

Anchor-based Bilingual Word Embeddings for Low-Resource Languages

Tobias Eder¹, Viktor Hangya² and Alexander Fraser²

¹Research Group Social Computing
Technical University of Munich

²Center for Information and Language Processing
LMU Munich

tobias.eder@in.tum.de,
{hangyav, fraser}@cis.lmu.de

Abstract

Good quality monolingual word embeddings (MWEs) can be built for languages which have large amounts of unlabeled text. MWEs can be aligned to bilingual spaces using only a few thousand word translation pairs. For low resource languages training MWEs monolingually results in MWEs of poor quality, and thus poor bilingual word embeddings (BWEs) as well. This paper proposes a new approach for building BWEs in which the vector space of the high resource source language is used as a starting point for training an embedding space for the low resource target language. By using the source vectors as anchors the vector spaces are automatically aligned during training. We experiment on English-German, English-Hiligaynon and English-Macedonian. We show that our approach results not only in improved BWEs and bilingual lexicon induction performance, but also in improved target language MWE quality as measured using monolingual word similarity.

1 Introduction

Bilingual Word Embeddings are useful for cross-lingual tasks such as cross-lingual transfer learning or machine translation. Mapping based BWE approaches rely only on a cheap bilingual signal, in the form of a seed lexicon, and monolingual data to train monolingual word embeddings (MWEs) for each language, which makes them easily applicable in low-resource scenarios (Mikolov et al., 2013b; Xing et al., 2015; Artetxe et al., 2016). It was shown that BWEs can be built using a small seed lexicon (Artetxe et al., 2017) or without any word pairs (Lample et al., 2018a; Artetxe et al., 2018) relying on the assumption of isomorphic MWE spaces. Recent approaches showed that BWEs can be built without the mapping step. Lample et al. (2018b) built FASTTEXT embeddings (Bojanowski et al., 2017) on the concatenated source and target

language corpora exploiting the shared character n-grams in them. Similarly, the shared source and target language subword tokens are used as a cheap cross-lingual signal in Devlin et al. (2019); Conneau and Lample (2019). Furthermore, the advantages of mapping and jointly training the MWEs and BWEs were combined in Wang et al. (2020) for even better BWEs.

While these approaches already try to minimize the amount of bilingual signal needed for cross-lingual applications, they still require a larger amount of monolingual data to train semantically rich word embeddings (Adams et al., 2017). This becomes a problem when one of the two languages does not have sufficient monolingual data available (Artetxe et al., 2020). In this case, training a good embedding space can be infeasible which means mapping based approaches are not able to build useful BWEs (Michel et al., 2020).

In this paper we introduce a new approach to building BWEs when one of the languages only has limited available monolingual data. Instead of using mapping or joint approaches, this paper takes the middle ground by making use of the MWEs of a resource rich language and training the low resource language embeddings on top of it. For this, a bilingual seed lexicon is used to initialize the representation of target language words by taking the pre-trained vectors of their source pairs prior to target side training, which acts as an informed starting point to shape the vector space during the process. We randomly initialize the representations of all non-lexicon target words and run Continuous Bag-of-Words (CBOW) and skip-gram (SG) training procedures to generate target embeddings with both WORD2VEC (Mikolov et al., 2013a) and FASTTEXT (Bojanowski et al., 2017). Our approach ensures that the source language MWE space is intact, so that the data deficit on the target side does not result in lowered source

embedding quality. The improved monolingual word embeddings for the target language outperform embeddings trained solely on monolingual data for semantic tasks such as word-similarity prediction. We study low-resource settings for English-German and English-Hiligaynon, where previous approaches have failed (Michel et al., 2020), as well as English-Macedonian.

2 Method

Previous mapping approaches rely on the alignment of two pre-trained monolingual word embedding spaces. In case one of the two languages has significantly fewer resources available, this will strongly affect the resulting mapping negatively. This is also an issue for joint approaches because the shared token representations are biased towards the language with more training samples. Our approach instead leverages the high resource language to improve performance on the low-resource language.

