Cross-lingual Word Embeddings beyond Zero-shot Machine Translation

Shifei Chen
Dep. of Linguistics and Philology
Uppsala University
shifei.chen.2701@student.uu.se

Ali Basirat
Dep. of Linguistics and Philology
Uppsala University
ali.basirat@lingfil.uu.se

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.

1 Introduction

The multilingual neural machine translation (NMT) aims at training a single translation model between multiple languages (Johnson et al., 2017; Aharoni et al., 2019). 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 (Johnson et al., 2017). 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 (Bojanowski et al., 2017; Ammar 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 Johnson et al. (2017) 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 preliminary 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.

2 Related 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 (Johnson et al., 2017; Lu et al., 2018). The practical importance is due to their effective use for the translation of low-resource languages (Zoph et al., 2016; Nguyen and Chiang, 2017). Johnson et al. (2017) shows that multilingual NMT trained on a massive training set can generalize reasonably well

to 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.

Cross-lingual word embeddings are important for the translation of low-resource languages. They could either be used as external auxiliary information (Lample et al., 2018; Lakew et al., 2019), or as the embedding layer directly (Neishi et al., 2017; Artetxe et al., 2017). Qi et al. (2018) explored how effective are 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 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 languages that are entirely unseen to the multilingual NMT setting.

3 Multilingual NMT

Multilingual Neural Machine Translation (NMT) (Johnson et al., 2017) extends the attentive encoder-decoder framework of NMT (Bahdanau et al., 2015) to translate between multiple languages simultaneously. Instead of training multiple bilingual NMT models, the multilingual NMT augments the input representations with a language indicator at the beginning of each training sentence, to be trained in an end-to-end fashion. Despite its simplicity, the multilingual NMT has shown significant improvement to the machine translation (Aharoni et al., 2019) and provides for 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 (Zoph et al., 2016; Nguyen and Chiang, 2017).

4 Cross-lingual Word Embeddings

Cross-lingual word embeddings represent lexicons from several different languages in a shared embedding space, providing for the word-level language transfer models and the cross-lingual study of words. Ruder et al. (2019) collects a survey of different approaches used to train cross-lingual word embeddings. One of the main approaches is to find a mapping between monolingual embeddings spaces based on a seed dictionary. The fastText cross-lingual embeddings are among the extensively used resources trained in this way, using the monolingual word embeddings of Bojanowski et al. (2017) and the mapping approach of Joulin et al. (2018).

5 Experiments

We study the transferability of multilingual NMT based on two disjoint sets of languages to train and test a translation model. We interpret the average BLEU score (Papineni et al., 2002), and individual accuracy scores on 1-3 grams of both translation directions from the test languages, as the transferability measure of the translation model to the unseen test languages. That is to say; we consider high values of BLEU score and individual 1-3 gram accuracy as the goodness of the model transfer from the training languages to the test languages. The test languages remain unseen to the translation model during the training phase.

We choose English (EN), German (DE), and French (FR) as the training languages together with Swedish (SV), Hungarian (HU), and Hebrew (HE) as the test languages. The languages are selected to analyze the effect of language similarity to the model transferability. Among the test languages, Swedish is the most similar one to the training languages. Hence, it is expected to obtain a higher BLEU score on Swedish than the two other test languages if language similarity plays any role in the models' transferability. On the other hand, Hebrew should get the lowest score since it is more different from the training languages.

We used the TED talk subtitle corpus (Qi et al., 2018) to train, validate, and test the multilingual NMT model.[1] We downcase all letters in the corpora and removed sentences longer than 60 words

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

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

Table 1: Number of sentences in each language combination after preprocessing

as well as less frequent words that appeared only once. Table 1 shows the corpus size of each language combinations after prepossessing. Each translation model is trained on the training data. The development data is used for early stopping the training procedure and the the test data is used to measure the transfer quality on the test languages.[2] Our 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 we have made is doubling the encoding layer to a 2-layer-bidirectional LSTM network in order to accommodate the additional information in the multilingual scenario. Everything else is the same as the original experiment settings, including the encoder-decoder model with attention (Bahdanau et al., 2015), beam search size of 5, batches of size 32, dropout at 0.1, the Adam optimizer (Kingma and Ba, 2015). The initial learning starts at 0.0002 and decays by 0.5 when the BLEU score on the development set decreases (Denkowski and Neubig, 2017).

