Cross-Lingual Word Embeddings for Morphologically Rich Languages

Ahmet Üstün **Gosse Bouma** **Gertjan van Noord**
University of Groningen
{a.ustun, g.bouma, g.j.m.van.noord}@rug.nl

Abstract

Cross-lingual word embedding models learn a shared vector space for two or more languages so that words with similar meaning are represented by similar vectors regardless of their language. Although the existing models achieve high performance on pairs of morphologically simple languages, they perform very poorly on morphologically rich languages such as Turkish and Finnish. In this paper, we propose a morpheme-based model in order to increase the performance of cross-lingual word embeddings on morphologically rich languages. Our model includes a simple extension which enables us to exploit morphemes for cross-lingual mapping. We applied our model for the Turkish-Finnish language pair on the bilingual word translation task. Results show that our model outperforms the baseline models by 2% in the nearest neighbour ranking.

1 Introduction

Cross-lingual word embeddings (CLEs) have drawn a lot of attention in recent times. CLE models learn vectors of words in two or more languages and represent them in a shared cross-lingual word embedding space, where words with similar meaning have similar vectors, independent of their language. Most popular approaches for CLEs are mapping-based approaches which are also called offline approaches. These kinds of approaches require only pre-trained monolingual embeddings and a small seed dictionary so that the CLE model learns a mapping that minimizes the distance between word pairs in the seed dictionary to align the pre-trained embedding spaces.

CLE models enable multi-lingual modeling which has direct applications on cross-lingual tasks such as unsupervised machine translation (Lample et al., 2017), and cross-lingual transfer for downstream NLP tasks and low-resource languages. Document classification (Klementiev et al., 2012), information retrieval (Vulić and Moens, 2015), dependency parsing (Guo et al., 2015), and sequence labelling (Zhang et al., 2016) are examples of downstream NLP tasks in which CLEs serve as a source of cross-lingual knowledge.

Although the existing models achieve high performance, agglutinative languages, such as Turkish, Finnish and Estonian, pose a challenge to learn cross-lingual word embeddings due to three main reasons. First, with respect to the monolingual aspects, morphological complexity causes high sparsity which decreases the quality of monolingual embedding spaces (Cao and Rei, 2016; Üstün et al., 2018). Second, in the context of CLEs, the rich morphology causes inaccurate mappings especially for complex words because the existing CLE models cannot access the sub-word level information to align complex words with the correct morphological counterparts. Søgaard et al. (2018) shows that the existing CLE models underperform on rich morphological complexity. On the bilingual dictionary induction task, while the baseline method achieves 82.62% score on English-Spanish, it performs very poorly on English-Finnish (28.01%), English-Estonian (31.45%) and English-Turkish (39.22%). In the Estonian-Finnish dictionary induction experiment in which both languages are morphologically complex, the baseline model performs even worse (24.35%). Last, in addition to this limitation, word-based CLE models are also unable to map an inflected word in the morphologically complex language to a counterpart which corresponds to a phrase in a language with simple morphology.

In this study, we propose a morphologically-

Proceedings of Recent Advances in Natural Language Processing, pages 1222–1228,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_140

sensitive cross-lingual word embedding model[1] in order to overcome the second limitation. We build a cross-lingual model to learn the morpheme representations in the source languages so that a word can be represented through its morphemes in the target space. We design a supervised learning setting as in the baseline model that contains a small bilingual dictionary consisting of morphologically complex word pairs. We perform experiments on Turkish and Finnish as a pair of morphologically complex languages and compare our approach with the baseline models.

2 The Morpheme-Based Alignment Model

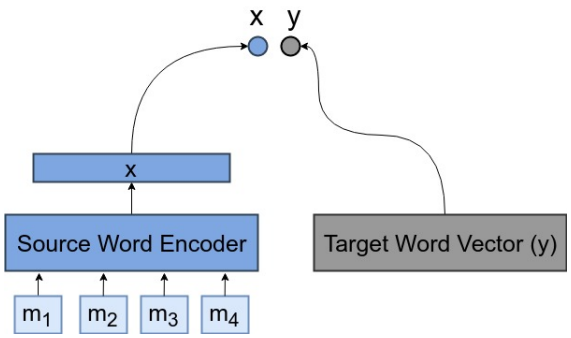

Figure 1: Morpheme-based cross-lingual alignment model that contains a source side word encoder for morphemes. The encoder is trained to learn morpheme representations in the target space

Baselines In this paper, we consider two baseline models. As the first baseline, we employ a simple projection-based CLE method which learns a mapping between embedding spaces by solving the Procrustes problem (Smith et al., 2017; Artetxe et al., 2016). This method first learns a linear transformation matrix to minimize the distance between vectors of word pairs in a seed dictionary by imposing the orthogonality constraint (Gower et al., 2004) and then it uses this matrix to transform the source language embedding space to represent both languages in a shared embedding space. The baseline method is denoted by Procrustes in this paper.