We pre-train MWEs for the source language and use the source MWEs to initialize the space of the low resource target language. Using a set of initial seed pairs, the representation of a seed word in the target space is replaced with the representation of its translation (anchor points). Then, training is performed on the initialized space using only monolingual data from the low resource language by only updating the representation of non-seed words which are initialized randomly. Through this method a BWE representation is directly induced from the anchor points of the fixed vectors.

In some cases there are multiple valid translations for a single target language word. We experiment with either initializing with the average over these possible translations or randomly selecting only one of them. The averaging helps by finding a common anchor for the different semantic nuances the token might represent in different target language contexts. Additionally, we experiment with enabling or disabling the updates of anchor vectors during training. We implemented the anchor point based initialization in both WORD2VEC and FASTTEXT with only complete token representations serving as potential anchors. In the case of FASTTEXT these initializations have no influence on the subword (character n-gram) embeddings which are still initialized randomly, which makes intuitive sense in the common case of morphologically different language pairs. Training is performed using standard hyperparameters included in the GENSIM

WORD2VEC and FASTTEXT packages (Řehůřek and Sojka, 2010). Unless stated otherwise, vectors are of dimensionality 300 with a context window of 5 words used during training. All models are trained for 5 epochs without further hyperparameter tuning utilizing a single desktop machine on a Intel Core i7-7700K CPU with 4.20Ghz, a NVIDIA GeForce GTX 1080 Ti graphics card and 32 GB of DDR4 SDRAM. The parameters of each trained model are equal to the standard implementation of the packages as listed above. Training time is largely dependent on input size, but corresponds to a few seconds up to roughly 5 minutes in the low resource setting.

2.1 Experimental Setup

First we conduct experiments on the German and English language pair, since large available corpora made it easier to test different sized dictionaries and corpora during training. The basic setup trains a MWE on the source language (English) up front. For this training the WMT 2019 News Crawl corpus in English, including approximately 532 million tokens, was chosen (Barrault et al., 2019). Similarly for the target language, we used the German WMT 2019 News Crawl from which we uniformly sample to obtain training sets of different sizes. All dataset are tokenized and lowercased before training.

To evaluate, we translate German words to English. We use the MUSE German-English dictionary (Lample et al., 2018a). There are 102K translation pairs with a total of roughly 68K unique German words. For each German word there might be multiple valid English translations, which are listed in the dictionary. For the initialization we select either randomly one translation option or the averaged word representations of all available translations, as discussed in section 2. However, many German words have only one valid translation. We used the MUSE test set containing roughly 3000 translation pairs in the frequency range 5000-6500, leaving 99K pairs as potential candidates for the initialization. In our experiments we mostly consider setups with much smaller training lexicon sizes, by taking the top- n most frequent source words and their translations from the lexicon.

In addition to the German experiments we test our system on two lower resource languages: Macedonian and Hiligaynon. For Macedonian we use data in the form of a Wikipedia dump, as well

as the MUSE dictionary for the language pair Macedonian-English for our test setup.¹ For Hiligaynon we use a corpus containing roughly 350K tokens as well as a corresponding dictionary containing 1100 translated terms between English and Hiligaynon and an additional test set of 200 terms released by Michel et al. (2020).

After training, bilingual lexicon induction (BLI) is done by taking the top n closest vectors measured by cross-domain similarity local scaling (CSLS) distance. For better comparability we use the evaluation method provided by MUSE (Lample et al., 2018a) for both the comparison baseline as well as our system. For Hiligaynon we use cosine to compare directly with (Michel et al., 2020).

3 Results

The following section evaluates different models quantitatively using $acc@5$ and $acc@1$ as a metric. The baseline runs MUSE tool in supervised mode using iterative procrustes refinement to obtain the mapping using default parameters as reported in Lample et al. (2018a). For the English embedding the full corpus size was used, while in the case of the (low-resource) languages the corpora sizes were varied to observe changes in performance.

