Does Size Matter?
Comparing Evaluation Dataset Size for the Bilingual Lexicon Induction

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 have been a popular approach for inducing bilingual lexicons. However, the evaluation of this task varies from paper to paper, and gold standard dictionaries used for the evaluation are frequently criticised for occurring mistakes. Although there have been efforts to unify the evaluation and gold standard dictionaries, we propose a new property that should be considered when compiling an evaluation dataset: size. In this paper, we evaluate three baseline models on three diverse language pairs (Estonian-Slovak, Czech-Slovak, English-Korean) and experiment with evaluation datasets of various sizes: 200, 500, 1.5K, and 3K source words. Moreover, we compare the results with manual error analysis. In this experiment, we show whether the size of an evaluation dataset impacts the results and how to select the ideal evaluation dataset size. We make our code and datasets publicly available [1]

Keywords: Cross-lingual word embeddings, bilingual lexicon induction, evaluation dataset's size.

1 Introduction

Cross-lingual word embeddings (CWEs) have drawn attraction among researchers due to their ability to connect meanings across languages. CWEs enable the alignment of two (or more) sets of independently trained monolingual word embeddings (MWEs) into one shared cross-lingual space, where similar words obtain similar vectors [18].

Given this property, they have proven useful in many NLP applications, for instance, document classification [16], machine translation [4,6] or language learning [1].

A broadly used way to evaluate these models is through the bilingual lexicon induction (BLI) task. In this task, the objective is to find top k target words for a source word whose word vectors are the closest in the aligned vector space. This is achieved typically by computing cosine similarity between the

[1] https://github.com/x-mia/Eval_set_size

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 47–56, 2023. © Tribun EU 2023

source and target word vectors. Then, the retrieved word pairs are compared to those occurring in the evaluation dataset, often referred to as gold-standard dictionary [18].

However, the evaluation process and evaluation datasets are not unified, and they vary from paper to paper, using various training parameters, evaluation criteria, and evaluation datasets [17,14,3]. This obstructs our ability to accurately assess the results, monitor any progress in new models, and compare models with each other.

Moreover, the most popular evaluation datasets Muse [7] are often criticised since they were compiled automatically and contain much noise in the form of occurring mistakes, such as inflected word forms (*wave*, singular - *Wellen*, plural, German dataset), a different part of speech (*darkness*, noun - *temné*, *dark*, adjective, Slovak dataset), same word translations (*android* - *android*, Korean dataset) [9]. They often have disproportional part-of-speech (POS) distribution, where a quarter of data consists of proper nouns that do not carry any lexical meaning and cannot indicate the performance adequately. For example, *Barack Obama*, *Skype*, *Bruno*, *Wisconsin*, etc. [15].

Some efforts have focused on uniting the evaluation by investigating how different training parameters influence the results [10]. Furthermore, some studies suggest consolidating the evaluation datasets through equal POS representations [15,13]. Nonetheless, other factors and properties of the evaluation dataset should also be considered.

One of them and one of the unifying steps is determining the size of the evaluation dataset. Muse evaluation datasets contain 1.5K source words, which have become a standard for the BLI.

In this paper, we investigate whether the number of source words in the evaluation dataset impacts the results. We explore how many source words are enough to assess the quality of the model. Our motivation is to study whether we can use fewer source words to create a high-quality evaluation dataset that reflects the model's performance precisely while minimising the time and effort of the human annotators to compile it.

We evaluate three popular baseline CWE models, i.e., Muse [7], VecMap [2,3], RCLS [14] on three diverse language pairs: distant language pair (Estonian-Slovak), close language pair (Czech-Slovak), and language pair that do not share a script (English-Korean). We utilise evaluation datasets of different sizes: 200, 500, 1.5K, and 3K source words, and observe how the results change. We compare the results against human performance to ensure the precise reflection of the resulting quality.

Our contribution is manifold:

- We provide an evaluation of three common baseline models with evaluation datasets of various sizes.
- We set the appropriate number of source words for the efficient, high-quality evaluation dataset that is less time-consuming to compile and indicate accurate results.

- We propose another unifying property for evaluation datasets to make the evaluation process comparable and reproducible for other researchers.