As the second baseline, we use relaxed cross-domain similarity local scaling (RCSLS) (Joulin et al., 2018). RCSLS optimizes the transformation matrix by maximizing the cross-domain similarity local scaling (CSLS) score, instead of minimizing the distance between word pairs in the training dictionary. CSLS is a modification of cosine similarity commonly used in information retrieval. In this way, RCSLS relaxes the orthogonality constraint used in Procrustes according to a retrieval criterion.

Note that for the both baseline models and our model, we use fastText (Bojanowski et al., 2017) to generate monolingual word embeddings. Fasttext represents words as sequence of character n-grams but in many cases this is suboptimal since not all character n-grams are morphemes (Üstün et al., 2018). Besides that the aim of this study is to incorporate morphology into cross-lingual training, whereas fastText is designed for monolingual training.

Morpheme-based Model In the morpheme-based model, we extend the projection-based baseline (Procrustes) in order to exploit sub-word (morpheme) level information for the cross-lingual mapping. Our model starts by splitting all words in the source language into morphemes by using a morphological analyzer. A vector is then computed for each generated morpheme, by using fastText (Bojanowski et al., 2017), as fastText is able to generate a vector for any sub-word since it is based on character n-gram representations.

After the resulting morpheme vectors are inserted into the source vector space, we apply a linear transformation based on the seed dictionary by using the Procrustes method to initialize source and target side vectors in a shared embedding space. Then, our model learns an encoder that encodes each word as a morpheme sequence and transforms them by aligning to their counterparts in the target language. In this way, the resulting encoder learns to represent source side morphemes in the target embedding space.

The model architecture is given in Figure 1. In the figure, x denotes the fixed length word representation generated by the word encoder through morphemes and y represents the target side word embedding. The source encoder is trained to mimic target word embeddings in the bilingual dictionary by minimizing the loss function:

$$L_{align} = dist(x, y) - \lambda(dist(x_c, y) + dist(x, y_c))$$

where (x, y) corresponds to the source and target word embeddings, (x_c, y_c) is a contrastive

[1]Code available at: https://bitbucket.org/ahmetustunn/morphology-sensitive-cle

term. λ^2 controls the effect of the negative samples in the alignment loss. We use the $cosine$ similarity for the distance measure.

For the encoder model, following (Conneau et al., 2017a), we use bidirectional LSTMs with max pooling. It encodes the words in both the forward and the backward direction to capture unidirectional information, then it combines the resulting numbers to form a fixed-size vector by selecting the maximum value over each dimension of the hidden units. Figure 2 shows the encoder model. In the figure, each word vector u is computed from morphemes m_n through the bidirectional LSTM encoder.

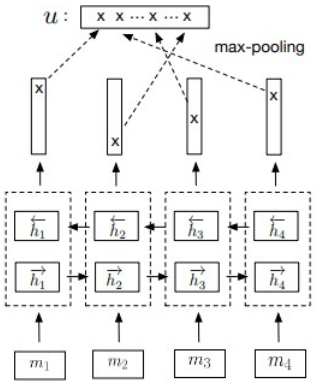

Figure 2: An overview of the word encoder which is used for the source language. It consists of bidirectional LSTMs and a max pooling layer. The inputs of the encoder are the morpheme sequences for each word

3 Morphologically Sensitive Bilingual Lexicon

In order to build a bilingual dictionary for the Turkish and Finnish word pairs, we use the MUSE dataset (Conneau et al., 2017b). Since the MUSE bilingual lexicon consists of translations to or from English for these two languages, we use the intersection of their translation to English. These bilingual lexicons are built with an automatic translation system and the dictionaries handle well the polysemy of words. However, the dictionaries mostly consist of morphologically simple word pairs since English is used as a pivot language.

Considering the morphological variations in these languages, we enriched the dictionary with morphologically complex word pairs. To this end, we first create a lookup table for the lemmas, their inflections and the corresponding morphological features for both languages by using the Universal Dependency Treebanks (Nivre et al., 2016) and the Universal Morphology (Sylak-Glassman, 2016) project.[3] Each word pair in the dictionary is then searched in these lookup tables to list their inflections. The inflected word forms which have the same morphological features for a pair are then added to the bilingual dictionary. Table 1 shows the inflected wordforms found in lookup table for the seed pair *gölge-varjo*.[4] The morphological features that occur in both languages are given in Table 2.

