Cross-lingual Word Embeddings beyond Zero-shot Machine Traslation

Shifei Chen
Department of Linguistics and Philology
Uppsala University
Shifei.Chen.2701@student.uu.se

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

1 Introduction

For translating low resource languages in Neural Machine Translation (NMT), cross-lingual word embeddings have shown their potential in this task. They could either be used as external auxiliary information (Conneau et al., 2017; Lakew et al., 2019), or as the embedding layer directly (Neishi et al., 2017; Artetxe et al., 2017). Furthermore, with Multilingual NMT, it is possible to leverage hidden information from high resource languages to low resource languages.

Learned from approaches like the Skip-gram model or the CBOW model, vectorized word representations tend to cluster words with similar semantics. 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. It is then interesting to see how far can aligned word embeddings could go beyond known languages. Zero-shot translation analyzed this question by testing the Multilingual NMT system on unseen language pairs — language in either 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.

Translating a completely unseen language 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 rest of 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 have conducted some preliminary experiments below.

2 Multilingual Neural Machine Translation

Multilingual Neural Machine Translation (MNMT) enables translation between multiple languages in the same encoder-decoder attentional model from NMT (Johnson et al., 2016; Ha et al., 2016). The only modification to NMT is introducing a target language indicator at the beginning of each training sentence. Despite its simplicity over maintaining several bilingual NMT systems, it also enables zero-shot translation – translating between an unseen language pair during the training time, which benefits low-resource languages by transferring knowledge from their high-resource relatives.

3 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 et al., 2013). Its ability to represent lexicons from several different languages in a shared cross-lingual space empowers cross-lingual transfer by providing word-level links between languages (Ruder et al., 2019).

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

Training Corpus	Test Corpus
EN+DE+FR	EN/DE/FR ↔ SV
EN+DE+FR	EN/DE/FR ↔ HU
EN+DE+FR	EN/DE/FR ↔ HE

Table 1: Experiment Settings

Language	BLEU	P@1	P@2	P@3
EN+DE+FR	29.22	57.30	34.06	24.09
SV	1.48	16.36	2.32	0.61
HU	1.12	17.65	1.65	0.44
HE	1.02	15.83	1.70	0.37

Table 2: Initial results for SV, HU and HE

we have conducted some preliminary experiments below.

4 Methodalogy

We picked up the training and test languages based on the following aspects:

- Language Similarity
- Shared Alphabets
- Word Order

We chose English (EN), German (De), and French (FR) to be the training languages. We trained a basic MNMT system using the training corpus of all three training languages for each experiment, including all six directions from the cartesian product without duplicates. The test languages are Swedish (SV), Hungarian (HU), and Hebrew (HE). All of the experiment languages combinations are shown in Table 1.

4.1 Experiment Settings

4.2 Corpus and Preprocessing

The author have used the TED talk subtitle corpus from Qi et al. (2018) [1] to train the MNMT. We removed sentences that are longer than 60 words for preprocessing, and less frequent words that appeared only once.

4.3 Neural Network

The neural network is a modified version of the one from Qi et al. (2018) which was built upon XNMT (Neubig et al., 2018). The only change is doubling the encoding layer to a 2-layer-bidirectional LSTM network in order 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).

4.4 Embeddings

The embeddings used in the experiments are fastText aligned word embeddings[2]. We concatenated different language files to build up multilingual word embedding files for the MNMT 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 $\vec{v_{mean}}$ will be the new vector.

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. We have also tried to distinguish every word by its origin, the result turned out to be much worse. Hence we will continue by using the averaged $\vec{v_{mean}}$.

5 Prelimentary Results

In the results shown in Table 2, all of the three languages got low BLEU scores in contrast with the baseline experiment. Thus we suspect that the slightly better result from Swedish was primarily due to the high similarity between Swedish and the three training languages. Besides, the low BLEU scores across all three experiments might be attributed to the fact that the output vector space has already been changed during the training process. It is no longer aligned with the input word embedding space.

5.1 Language Similarity

To validate our first assumption, we continue our experiment with more languages within the same Germanic languages branch. We have removed French and added Danish (DA), Dutch (NL), and

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

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

Language	BLEU	P@1	P@2	P@3
EN+DE+DA	4.05	28.17	6.56	2.07
above+NL	3.26	23.20	4.84	1.65
above+NO	4.69	31.04	7.41	2.50

Table 3: Results for langauge similarity. Three other Germanic languages DA, NL and NO were added one by one into the training corpus)

Norwegian (NO) as the training language one by one to see how language similarity would affect the Swedish result.

As shown in Table 3, the results confirm our assumption as the BLEU score generally grows when we add more similar languages into the training set. More specifically, we believe it is the large shared vocabulary between Swedish, Danish, and Norwegian that drove the performance increase. When we distinguish each word with its language origin, the result dropped to 1.66 BLEU score again (for the EN+DE+DA experiment, tested on language SV).

5.2 Transformed Vector Space

Furthermore, we have observed a large amount of error in the output was due to the incorrect output language. We designed our translation system on the hypothesis that it would learn the general 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, since their semantics groups every aligned word embeddings, the correct target word should also be around the wrong output word, as long as the input and output space is the same. We performed a target word replacement experiment where each target word in the wrong language w_t will be replaced by its nearest neighbor w'_t in the correct language, based on their Euclidean distance d

$$d(w_t, w'_t) = \sqrt{\sum_{i=1}^{n} (w_{t_i} - w'_{t_i})^2} \qquad (1)$$

The distance d is a variable here, and its value needs to be determined as well. Hence I have chosen to test the distance argument d by different experiments, ranging from $d = 0.25$ to $d = 4$. Results in Table 4 are based on the output from the previous EN+DE+DA experiment. Baseline stands for no substitution.

d	BLEU
0.25	4.05
0.5	4.05
1	4.12
2	4.12
3	4.12
4	4.12
baseline	4.05

Table 4: Initial results for SV, HU and HE

From Table 4 we saw minor improvements over the original translation output. The 0.07 BLEU score showed that our target word replacement is not effective. It also confirmed that the output vector space has already been transformed into one another and is no longer aligned to the original input word embedding space. Hence, to build a Multilingual NMT system for completely unseen languages using aligned word embeddings, we need to explore more sophisticated linear transformation on the output vector space in the future.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder.

Melvin Johnson, 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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Surafel M. Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Adapting multilingual neural machine translation to unseen languages.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation.

Masato Neishi, 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)*, pages 99–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Graham Neubig, 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.

Kishore Papineni, 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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ye Qi, 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?

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *JAIR*, 65:569–631.