Our paper is structured as follows. In Section 2, we present the details of baselines and training, our datasets and the metrics used. In Section 3, we evaluate the models with datasets of various sizes, show the outcomes and discuss the results. Finally, we offer concluding remarks in Section 4.

2 Bilingual Lexicon Induction

The BLI task includes several aspects, such as evaluation datasets used, evaluation metrics, and selected baselines and training. We introduce them in this section.

Evaluation Datasets. Since we wanted to assess only the impact of the size, the aim was to make each size group of source words as similar as possible.

The Estonian-Slovak (et-sk) evaluation dataset was compiled using the Estonian-Slovak dictionary from Denisová (2021) [8]. This dataset claims 40% accuracy; therefore, we post-processed the word pairs manually after selection. We randomly sampled 3K source words and then randomly split them into 200, 500, and 1.5K source words for the subsequent evaluation.

The evaluation dataset for Czech-Slovak (cs-sk) was constructed manually mostly from words that are different in both languages (e.g., *želva - korytnačka, turtle*). We applied the same procedure as for the Estonian-Slovak evaluation dataset, i.e., we compiled a 3K source-word dataset and randomly sampled 1.5K, 500, and 200 source words.

For English-Korean (en-ko), we used the open-source evaluation dataset MUSE, which consists of 1.5 source words (English-Korean test set). Afterwards, we randomly selected 500 and 200 source words for the subsequent evaluation. To extend this dataset, we randomly sampled another 1.5K source words from the full English-Korean MUSE dataset. Afterwards, we combined them with the MUSE evaluation dataset to create a 3K-source-words dataset.

Metrics. The most common reported metric in the BLI task is precision (%). Precision or P@*k* is the ratio of True Positives (TP) to the sum of the True Positives and False Positives (FP) defined by the following formula:

$$P = TP/(TP + FN)$$

Where *k* represents the number of top target words retrieved for a source word. In this paper, we compute P@1, i.e., we retrieve one closest target word for each source word.

Baselines. MUSE [7] is a generative-adversarial-network-based model in the unsupervised setting (MUSE-U). The supervised (MUSE-S) setting and setting that relies on identical strings (MUSE-I) uses iterative Procrustes alignment.

VecMap is a framework that encompasses various stages, including orthogonal mapping, re-weighting, and dimensionality reduction, within its supervised settings (VM-S, VM-I) [2]. In its unsupervised setting (VM-U) [3], VecMap employs a robust iterative self-learning procedure.

RCLS is an orthogonal-mapping-based method with implemented convex relaxation in the retrieval stage. We trained this method in the supervised setting only.

Seed Lexicons. The Seed lexicons used in supervised training were compiled the same way as evaluation datasets. For Estonian-Slovak, we randomly sampled 5K source words from Denisová (2021)'s dataset [8]. For Czech-Slovak, we automatically constructed a 5K-source-words dataset consisting of identical words from the MWE vocabularies. For English-Korean, we used Muse training dataset [7].

Training. The default settings closely adhere to the training outlined in [7] for the Muse model, and VM-S and VM-I are presented in [2]. The parameters for VM-U follow the training procedures from [3]. Additionally, RCLS training settings align with those described in [14].

During the training, we experimented with two MWEs. We used pre-trained FastText embeddings [11] for Estonian, Slovak, Czech, English, and Korean, which were trained on texts from Wikipedia[2] with dimension 300.

The second pre-trained embeddings were provided by SketchEngine [12].[3] These embeddings were trained with the same method [5] but on different data (web corpora), with dimensions 100 for Estonian-Slovak and English-Korean, and 300 for Czech-Slovak.

3 Evaluation

In the evaluation process, we assessed all three models on Estonian-Slovak, Czech-Slovak, and English-Korean with the split datasets into four groups: 200, 500, 1.5K, and 3K source words. We extracted one target word for each source word by computing the cosine similarity between the source and target word vector. Then, we calculated P@1. Tables 1, 2, and 3 show the results.

Tables 1 and 2 show that the difference between the precision for both groups fluctuates wildly within a margin of approximately 15%. The best results were achieved for the Estonian and Czech in combination with Slovak when the 3K-source-word datasets were used.

