KIT-Multi: A Translation-Oriented Multilingual Embedding Corpus

Thanh-Le Ha, Jan Niehues, Matthias Sperber, Ngoc Quan Pham and Alexander Waibel
Karlsruhe Institute of Technology, Karlsruhe, Germany
firstname.lastname@kit.edu

Abstract

Cross-lingual word embeddings are the representations of words across languages in a shared continuous vector space. Cross-lingual word embeddings have been shown to be helpful in the development of cross-lingual natural language processing tools. In case of more than two languages involved, we call them multilingual word embeddings. In this work, we introduce a multilingual word embedding corpus which is acquired by using neural machine translation. Unlike other cross-lingual embedding corpora, the embeddings can be learned from significantly smaller portions of data and for multiple languages at once. An intrinsic evaluation on monolingual tasks shows that our method is fairly competitive to the prevalent methods but on the cross-lingual document classification task, it obtains the best figures. We are in the process to produce the embeddings for more languages, especially the languages which belong to the same family or sematically close to each others, such as Japanese-Korean, Chinese-Vietnamese, German-Dutch, or Latin-based languagues. Furthermore, the corpus is being analyzedd regarding its usage and usefulness in other cross-lingual tasks.

Keywords: multilingual embeddings, cross-lingual embeddings, neural machine translation, multi-source translation

1. Introduction

Inducing cross-lingual word embeddings is essentially acquiring word embeddings in different languages. The cross-lingual word embeddings can then be used as pre-trained models in cross-lingual applications such as cross-lingual document classification, information retrieval, textual entailment and question answering. Cross-lingual word embeddings can also help to perform transfer learning from a well-resource language to another low-resource language on various tasks, e.g. in building WordNet or annotating semantic relations.

There have been various methods of cross-lingual embedding induction being proposed, but most of them are essentially bilingual in the perspective that they learn to induce *bilingual embeddings* from *bilingual* data [1]. Basically these methods optimize some cross-lingual constraints so that the semantic similarity between words corresponds to the closeness of these representations in a common vector space. Consequently, if they need cross-lingual embeddings for a new language pair, they must apply their inducing method on that new bilingual data. Furthermore, there would be some domain mismatch between the new acquired embeddings and the others if the new bilingual data are from different domain. The aforementioned limitations of those cross-lingual corpora motivates us to design a *multilingual embedding* inducing method from a *single* corpus which is available in as many languages as possible.

In this paper, we propose such an approach utilizing a multilingual neural machine translation (NMT) system to constrain the embeddings from n source languages while translating into the same target language (as we call it multi-source NMT). The source embeddings employed in this model are implicitly forced to learn the common semantic regularities in order to maximize the translation quality of every language pair in the system. Once the multi-source NMT model is trained to a good state, the source word embeddings can be simply extracted from the model and used as a multilingual word embeddings.

The contribution of this work is the introduction of a method and its product corpus, KIT-Multi[2], consisting multilingual word embeddings of English-German-French. Other languages such as Chinese, Japanese, Korean, Vietnamese, Dutch, Italian, Romanian, Spanish or Portuguese are being added. We conducted some preliminary evaluations on KIT-Multi and compares to other cross-lingual embedding corpora. It has been shown that our multilingual corpus achieves competitive performances in standard evaluations as well as it has better coverage while using much less data for the training process. The evaluations on other languages would be pulished in the final version of the paper.

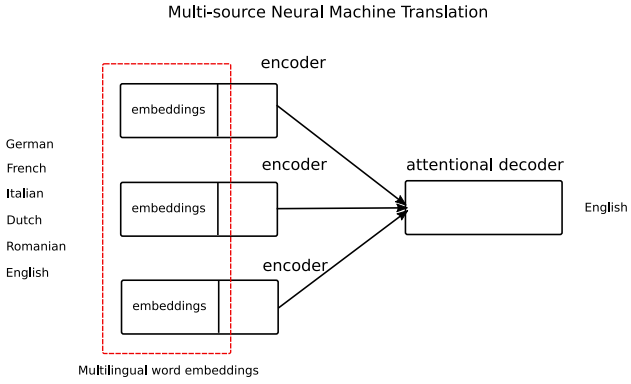

Figure 2: Multi-source Neural Machine Translation system and how to get multilingual word embeddings from it.

