Evaluating bilingual word embeddings on the long tail

Fabienne Braune[1,2], **Viktor Hangya**[1], **Tobias Eder**[1], **Alexander Fraser**[1]
[1]Center for Information and Language Processing
LMU Munich, Germany
[2]Volkswagen Data Lab Munich, Germany
fabienne.braune@volkswagen.de
{hangyav, fraser}@cis.uni-muenchen.de
tobias.eder@germanistik.uni-muenchen.de

Abstract

Bilingual word embeddings are useful for bilingual lexicon induction, the task of mining translations of given words. Many studies have shown that bilingual word embeddings perform well for bilingual lexicon induction but they focused on frequent words in general domains. For many applications, bilingual lexicon induction of rare and domain-specific words is of critical importance. Therefore, we design a new task to evaluate bilingual word embeddings on rare words in different domains. We show that state-of-the-art approaches fail on this task and present simple new techniques to improve bilingual word embeddings for mining rare words. We release new gold standard datasets and code to stimulate research on this task.

1 Introduction

Bilingual lexicon induction (BLI) is the task of generating accurate translations for each word in a list of source language words. Being able to perform BLI without parallel data is critical in many low resource scenarios. Bilingual word embeddings (BWEs) represent words from two different languages in the same vector space. BWEs have been shown to be very effective for BLI given a *small* seed lexicon (around 5000 word-pairs) as the only bilingual signal. Until now, BWEs have been evaluated on frequent words from parliament proceedings or Wikipedia articles and reached good accuracies on these datasets. However, evaluations on rare and domain-specific words have not yet been provided even though such evaluation scenarios are critical for applications like machine translation (e.g., mining of translations for OOV (out-of-vocabulary) items) or bilingual terminology mining. In this paper, we design a novel evaluation scenario for BWEs: given (i) large amounts of monolingual data and (ii) a seed lexicon of frequent word-pairs, the goal is to create BWEs that enable accurate mining of rare words. As gold standard data, we release manually annotated pairs of rare words and their translations from three domains: (i) web crawls (ii) news commentaries (iii) medical texts. We show that state-of-the-art BWEs perform poorly on these data sets. We present simple techniques to build and combine BWEs that yield strong performance improvements. We study using fasttext to build BWEs, using ensembles of BWEs, and dealing with orthographic distance in BWEs, all of which improve results for the new task of rare word translation mining. A secondary contribution is improvements over state-of-the-art approaches on frequent words (which have been already extensively studied in previous work). We make our datasets and code publicly available[1].

2 Bilingual Induction of Rare Words

We briefly present how BLI is performed using BWEs and then introduce our new datasets.

Bilingual Lexicon Induction. The goal is to generate translations t in target language V_t of provided words s from source language V_s. Given a BWE representing V_s and V_t, an n-best list of translations for each word $s \in V_s$ can be induced by taking the top n words $t_i \in V_t$ whose representation $\vec{x_{t_i}}$ in the BWE is closest to the representation $\vec{x_s}$ according to cosine distance.

Datasets. To create BWEs we use *post-hoc mapping* which requires only monolingual texts and a small seed lexicon (see §3). Our training set consists of two large monolingual corpora:

- GENERAL: 4,400,309 English and German sentences from parliament proceedings, news

[1]https://github.com/braunefe/BWEeval

Proceedings of NAACL-HLT 2018, pages 188–193
New Orleans, Louisiana, June 1 - 6, 2018.

commentaries and web crawls taken from the WMT 2016 shared task (Bojar et al., 2016).

- MEDICAL: 3,108,183 English and German sentences from titles of medical Wikipedia articles, medical term-pairs, patents, documents from the European Medicines Agency.[2]

Seed Lexicons. Throughout the paper, we work with two lexicons. For each lexicon, we take the most common words and translate these by taking the top-ranked translation from a probabilistic dictionary.[3] BWEs trained using this data are evaluated on our gold standards containing pairs of rare words (we will also report results on frequent words, as in previous work, see below).

- GENLEX: 4955 most frequent words from GENERAL
- MEDLEX: 6079 most frequent words from MEDICAL

Gold Standards for Rare Words. We created gold standard data for rare words by randomly sampling words occurring between 3 and 5 times[4] in GENERAL and MEDICAL. For GENERAL we sample rare words from news commentaries and web crawls separately, so we have two rare word data sets here. For each (English) sampled word, a German native speaker generated a German translation. We indicate the division into validation and test sets:

- CRAWLRARE: 1000 rare words from web crawls of GENERAL (250 validation, 750 test)
- NEWSRARE: 1144 rare words from news commentaries of GENERAL (369 validation, 775 test)
- MEDRARE: 2109 rare words from MEDICAL (1000 validation, 1109 test)

As English-German BLI of frequent words has not been studied before, following previous work, we annotated 2000 frequent English words taken from each of the General and Medical corpora with their German translations using the same probabilistic dictionary as was used to generate the Lexicon sets. These two silver standard datasets will also be released with the paper:

- GENFREQ: 2000 frequent words from GENERAL (1000 validation, 1000 test)
- MEDFREQ: 2000 frequent words from MEDICAL (1000 validation, 1000 test)

3 Bilingual Word Embedding Creation

To create bilingual word embeddings, we use *post-hoc mapping* (PHM), a method that projects monolingual words embeddings (MWEs) into a shared space using a linear transformation trained with a small seed lexicon (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Xing et al., 2015; Lazaridou et al., 2015; Vulić and Korhonen, 2016). Among methods to generate BWEs, PHM uses a very cheap bilingual signal.[5]

