Bilingual Lexicon Induction From Comparable and Parallel Data: A Comparative Analysis

Michaela Denisová[1][0009−0001−8402−504X] and Pavel Rychlý[1,2][0000−0001−5097−4610]

[1] Natural Language Processing Centre, Masaryk University, Brno, Czech Republic
[2] Lexical Computing, Brno, Czech Republic
{x449884,pary}@fi.muni.cz

Abstract Bilingual lexicon induction (BLI) from comparable data has become a common way of evaluating cross-lingual word embeddings (CWEs). These models have drawn much attention, mainly due to their availability for rare and low-resource language pairs. An alternative offers systems exploiting parallel data, such as popular neural machine translation systems (NMTSs), which are effective and yield state-of-the-art results. Despite the significant advancements in NMTSs, their effectiveness in the BLI task compared to the models using comparable data remains underexplored. In this paper, we provide a comparative study of the NMTS and CWE models evaluated on the BLI task and demonstrate the results across three diverse language pairs: distant (Estonian-English) and close (Estonian-Finnish) language pair and language pair with different scripts (Estonian-Russian). Our study reveals the differences, strengths, and limitations of both approaches. We show that while NMTSs achieve impressive results for languages with a great amount of training data available, CWEs emerge as a better option when faced less resources.

Keywords: Bilingual lexicon induction · Cross-lingual word embeddings · Neural machine translation systems.

1 Introduction

Bilingual Lexicon Induction (BLI) is an intrinsic evaluation task focusing on retrieving translations of individual words. This task has been widely adopted for evaluating cross-lingual word embeddings (CWEs). The advantage of CWEs lies in their ability to align two sets of monolingual word embeddings (MWEs) into a shared cross-lingual space while exploiting comparable data and only a few or no bilingual supervision signals. [18]

Leveraging this property, they have proven to be useful in many natural language processing (NLP) applications, including machine translation [4,8], cross-lingual information retrieval [21], language acquisition and learning [22].

The BLI task from comparable data offers a promising alternative for low-resource or rare language pairs with insufficient parallel data. Traditionally, in lexicography, exploiting parallel data for retrieving translations has been a preferred method for many years. However, in the NLP field, while word-level extraction from parallel data was central during the era of statistical machine translation [13], the specific task of BLI has not received as much attention.

E. Nöth, A. Horák, and P. Sojka (Eds.): TSD 2024 LNAI 15048, pp. 28–39, 2024.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

In NLP, neural machine translation systems (NMTSs) that utilise parallel data present another solution for retrieving translations. Although they have been proven effective for translating sentences or texts, yielding state-of-the-art results, their potential in the BLI task has not been fully explored yet, and to our knowledge, there are no experiments using NMTSs for the BLI task.

In this paper, we comparatively study the BLI task from comparable and parallel data. We select MarianMT [20] to represent NMTSs using parallel data, and the three most cited state-of-the-art CWE methods using comparable data, i.e., Muse [5], VecMap [2,3], and RCLS [11]. We evaluate all models across three diverse language pairs: distant language pair (Estonian-English), close language pair (Estonian-Finnish), and language pair with different scripts (Estonian-Russian). Our motivation is to study the differences, similarities, strengths, and limitations of both approaches. Moreover, the discrepancy in training data volumes between these two approaches motivated us to understand how models perform under such different conditions on the same task. On top of that, we aim to investigate whether recent trends favouring comparable data for the BLI task can compete with standard, widely used parallel data.

Our contribution is threefold.

1. We provide a thorough comparison of the advantages and disadvantages of the BLI task from comparable and parallel data.
2. We comprehensively evaluate three CWE models and one NMTS across diverse and rare language pairs.
3. We make our code and datasets publicly available. [3]

This paper is structured as follows. In Section 2, we explain the background behind the BLI task from comparable and parallel data. In Section 3, we present the metrics, data and training details. In Section 4, we evaluate the baseline models exploiting comparable and parallel data and discuss the results. In Section 5, we offer concluding remarks.