[1] For a thoroughly review of the most popular and advantageous techniques of cross-lingual word embedding induction, please refer to Upadhyay et al. (2016). For even more detailed and broader survey, please refer to Ruder (2017).

[2] The corpus is published and constantly updated at http://i13pc106.ira.uka.de/~tha/KIT-Multi/

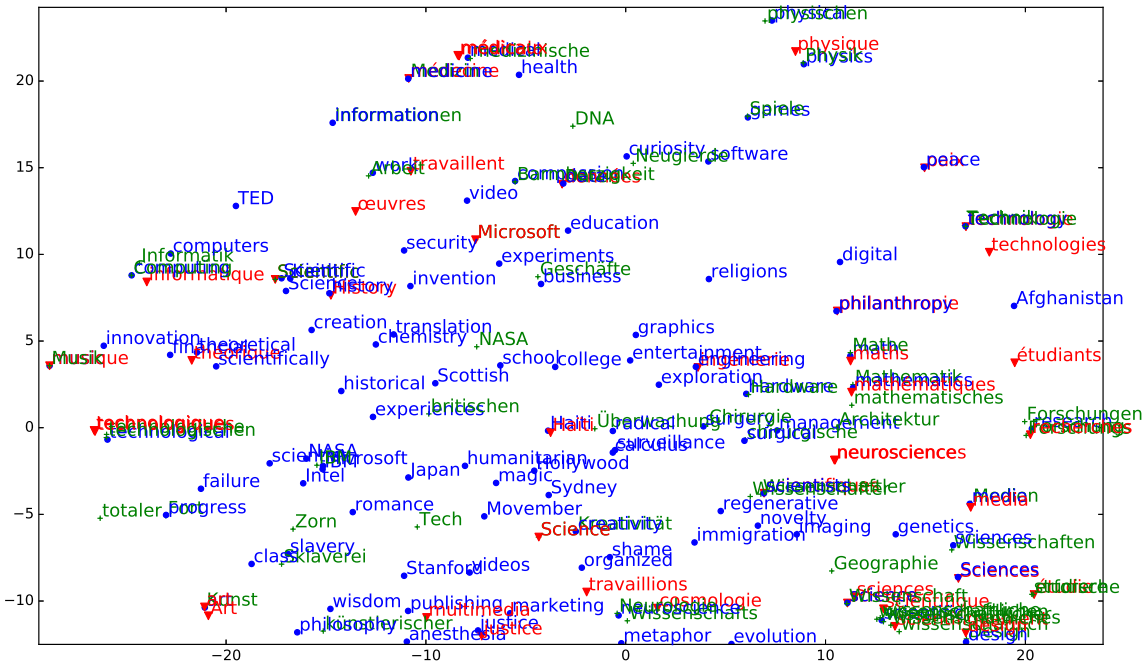

Figure 1: The multilingual word embeddings from the shared representation space of the source. To illustrate more clearly, only the word vectors of the words related to "*science*" are projected and visualized. The blue words are the English words, green for German and the red ones are the French words. Please zoom in to see more detailed.

2. Multilingual Word Embedding Corpus

2.1. Embedding Induction Method

A neural machine translation system (Bahdanau et al., 2014) consists of an encoder representing a source sentence and an attention-based decoder that produces the translated sentence. One of the most notable differences of NMT compared to the conventional statistical approach is that the source words can be represented in a continuous space (i.e. word embeddings) in which the semantic regularities are induced automatically. Being applied to multilingual settings, NMT systems have been proved to be benefited from additional information embedded in a common semantic space across languages (Johnson et al., 2016; Ha et al., 2016; Currey et al., 2017). An interesting and positive side effect of such a system is the simultaneous induction of multilingual embeddings from the source side.

