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

November 3, 2020

Supervisors:
Ali Basirat, Uppsala University

Abstract

Zero-shot translation is a transfer learning setup that refers to the ability of neural machine translation to generalize translation information into unseen language pairs. It provides an appealing solution to the lack of available materials for low-resource languages by transferring knowledge from high-resource languages.

So far, zero-shot translation mainly focuses on unseen language *pairs* whose individual component is still known to the system. There are fewer reports on transfer learning in machine translation being carried out on completely unknown test languages. This thesis pushes the boundary of zero-shot translation and explores the possibility of transferring learning from training languages to unknown test languages in a multilingual Neural Machine Translation (NMT) system.

Based on the fact that zero-shot translation systems primarily learn language invariant features, we use cross-lingual word embeddings as the only knowledge source since they are good at capturing the semantic similarity of words from different languages in the same vector space. By conducting experiments on an encoder-decoder multilingual NMT model with an attention module, we have examined the relationship of language similarity and the transferability of unseen languages.

We hypothesize that our multilingual NMT model with cross-lingual word embeddings should transfer reasonably even to completely unknown languages. However, we observe little transferability from the training languages to unseen test languages due to the transformed output vector space. Such minor transferability only happens between highly-related languages with a large number of shared vocabularies.

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 owned by UNINETT Sigma2 hosted all of the experiments in this thesis.[1] 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

Multilingual Neural Machine Translation (NMT) aims to train a single translation model for multiple languages (Aharoni et al., 2019; M. Johnson et al., 2016). One of its appealing points is zero-shot translation, which enables translations between unseen language pairs when knowledge from one language is transferred to another language by the model's shared parameters. Even though both source and target languages in such an unseen pair should still be in the set of training languages, a multilingual NMT system with zero-shot learning is still attractive as it lowers the cost of obtaining expensive parallel data, particularly when translating low-resource languages.

The success of zero-shot translation depends on the model's ability to learn language invariant features (Arivazhagan et al., 2019). Kim et al. (2019) believes the embedding layers is one of the critical components responsible for learning such generalized features in a multilingual NMT system. By contrast, Aji et al. (2020) concluded that sharing the embedding layer alone is not enough to transfer learning in zero-shot machine translation. No matter embedding layers are essential to zero-shot learning or not, both researches showed that the embedding layers would positively impact the multilingual model's transferability as long as they are aligned between the source language and the target language.

A special form of such aligned embedding layers is cross-lingual word embeddings, which are often pre-trained and remain frozen during the whole training process. By far, researches on zero-shot learning in multilingual NMT have mostly been restricted to the limited scope of unseen language pairs. A majority of the previous studies of cross-lingual word embeddings in multilingual translation target zero-shot language pairs, not completely unseen languages. People exam on test sets whose languages on both the source or the target side of the translation that are known to the system, but the paired combination remains unknown. For language pairs $A \rightarrow \text{EN}$ and $\text{EN} \rightarrow B$, they are all interested in the unseen language pair $A \rightarrow B$. There is less discussion about the multilingual NMT transferability on completely unknown languages that have never been in the training data.

1.1. Purpose

This thesis studies the importance of word representation in a multilingual NMT transfer model based on pre-trained cross-lingual word embeddings (Ammar et al., 2016; Bojanowski et al., 2016; Joulin et al., 2018; Ruder et al., 2019) and pushes it further than unseen language pairs — examining the transferability of a multilingual NMT system when it is applied to a new test language. Such test language has never been visible to the multilingual NMT system during its training time.

Despite the debate on whether cross-lingual word embeddings are vital when transferring information in zero-shot translation, it is generally acknowledged that cross-lingual word embeddings are beneficial for the model's transferability. Thus we will 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.

Regardless of whether there is transferability to unknown languages on a multilingual NMT system, we would also try to explore the role of language similarity in the transfer learning of unknown languages. As previous results reported (Qi et al., 2018), we anticipated considerable transferability to the unknown language by using cross-lingual word embeddings. If our experiments reproduce such moderate performance, we would like to know how language similarity worked, especially to understand whether similar vocabularies of different languages are essential to transfer learning of unknown languages.