2 Background

The objective behind the BLI task is to find the most suitable target word (or words) w_i^t for each source word w_i^s, given a list of P source words, where $P = \{w_1^s, w_2^s \ldots, w_n^s | n \in \mathbb{N}\}$. Afterwards, the output L, i.e., the list of the source and target words $L = \{(w_1^s, w_1^t), (w_2^s, w_2^t), \ldots, (w_l^s, w_l^t) | l \in \mathbb{N}\}$, is compared to a gold-standard evaluation dataset.

To achieve this objective, various approaches are available, often leveraging the two most common data types: comparable and parallel data. In the following Subsections 2.1 and 2.2, we outline the background of methods exploiting both data types, focusing on their advantages and limitations.

[3] https://github.com/x-mia/marianmt-bli

2.1 Comparable data

Comparable data or comparable corpus consists of texts in two or more languages that share a common domain or were collected under identical conditions. It is characterised as non-aligned and, most importantly, similar in genre. Additionally, these texts can be similar in size. [14]

The advantage of the models using comparable data lies in their availability for low-resource or untypical language combinations. By contrast, parallel corpora also often skew the actual distributions of lexical items in the target language, artificially elevating the occurrence of frequent words and cognates while disproportionately diminishing the presence of other, potentially more natural equivalents. Additionally, the texts in the parallel corpora are typically limited to the legislative or public domain, while comparable corpus tends to be more diverse.

In NLP, the BLI task from comparable corpora typically evaluates CWE models, where the aim is to find the closest target word vector to the source word vector in the aligned cross-lingual space, usually by computing cosine similarity between the source and target word vectors. The seminal study by Mikolov et al. (2013) [16] introduced this trend to NLP, followed by a plethora of research papers ranging from most cited baseline methods [1,5,11], comprehensive evaluation studies and surveys [18,9] to recent experiments with dynamic embeddings [15]. In this paper, we demonstrate the results across three CWE models, which are cited as baseline models in many research papers: Muse [5], VecMap [2], and RCLS [11]. We selected these models as they are publicly available and straightforward to use, and the performance gap compared to newer methods is not substantial. On top of that, they are more accessible and computationally less demanding to train than NMTSs.

Muse was released as a strong baseline model along with the evaluation and training datasets for over 110 languages. They employed a two-step process: adversarial training that develops a linear mapping between the source and target embedding spaces, challenging a discriminator to distinguish between them, and Procrustes refinement that optimises this mapping, leveraging a synthetic dictionary derived from the initial alignment. The introduction of the Cross-Domain Similarity Local Scaling (CSLS) metric aimed to address the high-dimensional space's hubness issue [4], significantly improving nearest-neighbour searches.

VecMap presented a multi-step framework for learning bilingual word embeddings while generalising and refining a wide array of previous approaches. Central to this framework is an orthogonal transformation, which allows for a detailed reinterpretation and improvement alongside additional steps such as normalisation, whitening, re-weighting, de-whitening, and dimensionality reduction. They also proposed a method in an unsupervised mode [3], relying on an unsupervised initialisation that exploits structural similarities between monolingual embeddings, coupled with a self-learning algorithm that iteratively refines the mappings.

RCLS aligned word embeddings from different languages by optimising the CSLS criterion, using convex relaxations for efficient optimisation, in contrast to traditional

[4] Hubness is an issue observed in high-dimensional space where some points are the nearest neighbours of many other points. [17]

approaches that typically solve a quadratic problem. It incorporated unsupervised data to enhance alignment, addressing the hubness problem by ensuring consistency between the loss used in training and inference.

2.2 Parallel data

The opposite of comparable data is parallel data or parallel corpora. It is a type of corpus that comprises two or more monolingual text collections that are aligned at the word, phrase, or sentence level. [14]

Exploiting parallel corpus for retrieving translations has been a preferred method mainly in lexicography, for instance, in the statistical-based method that computes the probabilities of word pair candidates based on their occurrences and co-occurrences presented in Kovář et al. (2016) [14]. In NLP, the parallel-data-based methods are often represented by NMTSs, which are not typically used for the BLI task. However, NMTSs could be used for compiling gold-standard dictionaries, as in the case of the widely used evaluation datasets Muse for evaluating CWEs.

The main advantage of the parallel corpus is that it contains rich context information while offering many target word candidates and performing well for polysemous words and multi-word expressions in contrast to the CWE models, which focus on pure word-to-word alignment. Moreover, the NMTSs can translate words that were unseen in training data, whereas CWEs are limited to the vocabulary in MWEs.

