Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022 , pages 8–14
Marseille, June 25, 2022

About Evaluating Bilingual Lexicon Induction

Martin Laville[1], Emmanuel Morin[1], Philippe Langlais[2]
[1]LS2N, UMR CNRS 6004, Universite de Nantes, France
[2]RALI-DIRO, Montreal, Canada
[1]firstname.lastname@ls2n.fr, [2]felipe@iro.umontreal.ca

Abstract

With numerous new methods proposed recently, the evaluation of Bilingual Lexicon Induction have been quite hazardous and inconsistent across works. Some studies proposed some guidance to sanitize this; yet, they are not necessarily followed by practitioners. In this study, we try to gather these different recommendations and add our owns, with the aim to propose an unified evaluation protocol. We further show that the easiness of a benchmark while being correlated to the proximity of the language pairs being considered, is even more conditioned on the graphical similarities within the test word pairs.

1. Introduction

Bilingual lexicon induction (BLI) is a long studied task (Rapp, 1995; Fung, 1998) that received a lot of attention recently (Gouws and Søgaard, 2015; Artetxe et al., 2016; Ruder et al., 2019; Hakimi Parizi and Cook, 2020). Thanks to the push of deep learning and so-called word-embedding models such as word2vec (Mikolov et al., 2013a), many new approaches vivified this task.

Many methods have emerged with the goal of computing accurate representations for cross-lingual word embeddings (CLWE). Mikolov et al. (2013b) used a linear transformation to project the source language into the target one, an approach known as *mapping*. In line, Faruqui and Dyer (2014) project the source and target embeddings in a new shared vector space. Artetxe et al. (2016) proposed several constraints (orthogonality, normalization, whitening etc.) to improve the quality of mapping.

More recently, unsupervised mapping methods (Conneau et al., 2017; Artetxe et al., 2018b) have been proposed which are nowadays starting to compete with supervised one. However, as noted in Artetxe et al. (2020), unsupervised methods, although interesting from a research point of view is not a realistic setup, as it is highly unlikely to have enough data to train CLWE without the existence of a seed lexicon.

A recent trend in BLI, known as *joint-training* consists in training the source and target word embeddings at the same time. Gouws and Søgaard (2015) proposed to concatenate the source and target corpora into which they randomly selected words (source or target) that they translated, thus producing a mixed corpus used to train a single embedding space. Following this, Duong et al. (2016) used a classic CBOW (Mikolov et al., 2013a) architecture and while training select the most appropriate translation of the context word based on a seed lexicon. Also Hakimi Parizi and Cook (2020) improved this by using the fastText model (Bojanowski et al., 2016). Finally, (Wang et al., 2020) mixed joint-trained embeddings with a mapping method.

While people have been working on the BLI task for many years, and even more so recently, the evaluation of BLI has been somehow surprisingly overlooked. (Conneau et al., 2017) created (making use of an internal translation tool) the MUSE dataset: over a hundred automatically collected bilingual lexicons of up to 100k pairs of words. This dataset rapidly became the defacto benchmark for BLI.

While MUSE is an invaluable resource per see, a number of concerns about it has surfaced. For instance, Czarnowska et al. (2019) observed that MUSE mainly gathers high frequency words, while Kementchedjhieva et al. (2019) indicate that about a quarter of the content of the lexicons consists of proper nouns, often perfectly identical graphically. Arguably, translating such entities is not of the utmost practical interest and focusing on less frequent words, for which translation are likely less listed in bilingual lexicons, is of more practical value.

In this paper, we review (Section 2) the different concerns already made about the evaluation in BLI (regarding the process itself or the data used) to which we add our own observations. We describe in Section 3 the data and the BLI systems we use to illustrate the concerns from Section 2. We then present in Section 4 the results of the different experiments made and analyze them. We finally conclude in Section 5.

2. Evaluation in BLI

The MUSE dataset is a collection of multiple bilingual lexicons in different languages: German, English, Spanish, French, Italian, and Portuguese languages all paired to each others. Lexicons from 39 other languages are also paired with English, in both directions. 108 language pairs are available in total, all with train and test sets already prepared.