1.2. Structure

The rest of the thesis is organized as below:

Chapter 2 talks about the background and previous works when working in the transferability of cross-lingual word embeddings scope, including information about cross-lingual word embeddings and multilingual NMT. After that, Chapter 3 introduces our experiment method, whose results are discussed and analyzed in Chapter 4. Finally, Chapter 5 gives out our conclusion. We also show sample outputs of our multilingual NMT model in Appendix A.

2. Background

2.1. Word Embeddings

2.1.1. Representing Words by Vectors

In Natural Language Processing, people need to represent words in the forms that are more efficient for computers to process. The idea started with statistical language modeling, which was introduced to machine translation in the early nighties (Brown et al., 1990), followed by Bengio et al. (2003) who powered statistical language modeling with neural networks.

To turn words into vectors, one could use a simple one-hot encoding. For a vocabulary V whose $|V| = n$, each word in the vocabulary can be uniquely represented in an n dimensional vector consists of a single high value one and $n - 1$ low value zeroes. Like in the example of $V = \{\text{water}, \text{hydrogen}, \text{oxygen}\}$ we could make $\vec{w} = [1, 0, 0]$, $\vec{h} = [0, 1, 0]$ and $\vec{o} = [0, 0, 1]$. However, these one-hot vectors cannot capture any latent semantic information between different words, nor to reflect the inflections between stems and their variants.

Recent vectorized word representations (word embeddings) were learned by neural networks. Compared to the naive one-hot vectors, word embeddings contain affluent information that links words together. They are able to transfer the semantical similarity between words to the numerical similarity between vectors. Given two sentences with the target words *oxygen* and *hydrogen*,

> *Oxygen* is a kind of gas.
> *Hydrogen* is also a kind of gas.

a neural network for learning word embeddings would see from the similar context words in both sentences and learn that the two target words *oxygen* and *hydrogen* have similar meanings. Hence it will produce two word embedding $\vec{o}$ and $\vec{h}$ with similar values represented in Figure 2.1.

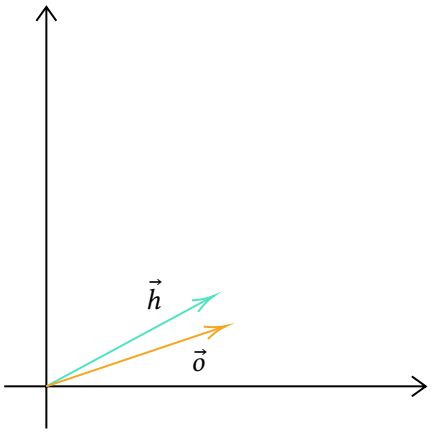

Figure 2.1.: Illustraion of a vector space where $\vec{o}$ = vec(oxygen) and $\vec{h}$ = vec(hydrogen) are learned from two similar sentences.

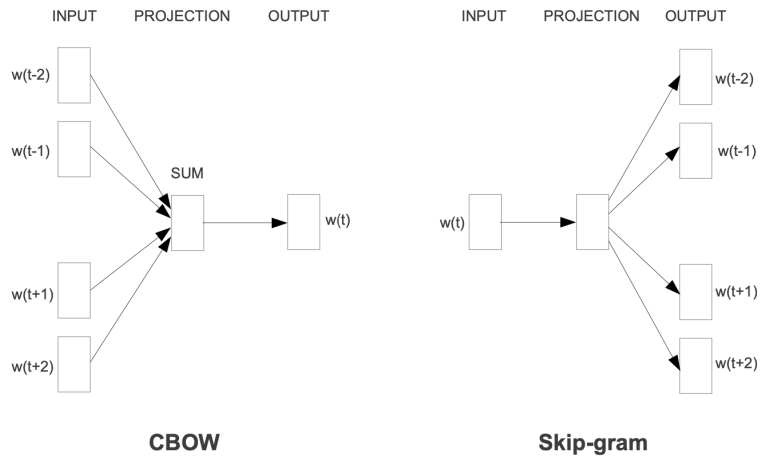

Figure 2.2.: The Skip-gram and the CBOW model (Mikolov, Le, et al., 2013). The Skip-gram model predicts contextual words based on the given center word, while the CBOW model predicts the center word from given context words.