This could mean that we get more precise results with datasets containing more source words or that the underlying distribution varies significantly after splitting the dataset.

The exemption was Table 3, the English-Korean language pair. In the majority of cases, the best results were gained with the 1.5K-source-word dataset,

[2] https://www.wikipedia.org/

[3] https://embeddings.sketchengine.eu/

Table 1: The results for the Estonian-Slovak language combination.

et-sk (%)	FastText				SketchEngine			
	200	500	1.5K	3K	200	500	1.5K	3K
Muse-S	17.34	18.93	21.37	**23.18**	26.53	27.02	32.30	**36.14**
Muse-I	10.20	13.40	15.30	**16.65**	27.55	24.68	28.82	**32.03**
Muse-U	11.73	12.76	13.52	**15.64**	20.40	20.85	23.80	**27.14**
VM-S	19.89	25.53	26.88	**30.72**	28.57	28.72	34.81	**38.85**
VM-I	17.34	18.29	22.18	**24.60**	21.93	22.76	26.63	**30.15**
VM-U	15.30	16.17	19.67	**21.72**	22.95	22.12	26.63	**29.80**
RCLS	16.83	19.78	22.99	**27.05**	27.55	26.59	34.73	**38.28**

Table 2: The results for the Czech-Slovak language combination.

cs-sk (%)	FastText				SketchEngine			
	200	500	1.5K	3K	200	500	1.5K	3K
Muse-S	58.08	62.10	64.99	**68.72**	62.26	65.89	71.50	**75.72**
Muse-I	59.59	61.68	64.92	**68.93**	61.61	65.68	70.97	**75.48**
Muse-U	60.60	62.31	65.51	**69.25**	61.00	65.68	70.97	**75.44**
VM-S	59.09	60.63	66.41	**69.13**	62.62	65.47	71.50	**75.84**
VM-I	59.09	64.42	68.66	**72.10**	61.61	65.89	71.42	**75.52**
VM-U	59.09	64.21	68.58	**72.10**	61.61	65.89	71.50	**75.60**
RCLS	57.57	61.05	64.32	**68.04**	64.14	67.36	72.70	**76.48**

Table 3: The results for the English-Korean language combination.

en-ko (%)	FastText				SketchEngine			
	200	500	1.5K	3K	200	500	1.5K	3K
Muse-S	13.91	13.57	**17.44**	15.91	16.49	19.82	**21.23**	19.00
Muse-I	11.34	14.22	**17.16**	15.80	10.30	**15.51**	14.64	13.90
Muse-U	10.30	11.42	**13.94**	12.78	12.37	**13.36**	12.05	11.63
VM-S	29.38	29.52	**35.31**	33.80	21.13	20.90	**23.75**	21.58
VM-I	20.61	17.67	**21.72**	19.03	13.91	15.30	**15.41**	13.43
VM-U	12.37	14.22	**16.53**	14.51	**6.70**	5.81	6.51	5.63
RCLS	30.92	27.80	**34.40**	32.54	21.13	20.90	**22.91**	20.25

within a margin of approximately 5%, which is not as distinct as in the previous two groups. This dataset came from a different source than the other ones, and as the only one, it was compiled automatically, which might be the reason for various outcomes.

Moreover, when comparing the results with different MWEs, the models trained with SketchEngine MWEs outperformed the models trained with FastText MWEs in most cases. We explain the differences in Subsection 3.1 in further detail.

In the next step, we split all three 3K-source-word datasets into six random groups of 500 source words. Afterwards, we evaluated VM-S with these datasets as an example. The objective was to observe whether the large gaps would be preserved in a different setup or whether the changed word distribution would bring more balance. Table 4 shows the development of this experiment.

Table 4: The results of 3K-headword evaluation datasets split into groups of 500.