2.1. Part-of-Speech (PoS) and Proper Nouns

Kementchedjhieva et al. (2019) conducted a study of the composition of MUSE. They manually annotated the English to/from German, Danish, Bulgarian, Arabic and Hindi lexicons[1]. We report in Table 1 the detail

[1] https://github.com/coastalcph/MUSE_dicos

of their annotations and the comparison made with the English Web Treebank (EWT)[2], which contains gold-standard PoS tags.

	Noun	PNoun	Verb	Adj/Adv
MUSE	49.6	24.9	12.5	13.0
EWT	35.6	15.1	23.3	25.9

Table 1: English PoS percentage of 4 categories for the MUSE dataset in comparison with the EWT. After Kementchedjhieva et al. (2019).

This table indicates that the proportion of these four categories in EWT — a representative set of sentences — is not respected in MUSE; the main problem being the high proportion of proper nouns. Moreover, Kementchedjhieva et al. (2019) note that proper nouns can reference totally different entities (for example first names or surnames) making it hard to establish a real sense (Pierini, 2008) and thus, questioning the pertinence of their presence in a BLI test set. In order to correct this issue, Kementchedjhieva et al. (2019) suggest as a first step to get rid of these pairs of words to use gazetteers to filter them out.

We also point in the next section that pairs of proper nouns are made of a lot of identical words and thus propose a simple solution to correct this.

2.2. Graphical Similarities of Word Pairs

We first focus on graphically identical word pairs. We suggest that these pair of words, present in high quantity in the MUSE dataset, are for the most part not of great interest, if not incorrect (*alignbars* or *wehrmacht* as the source and target word in the French-Spanish lexicon), and propose a simple solution to solve this. We then extend on the graphically close word pairs.

2.2.1. Identical word pairs

We report in Table 2 the percentage of identical word pairs in MUSE lexicons involving the German, English, Spanish, French, Italian, and Portuguese languages. We also add some languages linked only with English such as Czech, Norwegian and Russian.

Among the different bilingual lexicons we consider, many have over 30% of identical word pairs. In particular, German-French and German-Italian with over 49%, which is clearly worrisome. However, we note that with lexicons involving English, we have the lowest percentage, suggesting either a better control has been made on the English lexicons or the greater quality/quantity of the English corpora used to generate the dataset allowed a better quality in the automatically generated lexicons. Despite this, we still find some graphically identical word pairs in the *English-Russian* lexicon whereas the two languages have a different writing system (for instance, *motors* or *teen*).

	de	en	es	fr	it	pt	avg
de	-	18.5	29.4	49.2	49.8	46.1	38.6
en	16.0	-	16.5	21.0	21.1	18.4	18.6
es	20.3	18.4	-	30.3	31.3	47.9	29.6
fr	41.8	27.5	30.7	-	29.2	24.8	30.8
it	45.8	24.1	32.1	30.8	-	38.0	34.2
pt	40.9	21.6	47.5	27.4	41.2	-	35.7
avg	33.0	22.0	31.2	31.7	34.5	35.0	**31.3**
	en-cs		en-no		en-ru		-
	→	←	→	←	→	←	-
	16.1	17.6	26.1	36.8	2.4	0.0	-

Table 2: Percentage of pairs of graphically identical words in selected MUSE lexicons.

Taking advantage of this characteristic of MUSE is easy. For instance, Laville et al. (2020) reported that a simple approach to BLI based on this property could easily outperform mapping-based methods.

In order to understand why so many word pairs involve identical words and whether it makes sense to gather gather them in a test lexicon, we inspected the German-French and French-Spanish lexicons.

We sampled identical pairs of words and manually separated them in 4 different categories: First Names (FN), Named Entities (NE) (brand, geographical entities or names such as ”Roosevelt”), Doubtful (D) (*e.g.*, *#ffffff* or words from other languages, mostly English: *spirit* or *biography*). The remaining pairs being categorized as correct (C). The results of this annotation are presented in Table 3.