3.1 Bilingual experiments

BLI was performed using the method from section 2. Since Word2Vec SG and FastText embeddings performed much worse with the anchored training, all following numbers report Word2Vec CBOW embeddings.

Table 1 shows the comparison between four different possible setups for the proposed method as explained in Section 2: Either fixing anchor-vectors or allowing them to train or initializing with single word vectors or averaged ones. The overall best performing model utilizes averaged and non-fixed anchor vectors. Table 1 also shows the baselines at varying corpora sizes. Overall the anchor method performs much better than the baseline at lower corpora sizes and stays competitive as corpus size increases. Results are similarly consistent when looking at either $acc@5$ or $acc@1$.

One important parameter for the proposed method is how it scales with the available amount of anchor-vectors used for training. In a range of experiments, different initialization sizes were com-

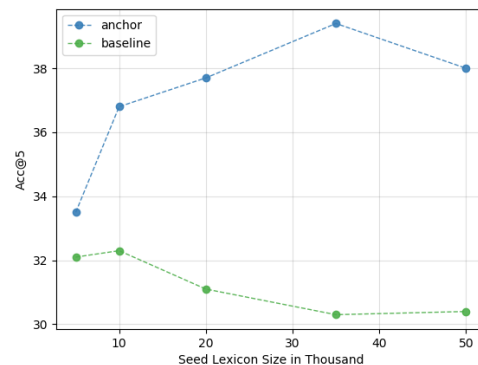


Figure 1: Anchor method for English-German with fixed vectors and baseline with varying training-dictionary sizes at corpus size 20 million

pared. Figure 1 shows the result for varying the number of anchor vectors. The general trend is that the more anchor vectors, the better the performance, which slowly caps off at the higher end, as more vectors of lower frequency words start introducing noise. The same development is not true for the baseline, which does not benefit equally from increasing the potential seed lexicon vocabulary and even starts losing performance at larger dictionary sizes.

This difference is likely rooted in the inclusion of less frequent word pairs in the larger dictionaries. These words have worse quality representations which introduces noise in the mapping process, thus restricting the precision of their orthogonal alignment as described by (Søgaard et al., 2018). In contrast, our method initializes all target language word embeddings given their pairs, i.e., perfectly aligning all words in the training dictionary, which serve as high quality anchor points for the remaining words.

3.2 Macedonian and Hiligaynon

Another set of experiments was done on the language pair English-Macedonian, a language that already offers less resources than German and is also more dissimilar from English.

Results for experiments comparing between the MUSE baseline and the anchor method are shown in Table 2. The best performing model again combines averaged initialization with trainable anchors.

Compared to the previous experiments with German, results for Macedonian are similar, while the baseline model is overall weaker than before, suggesting the anchor method benefits more strongly

¹<https://dumps.wikimedia.org/mkwiki/> (downloaded on 01/31/21)

Model // Corpus Size	100K	300K	500K	1M	2M	5M	10M	20M	50M
Baseline Word2Vec (CBOW)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.3 (0.0)	0.6 (0.2)	6.1 (1.4)	15.9 (4.2)	26.7 (12.8)	40.2 (21.3)
Baseline FastText (SG)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.6 (0.2)	1.9 (0.5)	8.9 (3.9)	19.7 (8.6)	32.1 (18.1)	45.1 (28.9)
Fixed not Averaged (CBOW)	1.3 (0.9)	0.8 (0.3)	0.8 (0.3)	1.5 (0.5)	4.1 (1.7)	13.6 (5.6)	23.9 (12.2)	26.3 (13.4)	35.7 (21.0)
Fixed and Averaged (CBOW)	1.3 (0.4)	1.2 (0.1)	1.8 (0.2)	1.8 (0.6)	4.3 (1.9)	13.1 (5.5)	23.3 (11.9)	33.2 (18.4)	44.1 (26.3)
Trained not Averaged (CBOW)	1.3 (0.9)	0.8 (0.4)	1.4 (0.8)	3.3 (0.5)	7.3 (1.7)	18.0 (5.6)	27.0 (12.2)	35.8 (21.0)	45.2 (28.4)
Trained and Averaged (CBOW)	1.7 (0.4)	1.5 (0.5)	2.9 (1.4)	4.5 (1.3)	10.7 (4.8)	22.3 (11.9)	32.2 (18.3)	40.5 (25.0)	48.5 (31.5)

