Parallel, or Comparable? That Is the Question

The Comparison of Parallel and Comparable Data-based Methods for Bilingual Lexicon Induction

Michaela Denisová

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

Abstract. Extracting translation equivalents from parallel data has been considered the main and most efficient method in the lexicography field. However, parallel data are not always available or sufficient, especially for rare and low-resource language pairs. Translation equivalents obtained from comparable data offer a solution for this problem. This paper compares the performance of some methods that utilize either parallel or comparable data and demonstrates the results on the Estonian-Slovak language combination. We show that comparable data usage aspires to be a viable alternative for low-resource languages or rare language pairs, and we propose a new equation for more effective translation equivalents' induction.

Keywords: Parallel data, Comparable data, Translation equivalents' extraction, Estonian, Slovak

1 Introduction

Over many years, inducing translation equivalents through parallel data has been a preferred method among lexicographers. Parallel data often means parallel corpora or, in some cases, bilingual dictionaries. Current lexicographic tools utilizing either of those provide an effective and reliable method for obtaining translation equivalents as they include a lot of context information.[5]

However, rare language pairs or low-resource languages often lack parallel data, which could mirror in the quality and amount of the resulting translation equivalents. An alternative offer quickly developing modern approaches from the NLP field that claim using only comparable data for this task as sufficient. Among these methods are cross-lingual embedding models requiring monolingual word embeddings and only a few or no supervision signals at all. These models are evaluated on various tasks, such as cross-lingual named entity recognition, information retrieval, etc. This paper focuses mainly on the bilingual lexicon induction task, i.e., the BLI task.

The drawback of the comparable data-based methods is that they do not involve any context information as they are one-to-one or one-to-many alignments. Therefore, in contrast to the parallel data-based methods, they exclude

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 3–14, 2022.

any phrases or multi-word expressions that are valid parts of bilingual dictionaries.

This paper aims to compare the quality of the resulting translation equivalents obtained by chosen methods that utilize either comparable or parallel data. We show the results on a rare language combination, namely Estonian-Slovak. We induce a certain amount of translation equivalents with each method and evaluate them manually.

Our motivation is to explore whether recent trends favouring comparable data can compete with standard, widely used parallel data.

In this case, the bilingual dictionary-based parallel data method is represented by a pivot Estonian-Slovak dictionary obtained from English-Estonian and English-Slovak dictionaries [8]. Regarding parallel corpus, we manually evaluated the Estonian-Slovak dictionary extracted from the Estonian-Slovak parallel corpus EUR-Lex from SketchEngine.[6]

For the comparable data, we picked three currently most cited cross-lingual embedding models that are often used as benchmarks, MUSE [7,14], VecMap [1,2,4,3], and FastText for bilingual alignment [11]. Moreover, we experiment with different levels of supervision.

This paper is structured as follows. In Section 2, we introduce the chosen methods for our comparison and divide them into comparable and parallel-data-based. In Section 3, we explain the data we used for the evaluation in further detail. In Section 4, we present our results and provide a thorough comparison of the evaluated models. In Section 5, we offer concluding remarks and outline new ideas for future work.

2 Related Work

This section provides insight into methods used in this paper for obtaining translation equivalents from either parallel or comparable data.

2.1 Comparable Data

One of the solutions for extracting translation equivalents using comparable data provides cross-lingual embedding models. Cross-lingual embedding models have recently become a popular research topic as they can connect meanings across languages. The monolingual word embeddings of two or multiple languages are projected into a shared joint space where words with similar meanings obtain similar vectors. Afterwards, translation equivalents' candidates are extracted by computing cosine similarity.[15]

Frequently, the methods use various levels of supervision; they can be strongly supervised, semi-supervised or unsupervised. Supervision signals are represented by word-to-word dataset and vary from 5,000 words or can be comprised of similar strings and numerals.

These models offer a good solution for low-resource languages or rare language pairs because they do not require extensive parallel data. On the other

hand, they are still behind their parallel data-based counterparts concerning multi-word expressions or phrases as they do not include any context information.

In the following subsections, we describe three approaches we choose for our experiment. The reason behind this is that these approaches are stated in many papers as benchmarks and considered state-of-the-art models among cross-lingual models.