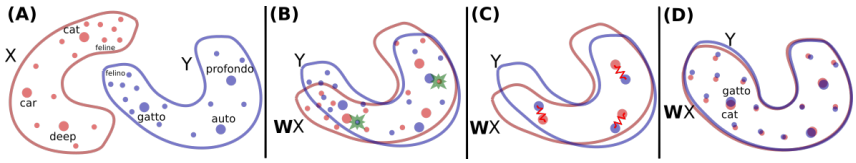

Figure 2.3.: Aligning bilingual vector spaces. (Conneau et al., 2017)

Also, from a geometric prespective, we can observe a small angle between $\vec{h}$ and $\vec{o}$. From the cosine similarity definition,

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

the smaller the angle θ between $\vec{h}$ and $\vec{o}$ is, the higher their cosine similarity is. In other words, when θ is zero, the cosine similarity $\text{sim}(x, y) \in [0, 1]$ will also approach its upper bound one.

One example of such neural networks to learn word embeddings from a large corpus of text is Word2Vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). It learns word representations through a Skip-gram model or a Continuous Bag of Words (CBOW) model. Both models are shown in Figure 2.2 and they have been continuously driving other word embedding algorithms such as fastText (Joulin et al., 2018).[1]

2.1.2. Cross-lingual 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. Multilingual word embeddings aligned in the same vector space are called cross-lingual word embeddings.

[1] https://fasttext.cc/docs/en/unsupervised-tutorial.html

In the multilingual scenario, alignment in two different vector spaces is vital in order to make word embeddings from different languages comparable. Figure 2.3 illustrated the alignment method from Conneau et al. (2017). Suppose there is a set of word pairs $\{x_i, y_i\}$ where $i \in \{1, ..., n\}$, the two vector spaces were aligned by a rotation matrix $W \in \mathbb{R}^{d \times d}$ as shown in process **(B)**, to optimize the objective function

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

Here ℓ is the loss function and it is usually the square loss function $\ell(x, y) = ||x - y||^2$. 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

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

In practice, people use lexicon as the most common parallel data to learn the alignment of multiple vector spaces, though there are other kinds of alignment using sentence or documents level data (Ruder et al., 2019). By using word-level information like a dictionary, we can start with a pivot language (usually English) and map each other monolingual word embeddings by looking the same word up in the dictionary (Mikolov, Le, et al., 2013). Sentence-level parallel data usually contains aligned sentence pairs from different languages (Hermann and Blunsom, 2013). Document-level information is usually a group of topic-aligned or class-aligned documents, such as articles from the same Wikipedia item in different languages (Vulić and Moens, 2013).

Recent development in word alignment has shown possibilities for less supervision and smaller seed data size to initiate the alignment process (Ruder et al., 2019). Though it is still unclear if cross-lingual word embedding alignment can run without parallel data or supervision (Dyer, 2019), there is evidence that even word embeddings from distant language pairs have similar geometric arrangements in vector spaces (Mikolov, Le, et al., 2013).

2.2. Multilingual Neural Machine Translation (NMT) 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. It has outperformed the traditional statistical machine translation in some machine translation settings and has enabled some new possibilities in the field. One of these new applications is multilingual NMT. As shown in Figure 2.4, the system has an encoder and a decoder consisting of several layers of LSTM network spread parallelly on multiple GPUs. An attention module serves as the connection bridge between the encoder and the decoder, emphasizing which part of the source sentence is more relevant to the current translation context especially when translating long sentences (Wu et al., 2016). Multilingual NMT uses the same attentional encoder-decoder model but it is trained on a multilingual corpus with additional artificial tokens to indicate the target language (M. Johnson et al., 2016).

