacement on the output vector space.

5. Conclusion and Future Work

From the observation in the last section, we hypothesis that a large amount of error in the output is because the output layer of the model's decoder does not provide any mechanism to transfer the translation knowledge between languages. The layer has one entry per each word in the entire vocabulary set. The entries are activated independently, one at each time. Since the model does not see any example from the unseen test languages during the training phase, the output connections corresponding to the unseen words are down-weighted during the training phase. Hence, it is improbable for the model to output words from the unseen languages, except for those shared with the training languages.

The above discussion suggests that the multilingual NMT architecture of M. Johnson et al., 2016 might transfer better to unseen languages if the decoder and the encoder embeddings are in the same vector spaces. A simple way to reach such an alignment between the two embedding spaces is to add a regularization cost based on the divergence of the two spaces from each other to the loss function of the multilingual model, just like what we have demostrated in our lexicon replacement experiments. We consider this as a future step for this research. Another potential step to be explored in the future is to provide the translation model with information about the language relatedness e.g., to use language embeddings (Littell et al., 2017) together with the cross-lingual embeddings. Moreover, a more in-depth error analysis is required to address the other potential limitations of the multilingual model transfer based on the cross-lingual embeddings.

Finally, we would like to emphasize that this is still an ongoing research and some caveats about the results. We examine the translation transfer on a relatively small set of languages with a more in-depth analysis of only one language. It will be interesting to consider a more extensive language set and study how the model transfer perform in different language families. We have tested only one set of cross-lingual embeddings. Although we believe it is unlikely to see marginally different results from other sets of embeddings (especially when it comes to conventional word vectors), it is still worth exploring how other sets of embeddings (e.g., the multilingual contextualized cross-lingual embeddings Devlin et al., 2019 and the multilingual sub-word embeddings Heinzerling and Strube, 2017) perform in this scenario.

A. Example Output from the Multilingual NMT Model

A.1. Results from the Initial Experiment

This sample output is taken from the SV ↔ EN+DE+FR experiment, described in Section 3.2. Its perfromance result is in Table 4.1.

We haved sampled 20 sentences from the output file. Each of the languages (SV, EN, DE and FR) as the target translation langauge has 5 examples.

Seg. id	Score	Segment comparison: Deletion Insertion Shift
1	211/273= 77%	Src: __sv__ by the end of this year , there 'll be nearly a billion people on this planet that actively use social networking sites . MT: en fin de cette année , il y aura presque un milliard de personnes sur cette planète qui utilisent activement les sites de réseaux sociaux . Ref: i slutet av detta år kommer det finnas närmare en miljard människor på denna planet som aktivt använder sociala nätverk på internet .
2	123/141= 87%	Src: __sv__ the one thing that all of them have in common is that they 're going to die . MT: the one thing that all of them are in common is that they 're going to . Ref: det enda alla dessa människor har gemensamt är att de kommer att dö .
3	208/226= 92%	Src: __sv__ while that might be a somewhat thought , i think it has some really profound implications that are worth exploring . MT: je pense que c' est un peu , je pense qu' il y a des implications très profondes qui sont . Ref: trots att det kan vara en något morbid tanke , tycker jag det har några djupgående konsekvenser som är värda att utforska .
4	269/355= 76%	Src: __sv__ what first got me thinking about this was a blog postearlier this year by derek k. miller , who was a science and technology journalist who died of cancer . MT: ce qui m' a dit , ce qui m' a fait réfléchir , c' était un blog de blog avant cette année par derek 26 miller , qui était un journaliste scientifique et de technologie qui est mort du cancer . Ref: det som först fick mig att tänka på detta var ett blogginlägg från tidigare i år , av derek k miller , en vetenskaps- och teknikjournalist som dog i cancer .
5	175/219= 80%	Src: __sv__ and what miller did was have his family and friends write a post that went out shortly after he died . MT: et ce que miller a fait , il y avait sa famille et amis , écrire un post qui a duré peu après qu' il est mort . Ref: det miller gjorde var att han bad familj och vänner skriva ett inlägg som publicerades kort efter hans död .
6	244/305= 80%	Src: __en__ i slutet av detta år kommer det finnas närmare en miljard människor på denna planet som aktivt använder sociala nätverk på internet . MT: now , i 've been able to the following line of conventional income , which is , which are of the lower planet , including , , internet . Ref: by the end of this year , there 'll be nearly a billion people on this planet that actively use social networking sites .
7	139/163= 85%	Src: __en__ det enda alla dessa människor har gemensamt är att de kommer att dö . MT: now , as common as the , , , , , , , . Ref: the one thing that all of them have in common is that they 're going to die .
8	223/271= 82%	Src: __en__ trots att det kan vara en något tanke , tycker jag det har några djupgående konsekvenser som är värda att utforska . MT: can actually be able to make , making sure that it would be used to think of the , , , . Ref: while that might be a somewhat morbid thought , i think it has some really profound implications that are worth exploring .