Table 1: Evaluation of models at varying data-sizes for English-German. Baselines (MUSE) and proposed methods, reporting acc@5 (acc@1).

Model // Corpus Size	1M	2M	5M	10M	20M	37M
Baseline Word2Vec (CBOW)	0.0 (0.0)	0.7 (0.0)	6.0 (1.9)	18.1 (7.9)	28.0 (15.1)	37.0 (20.4)
Trained and Averaged (CBOW)	1.6 (0.4)	4.7 (1.6)	13.1 (5.6)	26.3 (13.4)	35.5 (18.4)	44.7 (24.2)

Table 2: Anchor method vs. baseline (MUSE) at varying data sizes reporting acc@5 (acc@1) for English-Macedonian.

Algorithm	Acc@5
Baseline Michel et al. 50D	0.5
Baseline Michel et al. 300D	0.0
Anchor fixed not averaged CBOW 300D	2.6
Anchor trained not averaged CBOW 300D	4.3

Table 3: BLI on Hiligaynon-English.

from a high-resource embedding, even when language pairs become more dissimilar.

For English-Hiligaynon previous approaches failed due to limitations of the available monolingual training data. Table 3 shows the performance for translating Hiligaynon words to English. The evaluation was done using cosine-distance for better comparability between the Michel et al. (2020) paper and our results. Since there were only single translations of words in the provided dictionary, the method of averaging vectors for initialization was not used. Similarly, during evaluation only one valid term per word was possible. While Michel et al. (2020) reported 0.5% for 50 dimensional vectors, in our baseline the 50 dimensional vectors achieved a constant 0 (not shown). The numbers are comparable to the low frequency experiments between German and English as seen in Table 1.

3.3 Monolingual experiments

In addition to better BWEs, our approach also improves the low-resource embedding for purely monolingual tasks. To confirm this, the anchor-vector trained embeddings for German were evaluated on monolingual word similarity and compared to the results achieved by regular training of the embedding space. We evaluate on multiple datasets: GUR350 and GUR65 (Gurevych, 2005), SEMEVAL17 (Cer et al., 2017), SIMLEX-999 (Leviant and Reichart, 2015), WS-353 (Agirre et al.,

2009) and ZG22 (Zesch and Gurevych, 2006), and report the averaged Spearman’s rho correlation between cosine similarity of vector pairs and human annotations. Similar monolingual datasets are not available for Macedonian and Hiligaynon. In Figure 2 the effect of employing the anchor method on monolingual word similarity performance is compared against Word2Vec CBOW trained without anchor initialization. The improvements across different training corpora sizes are in favor of the proposed method, suggesting that it can be employed to improve performance on monolingual tasks. Overall this serves to demonstrate the advantage of the anchor-method on small datasets and allows to learn better monolingual representation from the same amount of data by utilizing the information from a pretrained embedding for a completely different language with more readily available training data. The thus learned representation can not only serve as an already aligned space for translation tasks as shown above, but is also the better performing representation of the monolingual space.