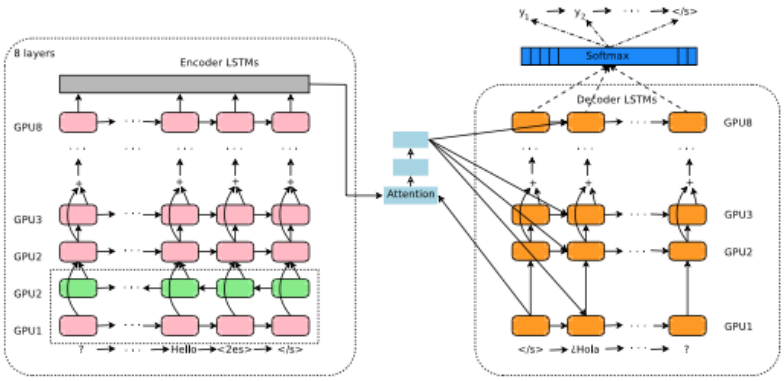

Figure 2.4.: Google's MNMT Architecture (M. Johnson et al., 2016; Wu et al., 2016)

.

The benefit of such a multilingual NMT system does not necessarily stop at better translation performance between common languages like English, French, or Spanish; it also leverages additional information from high resource languages to low resource languages, as a special form of transfer learning (Zoph et al., 2016).

Lakew et al. (2019) claim such form of transfer learning may happen both in the horizontal or vertical direction: In the horizontal direction, knowledge transfers from pre-trained data (such as word embeddings or language models) and is fine-tuned on the test data in known languages; in the vertical direction, knowledge transfers from known language pairs to the test data in an unknown language.

In our experiments, we use information from our training languages and examine its transferability to the unseen test language. We do not apply any fine-tuning based on our test data and keep our test data entirely untouched by the multilingual NMT model. Our Methodology is thus vertical transfer learning, the same as zero-shot machine translation.

2.2.2. Zero-shot Machine Translation

Zero-shot translation stands for translation between language pairs invisible to the multilingual NMT system during the training time. For example, building a multilingual NMT system with German-English and French-English training language pairs while testing 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 includes an encoder, a decoder, and an attention module. It requires no change to a standard NMT system except introducing an additional 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. Translation between unseen language pairs is attractive, especially for low-resource language pairs. Compared with a pivot based system, zero-shot translation eliminates the need for a bridging language, or *interlingua*, as an intermediary of the source and target language. However, zero-shot translation still underperforms than pivot-based translation.

Two reasons 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 the MT system tends to decode the target sentence into the wrong language during inference, 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, making the model difficult to learn to translate to the desired target language (Arivazhagan et al., 2019).

The other potential reason for the worse performance of a zero-shot system is poor generalization. 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. There is no guarantee that the model would discover language invariant representations as there is no explicit incentive to learn language invariant features, resulting in the intermediate encoder representations are too specific to individual languages (Arivazhagan et al., 2019).

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 of the improvements, zero-shot system could achieve better performance than pivot based systems.

2.3. Cross-lingual Word Embeddings in Multilingual NMT

In terms of the application of cross-lingual word embeddings in a multilingual NMT system, there are some successful applications such as using the cross-lingual word 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 an MT system using very little or none parallel data (Conneau et al., 2017) and heavily rely on cross-lingual word embeddings.

Qi et al. (2018) look at cross-lingual word embeddings and their performance in a multilingual setting in detail. They find that cross-lingual word embeddings are useful in a multilingual NMT system, while in bilingual NMT systems, pre-trained word embeddings do not necessarily need to be aligned. Qi et al. (2018) attribute the BLEU score increase to the architecture design of the attentional encoder-decoder multilingual NMT system. As such multilingual NMT system uses a single encoder for all source languages, alignment in the word embeddings condenses various vector spaces of the source language into a unified vector space, helping the system learn a much-simplified transform from its input vector space to its output vector space.

3. Methodology

3.1. Experiment Settings

We chose English (EN), German (De), and French (FR) to be the training languages as these three languages are all considered to be high-resource languages. Selecting them as our training languages would enable a wider choice of available training resources for our experiments.

Let Z denote the set of corresponding candidate languages, the training language set is then Z_{TRAIN} = EN ∪ DE ∪ FR. For the test language set, we picked up Swedish (SV) as a low-resource language representation within the same language family of our training languages; Hungarian (HU) and Hebrew (HE) as another two low-resource language examples that do not belong to the same language group as our training languages. Therefore Z_{TEST} = SV ∪ HU ∪ HE.