9	391/427= 92%	Src:	*__en__ det som först fick mig att tänka på detta var ett blogginlägg från tidigare i år , av derek k miller , en vetenskaps- och som dog i cancer .*
		MT:	now , you know , you know , you know , you know , you know , you know , you know , you know , you know , it 's common about the that i 've been doing right now in terms of that i do right now , and av : dan k kong , he said , " " .
		Ref:	what first got me thinking about this was a blog post authored earlier this year by derek k. miller , who was a science and technology journalist who died of cancer .
10	225/271= 83%	Src:	*__en__ det miller gjorde var att han bad familj och vänner skriva ett inlägg som publicerades kort efter hans död .*
		MT:	so miller , he saw 's 's bad samaritan , and he said , " he said , " he said , " i had an of earlier earlier earlier , by the end of hans .
		Ref:	and what miller did was have his family and friends write a post that went out shortly after he died .
11	194/239= 81%	Src:	*__de__ i slutet av detta år kommer det finnas närmare en miljard människor på denna planet som aktivt använder sociala nätverk på internet .*
		MT:	ich ford , , die , die planet , die die im internet .
		Ref:	zum ende dieses jahres wird es nahezu eine milliarde menschen auf diesem planeten geben , die aktiv soziale netzwerke nutzen .
12	102/106= 96%	Src:	*__de__ det enda alla dessa människor har gemensamt är att de kommer att dö .*
		MT:	.
		Ref:	was sie alle eint , ist , dass sie sterben werden .
13	228/263= 87%	Src:	*__de__ trots att det kan vara en något tanke , tycker jag det har några djupgående konsekvenser som är värda att utforska .*
		MT:	können auch sein , die .
		Ref:	das scheint ein morbider gedanke zu sein , doch ich finde , er liefert einige ziemlich profunde implikationen , die eine nähere betrachtung rechtfertigen .
14	270/328= 82%	Src:	*__de__ det som först fick mig att tänka på detta var ett blogginlägg från tidigare i år , av derek k miller , en vetenskaps- och som dog i cancer .*
		MT:	, die , die ich , gestern derek miller , im .
		Ref:	ich denke darüber nach , seit ich dieses jahr einen blog-eintrag von derek k. miller gelesen habe , der wissenschafts- und technologiejournalist war und der an krebs gestorben ist .
15	184/244= 75%	Src:	*__de__ det miller gjorde var att han bad familj och vänner skriva ett inlägg som publicerades kort efter hans död .*
		MT:	miller machte die schlechte frau , die , die , die ein .
		Ref:	und miller brachte seine familie und freunde dazu , einen eintrag zu schreiben , der kurz nach seinem tod veröffentlicht wurde .
16	265/336= 79%	Src:	*__fr__ i slutet av detta år kommer det finnas närmare en miljard människor på denna planet som aktivt använder sociala nätverk på internet .*
		MT:	j' ai pris la place de la ligne de ligne de la ligne de données , qui les choses en ligne , en gros la planète de la planète , qui l' internet de l' internet .
		Ref:	d' ici la fin de cette année , il y aura près d' un milliard de personnes sur cette planète qui utilisent activement les sites de réseaux sociaux .