	FN	NE	D (EN pairs)	C	Total
de-fr	17.1	28.8	48.9 (21.0)	5.2	767
fr-es	19.6	33.5	40.9 (20.9)	6.0	465

Table 3: Sample of graphically identical word pairs in the German-French and French-Spanish lexicons and their manual classification.

The FN and NE categories can be seen as sub-parts of the PNoun PoS tag, however, we decided to separate them because of what they really represent. As exposed earlier, FN (such as *Federico* or *Bryan*) do not represent much interest in a BLI task because they do not convey any real sense. However, for the NE part, if obtaining the equivalent in an other language (we can not say translation here) for a named entity can be of interest in some scenario, it seems more suited to a bilingual version of a Named-Entity Recognition task than to BLI. We add that a major part of this category is made of cities or regions from Germany (*Gelsenkirchen*), France (*Orléans*) or other countries (*Lugano*, *Nebraska*). The pairs of words we classified as Doubtful are mostly made of words from other languages (for instance *freedom*, or *musica*) but also acronyms such as *nva* (a Belgium political party), and thus are arguably of no compelling interest for evaluating BLI. Finally, we note some pair of words made of real perfect cognates (for instance *terminal* is present in

[2] https://universaldependencies.org/treebanks/en_ewt/index.html

both the German-French and French-Spanish lexicons) but they only represent 5% of identical word pairs we sampled.

Thanks to the available proper nouns lists created by (Kementchedjhieva et al., 2019) on three language pairs with identical writing system (*English* to and from *Danish, German and Spanish*), we measure that 86% of the proper noun pairs are made of identical words (*Tennessee* or *Georges*).

Thus, we argue that a major part of graphically identical words are mainly of no interest in a BLI evaluating setting. Since we measured that only 5% of identical word pairs present a real interest, we suggest to getting rid of them while evaluating BLI, which will incidentally correct the problem of the proportion of proper nouns we discussed in Section 2.1.

2.2.2. Graphically close word pairs

We now take a look at graphically close pairs. After the removal of the identical word pairs, there is still an average of 40.1% word pairs with a Levenshtein distance of at most 3[3]. If we can logically note the proportion being higher between romance language (*Portuguese-Spanish*; 69.8% or *Italian-French*: 57.2%), it is surprising to see pairs such as *Italian-English* (46.5%) or *French-English* (44.4%) sharing that much similarities in their vocabulary, despite French and Italian being Romance languages while English is a Germanic one.

As the lexicons are made of a lot of graphically close words, we suggest, in addition to the evaluation on the lexicons without identical pairs, to split the lexicons in two sublists using the Levenshtein distance. We show later in Section 4 that the graphic proximity of the pair of words is a major factor in the success of the systems.

2.3. The *Morph* Dataset

Czarnowska et al. (2019) points three main problems with the existing datasets and MUSE: the lack of diversity in the frequency of the words, the fact that a word and its inflections can appear in both the train and test set (semantic leakage), and finally the lack of morphological diversity in most of the existing datasets. With the objectives of solving those problems, Czarnowska et al. (2019) introduce a new dataset to evaluate BLI, containing morphologically complete lexicons for 5 Slavic (Polish, Czech, Russian, Slovak, and Ukrainian) and 5 Romance (French, Spanish, Italian, Portuguese, and Catalan) languages. The lexicons are in every directions for both Slavic and Romance separately (meaning there is no dictionary from a Romance language to a Slavic and vice versa). We refer to them as *Morph* in the following.

Frequency Range: historically, BLI has mostly been focused on high frequency words. For instance, Mikolov et al. (2013b) used the $6k$ most frequent words to construct their training and test lexicons. Similarly, Czarnowska et al. (2019) reports that the pairs of words in the test lexicon of the MUSE dataset are all coming from the $10k$ most frequent source words. As Jakubina and Langlais (2017) empirically showed, it is far more difficult to identify translations of less frequent words, while we argue is a more sensible task (translations of common words are likely already listed in existing dictionaries). The *Morph* dataset is far more diverse on the frequency of its word pairs, containing, for the French-Spanish pair, 1 163 pairs of words with a source word from the top $10k$ of the vocabulary, but also (for instance) 1 126 pairs in the $500 - 600k$ range.