In this work, we opted for NMTS called MarianMT [20] for evaluation. MarianMT was trained using the Marian C++ library[5] on OPUS parallel corpora[6] [19], which have various domains, such as subtitles, public texts, web texts, etc. It contains over 1,000 models, of which all are transformer encoders-decoders with six layers in each component. Additionally, it supports a wide diversity of languages and language combinations, including European, non-European, endangered languages, etc.

3 Experimental Setup

In this Section, we introduce four key aspects of this experiment: the evaluation and training datasets used, training details of the CWE models, the setup details of MarianMT, and the evaluation metrics and procedures we employed during the evaluation.

3.1 Data

Training data. To train CWEs, we utilised pre-trained fastText MWEs for English, Estonian, Finnish, and Russian. These were trained on Wikipedia with dimension 300 and contain over 9.2 billion words in English and under 10 million tokens for the other languages. [10] For supervised mode, we selected the training datasets Muse [5] for Estonian-English. For Estonian-Finnish and Estonian-Russian, we compiled new training datasets by aligning Estonian-English with English-Finnish and English-Russian

[5] https://marian-nmt.github.io/

[6] https://opus.nlpl.eu/

Muse training datasets while using English as the pivot language. All training datasets contain 5K source words.

Regarding the training data for MarianMT, OPUS parallel corpora contain 115,564,910 sentences for Estonian-English, 42,353,565 sentences for Estonian-Finnish, and 29,699,112 sentences for Estonian-Russian.

Evaluation data. In the evaluation part, we exploited the Estonian-English evaluation dataset Muse. Since the evaluation datasets Muse are often criticised for uneven part of speech distribution [12] and containing errors in translations [7], we included Estonian-English [7], Estonian-Finnish [8], and Estonian-Russian [9] dictionaries that were manually post-edited by lexicographers from the Institute of the Estonian Language (EKI). All of these dictionaries are published under a CC BY 4.0 Deed licence. [10]

3.2 Training details of CWEs

For our comparison, we selected three state-of-the-art CWE methods, Muse, VecMap (VM), and RCLS. All three models are trained in a supervised mode (Muse-S, VM-S, RCLS), while only Muse and VM in an unsupervised (Muse-U, VM-U) mode and mode that relies on identical strings (Muse-I, VM-I).

The default settings closely followed the Muse training described in [5], RCLS setting in [11], and VM-S and VM-I in [2], and VM-U settings in [3]. The results are computed from the first 200K aligned embeddings.

3.3 MarianMT

We experimented with three pre-trained MarianMT models: Helsinki-NLP/opus-mt-et-en, Helsinki-NLP/opus-mt-et-fi, and Helsinki-NLP/opus-mt-et-ru, using Python programming language with PyTorch framework and HuggingFace library. [11]

We employed a series of parameters during the translation generation. We set the beam search to 20, disabled random sampling, specified to return ten target words, restricted the maximum number of new tokens that the model can generate in the response to 10, and enabled the output of scores.

3.4 Metrics

The most common evaluation metric in the BLI task is precision@k where k represents the number of target words retrieved for a single source word. In this paper, we report, in addition to precision, also recall and F1 scores using fixed and dynamic k.

We calculate the precision (P) as the ratio of the positive target words to the number of all target words that the model found (positive and negative). The recall (R) and

[7] http://www.eki.ee/dict/ies/

[8] http://www.eki.ee/dict/efi/

[9] https://portaal.eki.ee/dict/evs/

[10] https://creativecommons.org/licenses/by/4.0/

[11] https://huggingface.co/

F1 score representing the balance between precision and recall are computed using the standard formula.

In the case of the fixed *k*, we set it to 1, i.e., we report P@1, R@1, and F1@1. When evaluating CWEs using dynamic *k*, instead of limiting the retrieved target words based on top-*k* nearest neighbours, we restrict cosine similarity scores with the following formula adopted from Denisová (2022) [6]:

$$limit = S_C(x_i^s, x_j^t) + j * 0.01$$

, where $S_C(x_i^s, x_i^t)$ represents cosine similarity between the source word vector x_i^s and target word vector x_i^t, and j denotes the position of the target word, i.e., the target word with the closest target word vector has a position 0, etc. The value of $S_C(x_i^s, x_j^t)$ was adjusted for each model and language pair individually.

