Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction

Yova Kementchedjhieva
University of Copenhagen
yova@di.ku.dk

Mareike Hartmann
University of Copenhagen
hartmann@di.ku.dk

Anders Søgaard
University of Copenhagen
soegaard@di.ku.dk

Abstract

The task of bilingual dictionary induction (BDI) is commonly used for intrinsic evaluation of cross-lingual word embeddings. The largest dataset for BDI was generated automatically, so its quality is dubious. We study the composition and quality of the test sets for five diverse languages from this dataset, with concerning findings: (1) a quarter of the data consists of proper nouns, which can be hardly indicative of BDI performance, and (2) there are pervasive gaps in the gold-standard targets. These issues appear to affect the ranking between cross-lingual embedding systems on individual languages, and the overall degree to which the systems differ in performance. With proper nouns removed from the data, the margin between the top two systems included in the study grows from 3.4% to 17.2%. Manual verification of the predictions, on the other hand, reveals that gaps in the gold standard targets artificially inflate the margin between the two systems on English to Bulgarian BDI from 0.1% to 6.7%. We thus suggest that future research either avoids drawing conclusions from quantitative results on this BDI dataset, or accompanies such evaluation with rigorous error analysis.

1 Introduction

Bilingual dictionary induction (BDI) refers to retrieving translations of individual words. The task has been widely used for intrinsic evaluation of cross-lingual embedding algorithms, which aim to map two languages into the same embedding space, for transfer learning purposes (Klementiev et al., 2012). Recently, Glavas et al. (2019) reported limited evidence in support of this practice—they found that cross-lingual embeddings optimized for a BDI evaluation metric were not necessarily better on downstream tasks. Here, we study BDI evaluation in itself, as has been done for other evaluation methods in the past (cf. Faruqui et al., 2016's work on word similarity), with concerning findings about its reliability.

A massive dataset of 110 bilingual dictionaries, known as the MUSE dataset, was introduced in early 2018 along with a strong baseline (Conneau et al., 2018). Subsets of the MUSE dictionaries have been used for model comparison in the evaluation of numerous cross-lingual embedding systems developed since (cf. Grave et al., 2018; Jawanpuria et al., 2018; Hoshen and Wolf, 2018a,b; Wada and Iwata, 2018; Joulin et al., 2018). Even though the field has been very active, progress has been incremental for most language pairs. Moreover, there have been very few attempts at a linguistically-informed error analysis of BDI performance as measured on MUSE (cf. Kementchedjhieva et al., 2018). This is problematic for two reasons: on one hand, most systems greatly vary in their approach and architecture, so it is difficult to identify the source of the reported performance gains; on the other hand, the MUSE dataset was compiled automatically, with no manual post-processing to clean up noise, so the real impact of the performance gains is unclear.

In this work, we study the composition and quality of the MUSE data for five diverse languages: German, Danish, Bulgarian, Arabic and Hindi. A manual part-of-speech annotation of the test sets for these languages reveals a strikingly high number of proper nouns. We refer to linguistic literature to argue that proper nouns, having no lexical meaning but rather just a referential function, cannot reliably be used in the evaluation of word-level translation systems. We find that excluding proper noun pairs from the test dictionaries for the aforementioned languages affects the ranking and degree of performance gaps between five of the most influential recent systems for BDI.

With a new, more reliable ranking at hand, we

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3336–3341,
Hong Kong, China, November 3–7, 2019.

perform qualitative analysis on the performance gap between the best and second best systems for Bulgarian. This reveals another major issue with the data: limited coverage of morphological variants for the target words. Through manual verification of the models' predictions, we find that the gap in performance between the two systems is far smaller than previously perceived.

The uncovered issues of high noise levels (proper nouns) and limited coverage (missing gold standard targets) clearly have a crucial impact on BDI results obtained on the MUSE dataset, and need to be addressed. Filtering out proper nouns could be achieved automatically, by checking against gazetteers of named entities. We find that an automatic procedure for the filling of missing targets, however, yields only minor improvements. Until an alternative solution to the latter problem is found, we urge researchers to be cautious when reporting quantitative results on MUSE, and to account for the problems presented here through manual verification and analysis of the results. We share our part-of-speech annotations, such that future work can use this resource for analysis purposes. [1]

2 Bilingual Dictionary Induction

Improvements on BDI mostly stem from developments in the space of cross-lingual embeddings, which use BDI for intrinsic evaluation.