Semantic Leakage: Czarnowska et al. (2019) indicate that MUSE suffers of semantic leakage, meaning it is common for a word to appear in the training part of the lexicon as well as in the test part with a different inflection. In the *Morph* dataset the separation is done cleanly between the training and the testing part of the lexicons, because it is done on the lemmata, preventing the possibility of having two different inflections of a same word in the two lexicons.

Morphological Diversity: finally, the authors indicate that most words in MUSE has only one inflection form, while their dictionary is looking to have the best possible coverage for each lemmata. For instance, in the French-Spanish lexicon, the French verb *injecter* have 46 different inflections (from the first-person present tense *injecte* to the very seldom simple past form *injectâtes*).

The *Morph* lexicons present many interesting characteristics, however we point some problems. First, they do not come usable as is: if the presence of multiple inflections for each lemmata is an interesting feature, we think that being able to find them all, and particularly when there is that many (often out of use), is not the first objective of BLI. Thus, we recommend the usage of lemmata only.

In a similar vein, the high quantity of proposed translation lemmata per source lemmata is not really suited to a BLI task. For instance, the verb *abandonner* in French has 21 different candidates lemmata in Italian (*abortire, allentare, arrendere, bandire, cedere, concedere, defezionare, demordere, desistere, disertare, fermare, interrompere, liberare, mollare, piantare, recedere, rinunciare, rinunziare, sfollare, sgomberare, sgombrare*), and we think that finding 21 different translations for a single word is not what BLI is about.

About semantic leakage, we also point that, as the author indicate, a human translator is able to find more complex forms such as a first-person plural future form *hablarámos* thanks to their knowledge of the canonical form *hablar*. Thus, we argue that semantic leakage should not be seen as problematic in BLI as it is very similar to this case.

While *Morph* presents less languages pairs than MUSE, we strongly recommend its use whenever possible, as we do next. Last, we note that in their work, Czarnowska et al. (2019) only evaluate *Morph* using

[3] A threshold we found empirically as the best way to separate pairs of cognates.

P@1, while we show in the next section that MAP would be much more relevant.

2.4. Mean Average Precision (MAP) vs Precision at rank k (P@k)

While most works in BLI use P@k (typically with $k \in \{1, 5, 10\}$) to evaluate the quality of their method, Glavaš et al. (2019) advocate for the use of MAP instead. They point that MAP is more informative, because in P@k, a model that ranks a correct translation at $k + 1$ is equally penalised as the model that ranks it at rank $k + 1000$, while MAP gives a reward based on the rank.

In addition to that, they point that using MAP with only one correct translation per query is equivalent to the Mean Reciprocal Rank. However, we stress that MUSE proposes multiple valid translations per source word and therefore, their remark does not apply here. To show this, we report the ratio of target word per source word in Table 4. We indicate this for the lexicons from and to English, but also for the lexicons that do not include English in addition to the average per lexicon.

	en-x	x-en	incl. en	no en	avg
ratio	1.73	1.61	1.67	1.09	1.58

Table 4: Ratio of target words per source word in the MUSE dataset.

When using P@k, the evaluation system is just looking for the best ranked correct translation, leaving aside all the other ones. For instance, for a source word with 2 proposed translations, a system ranking one translation at top 1 and the other at top 2 $\{1, 2\}$ will be rewarded the same as a system ranking $\{1, 1000\}$ in P@1, while it will only be fully rewarded on the first case using MAP. Thereby, while using P@k, the presence of multiple translations in the lexicons does not become the assurance of a system of quality that takes into account polysemy as it will only look for one translation, which is obviously easier than finding them all.

We elaborate more on this problem by indicating that the ignored words in the case of multiple correct translations amplifies the problem of low frequency words or graphically distant pairs, as most systems are likely to find the higher frequency or the graphically closer translations first[4].

Thus, we strongly agree with Glavaš et al. (2019), and highly recommend the usage of MAP over P@k when evaluating BLI.