For each experiment, we trained a basic multilingual NMT system using a training corpus C_{TRAIN} with all three training languages, including all six combinations from the cartesian product without duplicates. Here x and y are both training languages in the training set C_{TRAIN}.

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

During testing, 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 combinations of three different training language and that only test language.

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

We use BLEU score (Papineni et al., 2002) as one of our evaluation metrics, together with 1-4 word gram precision scores to better understand if the transferability only happens at the word-level or at the phrase-level.

3.1.1. Corpus and Preprocessing

we use the TED talk subtitle corpus from Qi et al. (2018) to train the multilingual NMT.[1] The whole corpus is split into three parts: train, dev and test, at the ratio of 0.95 : 0.025 : 0.025.

For preprocessing, since the original TED corpus is already tokenized by Moses (Koehn et al., 2007), we did not add additional tokenization steps. We turned all of the text into lower cases and applied a sentence length filter to remove all long sentences with more than 60 words. After that, when building the index to word (i2w) and the word to index (w2i) table for the pre-trained embeddings, we have also removed words that are less frequent than two times to prevent the system from overfitting to rare words.

All of the preprocess functions are built upon the built-in XNMT preprocess features (Neubig et al., 2018). Table 3.1 contains the corpus size for all of the training and test language sets after proprocessing.

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

Language	train	dev	test
EN+DE+FR	1013478	N/A	N/A
+SV	N/A	9390	12423
+HU	N/A	20332	25606
+HE	N/A	24554	28546
EN+DE+DA	491537	N/A	N/A
+SV	N/A	8037	9344
EN+DE+DA+NL	1225511	N/A	N/A
+SV	N/A	11126	13378
EN+DE+DA+NL+NO	1322133	N/A	N/A
+SV	N/A	12304	14430

Table 3.1.: Number of sentences in each language combination after preprocessing

3.1.2. Neural Network

Our neural network is a modified version of the one from Qi et al. (2018), which is built by XNMT (Neubig et al., 2018). The only change is doubling the encoding layer to a 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: we set the beam size to 5, the batch size to 32 and the dropout rate to 0.1. The optimizer used in our experiments is the Adam Optimizer (Kingma and Ba, 2014). The initial learning starts at 0.0002 and decays by 0.5 when the development BLEU score decreases (Denkowski and Neubig, 2017).

3.1.3. Embeddings

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

fastText is an extension of the original Word2Vec methods that uses sub-words to augment low-frequency and unseen words. It also has an extensive collection of pre-trained embeddings for multiple languages out of the box, making it attractive to cross-lingual word embedding experiments like ours, since researchers can save effort on training and aligning cross-lingual word embeddings while concentrating on experiments themselves.

Each of the fastText word embedding file is language-specific and contains word embeddings of 300 dimensions. We concatenated different language files to build up our cross-lingual word embedding files for the multilingual NMT system. The embeddings are frozen during the whole training to remain the original alignment throughout the whole experiments. If there is a shared word with two different vector values in different language embedding files, both vectors' average value will be used.

There will also be a different attempt in our experiments where the system treats each word as a unique word even though they might have the same spelling. Both of the results will be available below in Section 4.1.

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

4. Results and Analysis

Table 4.1 shows our initial result from a multilingual machine translation model trained on the combined EN+DE+FR corpus. Swedish, Hungarian and Hebrew all got unanticipated low BLEU scores since all of the three languages only achieved around 1 BLEU score. Unlike the reported results from Qi et al. (2018), even though Swedish is closer to the training languages, the expected high performance based on its high similarity to the training languages did not appear in our experiment result. 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. However, when it comes to randomly initialized embedding layers without pre-trained embeddings, cross-lingual word embeddings still provides better initialization than random settings. We see the translation model trained with cross-lingual embeddings performs substantially better (Avg BLEU=1.2) than a model trained with randomly initialized embeddings (BLEU=0.1).

By looking at individual 1-4 gram precision scores, all three languages had a significantly better unigram precision score than their bigram, trigram, and quadgram precision score. For example in Swedish, its bigram precision score in Swedish was about half of its unigram scores. The precision score on trigrams and quadgrams are close to zero 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, and that transferability only happens at word-level.