Systems Five influential recent systems for cross-lingual embeddings are MUSE (Conneau et al., 2018), which can be supervised (**MUSE-S**) or unsupervised (**MUSE-U**); VecMap, which also can be supervised (**VM-S**) (Artetxe et al., 2018a) or unsupervised (**VM-U**) (Artetxe et al., 2018b); and RCSLS (Joulin et al., 2018), a supervised system (**RCSLS**), which scores best on BDI out of the five. We refer the reader to the respective publications for a general description of the systems.

Metrics Performance on BDI in these works is evaluated by verifying the system-retrieved translations for a source word against a set of gold-standard targets. The metric used is Precision at k (P@k), which measures how often the set of k top predictions contains one of the gold-standard targets, i.e. what is the ratio of True Positives to the sum of True Positives and False Positives.

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

Data All systems listed above report results on one or both of two test sets: the MUSE test sets Conneau et al. (2018) and/or the Dinu test sets (Dinu et al., 2015; Artetxe et al., 2017). Similarly to MUSE, the Dinu dataset was compiled automatically (from Europarl word-alignments), but it only covers four languages. Due to the bigger size of MUSE (110 language pairs), we deem its impact larger and focus our study entirely on it.

3 Annotation-based observations

In order to gain insights into the linguistic composition of the MUSE dictionaries, we employ annotators fluent in German, Danish, Bulgarian, Arabic and Hindi (hereafter, DE, DA, BG, AR, HI) to annotate the entire dictionaries from English to one of these languages (hereafter, from-EN) and the entire dictionaries from these languages to English (hereafter, to-EN) with part-of-speech (POS) tags. Details on the annotation procedure can be found in Appendix A. Below, we discuss our findings on the POS composition of the data, and we evaluate the performance of RCSLS per POS tag.[2]

3.1 Analysis of POS composition

The average percentage of common nouns, proper nouns, verbs, and adjectives/adverbs in the dictionaries to-EN was respectively 49.6, 24.9, 12.5, and 12.9.[3] Nouns constitute half of the dictionaries' volume, while verbs and adjectives/adverbs collectively make up only about a fourth of the average dictionary. A skewed ratio between these three categories is not surprising: in the EWT dependency treebank, for example, which contain gold-standard POS tags, the proportion of noun, verb and adjective/adverb types is 34, 17 and 14 percent, respectively. Notice, however, that in the case of the MUSE data, the ratio is even more skewed in favour of nouns over the other two categories.

The large number of proper nouns in the dictionaries seems even more problematic. Proper nouns are considered to have no lexical meaning, but rather just a referential function (Pierini, 2008). Personal names usually refer to a specific referent in a given context, but they can, in general, be attributed to different referents across different contexts, and they are almost univer-

[2] For all experiments, we use the pretrained embeddings of Bojanowski et al. (2017), trained on Wikipedia.

[3] The numbers were similar across from-EN dictionaries.

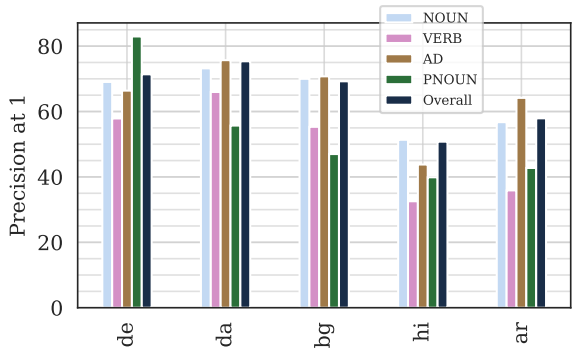

Figure 1: Precision of RCSLS by POS tag on to-EN data.

sally interchangeable in any given context. Some personal names and most place and organization names may have a unique referent, e.g. *Barack Obama, Wisconsin, Skype*, but these names still do not carry a *sense*, their referent is resolved through access to encyclopedic knowledge (Pierini, 2008). Considering that the pretrained embeddings which we use were trained on Wikipedia, we can expect that such encyclopedic information would indeed appear in the context of certain unique names, but importantly, the alignability of the embeddings for such entities would depend on the level of parallelism between the contents of Wikipedia articles in the different languages.