3. Protocol

In this section we briefly present the data and the two BLI methods we use to support the points discussed in Section 2.

3.1. Data

We use five different Wikipedia corpora as our training data: English, French, Italian, Russian and Spanish. We extracted the corpora using the WikiExtractor tool (Attardi, 2015).

We used the MUSE training part of the dataset when a training lexicon was needed.

3.2. BLI Methods

We compare two representative BLI methods that we now describe.

Mapping method Mapping (or alignment) methods consist in two steps. First, an embedding space is learnt separately for the source and target languages. We use fastText to train embeddings on the Wikipedia corpora. Second, a projection matrix is learned to map one language embedding space into the second one, allowing the comparison between languages. We use the VecMap tool (Artetxe et al., 2018a) as our mapping method.

Joint-training method Joint-training methods consist in the following steps. First, a bilingual corpus is build by concatenating both the source and target ones in order to create a shared vocabulary across languages. Then, the training of the embeddings for the two languages at the same time on the concatenated bilingual corpus, followed by the separation of embeddings into their original vocabulary. We use the *joint_align* framework (Wang et al., 2020) to do so. It also uses fastText to train the embeddings.

Wang et al. (2020) improved joint-training by adding a vocabulary reallocation phase such that, if an anchor word (i.e. a word graphically identical that appear in both part of the corpus and thus is only represented by one vector in the shared vocabulary) appears mostly in a language it is removed from the shared vocabulary in order to obtain a more precise representation during the mapping phase. For the alignment method, they use RCSLS (Joulin et al., 2018), which we follow.

3.3. Ranking of Candidates

Once the embeddings have been trained and projected in a shared space and in order to rank the candidates, we measure the similarity between every source word of the test dictionary with every target vocabulary word. We use the $CSLS$ (Conneau et al., 2017), an adaptation of the cosine similarity which reduces hubness[5], to order them:

$$CSLS(w_s, w_t) = 2\cos(w_s, w_t) - \mathrm{knn}(w_s) - \mathrm{knn}(w_t) \tag{1}$$

where w_s and w_t are the source and target word vectors, and knn(x) is a function that measures the mean cosine similarity between x and its k nearest neighbors.

[4] We back this claim with experiments in Section 4

[5] Words that tend to be the translation of many others.

		MUSE						*Morph*			
		es-fr	fr-it	it-es	en-ru	en-fr	avg	es-fr	fr-it	it-es	avg
Mapping	P@1	84.6	80.5	87.1	44.9	78.8	75.2	57.6	61.9	55.9	58.5
	MAP	87.9	84.4	87.3	51.3	72.8	76.7	45.0	48.7	45.8	46.5
Joint-Training	P@1	65.9	62.6	70.6	34.7	64.5	60.0	43.5	55.3	46.2	48.3
	MAP	71.8	67.5	73.7	39.8	61.1	62.8	37.3	44.8	41.4	41.2
Ratio target / source words		1.02	1.02	1.16	1.63	1.96	1.36	3.37	3.68	2.63	3.22

Table 5: Detailed results of the mapping and joint-training methods with MAP and P@1 metrics.

4. Experiments

From the *Morph* dataset, we considered the *Italian-Spanish*, *Spanish-French*, and *French-Italian* lexicons which have respectively 1 761, 1 173 and 2 273 source words. We selected the same language pairs from the MUSE dataset, as well as the *English-Russian* lexicon where the two languages have a different writing system, and finally the *English-French* pair. Each MUSE lexicon gathers around 1 500 source words.

4.1. P@1 vs MAP

We report in Table 5 the results obtained when using P@1 or MAP, the last row of the table indicates the ratio of target words per source word.

In Section 2.4, we reported that Glavaš et al. (2019) advocate for MAP because it is more informative, essentially because it takes into account all the proposed valid translations, and not just the highest ranked. This table confirms this claim and shows that the results in P@1 are higher than MAP when there is multiple possible translations, while MAP becomes higher whenever the target-to-source ratio tends to 1.