When evaluating MarianMT using dynamic *k*, we retrieved scores stored in the model to determine the reliability of each target word candidate. Then, we excluded each target word candidate with a score < 0.05.

4 Evaluation

Overall results for Estonian-English (et-en) are displayed in Tables 1 and 2, where Table 1 presents dynamic *k* and Table 2 fixed *k*, both using two different evaluation datasets. General results for Estonian-Finnish (et-fi) are outlined in Table 4 and for Estonian-Russian (et-ru) in Table 5.

Table 1. The results for the Estonian-English language pair evaluated using dynamic *k*.

et-en (%)	Muse dataset			EKI dataset		
	P	R	F1	P	R	F1
MarianMT	**50.39**	49.64	**50.01**	**32.17**	29.70	**30.89**
Muse-S	17.42	45.02	25.13	9.78	32.85	15.07
Muse-I	17.60	38.75	24.20	9.17	23.43	13.18
Muse-U	0.00	0.00	0.00	0.00	0.00	0.00
VM-S	21.60	50.36	30.23	8.31	**40.11**	13.77
VM-I	17.67	50.96	26.24	6.90	37.29	11.65
VM-U	15.15	46.95	22.91	6.74	36.84	11.40
RCLS	20.84	**53.82**	30.04	19.12	30.79	23.58

When looking at Tables 1 and 2, MarianMT outperformed CWE models in almost all metrics measured across Estonian-English language pair by a margin approximately ranging from 3% to 51%. Generally, the results for EKI evaluation dataset were worse than Muse by around up to 20%.

We examined some examples from both evaluation datasets, and the reason behind such a decrease in performance is that the Muse dataset is polluted by English to English equivalents, such as *act* - *act*, *ever* - *ever*, *girls* - *girls*, etc. Moreover, it contains a lot of English proper nouns which are identical in both languages, e.g., *adelaide* -

Table 2. The results for the Estonian-English language pair evaluated using fixed $k = 1$.

et-en (%)	Muse dataset			EKI dataset		
	P@1	R@1	F1@1	P@1	R@1	F1@1
MarianMT	**56.33**	**46.51**	**50.95**	**33.85**	**33.83**	**33.84**
Muse-S	40.33	33.30	36.48	25.11	23.49	24.28
Muse-I	36.80	30.38	33.28	19.42	18.17	18.77
Muse-U	0.00	0.00	0.00	0.00	0.00	0.00
VM-S	49.20	40.62	44.50	28.60	26.76	27.65
VM-I	49.07	40.51	44.38	24.26	22.70	23.46
VM-U	42.13	34.78	38.11	23.75	22.23	22.96
RCLS	45.53	37.59	41.18	30.74	28.76	29.71

adelaide, *hannah* - *hannah*, *selma* - *selma*, etc. We naturally get better outcomes when generating the target words in English for an English word. The same applies to fastText embeddings that are often noisy and contain English words.

Furthermore, we compared examples from MarianMT and RCLS models evaluated with the Estonian-English EKI evaluation dataset. Table 3 exemplifies the main findings. The performance of the model RCLS was worse than that of the model MarianMT, which confirmed the examination of their outputs. The main error which we discovered in RCLS model was that it aligned mainly word pairs with similar lexical-semantic relationships instead of translations. In MarianMT, the errors were various. Firstly, it often generated a target word with a capital letter (Type A), but the evaluation datasets were lowercase. The model appended extra numbers or punctuation to some target words (Types A and E) and, in some cases, generated complete sentences (Type C). Additionally, some target word candidates exhibited part-of-speech mismatches with the source word, for instance, a verb (jutustama) was translated as an adjective (narrated) (Type B). Finally, the model demonstrated the capability to handle multi-word expressions (Type D), offering a distinct advantage over CWE models that typically map single-word units to corresponding single-word units.

On the other hand, Tables 4 and 5 indicate a significant decrease in MarianMT performance. Although MarianMT still surpassed nearly all CWE models in the Estonian-Finnish evaluation by a margin of around 3% to 14%, it did not perform as well as the CWE model VM-U in the Estonian-Russian evaluation, where VM-U achieved the best performance.