With these considerations in mind, one should wonder how stable the representation of names can be in an embedding space. This question has previously been raised by Artetxe et al. (2017). We address it empirically below.

3.2 Evaluation by POS

Figure 1 shows the precision of the RCSLS embedding alignment method on different POS segments of the test data in mapping to-EN (results from-EN were similar and are shown in Appendix B). Verbs pose a greater challenge to BDI systems than nouns and adjectives do. Generally, we can attribute this observation to the higher abstraction of concepts described by verbs. This is a known problem for word embedding methods in general (Gerz et al., 2016), which BDI systems naturally inherit.

With respect to proper nouns, we observe that they indeed introduce a level of instability in the evaluation of BDI systems. Notice that while the other parts of speech follow a similar pattern across languages, with higher precision obtained for nouns and adjectives/adverbs than for verbs,

Corpora	NOUN	VERB	AD	PNOUN
Wikipedia	69.0	57.9	66.4	83.0
Mixed*	64.0	55.5	59.4	37.6

Table 1: Comparison in performance by POS category with two different embedding sets. * The out-of-vocabulary rate for items in the dictionaries is negligible: 2, 0, and 1 for NOUN, VERB , and AD , respectively.

relative precision on proper nouns is highly variable. For DE, proper nouns are easier to translate than other parts of speech by a margin of 15%, for HI and AR they are easier than nouns and adjectives/adverbs, but harder than verbs, and for DA and BG they are hardest out of all four categories. We looked into the individual word pairs marked as proper nouns in the DE and DA data, as these languages are related and RCSLS performs comparably on them otherwise, and did not find any patterns that could explain the large differences. In fact, between the 384 proper noun pairs in the EN-DE dictionary and the 330 proper noun pairs in the EN-DA dictionary, there was an overlap of 279 pairs, retrieved with precision of 89.21% in the EN-DE setting and 51.30% in the EN-DA setting. We conjecture that this result relates to the level of parallel content between the Wikipedia dumps for the different language pairs, which is likely higher for EN-DE , since the dumps for these languages are also closer in size: 5.8M articles in EN, 2.3M in DE (and only 0.2M in DA).[4]

We evaluate this hypothesis through an experiment where we train an RCSLS alignment for DE-EN using the DE embeddings of Artetxe et al. (2017), trained on SdeWaC (Baroni et al., 2009) and the EN embeddings of Dinu et al. (2015), trained on ukWaC (Baroni et al., 2009), Wikipedia and the BNC [5] corpora. The level of parallel content between the data used to train the two sets of embeddings is thus far more limited in this case, and the DE embeddings are not explicitly trained on Wikipedia data. Table 1 summarizes the results: while with the new embeddings performance is somewhat reduced for nouns, verbs and adjectives/adverbs, precision at 1 for proper nouns, in particular, drops by over 50%, indicating that this category of test word pairs is indeed highly sensitive to the nature of the training data.

[4] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[5] Available at http://www.natcorp.ox.ac.uk

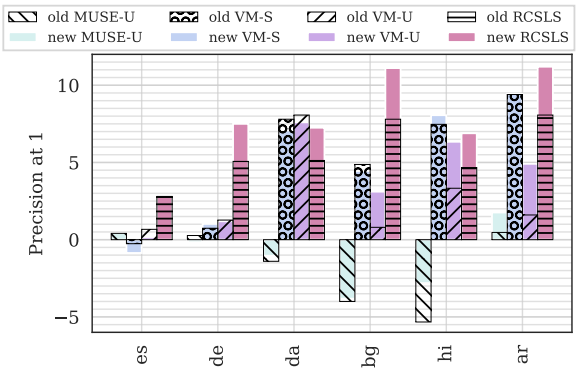

Figure 2: Absolute difference in performance on from-EN BDI, relative to MUSE-S. Pattern-filled bars show results as estimated on the original data (*old*), while colored bars show results as estimated on the cleaned data (*new*).