We have tried increasing the dropout rate to 0.3 and observed small improvements (average 0.5 BLEU score increase). As Arivazhagan et al. (2019) have pointed out, this technique improves the zero-shot performance at the cost of supervised translation directions. Thus we decided to explore other approaches below.

4.1. The Effect of Language Similarity

In the above results from Table 4.1, even though all of the BLEU scores from three test languages are relatively low, Swedish achieves the best result. To better understand whether and how Swedish benefits from its language similarity to our training languages, we have further designed experiments to see the effect of language similarity.

The additional experiments will still use Swedish as the test language while removing 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

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.)

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
+NL+NO	4.7	0.31	0.07	0.03	0.01

Table 4.2.: Results for language similarity tested on the Swedish language. Three other Germanic languages DA, NL and NO were added one by one into the training corpus.

are now Germanic, as well as the test language Swedish. We have also included two more Germanic languages to the training language set, Dutch (NL) and Norwegian (NO). We began with an experiment trained on English, German and Danish, and added the additional training languages one by one in the next two experiments. Everything else is the same. The results are shown in Table 4.2.

As the results show, the system gained the most improvements when Danish and Norwegian were added. Although Dutch does not help the multilingual NMT system learn how to translate from Swedish or into Swedish a lot, compared with the original baseline EN+DE+FR experiment, adding Dutch in the training language still increases the BLEU score by more than 2 points. The result from our language similarity experiments confirms that close languages would benefit each other more than distant languages in a multilingual NMT system using pre-trained cross-lingual word embeddings (Qi et al., 2018).

Swedish, Danish and Norwegian have deep historical relationships with each other. Therefore, these languages share many vocabularies as well as grammar and syntax rules. To study how much of such kind of benefit was brought by shared vocabularies or similar syntax, we conducted a further experiment by differentiating the word origins: tagging each word in the training corpus by its source language token to distinguish its origin. Punctuations are not distinguished among languages, which means they do not receive language-specific tokens. Word embeddings are also tagged to point to the correct source words. 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 languages; if its results still hold after the modification, it would primarily learn translation from information other than shared vocabularies.

The system has obtained a 1.7 BLEU score on the EN+DE+DA to SV experiment. It showed that if each word is no longer allowed to be shared between languages, the models' performance would dramatically decrease. Hence most of the improvements were brought by the fact that Swedish, Danish and Norwegian have a large number of common vocabularies. On the other side, it also indicates that the system did not learn too much non-vocabulary information during training, e.g., similar grammar structures. Otherwise, we would see a smaller BLEU score gap between the results as such non-vocabulary information will be preserved in the embedding layer. We conclude that our multilingual NMT system here primarily learns lexicon translation.

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

4.2. The Effect of the Transformed Vector Space

In addition to the role of language similarity between the training languages and the test languages, we also hypothesize that our multilingual NMT model's poor transferability to unseen languages is due to transformed vector space in the translation model. After a series of linear operations from the neural network onto the word embeddings, the output vector space is no longer aligned with the input vector space.

Before our experiments, we predicted that our NMT system should have learned the generally mapping between words in the source vector space and the ones in the target vector space, even though the system has not seen the correct word in the target word space during training. However, by looking closely at the output translation in Table 4.1, we have observed the contrary — almost none of the words in the output text are translated to the correct word in the desired languages. They were either incorrectly translated into one of the training languages, or were entirely copied from the source text. 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 analyzing both translation directions, there are other traces to support our transformed vector space suspicion. We have conducted comparisons in both directions on the Swedish language experiment, as shown in Table 4.3. When translating 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 4.2, the results of the translation direction EN+DE+DA → SV contributed almost nothing to the combined translation performance. We present the sample output of both directions in Appendix A.1.

Thus, the model's decoder's output vectors may 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 neighbors close to the predicted output vector in the output vector space. However, there opens a new possibility to translate from completely unknown languages to known languages — We could perform lexicon replacements for words in the unknown language by other words in one of the known languages based on their Euclidean distance between, then feed the processed text into a translation model that has already been trained on known languages. During the whole process, the unknown language remains untouched by the translation model, therefore it still qualifies as zero-shot translation.