One notable exception however is for *English-Russian*, where the MAP is above P@1 despite a ratio of 1.63. This can be explained by a P@5 of 72.0 (+27 points from P@1), meaning that the system find a good part of the correct translations between the second and fifth rank, which is rewarded by the MAP. While for other languages, the P@5 is usually better than P@1 by at most 10 points.

And thus, it shows that having multiple possible translations artificially improves the P@k whereas intuitively, the introduction of polysemy should make it harder to find all the translations. Following this, we report only MAP results next.

4.2. Graphically Close Words

In Table 7, we report the results on different lexicons. In the first sublists (*not id.*), we remove all the graphically identical word pairs, as we suggested in Section 2.2. Then, we split these sublists based on Levenshtein distance: *Far* contains pairs of words with a distance over 3, while the sublist *Close* gathers close word pairs (distance less than 4).

This table clearly indicates that for both methods, it is much easier to conduct BLI on graphically close word pairs. If we let aside the *English-Russian* lexicon[6], the difference between the *Far* and *Close* sublists goes from 8 points (*es-fr* with joint-training on MUSE) up to 50 points (*it-es* with mapping on *Morph*).

Since popular reference lexicons such as MUSE are built largely from similar word pairs, performances reported on this dataset are in a way optimistic, and reporting results on both *Far* and *Close* lists as we did here is we believe a good practice.

4.3. Analysis

We show in Table 6 some output of the VecMap system for three hand-picked source words, along their rank in the list of proposed candidates, as well as their number of occurrences in the target corpus. This table supports the idea that in the case of multiple possible translations, the first target word found will likely be the graphically close or very frequent; and thus with P@1, the system will not be evaluated much on its ability to handle rare or graphically distant words.

On the *English-French* lexicon, 802 source words have at least 2 candidate translations. For 69% of the source word, the best ranked candidates was the most frequent one, for 74% it was the graphically closest with the source word and it was the most frequent and graphically closest one for 51% of the source words.

Source word	Target word	Rank	#occ.
customs	coutumes	1	7221
	douanes	2	4165
arch	arche	1	7407
	voûte	3	541
reveal	révéler	1	7577
	dévoiler	5	1858

Table 6: Some candidates proposed by the mapping method.

Figure 1 shows the correlation between the MAP and the average Levenshtein distance between word pairs of the test lexicon. It shows that the difficulty of the task does not only correlate with the diversity of the pair of languages considered, but also from the graphical proximity of word pairs. *English-Russian* are two languages that present many more differences than

[6] Those languages have different writing systems and thus variations in the Levenshtein distance mainly come from the length of the words.

		MUSE						*Morph*			
		es-fr	fr-it	it-es	en-ru	en-fr	avg	es-fr	fr-it	it-es	avg
Mapping	*not id.*	88.5	84.0	84.2	50.9	63.8	74.3	41.4	47.5	36.0	41.6
	Far	78.9	71.3	63.3	51.4	46.8	62.3	16.2	19.7	11.5	15.8
	Close	91.2	88.7	87.6	36.4	68.3	74.4	62.9	71.4	58.9	64.4
Joint-Training	*not id.*	68.6	64.0	67.1	39.3	48.9	57.6	33.3	43.4	30.5	35.7
	Far	62.4	55.1	52.8	40.1	35.4	49.2	13.9	19.7	10.6	14.7
	Close	70.4	67.3	69.4	33.7	53.1	58.8	49.0	63.5	45.9	52.8

Table 7: MAP results when test lexicons are split based on the graphical proximity of their word pairs.

French-Italian, but as the *Morph* lexicons are made of very few graphically close word pairs (and thus have a high average of Levenshtein distance), the systems does not perform well in both case.