17	151/193= 78%	Src:	__fr__ det enda alla dessa människor har gemensamt är att de kommer att dö .
		MT:	la plupart des gens qui ont eu lieu en train de départ , les l' utilisation des de .
		Ref:	la chose qu' elles ont toutes en commun est qu' elles vont mourir .
18	265/313= 85%	Src:	__fr__ trots att det kan vara en något tanke , tycker jag det har några djupgående konsekvenser som är värda att utforska .
		MT:	les peuvent être en , en fait , en fait , en fait , les taux de les d' une certaine façon d' une autre manière .
		Ref:	alors que ça pourrait être une pensée quelque peu morbide , je pense qu' elle a de très profondes implications qui valent d' être explorées .
19	346/419= 83%	Src:	__fr__ det som först fick mig att tänka på detta var ett blogginlägg från tidigare i år , av derek k miller , en vetenskaps- och som dog i cancer .
		MT:	une fois , l' automne ordinaire , en fait , , en train de faire face à l' origine des classiques , des , j' ai parlé , le titre d' un chien , en le chien connu du cancer .
		Ref:	ce qui m' a fait penser à ça au départ est une entrée d' un blog écrite il y a quelques mois par derek k. miller , qui était un journaliste scientifique et technologique qui est mort du cancer .
20	238/286= 83%	Src:	__fr__ det miller gjorde var att han bad familj och vänner skriva ett inlägg som publicerades kort efter hans död .
		MT:	le miller a fait de l' origine des de mauvaises , qui les gens qui ont suivi l' impression d' une étude d' origine de l' origine de hans .
		Ref:	miller a demandé à sa famille et ses amis d' écrire une entrée qui a été mise en ligne peu après sa mort .

A.2. Results from the Language Similarity Experiment

This sample output is taken from the SV $\leftrightarrow$ EN+DE+DA experiment, described in Section 4.2. Its perfromance result is in Table 4.3.

We haved sampled 20 sentences from the output file. Each of the languages (SV, EN, DE and DA) as the target translation langauge has 5 examples.

Seg. id	Score	Segment comparison: Deletion Insertion Shift
1	196/252= 78%	Src: *__sv__ by the end of this year , there 'll be nearly a billion people on this planet that actively use social networking sites .* MT: am ende dieses jahres wird es fast eine milliarde menschen auf diesem planeten geben , die aktiv soziale nutzen . Ref: i slutet av detta år kommer det finnas närmare en miljard människor på denna planet som aktivt använder sociala nätverk på internet .
2	72/119= 61%	Src: *__sv__ the one thing that all of them have in common is that they 're going to die .* MT: det ene af dem alle har i fælles er at de vil dø . Ref: det enda alla dessa människor har gemensamt är att de kommer att dö .
3	157/237= 66%	Src: *__sv__ while that might be a somewhat thought , i think it has some really profound implications that are worth exploring .* MT: mens det måske er en smule tanke , tror jeg , at det har nogle virkelig dybe der er værd at udforske . Ref: trots att det kan vara en något morbid tanke , tycker jag det har några djupgående konsekvenser som är värda att utforska .
4	186/323= 58%	Src: *__sv__ what first got me thinking about this was a blog postearlier this year by derek k. miller , who was a science and technology journalist who died of cancer .* MT: hvad først fik mig til at tænke over dette var en tidligere dette år af derek kenobi miller , som var en videnskab og journalist der døde af kræft . Ref: det som först fick mig att tänka på detta var ett blogginlägg från tidigare i år , av derek k miller , en vetenskaps- och teknikjournalist som dog i cancer .
5	117/214= 55%	Src: *__sv__ and what miller did was have his family and friends write a post that went out shortly after he died .* MT: og hvad miller gjorde var , at hans familie og venner skriver en der gik ud efter , efter han døde . Ref: det miller gjorde var att han bad familj och vänner skriva ett inlägg som publicerades kort efter hans död .
6	169/240= 70%	Src: *__en__ i slutet av detta år kommer det finnas närmare en miljard människor på denna planet som aktivt använder sociala nätverk på internet .* MT: in the early years , it happens to be a of planet as active in the internet . Ref: by the end of this year , there 'll be nearly a billion people on this planet that actively use social networking sites .
7	93/167= 56%	Src: *__en__ det enda alla dessa människor har gemensamt är att de kommer att dö .* MT: it like regular beings who have the evidence that they 're going to be . Ref: the one thing that all of them have in common is that they 're going to die .
8	204/284= 72%	Src: *__en__ trots att det kan vara en något tanke , tycker jag det har några djupgående konsekvenser som är värda att utforska .* MT: the of that can prove a thought , i 'm sure that this has consistent consequences , which represents the of the . Ref: while that might be a somewhat morbid thought , i think it has some really profound implications that are worth exploring .