Turkish	Finnish	
gölgem	varjoni	`N;SG;PSS1S`
gölgenin	varjojen	`N;SG;GEN`
gölgelerin	varjoja	`N;PL;PSS2S`
gölgelerde	varjoissa	`N;ESS;PLS`

Table 1: The inflected wordforms with the same morphological features for the word pair *gölge-varjo* which mean *shadow*

Attribute	Morphological Classes
Number	Sing, Plu
Polarity	Neg, Pos
Person	{Pss1,Pss2,Pss3}+{Sg,Pl}
Case	{in,on,at}+{Ess,Abl}, Gen, Prt
Tense	Pst, Prs, Imp
Agreement	P1,P2,P3
Voice	Pass
Mood	Ind, Imp, Cond

Table 2: Morphological features which are common in both Turkish and Finnish

The training dictionary comprises the first 5000 Turkish words and their Finnish counterparts while the test set is composed of the following 1500 word pairs in the lexicon.

4 Experiments

We compare our morpheme-based model with Procrustes (Smith et al., 2017; Artetxe et al., 2016) and relaxed cross-domain similarity local scaling (RCSLS) (Joulin et al., 2018), as explained in Section 2.

[2]Following Conneau et al. (2018), we set λ to 0.25

[3]During the preprocessing step, the default morphological features which are language specific are removed from the datasets.

[4]Both words mean *shadow* in English

Model	NN	CSLS
Turkish-Finnish (TR-FI)		
Procrustes	16.54	17.89
RCSLS	18.26	**21.06**
Our model	**20.35**	20.40
TR-FI on English		
Procrustes	12.72	14.89
RCSLS	15.10	17.05

Table 3: Bilingual word translation performance of the models at P@1 (%). First three rows show the results after training with Turkish-Finnish morphologically sensitive seed dictionary. The last two rows present the results when English is used as a pivot language.

Evaluation Task In order to evaluate the models, we used the bilingual word translation task. Bilingual word translation has become the standard evaluation task for mapping-based CLE models. Given a shared embedding space which is learned by a CLE model, the task is to translate source language words to the target language by retrieving a word in the target language. As the retrieval criterion, either nearest neighbor search (NN) or cross-domain local scaling (CSLS) can be used.

Implementation Details Our evaluation comprises Turkish and Finnish which are both morphologically complex languages. We use the l_2-normalized fastText word and morpheme vectors (Bojanowski et al., 2017) trained on Wikipedia for these languages. We initialize source side embeddings with a linear transformation defined by a Procrustes operation based on the seed dictionary. In order to split words into morphemes, we use the Zemberek toolkit (Akın and Akın, 2007)[5] for Turkish and the Omorfi project (Pirinen, 2015)[6] for Finnish. Both morphological analyzers are rule-based and run with high accuracy. All models are trained with the same seed dictionary and evaluated on the same test set. We evaluated the model by the scores of precision at rank 1 (P@1) so that the results can be morphologically sensitive.

5 Results

Table 3 shows the results on the bilingual word translation performance of the models for the Turkish-Finnish language pair. According to the CSLS scores, RCSLS (Joulin et al., 2018) outperforms our models by a slight margin (0.66%). This is expected because the RCSLS model is explicitly designed to maximize the CSLS objective which causes better performance on the bilingual word translation task according to the CSLS score. However, according to nearest neighbor ranking, our model displays the strongest performance compared to Procrustes and RCSLS with a 2.09% score difference. Table 5 show examples of the nearest neighbour predictions of different models including our model.

We also run Procrustes and RCSLS on the Turkish-English and Finnish-English language pairs so that all three languages share the monolingual English embedding space. We use the MUSE (Conneau et al., 2017b) training dictionaries for both language pairs. In this setting, both Procrustes and RCSLS perform worse on Turkish-Finnish bilingual word translation suggesting that a third language (as a pivot language) does not provide benefit for word translation across morphologically rich language pairs even if it has high-quality word vectors. As our model requires the translations of inflected (morphologically complex) words in the target language, we can not run our model on Turkish-English or Finnish-English pairs because the translations mostly correspond to phrases instead of words.

Model	Spearman
Morph2Vec (Üstün et al., 2018)	**52.90**
Our model	42.05
Fasttext (Bojanowski et al., 2017)	20.80

Table 4: The comparison of the Spearman correlation between human judgments and the word similarities obtained by computing the cosine similarity between the learned word embeddings for Turkish.

Monolingual Impact Similar to the RCSLS, our model changes the cosine distance between word vectors in the same language, that is, it also has an impact on the monolingual embedding space. We evaluate this impact on the Turkish morphologically complex wordlist (Üstün et al., 2018). Results are given in the Table 4.

