When Word Pairs Matter
Analysis of the English-Slovak Evaluation Dataset

Michaela Denisová and Pavel Rychlý

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{449884,pary}@mail.muni.cz

Abstract. Cross-lingual word embeddings facilitate the transfer of lexical knowledge across languages, and they are mainly used for finding translation equivalents. Translation equivalents obtained in this way are usually evaluated with the help of ground truth dictionaries. However, the evaluation process, including the ground truth dictionaries, differs from model to model, impeding the correct interpretation of the results. Therefore, in this paper, we provide a thorough analysis of the English-Slovak ground truth dictionary and employ our analysis in evaluating two cross-lingual word embedding models. We show that word pairs choice is an important factor when accurately reflecting the model's performance.

Keywords: Cross-lingual word embeddings · Ground truth dictionary · Evaluation · English · Slovak

1 Introduction

In recent years, the popularity of cross-lingual word embeddings has risen among researchers due to their ability to connect meanings across languages. Cross-lingual word embeddings enable us to align the word vector representations from two or several languages into a single vector space where similar words obtain similar vectors [10]. In most cases, cross-lingual embedding models are evaluated via finding translation equivalents known as bilingual lexicon induction task [9,4,1]. In the bilingual lexicon induction task, translation equivalents are obtained from the aligned vector space through nearest neighbor search and then compared to the ground truth dictionaries. However, there is no united evaluation procedure agreed upon, and many authors consider different evaluation strategies, starting with different ground truth dictionaries, which causes inconsistencies between the stated results [5].

In this paper, we want to thoroughly analyze the English-Slovak dataset with 2,739 word pairs (1,500 English headwords) used as a ground truth dictionary to evaluate the MUSE model [4] and assign weight to each word pair accordingly. We aim to evaluate MUSE and VecMap [1] models with and without weighted word pairs to see how the model's performance changes. We think that current ground truth dictionaries used for evaluation may contain mistakes and

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 141–149, 2021. © Tribun EU 2021

irrelevant word pairs. Usage of such an evaluation dictionary can distort the actual model's performance.

The reason for our experiment is not to penalize the model when it does not find word pairs with lower weight, and we want the model to achieve higher accuracy when it includes word pairs with higher weight. Also, we believe that having a good quality evaluation dataset can reflect the model's performance more realistically and be the first step to a united evaluation procedure.

This paper is structured as follows. In Section 2, we describe MUSE and VecMap models. In Section 3, we analyze the English-Slovak dataset, and in Section 4, we use this dataset for MUSE and VecMap model evaluation. In Section 5, we offer concluding remarks.

2 Related Work

2.1 MUSE

The English-Slovak dataset we used for the analysis originates from the MUSE project. MUSE is an open-source cross-lingual word embedding model published by Facebook research in 2018. Except for the model, there are available pre-trained multilingual word embeddings aligned into shared vector space for 35 languages and ground truth evaluation dictionaries for 6 European languages in every direction and for 47 languages more from and to English. The model could be trained in a supervised [4] or unsupervised way [7]. For the supervised training, the Procrustes iterative alignment is used. The unsupervised method uses adversarial training and iterative Procrustes refinement.

In our experiments, we used supervised pre-trained multilingual embeddings for English and Slovak that are available in the MUSE library.[1]

2.2 VecMap

VecMap is an open-source cross-lingual word embedding model[2] released by Artetxe et al. in 2016. It provides four types of training: supervised [1], semi-supervised or identical training (relying on identical strings) [2], and unsupervised training [3]. For all of them, are required pre-trained monolingual word embeddings. Additionally, for semi-supervised and supervised training is necessary to have a training dataset from 25 up to 5,000 word pairs, respectively.

In this paper, we trained the model under strong supervision using the English-Slovak training dataset obtained from MUSE with 5,000 word pairs. Moreover, we used fastText monolingual embeddings [8] for English and Slovak in the training, downloaded from fastText library.[3]

[1] https://github.com/facebookresearch/MUSE

[2] https://github.com/artetxem/vecmap

[3] https://fasttext.cc/

3 Analysis of the Dataset

In the analysis, we considered three aspects that can influence the quality of an evaluation dataset. The first one was the frequency of given word pair in the parallel corpus. We obtained the frequencies for each word pair from the parallel English-Slovak corpus OPUS2 [11] via SketchEngine API [6]. The corpus contained approximately 8,000,000 sentences derived from 8,000 documents.

Logarithmic Zipf's curve of the obtained frequencies in Fig. 1 shows that most of the word pairs in the dataset had a lower frequency than 2,500.