The reason behind this is twofold. Firstly, the Estonian-Russian evaluation dataset contains a lot of target word variants for each source word, which influences the result, especially for models that are not performing well for polysemous words. Table 6 shows the number of target words in each dataset. We can observe that both Estonian-English evaluation datasets contain a lot of targets with one equivalent, whereas Estonian-Finnish and Estonian-Russian are more spread out. This is consistent with the performance of our models, i.e., the recall is high for Estonian-English but significantly decreases for Estonian-Finnish and Estonian-Russian.

NMTSs are known for generating only one output, which is reflected in the recall performance. Some models, including MarianMT, are able to offer more than one output, but access to the model is necessary. Table 7 displays a few examples from

Table 3. Examples of the source (SRC) and target (TGT) words from the Estonian-English evaluation dataset EKI compared to the output from MarianMT and RCLS models. The target words in bold are correct.

Type	SRC	TGT	MarianMT	RCLS	Explanation
A	aasta	year	Year 4 Year 3 Year **year**	**year** month summer autumn	capital letter, numbers
	mänguasi	toy	Toy **toy**	game play boardgame	
B	jutustama	narrate	**narrate** narrated	tale story tell	part-of-speech mismatch
C	aluspüksid	panties	Terry towelling and similar woven terry fabrics	trousers pants shirts	nonsense/ sentence
	loodetavasti	hopefully	I hope so. Hopefully.	probably possibly hope greatly	
D	ristsõna	crossword	crossword puzzles crossword word Crossword Puzzle	-	multi-word expression
E	au	honor	(au) - (au)	**honor** honour ain	punctuacion, symbols, English
	hõlmama	encompass	cover:	covers covering **encompass**	

Table 4. The results for the Estonian-Finnish language pair evaluated using dynamic and fixed k = 1.

et-fi (%)	P	R	F1	P@1	R@1	F1@1
MarianMT	19.40	**21.21**	**20.27**	**50.57**	**13.62**	**21.46**
Muse-S	17.74	11.25	13.77	37.57	10.12	15.95
Muse-I	16.11	11.38	13.34	36.22	9.76	15.37
Muse-U	14.70	12.03	13.24	36.93	9.95	15.67
VM-S	15.43	15.80	15.62	38.28	10.31	16.25
VM-I	13.65	16.59	14.97	42.33	11.40	17.97
VM-U	13.66	16.59	14.98	42.26	11.38	17.94
RCLS	**28.76**	12.11	17.04	39.49	10.64	16.76

Table 5. The results for the Estonian-Russian language pair evaluated using dynamic and fixed *k* = 1.

et-ru (%)	P	R	F1	P@1	R@1	F1@1
MarianMT	9.97	4.05	5.76	17.73	2.91	5.01
Muse-S	22.77	6.40	10.0	35.17	5.79	9.94
Muse-I	21.92	5.51	8.81	31.98	5.26	9.04
Muse-U	0.00	0.00	0.00	0.00	0.00	0.00
VM-S	18.54	10.04	13.03	35.96	5.92	10.16
VM-I	18.04	11.09	13.73	39.58	6.51	11.19
VM-U	19.21	**11.93**	**14.72**	**44.21**	**7.28**	**12.5**
RCLS	**30.59**	8.29	13.04	35.89	5.90	10.14

Table 6. The number of target words (TGW) in the evaluation datasets. [1] Muse dataset. [2] EKI dataset.

TGW	**et-en**[1]	**et-en**[2]	**et-fi**	**et-ru**
1	1240	3150	381	343
2	211	2	275	230
3	43	0	226	186
4	4	0	152	160
5	2	0	118	93
6+	0	0	256	488

the models MarianMT and VM-U trained across Estonian-Russian. It can be observed that the MarianMT model typically generates fewer target words, and a majority of these words include symbols, punctuation, and capital letters, i.e., the output is in the form of a sentence since the NMTSs are trained to translate sentences and not words in isolation.