MUSE is a framework that combines domain-adversarial settings with applying the iterative Procrustes algorithm. The model can be trained in a supervised or unsupervised manner. Furthermore, it provides an option to rely only on identical strings. Code and pre-trained aligned word embeddings are available in an open-source GitHub library.[1]

VecMap is a robust framework that consists of multiple steps, including iterative refinement and bootstrapping techniques. Similarly to MUSE, it has multiple types of training, such as supervised, semi-supervised, training relying on identical strings and unsupervised training. Moreover, the library and code are available on GitHub.[2]

FastText utilizes orthogonal mapping and modified CSLS retrieving method.[11] Script is available on the GitHub repository[3] and pre-trained aligned word embeddings are available on FastText official website.[4]

2.2 Parallel Data

As mentioned above, this paper recognizes two types of parallel data: parallel corpora and bilingual dictionaries.

Parallel corpora usage has been the preferred approach among lexicographers as it produces high-quality dictionaries. The crucial argument is that parallel corpora contain rich context information, lowering human intuition in building a bilingual dictionary.[5]

The downside of parallel corpora is that it does not offer enough data for small languages or uncommon language pairs. The parallel corpora-driven bilingual dictionaries for such languages do not cover a sufficient amount of information for their users.

In this experiment, we manually evaluate Estonian-Slovak translation equivalents extracted from the Estonian-Slovak parallel corpus EUR-Lex with around 300,000,000 tokens. EUR-Lex is a multilingual corpus composed of texts from

[1] `https://github.com/facebookresearch/MUSE`
[2] `https://github.com/artetxem/vecmap`
[3] `https://github.com/facebookresearch/fastText`
[4] `https://fasttext.cc/`

the EUR-Lex database[5] that includes official documents and law and legislation-related texts.[6]

The statistics-based method for obtaining translation equivalents from the EUR-Lex Estonian-Slovak parallel corpus computes the probability that the current word pair is a translation equivalent by measuring the logDice association score.[16] This score considers the frequency of the current word pair (the higher frequency, the higher probability of being a translation equivalent) and the frequency of each word separately (the higher frequency, the lower probability of being a translation equivalent).[13]

Another option for inducing translation equivalents is to utilize existing bilingual dictionaries that share a common language. The idea is to connect meanings from two languages through a third, pivot language. The pivot language is usually well-resourced, for instance, English. This offers an alternative for rare language pairs with no parallel data. However, the resulting translation equivalents are often polluted by the pivot language causing incorrect alignments due to the polysemy of the words.

In this paper, we adopted the results from the Estonian-Slovak dictionary [8] that was obtained by merging English-Estonian and English-Slovak dictionaries. The dictionary was manually assessed on randomly sampled 1,000 translation equivalents, and the achieved accuracy with parallel data was around 40%.

3 Experimental Setup

In the training process, we experimented with two types of pre-trained monolingual word embeddings, FastText monolingual word embeddings[9] and SketchEngine monolingual word embeddings.[6][10]

Pre-trained FastText monolingual word embeddings for Estonian and Slovak contain around 300,000 words, and we included all of them in the models' training. In SketchEngine pre-trained embeddings for both languages were included around 1 million tokens. For our purposes, we worked only the first 300,000 and added embeddings for words with lower ranks that occurred in the evaluation dataset described in Section 4.

Moreover, we trained MUSE and VecMap models in a supervised (*model*-S), unsupervised (*model*-U) and semi-supervised mode that relies only on identical strings and numerals (*model*-I). FastText model, we trained in a supervised manner only. In the supervised training, we used our manually created word-to-word dataset with around 5,000 word pairs. The training dataset contained only words occurring in both monolingual word embedding files.

4 Evaluation

This section focuses on the data and methodology used in the evaluation process.

[5] `https://eur-lex.europa.eu/homepage.html`

[6] `https://embeddings.sketchengine.eu/`

To evaluate the Estonian-Slovak dictionary induced from parallel corpora, we randomly sampled 1,000 translation equivalents with two constraints: the achieved logDice score must be above 10, and the Estonian words' frequency must be above 1,000. This limited our choice to 35,528 translation equivalents. The aim was to eliminate noisy word pairs such as numbers, proper names from other languages, symbols or words from languages other than Estonian or Slovak.