The Morph2Vec model (Üstün et al., 2018) learns a morpheme-based encoder which is monolingually trained on the large Turkish wordlist which consists of 100K unique words. Although our model is trained on 5K Turkish-Finnish word

[5] https://github.com/ahmetaa/zemberek-nlp
[6] https://github.com/flammie/omorfi

No	Source Word *(Turkish)*	Target Translations *(Finnish)*		
		Procrestus	RCSLS	Our Model
1	öptüm	suutelit *(you kissed)*	suutelen *(I kiss)*	**suitelin** ***(I kissed)***
2	aileler	perhe *(a family)*	**perheet** ***(families)***	**perheet** ***(families)***
3	zamanımız	aikani *(my time)*	aikani *(my time)*	**aikamme** ***(our time)***
4	acemilerden	aloittelijoilla *(in the beginners)*	aloittelijasta *(from the beginner)*	**aloittelijoilta** ***(from the beginners)***
5	makineler	**koneet** ***(machines)***	**koneet** ***(machines)***	koneissa *(in machines)*

Table 5: Examples comparing the translations of different models which also includes the glosses in English. Bolding indicates the correct translation. In Examples 1-4, our model predicts correct word considering the morphological structure but in the Example 5, our model gives wrong translation.

pairs, it improves the monolingual quality of Turkish word vectors for morphologically complex words. The reason behind this impact is that our model also changes the cosine similarity among Turkish word vectors according to morphologically sensitive cross-lingual signals, during the cross-lingual transformation. However, the Morph2Vec model still outperforms our model by a high margin. The results demonstrate that even if training a morpheme-based encoder on cross-lingual word pairs improves the monolingual embedding quality, the same training strategy still performs substantially better on a monolingual wordlist.

Error Analysis Here we study the errors produced by our model on Turkish-Finnish word pairs. Although our model is motivated by morphology, a small portion of the wrong translations is caused by the prediction of wrong inflections of a correct root word. The model translates the Turkish word *santralin* (*of the power plant*) as *voimalaa* (*in the power plant*) instead of *voimalan*. However, the majority of errors have incorrectly translated root words with correct inflections. These observations can suggest two shortcomings. Firstly, our model over-focuses on morphemes so that in some cases it lost the meaning of the content word. Secondly, especially for the distant language pair, some morphological features have different meanings which depend on the sentence syntax and contextual meaning, even if they have the same label. This issue could be alleviated by modeling and processing sentence-level context.

Limitations Similar to the baseline models, the main limitation of our model is that it can not generate multi-word expressions such as phrases on the target side, although our model is able to represent a sequence of strings in the source encoder. However, a morphologically complex word in the source language such as Turkish or Finnish, in most cases corresponds to a phrase, containing more than one word, in morphologically simple target languages such as English. For this reason, our model does not have any direct benefit for the morphologically simple languages and this issue will be the focus of follow-up studies.

Another limitation is that, our model requires a morphological segmenter to split words into morphemes. A simple solution for this could be to employ an unsupervised morphological segmenter which is commonly used in the literature such as Morfessor (Creutz and Lagus, 2005).

6 Conclusion

In this work, we extend the simple mapping-based cross-lingual embedding (CLE) model to learn a morphology-sensitive transformation between embedding spaces for morphologically rich language pairs. We start with the baseline transformation to initialize the source and target embedding spaces and then our model learns an encoder based on morphological segments in the source side and their counterparts in the target space. Thus, the transition matrix which is computed to produce a shared cross-lingual embedding space, is learned through morpheme representations and their composition in the source language.

We evaluated our model on the bilingual word translation task and compare our results with Procrustes and RCSLS (Joulin et al., 2018) scores. Results show that our morpheme-based cross-lingual embeddings model learns slightly better

alignments for complex word pairs for languages having rich morphology compared to the baseline models. In this work, we have made the first step towards the comprehensive evaluation of CLE models according to the morphology of languages, however, our evaluation is limited to the bilingual word translation task. For further analysis, we are planning to evaluate our model on other language pairs which consists of both morphologically complex and simple languages and on downstream NLP tasks such as POS tagging and dependency parsing.

References

Ahmet Afsin Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source nlp framework for turkic languages. *Structure*, 10:1–5.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

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.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. *arXiv preprint arXiv:1606.02601*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017a. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017b. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology Helsinki.

John C Gower, Garmt B Dijksterhuis, et al. 2004. *Procrustes problems*, volume 30. Oxford University Press on Demand.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1234–1244.

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.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012*, pages 1459–1474.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Tommi A Pirinen. 2015. Omorfifree and open source morphological lexical database for finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. *arXiv preprint arXiv:1805.03620*.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). Technical report, Technical report, Department of Computer Science, Johns Hopkins University.

Ahmet Üstün, Murathan Kurfalı, and Burcu Can. 2018. Characters or morphemes: How to represent words? In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 144–153.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the*

7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), volume 2, pages 719–725.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. In *Proceedings of NAACL-HLT*, pages 1307–1317.