In a multi-source NMT systems where the sentences from *several sources languages* are translated to *one target language*, the source embeddings are tied to a common semantic space across languages. So the source embeddings has its inherent cross-lingual characteristics, which could be extremely helpful for the cross-lingual applications employing the embeddings. More specifically, in our previous work on multi-source NMT (Ha et al., 2016), the words in each source sentence are coded with the language of that sentence before feeding to the training process of a standard neural machine translation system. For example, the source sentence in English: `they have since abandoned that project` would become ***en***`_they` ***en***`_have` ***en***`_since` ***en***`_abandoned` ***en***`_that` ***en***`_project`. anguage coding is conducted in the preprocessing phrase. Our multilingual embeddings are the derived product of this multi-source system. The figure 2 describes the process.

2.2. `KIT-Multi` Corpus

Our corpus is induced from WIT3's TED subtitle corpus (Cettolo et al., 2012) including bilingual corpora from French, German, Dutch, Italian and Romanian to English. TED is a much smaller multilingual data compared to Europarl and contains other languages than European languages. The multi-source NMT is trained using the NMT framework `OpenNMT`[3] (Klein et al., 2016) to translate from aforementioned languages (including English) to the only target language English. The statistics of TED bilingual corpora and our multilingual embedding corpus are shown in Table 1 and Table 2, respectively.

Language pairs	Number of sentences
German-English	196794
French-English	195025
Dutch-English	230866
Italian-English	220812
Romanian-English	210402

Table 1: Statistics of pair-wise TED bilingual corpora

[3] `http://opennmt.net`

Languages	Number of entries
English	21001
French	25685
German	24182
Dutch	24167
Italian	23422
Romanian	25505

Table 2: The size of the KIT-Multi embedding corpus

Figure 1 illustrates the visualization of multilingual word embeddings extracted from the multi-source NMT system and projected to the 2D space using *t-SNE* (Maaten and Hinton, 2008). It shows how different words in different languages, i.e. English-German-French, can be close in the shared semantic space after being trained to translate into a common language (English).

Table 3 shows the closest words in the semantic space based on Cosine similarity with respect to some examples. We also include the language codes to clarify the origin of each word. From the table, we can see that the most close words are actually the words having the same meaning but in other languages.

@en@research	
Word	Cosine Similarity
@de@Forschung	0.727675
@fr@recherches	0.697122
@de@Forschungs	0.671166
@fr@recherche	0.643990
@de@geforscht	0.637604
@en@humanity	
Word	Cosine Similarity
@de@Menschlichkeit	0.691524
@fr@humanité	0.684639
@de@Menschheit	0.645123
@de@Menscheit	0.634902
@en@mankind	0.621472

Table 3: Top 5 closest words by Cosine similarity.

3. Preliminary evaluation of KIT-Multi

In this section, we describe some initial evaluation of our multilingual embedding corpus over some standard intrinsic and extrinsic evaluations, in comparisons with some other popular approaches for cross-lingual word embedding induction.

We mostly follow the experimental layout and settings of Upadhyay et al. (2016), conducting intrinsic and extrinsic evaluations on three European languages: English, French and German. The intrinsic evaluation is the *monolingual word similarity* task. The extrinsic evaluation focuses on the *cross-lingual document classification*. In this task, a document classifier is trained on a training set composed by a language L_1 and then predict the test set which is in the different language L_2. The process is then reversed for the language pair, and the classification accuracy is used to judge the quality of the cross-lingual embeddings. The corpora chosen to be compared are the corpora induced by Skip - Bilingual Skip-gram (Luong et al., 2015), CVM - Bilingual Compositional Model (Hermann and Blunsom, 2014) and VCD - Bilingual Vectors from Comparable Data (Vulic and Moens, 2015), which are all trained on much bigger Europarl v7 parallel corpora[4] (Koehn, 2005). To show the impact of the corpus size, we also train the Bilingual Skip-gram embeddings with the same corpora used to train our model, and name it Skip-TED. For the details of those methods, please refer to Upadhyay et al. (2016).