Despite the dataset limitation, the manual evaluation revealed many mistakes. For example, numbers ('558' : '558'), incorrect proper names ('engström' : 'alfonsi'), website links ('vormis' : 'http://eur-lex.europa.eu/en/index.htm'), different language words ('nuts' : 'nuts'). Thus, the resulting accuracy was 16.1%. The result is displayed in Table 3.

To evaluate the cross-lingual embedding models, we utilized a basic Estonian vocabulary word list. We extracted this word list from the Basic Estonian Dictionary[7] provided by the Institute of the Estonian Language, which covers basic vocabulary aimed mainly at A2 to B1 CEFR learners.

We assigned the frequency to each word based on its occurrences in Estonian National Corpora from 2017. Afterwards, we randomly picked 70 Estonian words with very high frequency and 30 words with low occurrence. The aim was to see how each model performs on high- and low-frequency words.

The metric picked for evaluation is Precision at k (P @ k), which is the proportion of the number of correct translation equivalents to the number of all extracted translation equivalents where k is the amount of extracted target words for each source word. [12]

In the evaluation process, we extracted the 10, 5, and 1 nearest neighbour, their position, and their scores from which we computed two relative scores: the difference between the highest and current scores and the ratio between the highest and current scores. We gained 1,000 translation equivalents for 10 nearest neighbours, 500 translation equivalents for 5 nearest neighbours, and 100 translation equivalents for 1 nearest neighbour. Then, we randomly sampled 100 translation equivalents from the first two groups for the manual evaluation. In the group with 1 nearest neighbour, we evaluated all of them. The results are stated in Table 3.

Additionally, we excluded all unknown words. These unknown words arise because they do not occur in the pre-trained monolingual word embeddings in the first place. FastText did not include 4 of the Estonian words we picked for the evaluation, i.e., 'dressid' (soccer jersey), 'kontoritarbed' (office supplies), 'lastevanemad' (parents), 'ujumisriided' (swimwear). SketchEngine contains all of the words from the evaluation dataset.

During the manual control, we labeled each translation equivalent in two categories: correctness, whether the translation equivalent is correct or not, and in the second category, we reasoned our decision. The motivation was to analyze occurred errors. Table 1 summarizes all labels.

[7] `http://www.eki.ee/dict/psv/`

Table 1: Manual controls' labels and examples.

Correctness	Label	Example
Yes	meanings match	*'ammu'* : *'dávno'*(long time ago)
Yes	inflected word form	*'demokraatia'* : *'demokracii'* (democracy)
Yes	adjective in different grade	*'õnnelik'* : *'najšťastnejší'* (the happiest, correct: happy)
Yes	near equivalent or synonym	*'sõitma'* : *'šoférovať'* (travel/drive)
Yes	additional word needed	*'kontoritarbed'* : *'kancelárske'* (office, correct: office supplies)
No	different part of speech	*'sõbralik'* : *'priateľstvo'* (friendship, correct: friendly)
No	antonym	*'kiire'* : *'pomalé'* (slow, correct: fast)
No	number	*'kell'* : *'17.00'* (clock)
No	shortcut	*'kilo'* : *'kg'*
No	symbol	*'kell'* : *'+'*
No	meanings do not match	*'linn'* : *'radnica'* (city hall, correct: city)

Apart from assessing the quality of the translation equivalents, we looked at the models' performance on the high- and low-frequency words. We divided the labelled dataset into these two groups and computed their precision separately.

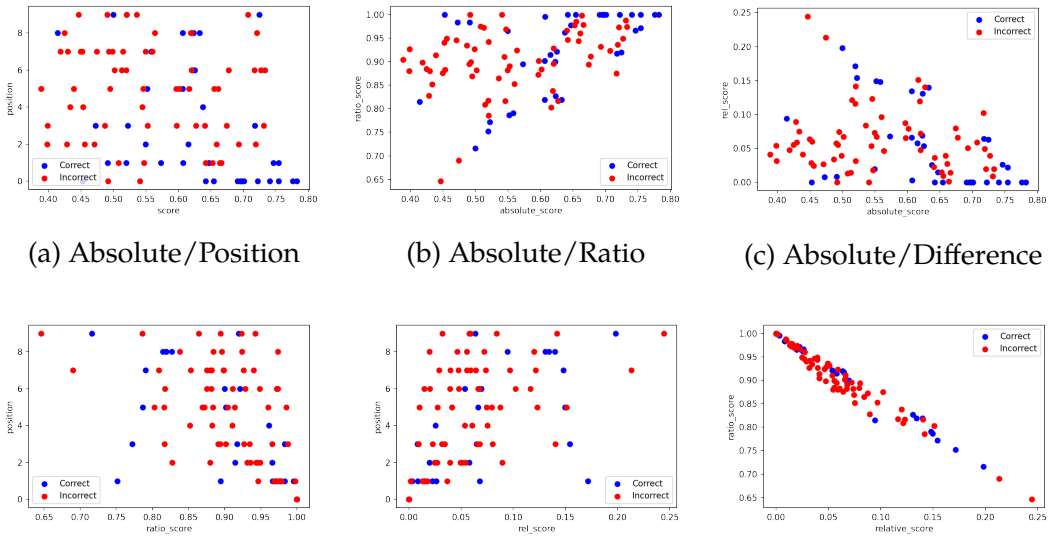
In most cases, models performed worse on low-frequency words; however, there are some exceptions, i.e., FastText model regardless the monolingual word embeddings, etc.