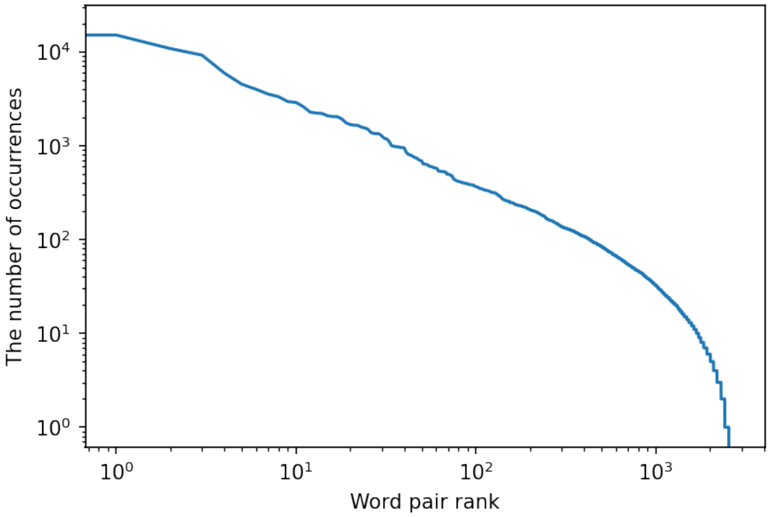

Fig. 1: Frequency distribution of each word pair in the parallel English-Slovak corpus OPUS2 represented by logarithm of Zipf's curve

In the following step of the analysis, we manually checked the word pairs, and according to the observed mistakes, we divided them into categories from A to J. The A category was for the correct translations, and the rest was for minor or major mistakes in the translations. For example, we found inflected word forms (‘*compiled*’: ‘*zostavujú*‘, ‘*advocacy*’: ‘*obhajobu*‘), words translated with the same word that is not in Slovak (‘*brook*’: ‘*brook*’), abbreviations (‘*bbc*’: ‘*bbc*’), proper names (‘*bruno*’: ‘*bruno*’) or even non-existing English words (‘*wwe*’: ‘*mozeme*’), etc. Each category and its explanation are shown in Table 1.

The bar chart in Fig. 2 outlines how many word pairs were in each category. Given the graph, most of the word pairs received category A. However, the translation was not always the most frequent one (e.g., ‘*customer*’: ‘*odberateľ*’).

In the last step, we proposed our Slovak translation for each incorrect word pair in the categories from B to J. All word pairs in the A category kept their original Slovak translations. After annotating the English headwords with our Slovak translations, we measured the cosine similarity between word vector

Table 1: Categories, their description, weights and an example of a word pair from the respective category.

Category	Description	Weight	Example
A	correct translation	1	*'admit'* : *'priznať'*
B	inflected word form	0.80	*'advocacy'* : *'obhajobu'*
C	different part of speech	0.30	*'darkness'* : *'temné'*
D	translated as same non-Slovak word, abbreviations	0.20	*'bbc'* : *'bbc'*
E	proper names	0.20	*'bruno'* : *'bruno'*
F	synonym or incorrect translation	0.10	*'intensity'* : *'svietivosť'*
G	incomplete word pair	0.20	*'brigadier'* : *'brigádny'* (generál)
H	non-existing English word	0.10	*'wwe'* : *'mozeme'*
I	interjection	0.80	*'boom'* : *'bum'*
J	missing diacritics	0.60	*'joy'* : *'radost'*

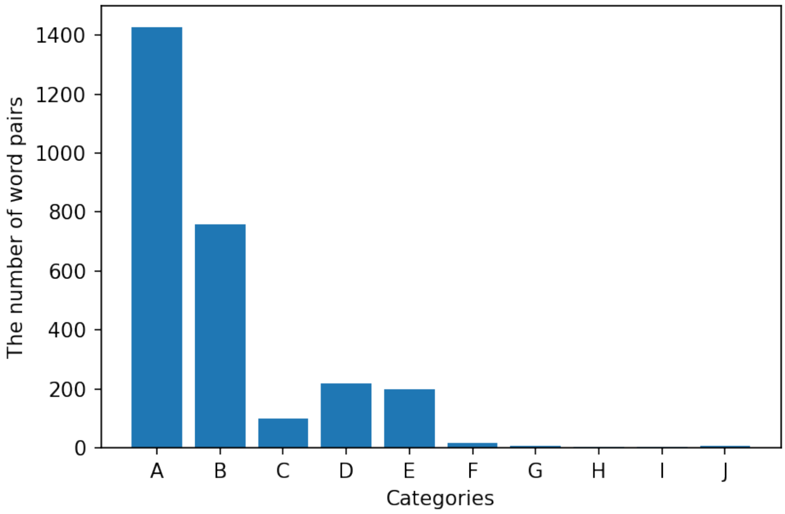

Fig. 2: The number of word pairs in each category.

representations of the original translation and our suggestion. To obtain these word vector representations, we used a pre-trained fastText word embedding model for the Slovak language. The results of this experiment are shown in Fig. 3.