3.3 Re-ranking on clean data

Based on the analysis presented above, we removed all pairs that were annotated as proper nouns and all pairs that were marked as invalid during the annotation process.[6] This clean-up resulted in a drop in the size of the test dictionaries of about 25% on average. A detailed size comparison between the old test dictionaries and their new cleaned versions is presented in the top rows of Table 4 in Appendix B. Figure 2 visualizes a re-evaluation of the five systems for BDI listed in Section 2, on the original test data and on the new clean versions of the test dictionaries from-EN.[7] The results are reported in terms of change in performance relative to MUSE-S (chosen as a baseline) as estimated on the original MUSE data (pattern-filled bars) and on the cleaned version of the data (colored bars). The absolute system performances before and after the clean-up can be found in Table 4 in Appendix B.

We see that the ranking between the models changes most notably for AR, where RCSLS appears inferior to VM-S on the original test data, but on the clean data it emerges as best. For BG, the evaluation on the clean test data reveals that RCSLS outperforms the next best system, VM-S, by a larger factor than it appeared on the original test data. Lastly, for DA, evaluation on the original test data makes RCSLS seem far inferior to VM-S and VM-U, but on the clean test data we see that it outperforms VM-S and matches the performance of VM-U. These observations show that the noise coming from proper nouns has a large impact on the perceived ranking and difference in performance between systems.

4 *False* False Positives

With a more reliable estimate of the models' performance at hand, we next manually study the remaining performance gap between RCSLS, the best-performing model overall, and VM-S, the second best model overall, for EN–BG.[8] We present some examples in Table 2 and more can be found in Table 5, Appendix C.

We find that there are 125 source words that RCSLS translated correctly and VM-S did not. Upon closer inspection, we find that for 54% of these words, both RCSLS and VM-S predicted a valid translation, but RCSLS predicted a more *canonical* translation, which was listed among the gold-standard targets, while VM-S predicted another word form that was missing from the list of gold-standard targets. By more canonical we mean, for example, indefinite instead of definite forms of nouns and adjectives (see Ex. A, Table 2, masculine instead of feminine or neuter forms of adjectives (see Ex. B), singular instead of plural forms. To the extent that a more canonical translation should be considered better, RCSLS is definitely showing superiority over VM-S. It is not clear, however, if that should be the case, since for some words, the test dictionary exhibits higher coverage than for others, i.e. the less canonical translations are not omitted by design, but appear to be accidental gaps.

Another 19% of the instances where RCSLS outperformed VM-S, we find to be clear cases of a missing translation in the test dictionary, i.e. not a missing form of a listed target, but a missing synonym or a missing sense altogether (see Ex. C and D).

The two types of errors in precision at 1 discussed above can be considered cases of *false* False Positives, because they really should have been True Positives. The remaining 27% of the gap between the two models' performance indeed illustrate that RCSLS provides better translations in some cases (see Ex. E).

Notice, however, that it is not the case that RCSLS outperformed VM-S in all cases–for 50 test words, VM-S predicted a correct translation and RCSLS did not. Among these, there are cases of missing translations from the dictionary as well

[6] The latter constitute less than 1% of the removed data.

[7] The to-EN results were similar, see Appendix B.

[8] We also analyzed EN–DE, with very similar results.

Ex.	SRC	TGT	RCSLS	VM-S	Description
A	joke	шега лаф виц	шега [INDEF]	шегата [DEF]	definite form missing from targets
B	remembered	запомнен	запомнен[MASC]	запомнена[FEM]	feminine form missing from targets
C	hide	скриване	скриване [NOUN]	скрива̀т[VERB]	*hide* as a verb vs. *hide* as a noun
D	bench	пейка пейката	пейка	скамейка	synonym missing from targets
E	depot	депо	депо	гара	VM-S predicted 'train station'
F	crowned	коронован	коронована[FEM]	коронован[MASC]	feminine form missing from targets
G	pond	езерце	къщичка	езерце	RCSLS predicted 'cottage'
H	grants	субсидии	стипендии	стипендии	synonym missing from targets
I	armies	армии	армиите	армиите	definite form missing from targets

Table 2: Example translations from EN to BG. Underlined forms are more canonical. Grey forms are incorrect.