The fastText word embeddings are used as the source of the knowledge transfer across languages. [3] We concatenated the embeddings of all languages in both the training and test sets into a single file, and we used it to initialize the embedding layers of the NMT model. The embedding layers of the networks are kept frozen during the training to preserve the training embeddings in their original embeddings space, the same space as the test embeddings. We use all embeddings with no change except those related to the words with the same form in multiple languages. The vector of these words is set to the mean vector of the individual word vectors in each language.

[2] We have not seen a significant difference between using and not using the development set for early stopping.

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

Language	BLEU	1gram	2gram	3gram
EN+DE+FR	29.2	0.57	0.34	0.24
SV	1.5	0.16	0.02	0.00
HU	1.1	0.17	0.02	0.00
HE	1.0	0.16	0.02	0.00

Table 2: The translation performance on the training languages (EN+DE+FR), and and each of the test languages.

Language	BLEU	1gram	2gram	3gram
EN+DE+FR	1.5	0.16	0.02	0.00
EN+DE+DA	4.1	0.28	0.07	0.02
+NL	3.3	0.23	0.05	0.02
+NL+NO	4.7	0.31	0.07	0.03

Table 3: The transfer results to Swedish.

6 Preliminary Results

Table 2 summarizes the translation performance on the test division of the training languages altogether (EN+DE+FR) and each of the test languages (SV, HU, and HE). 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).

The slightly better result on Swedish suggests that the model transfers better to similar languages. We continue our experiment with more languages to further study the effect of language similarity. For this purpose, we homogenize the training languages more toward Swedish by excluding French and adding Danish (DA), Dutch (NL), and Norwegian (NO) one-by-one to the training set.

Table 3 summarizes the transfer results from each of the training settings to Swedish. We see that the homogeneous setting performs better than the other setting that includes French in the training set. More specifically, when adding DA nad NO to the training language, both the BLEU score and individual accuracy scores are improved, while

adding NL to the training language worsened the results. We speculate that the large shared vocabulary between Swedish, Danish, and Norwegian drives the performance increase. When we distinguish each word with its language origin, the result dropped to 1.7 BLEU score again (for the EN+DE+DA experiment, tested on language SV). This observation further strengthens the effect of the shared vocabulary on transfer learning.

6.1 Discussion

Our preliminary results on the transferability of multilingual NMT models to unseen languages show that these models can transfer weakly to completely unseen languages if the transfer learning is grounded on the cross-lingual word embeddings. One reason for the weak translation transfer across languages is that the output space of the model's decoder is not aligned to the cross-lingual embeddings' space. This is because of the many transformations applied to the input vectors during the translation process. It can be seen if we compare the BLUE score for each side of the translations, say SV→EN+DE+DA versus EN+DE+DA→SV. We get a relatively higher BLEU score of 6.0, when translating from the unseen language (SV→EN+DE+DA), but almost no translation (BLUE=1.0) is performed on the opposite direction when translating to the test language (EN+DE+DA→SV).

Furthermore, 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 Johnson et al. (2017) 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. 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, either to use language embeddings (Littell et al., 2017) together with the cross-lingual embeddings, or to constrain the model's output to the desired target language by re-aligning the output vector space back to the input vector space. 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 on an attention-based encoder-decoder NMT architecture. 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, 2018) perform in this scenario. It will also be interesting too see if other NMT architectures, i.e., Transformers (Vaswani et al., 2017), would bring any improvements.

References

Roee Aharoni, 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)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings.

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. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver. Association for Computational Linguistics.

Jacob Devlin, 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Patrick Littell, 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*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Yichao Lu, 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*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

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.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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 Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.