VM-S	ET-SK		CS-SK		EN-KO	
	FastText	SketchEngine	FastText	SketchEngine	FastText	SketchEngine
I.	26.73	29.34	64.64	70.50	28.33	17.45
II.	21.42	25.10	61.89	68.00	28.45	17.78
III.	26.30	33.04	60.45	67.28	31.12	18.67
IV.	22.82	30.65	58.10	63.36	27.55	20.00
V.	22.73	30.31	57.74	64.22	29.35	19.07
VI.	21.61	29.55	53.52	57.64	30.06	18.88

Given Table 4, the gaps for each group of evaluation source words were reduced, remaining within the margin of approximately 8%. This suggests that random sampling seemingly might preserve the underlying distribution; however, the variance in the real scenario is more significant.

3.1 Error Analysis

Due to the inconsistencies in the outcomes, we performed manual error analysis for the Estonian-Slovak and Czech-Slovak language pairs while using the model VM-S as an example. Table 5 outlines the results.

Based on the results stated in Table 5, we can observe that the gaps reduced significantly, staying within the margin up to 4%. The best result was achieved twice with the 1.5K-source-word datasets, once with the 200- and 500-source-word datasets.

The reasons behind the large gaps between the results were twofold. Firstly, the top first target word that the model found was not in the evaluation dataset, although it was correct, e.g., *ajajärk* (*time period, era, epoch*) - *obdobie* (VM-S), *doba* (evaluation dataset).

Table 5: Manual error analysis of the results of the model VM-S for Estonian-Slovak and Czech-Slovak.

VM-S	ET-SK		CS-SK	
	FastText	SketchEngine	FastText	SketchEngine
3K	45.41	56.51	86.57	94.41
1.5K	**47.61**	59.67	**87.28**	94.83
500	46.59	58.51	86.31	**94.94**
200	46.93	**60.71**	86.86	94.44

This happened quite often because we did not include such a target word in the evaluation dataset, or the source words with multiple target words were randomly spread out in different datasets during the splitting. For example, the Estonian source word *puhuma* (*to blow*) had multiple target words such as, *pofúkať*, *fúkať*, *trúbiť*, *vanúť*, *zaviať*, *viať*, and in the 200-source-word dataset got the target word *zaviať* that was not the top one (which was *fúkať*) but the top second or third target word that the model found.

The second common reason was the uneven distribution of out-of-the-vocabulary (OOV) words. These were words that were not in the MWEs, low-frequency words, and words left out during the training. For example, *řeřicha* (*garden cress*), *pulec* (*tadpole*), *drobek* (*crumb*), *segisti* (*faucet*), *ahing* (*fish-spear*), etc.

On top of that, we analysed the gaps between SketchEngine and FastText MWEs. A closer look revealed that models trained with FastText MWEs were more likely to find a correct equivalent for proper nouns (e.g., *Clara*, *Emma*, *Erik*, *Phillip*, etc., see Table 6, type A) which have a bigger representation in the English-Korean datasets than in Czech-Slovak or Estonian-Slovak evaluation datasets.

Moreover, the English-Korean dataset contained a lot of noise in the form of words translated with the same word (e.g., *vms–vms*, *pgm-pgm*, etc., see Table 6, type B), for which the models trained with FastText were more likely to find a target word from the evaluation dataset.

On the other hand, models trained with SketchEngine MWEs were better at finding more accurate translation equivalents rather than words with lexical-semantic meanings (e.g., *hnědá*), see Table 6, type C). Additionally, they outperformed the FastText MWEs on the vocabulary, slang (e.g., *emps*) and low-frequency words (e.g., *stýskat*, *chasník*, etc., see Table 6, type D). The examples are displayed in Table 6

4 Conclusion

In this paper, we have investigated whether the standard 1.5K source words used in the evaluation datasets are enough to assess the CWE model accurately

Table 6: The differences between the models trained with FastText and SketchEngine MWEs (examples from EN-KO, ET-SK, and CS-SK trained with the VM-S model). (FT = FastText; SkeEng = SketchEngine)

Type	SRC	ED	FT	SkeEng	Description
A	Clara	클라라	클라라	외가에서	proper names
	Emma	엠마	엠마	희진	
	Erik	에릭	에릭	동료인	
	Phillip	필립	필립	옹은	
B	vms	vms	vms	램도	same word with same word
	pgm	pgm	pgm	변환하고	
C	hnědá	hnedá (brown)	žltohnedá	hnedá	precise translations
D	stýskat	cnieť (to miss)	-	cnieť	low-frequency words, slang
	chasník	mládenec (young man)	-	chasník	
	emps	mamka	-	mamka	

or whether we need more or fewer source words. We have experimented with evaluation datasets of various sizes: 200, 500, 1.5K, and 3K source words. Furthermore, we have split the 3K-source-words evaluation dataset into six random groups of 500 to observe how the outcomes would change.