Furthermore, we observed big gaps between the precision for the high- and low-frequency words in some models. For instance, model Muse trained with FastText embeddings in a supervised mode, etc. All results are stated in Table 2.

Table 2: The comparison of the precision P@10, P@5, and P@1 when separating words into high- and low-frequency words.

(high-/low-frequency)	P@10	P@5	P@1
FastText			
Muse-S (%)	26.86/12.12	40/8	58.57/30
Muse-I (%)	28.76/7.4	25/16.66	45.71/13.33
Muse-U (%)	27.27/30.43	43.05/32.14	60/26.66
VecMap-S (%)	43.66/17.24	45.2/29.62	74.28/43.33
VecMap-I (%)	28.04/22.22	46.05/37.5	60/33.33
VecMap-U (%)	28.57/20	36.23/41.93	61.42/ 30
FastText (%)	35.21/17.24	33.8/27.58	72.85/40
SketchEngine			
Muse-S (%)	47.05/34.37	39.39/41.17	70/66.66
Muse-I (%)	28.35/27.27	38.15/29.16	68.57/56.66
Muse-U (%)	30.12/23.52	48.64/38.46	71.42/53.33
VecMap-S (%)	39.70/12.5	48.57/46.66	74.28/66.66
VecMap-I (%)	38.88/17.85	47.14/23.33	75.71/63.33
VecMap-U (%)	40/20	39.72/40.74	77.14/60
FastText (%)	13.88/21.42	45.45/47.05	77.14/56.66

After the assessment of the translation equivalents, we visualized scores and positions of the correct and incorrect translation equivalents in 6 different graphs: absolute score and position, absolute score and relative scores (difference, ratio), relative scores and position, and finally, relative scores against each other. The graphs of the VecMap trained in a supervised mode with FastText word embeddings are displayed in Fig. 1.



(a) Absolute/Position (b) Absolute/Ratio (c) Absolute/Difference

(d) Ratio/Position (e) Difference/Position (f) Difference/Ratio

Fig. 1: Various graphs for correct and incorrect translation equivalents extracted from VecMap trained in a supervised mode with FastText embeddings

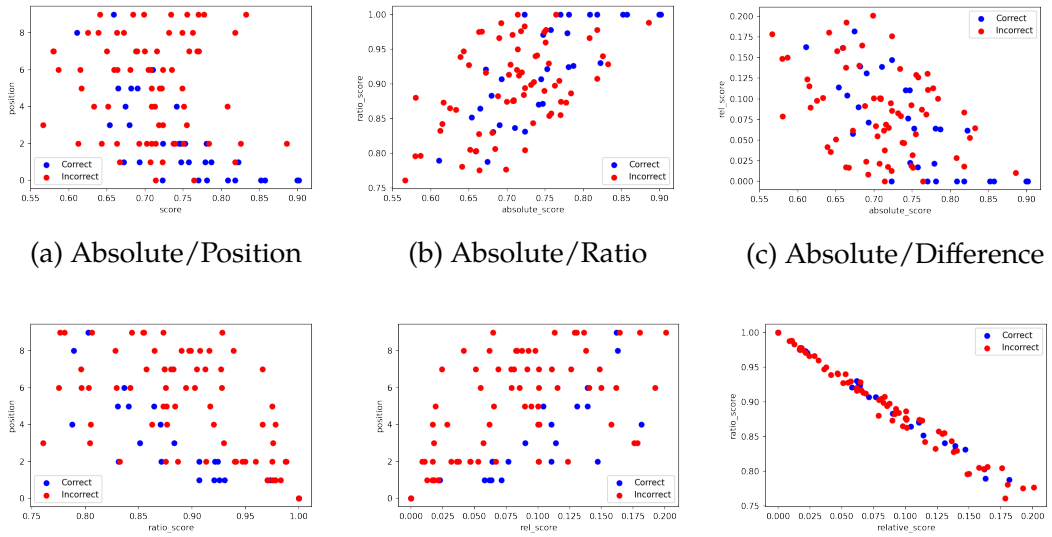
According to these graphs, the score line between correct and incorrect ranges between 0.4 - 0.5. This means that instead of extracting the 10, 5, or 1 nearest neighbours for each Estonian word, we can set the limit based on the current induced word's score and eliminate some incorrect translation equivalents. The limit can be expressed as follows:

$$limit = 0.45 + position * 0.01$$

However, given the Fig. 2 obtained scores with SketchEngine monolingual word embeddings were higher. Therefore, the score line rose to 0.6 - 0.7. In this case, the limit can be formulated like this:

$$limit = 0.65 + position * 0.01$$

In the next step, we decided to restrain the score limit of the translation equivalents, compute precision again, and see how the result changed. The results are shown in Table 3.



(a) Absolute/Position (b) Absolute/Ratio (c) Absolute/Difference

(d) Ratio/Position (e) Difference/Position (f) Difference/Ratio

Fig. 2: Various graphs for correct and incorrect translation equivalents extracted from VecMap trained in a supervised mode with SketchEngine embeddings

Given Table 3, the models' precision rose significantly after putting a limit constraint. Moreover, models trained with SketchEngine monolingual embeddings performed with and without limit in most cases better than with FastText embeddings.

Generally, all models achieved the best precision when only the closest nearest neighbour was considered. VecMap trained with SketchEngine monolingual word embeddings was able to ascend the precision up to 72%, which makes it the best model at P@1.

However, some inconsistencies among the results of the models' precision occurred. The reasons could be various. For instance, the random sampling picked more word pairs with a higher position, the model found better equivalents on higher positions, or we did not set the limit for extracting word pairs accurately.

The most remarkable gap between the monolingual word embeddings was in the model Muse-S and the FastText model.

Muse-S trained with SketchEngine embeddings found 30 word pairs that the model trained with FastText embeddings did not find. Reversely, FastText found 6 word pairs that were not in SketchEngine, and both matched in 11 word pairs. Table 4 displays some examples.

FastText trained with FastText monolingual word embeddings found 24 word pairs, SketchEngine 11, and both matched in 11 word pairs. In Table 5 are provided some examples.

Compared to the parallel data-based methods, the pivot dictionary significantly surpassed the Estonian-Slovak dictionary induced from a parallel corpus and is still a concurrence to the cross-lingual embedding models. On the other

Table 3: The precision P@10, P@5, and P@1 of comparable data-based models (MUSE, VECMAP, FastText) before and after applying a limit for the extraction of the translation equivalents compared to the performance of the parallel data-based methods.

	P@10/ Limit	P@5/ Limit	P@1
Comparable data			
FastText			
MUSE-S (%)	22/ 36.84	32/ 45.94	50
MUSE-I (%)	23/ 35.71	23/ 35.29	36
MUSE-U (%)	28/ 34.48	40/ 46.05	50
VECMAP-S (%)	36/ 45.2	41/ 46.06	65
VECMAP-I (%)	27/ 31.57	44/ 50.60	52
VECMAP-U (%)	26/ 32.81	38/ 42.35	52
FastText (%)	30/ 40.9	32/ 48.83	63
SketchEngine			
MUSE-S (%)	43/ 47.16	40/ 49.12	69
MUSE-I (%)	28/ 33.33	36/ 54.54	65
MUSE-U (%)	29/ 33.33	46/ 52.72	66
VECMAP-S (%)	31/ 35.36	48/ 52.17	72
VECMAP-I (%)	33/ 37.5	40/ 47.76	72
VECMAP-U (%)	33/ 33.76	40/ 58.33	72
FastText (%)	16/ 21.81	46/ 52.56	71
Parallel data			
Pivot dictionary (%)	40	-	-
Parallel corpus (%)	16.1	-	-

Table 4: Comparison of the word pairs that were found or were not found by MUSE either trained with FastText (MUSE-S-F) or SketchEngine (MUSE-S-S).