4 Conclusion

We proposed a novel approach to build BWEs to improve performance on language pairs with limited monolingual data on the target side. By utilizing pre-trained MWEs of resource rich languages and a seed lexicon to fix anchor points before training, a structurally similar embedding space can be learned for the low resource language which is aligned with the source representations. We evaluated our approach on the BDI task using English-German to test varying training parameters and corpora sizes, on English-

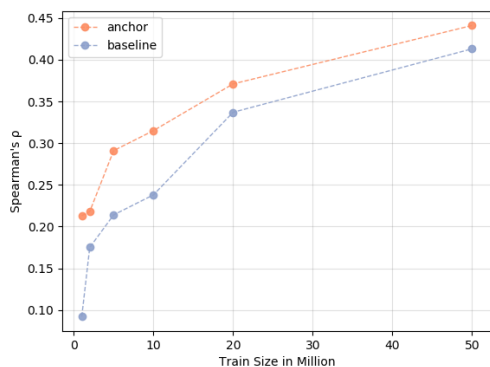


Figure 2: Spearman’s rho correlation on monolingual word similarity across corpora sizes for German.

Macedonian and the extremely low resource language pair English-Hiligaynon on which previous approaches failed. We showed that the performance of existing mapping approaches degrades drastically with lower monolingual data sizes, even when there are large seed lexicons available. In contrast, our proposed system outperformed previous mapping based approaches on these setups including English-Hiligaynon. On top of improved BWEs, we showed improved MWE quality as well for the target language by outperforming standard MWEs on the monolingual word similarity task showing that it is beneficial for monolingual tasks as well. We implemented our approach for both Word2Vec and FastText which we publicly release to promote reproducibility and further research.²

Ethical Considerations

The proposed system acts as a tool to specifically help in the low resource setting that predominantly affects less researched languages. Even though part of the experiments were done on the higher resource language pair English-German, the results were further confirmed for other pairs of languages.

As a word embedding based system, the resulting mappings and embedding spaces are highly affected by the kind of monolingual content that goes into their training, which is why we made sure to train the embeddings on texts that should adhere to a higher standard, such as verified news media and online articles, instead of a general web crawl. Additionally the seed lexicons used come from verified sources, such as the popular MUSE lexicons in the case of English-German and

English-Macedonian as well as from translations by a native speaker of Hiligaynon in the case of English-Hiligaynon.

We hope that in general the proposed methods can help alleviating some of the resource problems less researched languages are facing and thus to close the gap for language technology working with and on these languages.

As part of the ethical responsibility to ensure reproducibility and responsibility in terms of computational resources, we reported results with a set of standard hyperparameters instead of searching for the most optimal setting for our proposed method. Our models are as lightweight as regular training methods for word embeddings and are therefore not very demanding in terms of computation. This is especially true in the low-resource setting, where training time is reduced to just a fraction compared to the bigger corpora.

Acknowledgments

The work was supported by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No. 640550) and by German Research Foundation (DFG; grant FR 2829/4-1).

References

- Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. [Cross-lingual word embeddings for low-resource language modeling](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947.
- Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL-HLT 2009*.
- 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.
- 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.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. [A robust self-learning method for fully unsupervised](#)

²<http://cistern.cis.lmu.de/anchor-embeddings>

- cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.
- Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. [A call for more rigor in unsupervised cross-lingual learning](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(wmt19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.
- 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.
- Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.
- 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.
- Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *International Conference on Natural Language Processing*, pages 767–778.
- Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word Translation Without Parallel Data. In *International Conference on Learning Representations*.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018b. [Phrase-Based & Neural Unsupervised Machine Translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.
- Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.
- Leah Michel, Viktor Hangya, and Alexander Fraser. 2020. Exploring bilingual word embeddings for Hiligaynon, a low-resource language. In *LREC 2020, Twelfth International Conference on Language Resources and Evaluation*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.
- Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. [On the limitations of unsupervised bilingual dictionary induction](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.
- Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2020. [Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework](#). In *Proceedings of the 8th International Conference on Learning Representations*.
- Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.
- Torsten Zesch and Iryna Gurevych. 2006. Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the Workshop on Linguistic Distances*, pages 16–24.