9	203/322= 63%	Src:	__en__ det som först fick mig att tänka på detta var ett blogginlägg från tidigare i år , av derek k miller , en vetenskaps- och som dog i cancer .
		MT:	what struck me was the of the , the was the in the year , av derek 's miller , a , who was in cancer .
		Ref:	what first got me thinking about this was a blog post authored earlier this year by derek k. miller , who was a science and technology journalist who died of cancer .
10	185/245= 76%	Src:	__en__ det miller gjorde var att han bad familj och vänner skriva ett inlägg som publicerades kort efter hans död .
		MT:	the miller was doing his fear of his bad samaritan , and the meeting of the was the conversation as the beginning of his accident .
		Ref:	and what miller did was have his family and friends write a post that went out shortly after he died .
11	179/257= 70%	Src:	__de__ i slutet av detta år kommer det finnas närmare en miljard människor på denna planet som aktivt använder sociala nätverk på internet .
		MT:	ich habe jahre , die einen auf dem planeten , wie die aktiven des internets .
		Ref:	zum ende dieses jahres wird es nahezu eine milliarde menschen auf diesem planeten geben , die aktiv soziale netzwerke nutzen .
12	90/135= 67%	Src:	__de__ det enda alla dessa människor har gemensamt är att de kommer att dö .
		MT:	es ist die , wie menschen die haben , die sie werden .
		Ref:	was sie alle eint , ist , dass sie sterben werden .
13	232/277= 84%	Src:	__de__ trots att det kan vara en något tanke , tycker jag det har några djupgående konsekvenser som är värda att utforska .
		MT:	, dass es ein , ist , dass es eine solche folgen hat , die die des .
		Ref:	das scheint ein morbider gedanke zu sein , doch ich finde , er liefert einige ziemlich profunde implikationen , die eine nähere betrachtung rechtfertigen .
14	221/322= 69%	Src:	__de__ det som först fick mig att tänka på detta var ett blogginlägg från tidigare i år , av derek k miller , en vetenskaps- och som dog i cancer .
		MT:	es hat mich , die mich , dass die im jahr alt war , derek e. miller , ein als hund im krebs .
		Ref:	ich denke darüber nach , seit ich dieses jahr einen blog-eintrag von derek k. miller gelesen habe , der wissenschafts- und technologiejournalist war und der an krebs gestorben ist .
15	168/248= 68%	Src:	__de__ det miller gjorde var att han bad familj och vänner skriva ett inlägg som publicerades kort efter hans död .
		MT:	das miller machte sich , dass er schlechte , , wie die kurz nach seinem kopf .
		Ref:	und miller brachte seine familie und freunde dazu , einen eintrag zu schreiben , der kurz nach seinem tod veröffentlicht wurde .
16	166/236= 70%	Src:	__da__ i slutet av detta år kommer det finnas närmare en miljard människor på denna planet som aktivt använder sociala nätverk på internet .
		MT:	jeg år efter at en på planet som aktivt på internettet .
		Ref:	når dette år er ovre vil der være næsten én milliard mennesker på denne planet som aktivt bruger sociale netværkssider .
17	91/120= 76%	Src:	__da__ det enda alla dessa människor har gemensamt är att de kommer att dö .
		MT:	det som har de kommer til .
		Ref:	den ting de alle har til fælles er at de alle dør .

18	184/255= 72%	Src:	__da__ trots att det kan vara en något tanke , tycker jag det har några djupgående konsekvenser som är värda att utforska .
		MT:	, at det kan en , så vil det mig konsekvenser som .
		Ref:	selvom det måske er en temmelig morbid tanke , mener jeg at det har nogle meget vidtrækkende konsekvenser som er værd at undersøge .
19	170/306= 56%	Src:	__da__ det som först fick mig att tänka på detta var ett blogginlägg från tidigare i år , av derek k miller , en vetenskaps- och som dog i cancer .
		MT:	det som fik mig til at huske , var i år , derek k miller , en som hund i kræft .
		Ref:	det som først fik mig til at tænke over dette var et blog-indlæg forfattet af derek k. miller tidligere i år , en journalist indenfor videnskab og teknologi , som døde af kræft .
20	112/206= 54%	Src:	__da__ det miller gjorde var att han bad familj och vänner skriva ett inlägg som publicerades kort efter hans död .
		MT:	det miller gjorde var han dårligt , som kort efter hans .
		Ref:	hvad miller gjorde var at få hans familie og venner til at skrive et indlæg som udkom kort efter hans død .