Secondly, the amount of training data available in parallel corpus OPUS is larger and of better quality for Estonian-English than for Estonian-Finnish and Estonian-Russian. The Estonian-English corpora include over 115,500,000 sentences sourced from top-tier resources such as ParaCrawl, Europarl, DGT, and open subtitles. In contrast, the Estonian-Finnish corpora comprise over 42,300,000 sentences from similar high-quality sources like MultiParaCrawl, Europarl, DGT, and open subtitles. However, the Estonian-Russian corpora contain approximately 29,700,000 sentences, primarily from open subtitles and lower-quality sources like KDE4. Although over 29,000,000 sentences of training data for a language pair are not considered under-resourced, it has a major impact on the resulting quality of the translations. In the lower-data scenario and despite the training data volume discrepancy, the CWE models yield better results and prove to be a good supplement to the NMTSs.

5 Conclusion

In this paper, we have conducted a comparative analysis of the BLI task from comparable and parallel data. We have thoroughly discussed both data types and compared their advantages and limitations. From each group, we have selected models representing the specific data type, i.e., popular CWE models Muse, VecMap, and RCLS for the

Table 7. Examples of the source (SRC) and target (TGT) words from the Estonian-Russian evaluation dataset compared to the output from MarianMT and VM-U models. The target words in bold are correct.

SRC	TGT	MarianMT	VM-U	Explanation
inglane	англичанин	Англичанин.	**англичанин**	capital letters,
	англичанка	Англичанин	американец	sentence-form of
	британец		шотландец	the output
	бритт		**британец**	
			француз	
ületama	переходить	Преодолеть	**преодолевать**	capital letters
	проходить		**пересекать**	
	пересекать		доходить	
	переезжать		подниматься	
	переправляться		**переправляться**	
	преодолевать			
	превышать			
	превосходить			
	перекрывать			
patarei	батарейка	& Батарея	**батарея**	symbols,
	куча	Батарея…	**батареи**	punctuation
	батарейный	Батарея:	батарее	
	батареи		**батарейный**	

BLI task from comparable data, and NMTS MarianMT for the parallel data. We have evaluated these models across three diverse language pairs: distant (Estonian-English), close (Estonian-Finnish), and language pair with different scripts (Estonian-Russian), and we have analysed the results rigorously.

In conclusion, although NMTSs are still a competition to the CWE models due to their ability to capture context and handle multi-word expressions, their outcomes heavily depend on the amount of training data available. The CWE models represent a good alternative or can serve as a supplement data, especially for languages with fewer resources or when recall is favour over precision.

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. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/D16-1250
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. 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
6. Denisová, M.: Parallel, or comparable? That is the question: The comparison of parallel and comparable data-based methods for bilingual lexicon induction. In: Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022. pp. 3–13. Tribun EU (2022)
7. 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)
8. Duan, X., Ji, B., Jia, H., Tan, M., Zhang, M., Chen, B., Luo, W., Zhang, Y.: Bilingual dictionary based neural machine translation without using parallel sentences. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1570–1579. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.143
9. 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: Korhonen, A., Traum, D., Màrquez, L. (eds.) 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
10. 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)
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. pp. 2979–2984. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1330
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 (2019). https://doi.org/10.18653/v1/D19-1328
13. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition. pp. 9–16. Association for Computational Linguistics (Jul 2002). https://doi.org/10.3115/1118627.1118629
14. 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
15. Li, Y., Korhonen, A., Vulić, I.: On bilingual lexicon induction with large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 9577–9599. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.emnlp-main.595
16. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
17. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. The Journal of Machine Learning Research **11**, 2487–2531 (2010). https://doi.org/10.5555/1756006.1953015

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/10.1613/jair.1.11640
19. Tiedemann, J.: News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces, vol. V, pp. 237–248. Recent Advances in Natural Language Processing (2009)
20. Tiedemann, J., Thottingal, S.: OPUS-MT – building open translation services for the world. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. pp. 479–480. European Association for Machine Translation, Lisboa, Portugal (Nov 2020)
21. Vulić, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval pp. 363–372 (2015). https://doi.org/10.1145/2766462.2767752
22. Yuan, M., Zhang, M., Van Durme, B., Findlater, L., Boyd-Graber, J.: Interactive refinement of cross-lingual word embeddings. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5984–5996. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.482