In the intrinsic monolingual evaluation, we consider the word embeddings in one language at a time, i.e. the monolingual word embeddings, in order to conduct the *word similarity*. The Spearman's rank correlation coefficient (Myers et al., 1995) between system similarity and human is the measure to judge the quality of the induced word embeddings. The English evaluation datasets are SimLex999 (*En-999*) and WordSim353 (*En-353*), in which the former (Hill et al., 2016) is claimed to better capture the similarity rather than both similarity and relatedness like in the latter (Finkelstein et al., 2002). The German (*De*) and French (*Fr*) datasets are the WordSim353 counterparts (Camacho-Collados et al., 2015; Leviant and Reichart, 2015).

The scores in Table 4 show that our word embeddings are competent in term of monolingual aspect even though they are not trained to be adapted to monolingual quality. Moreover, our word embeddings perform better than the Skip embeddings trained on the same data by a large margin.

As shown in Table 5, the classifiers trained on our embeddings achieve highest accuracy on both directions of English⇔German, considerably better than other approaches. It is notable that, our model is trained on a substantially smaller corpus.

4. Related Work and Discussion

In (Upadhyay et al., 2016), the most popular and advantageous techniques for multilingual word embedding induction have been thoroughly evaluated. Corpora induced by Skip and VCD are the methods having the capability of monolingual adaptation by adjusting a hyper-parameter (in Skip models) or the portion of texts in each language (in VCD models). Furthermore, since they are designed based on the *skip-gram* models (Mikolov et al., 2013), it is unsurprising that they perform well on monolingual tasks. Corpora induced by CVM and our KIT-Multi, in contrast, are designed with cross-lingual orientation so that they focus more on similarity instead of relatedness. Aforementioned, our KIT-Multi corpus has shown its potential by achieving high accuracies on the task despite being induced from a significally smaller corpus. Compared to the corpora acquired by their method, our embedding inherently induced in multilingual settings, with an arbitrary number of source and target languages, instead of being limited to bilingual. Those advantages allow us to extend our corpus seamlessly to many languages using small multilingual corpus, ideally from TED talks.

[4] http://www.statmt.org/europarl/

Language	Skip	Skip-TED	CVM	VCD	*KIT-Multi*
En-999	0.34	0.22	**0.37**	0.32	***0.37***
En-353	0.53	0.39	0.43	**0.59**	*0.45*
De	0.52	0.40	0.40	**0.54**	*0.51*
Fr	**0.50**	0.09	0.38	0.43	*0.48*

Table 4: Monolingual evaluation tasks.

L_1	L_2	Skip	Skip-TED	CVM	VCD	*KIT-Multi*
En	De	85.2	84.3	85.0	79.9	***86.6***
De	En	74.9	73.5	71.1	74.1	***79.7***

Table 5: The accuracy of cross-lingual document classification task using the word embeddings.

5. Conclusion and Future Work

In this proposal, we introduce a method to extract multilingual embedding corpus and its production, `KIT-Multi`. We would like to extend it for more languages as well as more cross-lingual natural language processing applications. The corpus will be available in Japanese, Korean, Chinese, Vietnamese, English, German, Dutch, Italian, French, Spanish and Portuguese at the time of the conference. We welcome other groups download and use it in other tasks and discuss about its usefulness.

6. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2015). A framework for the construction of monolingual and cross-lingual word similarity datasets. In *ACL (2)*, pages 1–7.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Currey, A., Barone, A. V. M., and Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, WA, USA.

Hermann, K. M. and Blunsom, P. (2014). Multilingual Models for Compositional Distributed Semantics. *Acl*, pages 58–68.

Hill, F., Reichart, R., and Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2016). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Leviant, I. and Reichart, R. (2015). Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Bilingual Word Representations with Monolingual Quality in Mind. *Workshop on Vector Modeling for NLP*, pages 151–159.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Myers, J. L., Well, A., and Lorch, R. F. (1995). *Research design and statistical analysis*. Routledge.

Ruder, S. (2017). A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Upadhyay, S., Faruqui, M., Dyer, C., and Roth, D. (2016). Cross-lingual Models of Word Embeddings: An Empirical Comparison. *Acl 2016*, pages 1661–1670.

Vulic, I. and Moens, M.-F. (2015). Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. *Acl-2015*, pages 719–725.