(see Ex. F), but they can explain less of the lack in performance of RCSLS, i.e. 50% of the translations of RCSLS are indeed erroneous (see Ex. G).

To summarize, originally the performance gap between the two models appeared to be $(125 - 50)/1125 * 100 = 6.67\%$, while after the manual verification, it is $(27\% * 125 - 50\% * 50)/1125 * 100 = 0.1\%$.[9] Such a substantial narrowing in the gap between the two models clearly indicates that conclusions drawn on the original result, i.e. that RCSLS is far superior that VM-S for this language pair, is hardly supported by the updated result.

A surface analysis of the subset of words for which neither RCSLS nor VM-S retrieved correct translations revealed similar patterns of extensive false False Positives, due to gaps in the coverage of the dictionary (see Ex. H and I). Our takeaway from these observations is two-fold. Firstly, when RCSLS retrieves a correct target form, it also usually retrieves its most *canonical* form. More importantly, the evaluation of BDI systems on even the cleaned test dictionaries still does not represent accurately the differences in quality between them, due to major gaps in the coverage of the test dictionaries.

5 Concluding remarks

Our study of the MUSE dataset revealed two striking problems: a high level of noise coming from proper nouns, and an issue of *false* False Positives, due to gaps in the gold-standard targets. The former problem, we conjecture, can be solved by filtering names out with gazetteers. The quality of this solution would depend on the coverage of the gazetteers. The more challenging problem, however, is filling in the gaps, especially in terms of inflectional forms. We carried out preliminary experiments aiming to enrich the EN–BG and EN-DE dictionaries. We extracted additional word forms of verbal and nominal targets from the UniMorph inflectional tables (Kirov et al., 2018), according to a manually designed morphosyntactic correspondence map.[10] Unfortunately, due to limited coverage of the UniMorph data, and, in the case of BG, limited vocabulary of the pretrained embeddings, the impact of this procedure was almost negligible. Alternative approaches for enrichment exists, of course, but we wonder how worthwhile further efforts would be. That is, especially in light of Glavas et al. 2019's findings that BDI performance is not necessarily indicative of cross-lingual embedding quality. We therefore hope that our work adds weight to the call of Glavas et al. (2019) for more reliable evaluation methods in cross-lingual embedding research. When BDI performance is used for evaluation purposes, it should be accompanied by manual verification, of the type presented here.

6 Acknowledgements

We thank Marcel Bollmann, Matthew Lamm, Maria Barrett, Meriem Beloucif, Mostafa Abdou, Rahul Aralikatte and Victor Petrén Hansen for help with annotations. We would also like to thank Adam Lopez, Andreas Grivas and Sameer Bansal for useful feedback on drafts of the paper, and to the anonymous reviewers for their comments and suggestions. Anders Søgaard was supported by a Google Focused Research Award; Mareike Hartmann by the Carlsberg Foundation.

[9] 1125 is the total dictionary size.

[10] Details can be found in Appendix D.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations. In *Proceedings of AAAI 2018*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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 (Volume 1: Long Papers)*, pages 789–798.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *Proceedings of ICLR 2018*.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving Zero-Shot Learning by Mitigating the Hubness Problem. *ICLR 2015 Workshop track*, pages 1–10.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *Proceedings of EMNLP*.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *ACL*.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes. *arXiv preprint arXiv:1805.11222*.

Yedid Hoshen and Lior Wolf. 2018a. An iterative closest point method for unsupervised word translation. *CoRR*, abs/1801.06126.

Yedid Hoshen and Lior Wolf. 2018b. Non-adversarial unsupervised word translation. *arXiv preprint arXiv:1801.06126*.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2018. Learning multilingual word embeddings in latent metric space: a geometric approach. *arXiv preprint arXiv:1808.08773*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 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*.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing procrustes analysis for better bilingual dictionary induction. *arXiv preprint arXiv:1809.00064*.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. *arXiv preprint arXiv:1810.11101*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474. The COLING 2012 Organizing Committee.

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

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Takashi Wada and Tomoharu Iwata. 2018. Unsupervised cross-lingual word embedding by multilingual neural language models. *CoRR*, abs/1809.02306.