3.1 Assigning weights

Given the described aspects, we assigned a weight to each word pair to reflect its relevance. Another reason was to increase the accuracy when the model finds word pairs with higher weight and not penalize the model for not including word pairs with a lower weight.

The weight was determined to be in the range between 0 to 1, so the first necessary step was to scale frequencies of the word pairs to the same range. However, as shown in Fig. 1, the word's frequency is inversely proportional to

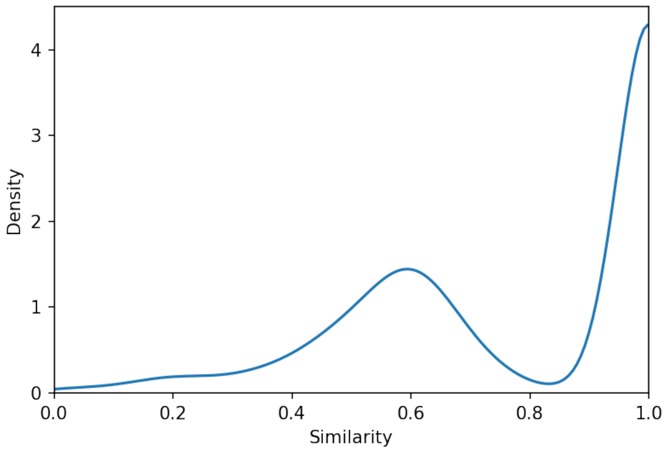

Fig. 3: Cosine similarity distribution from 0 to 1.

the word's rank, meaning that only a few word pairs have a very high frequency (the highest is 19,077), and the majority of the word pairs in the dataset have a frequency lower than 2,500. As a result, most word pairs would receive very small weight. The solution was to compute the logarithm of each weight first and then re-scale the numbers to the range between 0 to 1.

Furthermore, we added weights between 0 to 1 to each category, depending on whether the category represents a major or minor mistake. For example, category B or I was not considered a huge mistake, so it received higher weight while the weights for categories D and E were significantly lower. Categories, their explanations with an example, and assigned weights are shown in Table 1.

The cosine similarity was already in the range from 0 to 1, so it was not needed to process it.

Having frequencies scaled, weights for categories assigned, and cosine similarities computed, we multiplied these three values to obtain the weights for each word pair. Fig. 4. shows the overlapping histograms of weights distribution in each category.

However, the assigned categories and cosine similarity computed between the word vector representations of the original Slovak translation and proposed translation are subjective aspects. Thus we decided also to use only scaled frequencies (to the range from 0 to 1) obtained from the parallel corpora as weights for the word pairs when evaluating the models. The following sections discuss the results.

4 Evaluation

We chose models MUSE and VecMap, for the evaluation to see how the performance changes before and after applying weights on each word pair in the test dataset. We divided weights into two subcategories: first is weights computed

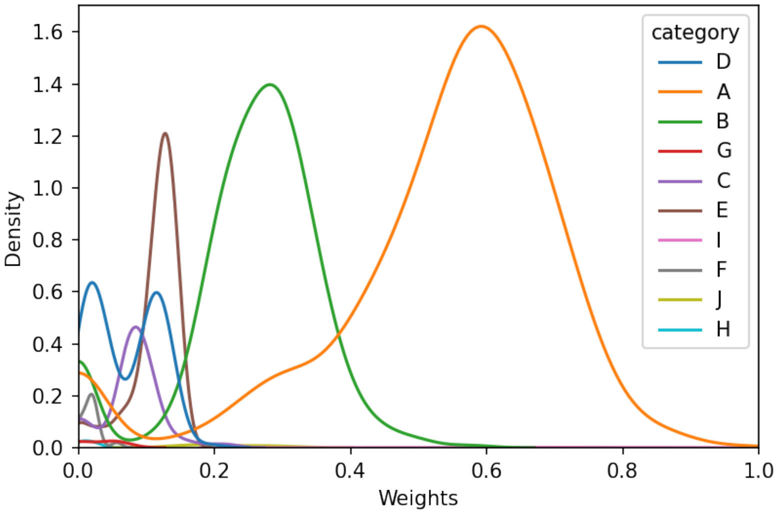

Fig. 4: Histograms of weights distribution in each category.

from weighted categories, frequencies, and cosine similarity, and the second one is scaled frequencies of the word pairs used as weights. Table 2 summarizes the results.