	ET	SK	Pos.	Score	Rank	Correct
MUSE-S-F	*laupäev*	*piatok*	1	0.621176	34506	**No**
	laupäev	*sviatok*	8	0.578794	34506	**No**
	suhkur	*cukry*	1	0.962491	28078	**Yes**
	samuti	*rovnako*	5	0.500055	108	**Yes**
MUSE-S-S	*laupäev*	*nedeľu*	4	0.756580	14506	**Yes**
	laupäev	*Nedeľa*	8	0.733683	14506	**Yes**
	laupäev	*víkend*	5	0.756557	14506	**No**
	suhkur	*škrob*	6	0.802334	7490	**No**
	samuti	*rovnako*	4	0.792155	190	**Yes**

Table 5: Comparison of the word pairs that were found or were not found by FastText either trained with FastText (FastText-F) or SketchEngine (FastText-S).

	ET	SK	Pos.	Score	Rank	Correct
FastText-F	*laul*	*pesnička*	0	0.253858	752	**Yes**
	lõplik	*konečné*	4	0.214931	5056	**Yes**
	õnnelik	*milovaný*	1	0.172903	9829	**No**
	sõitma	*cestovať*	0	0.264190	13698	**Yes**
	sõitma	*jazda*	2	0.180665	13698	**No**
FastText-S	*laul*	*hymna*	7	0.573021	4021	**No**
	lõplik	*presný*	8	0.522405	6906	**No**
	õnnelik	*šťastné*	6	0.517561	2381	**Yes**
	sõitma	*viezť*	3	0.628647	2734	**Yes**

hand, the parallel corpus-based method performed noticeably worse than most of the cross-lingual embedding models.

Importantly, we did not focus on words' senses and recall of the models. As we can see from the examples, the cross-lingual models could connect various word forms, but they perform poorly on different word senses. If we focused on models' recall, the parallel data-based models would perform better as they can capture context information.

5 Conclusion and Future Work

We have compared the precision of the translation equivalents induced by three approaches utilizing comparable data with two parallel data-based approaches. We have manually analysed the obtained translation equivalents and have provided insight into occurred errors. Additionally, we have introduced a new formula for extracting translation equivalents from cross-lingual embedding models more effectively.

Although the parallel data are still a competition to the comparable data, as they contain rich context information, in some disciplines, the comparable data outperformed parallel data significantly. Moreover, given the amount of research conducted in the cross-lingual embedding models' field, they represent a good alternative and show promising results for the future, either stand-alone or as supplement data, especially for low-resource languages or rare language pairs.

Finally, the formula for extracting translation equivalents was inferred manually based on the graphs' observations. However, every model and monolingual word embeddings are specific and require different weights for their limit. We propose for future work to implement an algorithm that would tailor the most appropriate limit for each model separately.

Acknowledgments The research in this paper was supported by the Internal Grant Agency of Masaryk University, project MUNI/IGA/1285/2021.

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.: 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)
4. 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
5. Atkins, B., Rundell, M.: The Oxford Guide to Practical Lexicography. OUP Oxford (2008)
6. Baisa, V., Michelfeit, J., Medveď, M., Jakubíček, M.: European Union language resources in Sketch Engine. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 2799–2803. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1445
7. 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
8. Denisova, M.: Compiling an estonian-slovak dictionary with english as a binder. In: Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. pp. 107–120 (2021)
9. 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)
10. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The tenten corpus family. In: 7th International Corpus Linguistics Conference CL 2013. pp. 125–127. Lancaster (2013), http://ucrel.lancs.ac.uk/cl2013/
11. 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 (2018)
12. 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, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1328, https://aclanthology.org/D19-1328

13. Kovář, V., Baisa, V., Jakubíček, M.: Sketch Engine for Bilingual Lexicography. International Journal of Lexicography **29**(3), 339–352 (07 2016). https://doi.org/10.1093/ijl/ecw029, https://doi.org/10.1093/ijl/ecw029
14. 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
15. 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
16. Rychlý, P.: A lexicographer-friendly association score. In: RASLAN (2008)