Afterwards, we provided a manual error analysis, focusing on gaps between the results with different evaluation datasets. And we analysed the differences between the models trained with FastText and SketchEngine MWEs. We explained them and presented examples.

In conclusion, when splitting the datasets randomly, the results oscillated intensely within a margin of up to 15%. However, the manual error analysis revealed that the actual results faintly varied between 1-4%, remaining approximately the same within all datasets.

This outcome suggests that the random splitting of datasets does not ensure an equal underlying distribution within all the datasets. Moreover, it shows that the result strongly depends on the appropriate vocabulary choice rather than on the size of the dataset. This confirms the exemptional results from the English-Korean language pair evaluated on a dataset from a different resource than the others.

Generally, when selecting the size of the evaluation dataset, comparable results could be achieved even with a smaller dataset when the focus is on the quality of the chosen vocabulary for the evaluation dataset.

Acknowledgments. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2023062.

References

1. Akyurek, E., Andreas, J.: Lexicon learning for few shot sequence modeling. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4934–4946. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.382
2. Artetxe, M., Labaka, G., Agirre, E.: Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. pp. 5012–5019 (2018). https://doi.org/10.1609/aaai.v32i1.11992
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. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/P18-1073
4. Artetxe, M., Labaka, G., Agirre, E.: Unsupervised statistical machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3632–3642. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1399
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017). https://doi.org/10.48550/arXiv.1607.04606
6. Chronopoulou, A., Stojanovski, D., Fraser, A.: Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 173–180. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.naacl-main.16
7. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., J'egou, H.: Word translation without parallel data. ArXiv **abs/1710.04087** (2017). https://doi.org/10.48550/arXiv.1710.04087
8. Denisová, M.: Compiling an Estonian-Slovak dictionary with English as a binder. In: Proceedings of the eLex 2021 conference. pp. 107–120. Lexical Computing CZ, s.r.o. (2021), `https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_06_pp107-120.pdf`
9. Denisová, M., Rychlý, P.: When word pairs matter: Analysis of the english-slovak evaluation dataset. In: Recent Advances in Slavonic Natural Language Processing (RASLAN 2021). pp. 141–149. Brno: Tribun EU (2021), `https://nlp.fi.muni.cz/raslan/2021/paper3.pdf`
10. 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 (2019). https://doi.org/10.18653/v1/P19-1070

11. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018), https://aclanthology.org/L18-1550
12. Herman, O.: Precomputed word embeddings for 15+ languages. RASLAN 2021 Recent Advances in Slavonic Natural Language Processing pp. 41–46 (2021), https://www.sketchengine.eu/wp-content/uploads/2021-Precomputed-Word-Embeddings.pdf
13. Izbicki, M.: Aligning word vectors on low-resource languages with Wiktionary. In: Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022). pp. 107–117. Association for Computational Linguistics (2022), https://aclanthology.org/2015.mtsummit-papers.27
14. Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., Grave, E.: Loss in translation: Learning bilingual word mapping with a retrieval criterion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2984. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1330
15. Kementchedjhieva, Y., Hartmann, M., Søgaard, A.: Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3336–3341. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1328
16. Klementiev, A., Titov, I., Bhattarai, B.: Inducing crosslingual distributed representations of words. In: International Conference on Computational Linguistics. pp. 1459–1474 (2012), https://aclanthology.org/C12-1089
17. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. ArXiv **abs/1309.4168** (2013). https://doi.org/10.48550/arXiv.1309.4168
18. Ruder, S., Vulić, I., Søgaard, A.: A survey of cross-lingual word embedding models. The Journal of Artificial Intelligence Research **65**, 569–631 (2019). https://doi.org/10.48550/arXiv.1706.04902