Table 2: The performance of Muse and VecMap models before and after applying weights and scaled frequencies used as weights on each word pair in the evaluation dataset.

	Without Weights	With Weights	Scaled frequencies
Muse (%)	30.41	**34.60**	32.82
VecMap (%)	38.15	48.43	**54.74**

Firstly, we downloaded from the muse library pre-trained word embeddings aligned into a single vector space for English and Slovak language. The English-Slovak evaluation dataset contained 2,793 word pairs and 1,500 English headwords, so we extracted the nearest neighbors of each English headword from the aligned vector space, depending on how many times the headword occurred in the evaluation dataset. For example, we extracted the first three nearest neighbors if there was an English headword with three different Slovak translations. Then we compared how many extracted word pairs using the muse model matched word pairs from the evaluation dataset. In the second evaluation, we included the weights from our analysis and scaled frequencies of the word pairs.

According to Table 2, the model's performance did not markedly change when using scaled frequencies as weights, but the numbers are slightly higher when considering weights from the analysis.

For the VecMap, we trained the model on English and Slovak FastText monolingual embeddings. The training was under strong supervision using 5,000 English-Slovak word pairs obtained from the MUSE training dataset. The result was embeddings for English and Slovak aligned to a single vector space. The evaluation part was the same as for the MUSE model. In comparison to the previous model, performance was significantly better when applying weights on each word pair. The best performance model achieved when considering only scaled frequencies as weights.

We examined and compared the word pairs that MUSE and VecMap models found through nearest neighbor search. MUSE looked up 294 word pairs from the evaluation dataset that VecMap was not able to find. Reversely, VecMap found 506 word pairs that MUSE did not include. Both models matched in 539 word pairs. Table 3 displays word pairs with the highest frequency and/or highest weight in which MUSE and VecMap models differ from each other.

Table 3: Comparison of the word pairs with the highest frequency (in hits per million) and/or highest weight that were found either by MUSE or VecMap model.

EN	SK	Frequency	Weight	Muse	VecMap
decrease	*zníženie*	**274**	**0.8709**	Yes	No
estonia	*estónsko*	42	0.7592	Yes	No
luxembourg	*luxembursko*	39	0.7555	Yes	No
euro	*eurá*	188	0.3957	Yes	No
vii	*vii*	254	0.1733	Yes	No
carefully	*starostlivo*	101	0.8115	No	Yes
decrease	*pokles*	253	0.8663	No	Yes
infection	*infekcia*	283	**0.8730**	No	Yes
hey	*hej*	1349	0.7728	No	Yes
tel	*tel*	**2384**	0.2000	No	Yes

5 Conclusion

Although applying weights in the evaluation of the MUSE model did not change the results remarkably, they helped to provide a more accurate picture of the VecMap model. VecMap outperforms the MUSE model in every evaluation discipline, and the evaluation proposes that VecMap is better when considering the most frequent word pairs in the parallel corpora.

Moreover, this analysis suggests that the choice of the word pairs and their frequency in corpus plays an important role in the evaluation and can reflect the model's performance more accurately.

Future work should focus on the analysis of the evaluation datasets for various language pairs. Especially we want to emphasize the morphologically rich languages to see to what extinct the inflected word forms influence the evaluation of the model's performance.

Acknowledgements. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

References

1. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2289–2294 (2016), https://aclanthology.org/D16-1250
2. Artetxe, M., Labaka, G., Agirre, E.: 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). pp. 451–462 (2017), https://aclanthology.org/P17-1042
3. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 789–798 (2018), https://arxiv.org/abs/1805.06297
4. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017), https://arxiv.org/abs/1710.04087
5. Glavaš, G., Litschko, R., Ruder, S., Vulić, I.: How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 710–721. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1070, https://aclanthology.org/P19-1070
6. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The sketch engine: ten years on. Lexicography pp. 7–36 (2014), http://dx.doi.org/10.1007/s40607-014-0009-9
7. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043 (2017), https://arxiv.org/abs/1711.00043
8. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018), https://arxiv.org/abs/1712.09405
9. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. ArXiv **abs/1309.4168** (2013), https://arxiv.org/abs/1309.4168

10. Ruder, S., Vulić, I., Søgaard, A.: A survey of cross-lingual word embedding models. J. Artif. Int. Res. **65**(1), 569–630 (May 2019), https://doi.org/10.1613/jair.1.11640

11. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Chair), N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012), https://aclanthology.org/L12-1246/