4.2.1. Lexicon Replacement By Euclidean Distance

Suppose we have a vector space S that contains cross-lingual 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 is within a defined radius of the original w_s. The distance should still be relatively small so that w_s and w_k are considered an adequate translation of each other.

In theory, to determine the nearby neighbor 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: Translation hypothesis H, source language embeddings E_s, target language embeddings E_t, distance threshold D
Result: Updated translation 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 substitution. Each word in the NMT output hypothesis that is not in the desired language will be replaced by its closest neighbor in that language.

Performing a distance query on a vector space with more than 3×10^6 vectors is slow, especially when all these vectors are high dimensional vectors. Our code was implemented with SciPy (Virtanen et al., 2019). There are algorithms like KD-tree (Maneewongvatana and Mount, n.d.) in SciPy that could reduce the calculation time for low-dimensional vectors, but for vectors higher than 20 dimensions, 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 anymore, or we are risking of not being able to distinguish each word. After all, the script is slow when substituting every word in the output hypothesis into the corresponding one in the desired language.

[1] As described 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 4.4.: 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).

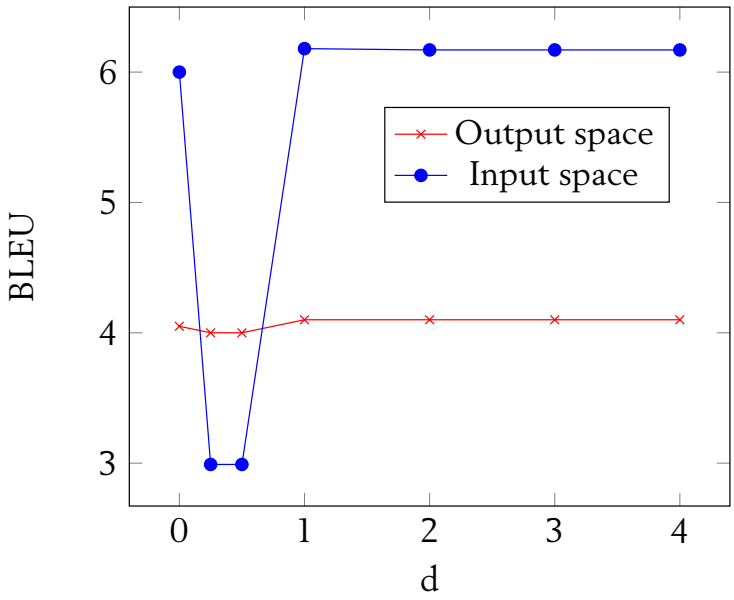

Figure 4.1.: BLEU scores for the lexicon replacement algorithm applied on both the input and the output vector space.

We have performed the lexicon replacement experiment on the SV ↔ EN+DE+DA text from the same TED text corpus (Qi et al., 2018), fed into an already trained translation 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 process all of the sentences in the source text without the target language token `__SV__`. In other words, we translate all of the source sentences in Swedish to Norwegian first, leaving all of the other source sentences in EN, DE and DA untouched. The target translation reference also remains as is.

We report our BLEU scores of the lexicon replacement experiment in Table 4.4. To demonstrate the difference of applying the same algorithm on the input and the output vector space, we have also selected results from $d = 0$ to $d = 4$ and compared them with the results when Algorithm 1 was performed on the output vector space. The comparison is in Figure 4.1.

From both Table 4.4 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 model, the lexicon replacement experiment's result still leads ahead

for about 2 BLEU scores. Our lexicon replacement hypothesis is effective within the input vector space.

Moreover, the value of d will affect the translation model. As we increase d from 0.5 to 1, we see a significant translation 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 dropped around 0.2 points, mostly due to the lowered bigram precision score. We conclude that there is a sweet spot for the d value and its accurate value needs to be fine-tuned to adapt to different source and pivot language combinations.

Finally, when comparing the results of lexicon replacement on the input vector space and the output vector space results, we confirmed the output vector space did change by the model's neural network during its training process. Unlike the input vector space, executing 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 performing operations like lexicon replacement on the output vector space.