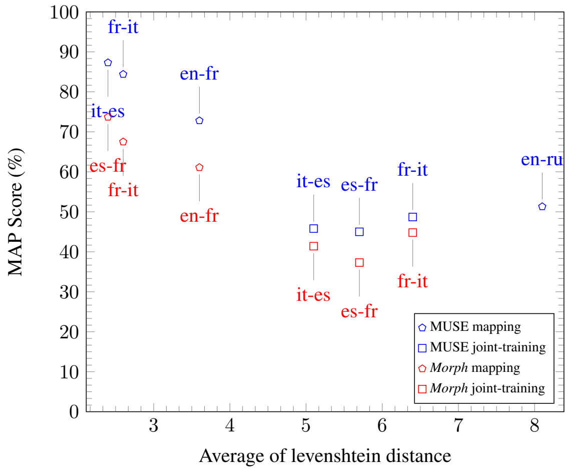

Figure 1: MAP versus Levenshtein distance of test word pairs.

5. Conclusion

In this work, we discuss different studies on BLI evaluation and add our own findings. We articulate a number of concerns that should guide BLI evaluation, leading us to formulate recommendations that are intended — we believe — to target what matters in practice; notably the ability to handle graphically distant pair of words.

First, using MUSE as an evaluation dataset, we recommend the removal of graphically identical pair of words. As we have seen in Section 2, they represent a major part of the MUSE lexicons and are often not interesting or even incorrect word pairs. Second, and if the language pairs allow it, we recommend an evaluation on both MUSE and *Morph*. Then, and for both dataset, we recommend that the lexicons should be evaluated as a whole but also in two groups based on the Levenshtein distance. The results presented in Section 4 show that for both type of methods (mapping or joint-training), the systems perform way better on close pair of words.

Also, we endorse the usage of MAP over P@k, especially if multiple candidate translations per source words are available, as it will be way more representative of the capacity a system to handle polysemy.

Finally, we highly recommend a more thorough evaluation than just looking at the MAP alone, and selecting a few pair of words with different characteristics can give great insights on the reality of the quality of the system and what are its strengths and weaknesses.

Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2289–2294, Austin, TX, USA.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). 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 (AAAI'18)*, pages 5012–5019, New Orleans, LA, USA.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). 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 (ACL'18)*, pages 789–798, Melbourne, Australia.

Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., and Agirre, E. (2020). A call for more rigor in unsupervised cross-lingual learning. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Attardi, G. (2015). Wikiextractor. `https://github.com/attardi/wikiextractor`.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *CoRR*, abs/1710.04087.

Czarnowska, P., Ruder, S., Grave, E., Cotterell, R., and Copestake, A. (2019). Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon 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)*, pages 974–983, Hong Kong, China, November. Association for Computational Linguistics.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 1285–1295, Austin, TX, USA, November.

Faruqui, M. and Dyer, C. (2014). Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, pages 462–471, Gothenburg, Sweden.

Fung, P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98)*, pages 1–17, Langhorne, PA, USA.

Glavaš, G., Litschko, R., Ruder, S., and Vulić, I. (2019). 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*, pages 710–721, Florence, Italy, July. Association for Computational Linguistics.

Gouws, S. and Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, pages 1386–1390, Denver, CO, USA, May–June.

Hakimi Parizi, A. and Cook, P. (2020). Joint training for learning cross-lingual embeddings with subword information without parallel corpora. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM'20)*, pages 39–49, Barcelona, Spain (Online), December.

Jakubina, L. and Langlais, P. (2017). Reranking translation candidates produced by several bilingual word similarity sources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 605–611, Valencia, Spain, April. Association for Computational Linguistics.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., and Grave, E. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium, October-November. Association for Computational Linguistics.

Kementchedjhieva, Y., Hartmann, M., and Søgaard, A. (2019). 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)*, pages 3336–3341, Hong Kong, China, November. Association for Computational Linguistics.

Laville, M., Hazem, A., and Morin, E. (2020). TALN/LS2N participation at the BUCC shared task: Bilingual dictionary induction from comparable corpora. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora (BUCC'20)*, pages 56–60, Marseille, France, May.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Pierini, P. (2008). Opening a pandora's box: Proper names in english phraseology. *Linguistik Online*, 36, 10.

Rapp, R. (1995). Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Wang, Z., Xie, J., Xu, R., Yang, Y., Neubig, G., and Carbonell, J. (2020). Cross-lingual alignment vs joint training: A comparative study and a simple unified framework.