Bibliography

Aharoni, Roee, Melvin Johnson, and Orhan Firat (2019). "Massively Multilingual Neural Machine Translation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3874–3884. DOI: 10.18653/v1/N19-1388. URL: https://www.aclweb.org/anthology/N19-1388.

Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith (2016). "Massively Multilingual Word Embeddings" (Feb. 2016). eprint: 1602.01925. URL: https://arxiv.org/pdf/1602.01925.pdf.

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey (2019). "The Missing Ingredient in Zero-Shot Neural Machine Translation" (Mar. 2019). eprint: 1903.07091. URL: https://arxiv.org/pdf/1903.07091.pdf.

Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho (2017). "Unsupervised Neural Machine Translation" (Oct. 2017). eprint: 1710.11041. URL: https://arxiv.org/pdf/1710.11041.pdf.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate" (Sept. 2014). eprint: 1409.0473. URL: https://arxiv.org/pdf/1409.0473.pdf.

Baziotis, Christos, Barry Haddow, and Alexandra Birch (2020). "Language Model Prior for Low-Resource Neural Machine Translation" (Apr. 2020). eprint: 2004.14928. URL: https://arxiv.org/pdf/2004.14928.pdf.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin (2003). "A Neural Probabilistic Language Model". *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435.

Blackwood, Graeme, Miguel Ballesteros, and Todd Ward (2018). "Multilingual Neural Machine Translation with Task-Specific Attention" (June 2018). eprint: 1806.03280. URL: https://arxiv.org/pdf/1806.03280.pdf.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2016). "Enriching Word Vectors with Subword Information" (July 2016). eprint: 1607.04606. URL: https://arxiv.org/pdf/1607.04606.pdf.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2017). "Word Translation Without Parallel Data" (Oct. 2017). eprint: 1710.04087. URL: https://arxiv.org/pdf/1710.04087.pdf.

Denkowski, Michael and Graham Neubig (2017). "Stronger Baselines for Trustable Results in Neural Machine Translation" (June 2017). eprint: 1706.09733. URL: https://arxiv.org/pdf/1706.09733.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Di Gangi, Mattia and Marcello Federico (2017). “Monolingual Embeddings for Low Resourced Neural Machine Translation”. In: Dec. 2017.

Firat, Orhan, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho (2016). “Zero-Resource Translation with Multi-Lingual Neural Machine Translation” (June 2016). eprint: 1606.04164. URL: https://arxiv.org/pdf/1606.04164.pdf.

Glorot, Xavier and Yoshua Bengio (2010). “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256.

Ha, Thanh-Le, Jan Niehues, and Alexander Waibel (2016). “Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder” (Nov. 2016). eprint: 1611.04798. URL: https://arxiv.org/pdf/1611.04798.pdf.

Ha, Thanh-Le, Jan Niehues, and Alexander Waibel (2017). “Effective Strategies in Zero-Shot Neural Machine Translation” (Nov. 2017). eprint: 1711.07893. URL: https://arxiv.org/pdf/1711.07893.pdf.

Heinzerling, Benjamin and Michael Strube (2017). “BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages” (Oct. 2017). eprint: 1710.02187. URL: https://arxiv.org/pdf/1710.02187.pdf.

Hermann, Karl Moritz and Phil Blunsom (2013). “Multilingual Distributed Representations without Word Alignment” (Dec. 2013). eprint: 1312.6173. URL: https://arxiv.org/pdf/1312.6173.pdf.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2016). “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation” (Nov. 2016). eprint: 1611.04558. URL: https://arxiv.org/pdf/1611.04558.pdf.

Johnson, William B and Joram Lindenstrauss (1984). “Extensions of Lipschitz mappings into a Hilbert space”. *Contemporary mathematics* 26.189-206, p. 1.

Joulin, Armand, Piotr Bojanowski, Tomas Mikolov, Herve Jegou, and Edouard Grave (2018). “Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion” (Apr. 2018). eprint: 1804.07745. URL: https://arxiv.org/pdf/1804.07745.pdf.

Kim, Yunsu, Yingbo Gao, and Hermann Ney (2019). “Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies” (May 2019). eprint: 1905.05475. URL: https://arxiv.org/pdf/1905.05475.pdf.