5. Conclusion and Future Work

In this thesis, we have explored the transferability and generalizability of cross-lingual word embeddings on unknown languages. From the achievement of zero-shot machine translation, we took a step further and expected moderate performance from those word embeddings as if they would transfer knowledge learned from training languages onto other completely unknown languages. However, our experiment results suggested that only slight knowledge transfer happened between closely related languages, which echoes back to some of the findings that the embedding layer along in a multilingual NMT system is not enough to handle the knowledge transfer (Aji et al., 2020).

Despite the weak performance, we analyzed whether language similarity is vital to transfer learning on unknown languages. We agree with the conclusion of Qi et al. (2018) that language similarity is the main factor of transferability between languages. When transferring knowledge to unknown languages, we only observe transferability between highly related languages, particularly between languages that share a large portion of their vocabularies, i.e., Swedish, Danish, and Norwegian. The multilingual NMT model in our settings primarily learns word-level translation. Without the help of similar words, the little transferability would disappear too.

To increase cross-lingual word embeddings' transferability in a multilingual NMT architecture similar to M. Johnson et al. (2016), there should be some additional alignment in the output vector space between the source and the target languages. During the training process, the neural network has never seen any positive examples from the test language. Therefore its output weight on the test language has been continuously deducted, which resulted in a transformed output vector space. Such output space is no longer aligned with the input vector space, hence connections between different languages solely rely on shared vocabularies. As a result, only very similar languages such as Swedish, Norwegian and Danish benefited in our experiments.

As we have demonstrated in our lexicon replacement experiments, a solution to align the deviated input and output embedding spaces is to add regularization to the multilingual model's loss function based on the two vector spaces' divergence. Another potential solution to be explored is to supplement the translation model with language level information such as language embeddings (Littell et al., 2017; Malaviya et al., 2017) together with the cross-lingual embeddings. Language level information could also answer more questions remained in our study. Together with some previous studies (Aji et al., 2020; Qi et al., 2018), we believe the transferability of cross-lingual word embeddings is related to the similarity between the source and the target language, but how to measure the language similarity and link it to the transferability of the embedding layers could be an exciting topic.

Finally, it is worth exploring how other sets of embeddings would enhance cross-lingual word embeddings' transferability. For example, to try the multilingual contextualized cross-lingual embeddings (Devlin et al., 2019) and see if it would benefit the transferability by adding contextual information, or the multilingual sub-word embeddings (Heinzerling and Strube, 2017) to see if it would perform better by aligning sub-words across different languages.

A. Example Output from the Multilingual NMT Model

A.1. Example Output from the Directional Translation Experiment

This sample output is taken from the SV ↔ EN+DE+DA experiment, described in Section 4.1. Its perfromance result is in Table 4.2.

We have sampled 20 sentences from the output file. Each of the languages (SV, EN, DE and DA) as the target translation language has five 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.

Aji, Alham Fikri, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich (2020). "In Neural Machine Translation, What Does Transfer Learning Transfer?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7701–7710. DOI: 10.18653/v1/2020.acl-main.688. URL: https://www.aclweb.org/anthology/2020.acl-main.688.pdf.

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.

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.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (1990). "A Statistical Approach to Machine Translation". *Computational Linguistics* 16.2, pp. 79–85. URL: https://www.aclweb.org/anthology/J90-2002.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.

Dyer, Andrew (2019). "Low Supervision, Low Corpus size, Low Similarity! Challenges in cross-lingual alignment of word embeddings : An exploration of the limitations of cross-lingual word embedding alignment in truly low resource scenarios". MA thesis. Uppsala University, Department of Linguistics and Philology, p. 53.

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.

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.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: As-

sociation for Computational Linguistics, June 2007, pp. 177–180. URL: https://www.aclweb.org/anthology/P07-2045.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.

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.

Malaviya, Chaitanya, Graham Neubig, and Patrick Littell (2017). "Learning Language Representations for Typology Prediction". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2529–2535. DOI: 10.18653/v1/D17-1268. URL: https://www.aclweb.org/anthology/D17-1268.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.

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.

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.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, and Jason Smith (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation" (Sept. 2016). eprint: 1609.08144. URL: https://arxiv.org/pdf/1609.08144.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.