Kingma, Diederik P. and Jimmy Ba (2014). “Adam: A Method for Stochastic Optimization” (Dec. 2014). eprint: 1412.6980. URL: https://arxiv.org/pdf/1412.6980.pdf.

Lakew, Surafel M., Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi (2019). “Adapting Multilingual Neural Machine Translation to Unseen Languages” (Oct. 2019). eprint: 1910.13998. URL: https://arxiv.org/pdf/1910.13998.pdf.

Lakew, Surafel M., Quintino F. Lotito, Matteo Negri, Marco Turchi, and Marcello Federico (2018). “Improving Zero-Shot Translation of Low-Resource Languages” (Nov. 2018). eprint: 1811.01389. URL: https://arxiv.org/pdf/1811.01389.pdf.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). “Improving Distributional Similarity with Lessons Learned from Word Embeddings”. *Transactions of the Association for Computational Linguistics* 3, pp. 211–225. DOI: 10.1162/tacl_a_00134. URL: https://www.aclweb.org/anthology/Q15-1016.pdf.

Littell, Patrick, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin (2017). “URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 8–14. URL: https://www.aclweb.org/anthology/E17-2002.pdf.

Lu, Yichao, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun (2018). "A neural interlingua for multilingual machine translation". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 84–92. DOI: 10.18653/v1/W18-6309. URL: https://www.aclweb.org/anthology/W18-6309.pdf.

Maneewongvatana, Songrit and David M. Mount (n.d.). "Analysis of approximate nearest neighbor searching with clustered point sets" (). eprint: cs/9901013. URL: https://arxiv.org/pdf/cs/9901013.pdf.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space" (Jan. 2013). eprint: 1301.3781. URL: https://arxiv.org/pdf/1301.3781.pdf.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever (2013). "Exploiting Similarities among Languages for Machine Translation" (Sept. 2013). eprint: 1309.4168. URL: https://arxiv.org/pdf/1309.4168.pdf.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality" (Oct. 2013). eprint: 1310.4546. URL: https://arxiv.org/pdf/1310.4546.pdf.

Neishi, Masato, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda (2017). "A Bag of Useful Tricks for Practical Neural Machine Translation: Embedding Layer Initialization and Large Batch Size". In: *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 99–109. URL: https://www.aclweb.org/anthology/W17-5708.pdf.

Neubig, Graham, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Singh Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang (2018). "XNMT: The eXtensible Neural Machine Translation Toolkit" (Mar. 2018). eprint: 1803.00188. URL: https://arxiv.org/pdf/1803.00188.pdf.

Nguyen, Toan Q. and David Chiang (2017). "Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation" (Aug. 2017). eprint: 1708.09803. URL: https://arxiv.org/pdf/1708.09803.pdf.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://www.aclweb.org/anthology/P02-1040.pdf.

Qi, Ye, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig (2018). "When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation?" (Apr. 2018). eprint: 1804.06323. URL: https://arxiv.org/pdf/1804.06323.pdf.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (2019). "A Survey Of Cross-lingual Word Embedding Models". *JAIR* 65, pp. 569–631. DOI: 10.1613/jair.1.11640. eprint: 1706.04902. URL: https://arxiv.org/pdf/1706.04902.pdf.

Smith, Samuel L., David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla (2017). "Offline bilingual word vectors, orthogonal transformations and the inverted softmax" (Feb. 2017). eprint: 1702.03859. URL: https://arxiv.org/pdf/1702.03859.pdf.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention Is All You Need" (June 2017). eprint: 1706.03762. URL: https://arxiv.org/pdf/1706.03762.pdf.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, and Denis Laxalde (2019). "SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python" (July 2019). DOI: 10.1038/s41592-019-0686-2. eprint: 1907.10121. URL: https://arxiv.org/pdf/1907.10121.pdf.

Vulić, Ivan and Marie-Francine Moens (2013). "A Study on Bootstrapping Bilingual Vector Spaces from Non-Parallel Data (and Nothing Else)". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1613–1624. URL: https://www.aclweb.org/anthology/D13-1168.pdf.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight (2016). "Transfer Learning for Low-Resource Neural Machine Translation" (Apr. 2016). eprint: 1604.02201. URL: https://arxiv.org/pdf/1604.02201.pdf.