Given MWEs in two languages V_s and V_t, the goal of post-hoc mapping is to find a matrix $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ that maps each representation $\vec{x_s} \in \mathbb{R}^{d_1}$ of a source word $s \in V_s$ to the representation $\vec{y_t} \in \mathbb{R}^{d_2}$ of its translation $t \in V_t$. Typically, $\mathbf{W}$ is learned using a seed lexicon $L = \{(\vec{x_1}, \vec{y_1}), \ldots, (\vec{x_n}, \vec{y_n})\}$, where each pair $(\vec{x_i}, \vec{y_i}) \in V_s \times V_t$ are mutual translations. A common objective for cross-lingual mapping is ridge regression (Mikolov et al., 2013b) (RIDGE), where $\mathbf{W}$ is estimated by:

$$\mathbf{W}^* = \underset{\mathbf{W} \in \mathbb{R}^{\mathbf{d_1} \times \mathbf{d_2}}}{\arg\min} \; || \mathbf{XW} - \mathbf{Y} || + \lambda || \mathbf{W} || \quad (1)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are stacked vectors of $\vec{x_i}$ and $\vec{y_i}$ respectively. Lazaridou et al. (2015) use a max-margin ranking loss (MAX-MARG) to estimate $\mathbf{W}$. For each $(\vec{x_i}, \vec{y_i}) \in L$, a candidate $\vec{y_i^*} = \mathbf{W} \cdot \vec{x_i}$ is computed. The ranking loss is:

$$\sum_{j \neq i}^{k} \max\{0, \gamma + sim(\vec{y_i^*}, \vec{y_i}) - sim(\vec{y_i^*}, \vec{y_j})\} \quad (2)$$

where $\vec{y_j}$ is a randomly selected negative example, i.e., it is not a translation of $\vec{x_i}$, k is the number of negative examples and $sim(\vec{x}, \vec{y})$ computes cosine similarity between $\vec{x}$ and $\vec{y}$. Hyperparameters γ and k are tuned on held-out validation data.[6]

[2] This is taken from the in-domain part of: https://ufal.mff.cuni.cz/ufal_medical_corpus.

[3] This word-level dictionary is taken from a standard phrase-based SMT system trained on WMT 2017 data.

[4] Words with frequencies 1 and 2 are very often tokenization errors or borrowings from other languages, therefore we start at frequency 3. We did not consider tokenization errors as rare words and removed those from our data.

[5] Gouws and Søgaard (2015) and Duong et al. (2016) also leverage seed lexicons. However, in order to generate high quality BWEs, these approaches leverage much larger bilingual dictionaries.

[6] Ideally, the sum in Equation 2 should be computed over the complete target vocabulary (i.e., $k = | V_t |$). Since this is not feasible in practice, Lazaridou et al. (2015) treat k as another hyperparameter tuned together with γ.

Domain	mapping	W2V skip	W2V cbow
genFreq	ridge	27.1 (43.7)	24.0 (41.4)
	max-marg	**32.1 (47.7)**	22.8 (40.0)
medFreq	ridge	14.9 (24.0)	**18.1 (30.1)**
	max-marg	16.0 (27.2)	16.8 (27.2)

Table 1: Bilingual lexicon induction of **frequent** word-pairs on **general and medical domain** data. We report top-1 and (top-5 in brackets) percentage accuracy. In this paper, bolding indicates a best result so far for a particular dataset.

3.1 Testing Previous Work

We reimplement Mikolov et al. (2013b) as well as Lazaridou et al. (2015). To replicate their setup on English-German texts we first evaluate these on two standard tasks, mining frequent words from GENERAL and MEDICAL. We follow the approach of (Heyman et al., 2017) and use English as the source language. First, we train 300 dimensional MWEs on the monolingual data using W2V[7] with default parameters except that we lowered the minimum word frequency threshold to 3 (Mikolov et al., 2013a). To generate BWEs, we use MEDICAL and MEDLEX for MEDRARE and MEDFREQ, while we use GENERAL and GENLEX for the rest of the test sets. We report results with the combination of skip-gram (W2V SKIP) or cbow (W2V CBOW) and RIDGE or MAX-MARG. As in previous work, we use top-1 (translation is the closest neighbor) and top-5 (translation is one of 5 closest neighbors) accuracies. The results in Table 1 show that the best performing setups are W2V SKIP with MAX-MARG for GENFREQ and W2V CBOW with RIDGE for MEDFREQ. Accuracies are comparable to previous work (which was on different language pairs). The poor performance on MEDFREQ is consistent with Heyman et al. (2017), who introduced the task of mining frequent medical terms.

4 Applying BWEs for Mining Rare Word-Pairs

We use the exact same BWEs training setup as above (§3) and perform BLI on our new test sets of rare words. The results in Table 2 show that on **low frequency** word-pairs BWEs perform very poorly. Compared to standard evaluation scenarios (see Table 1) a massive performance decrease is observed. Low accuracy is clearly caused by the inability of context-based models (W2V) to build accurate embedding vectors for words occurring

[7] https://github.com/dav/word2vec

Domain	mapping	W2V skip	W2V cbow
crawlRare	ridge	**2.3** (3.2)	2.0 (2.4)
	max-marg	2.1 **(3.3)**	1.7 (2.3)
newsRare	ridge	4.6 (9.4)	1.9 (4.9)
	max-marg	**5.5 (11.0)**	2.3 (4.8)
medRare	ridge	0.1 (0.2)	0.1 (0.1)
	max-marg	**0.1 (0.4)**	0.1 (0.1)

Table 2: Bilingual lexicon induction of **low frequency** word-pairs in different domains.

Domain	mapping	FTT skip	FTT cbow
crawlRare	ridge	10.1 (14.7)	4.7 (6.7)
	max-marg	**11.5 (15.9)**	7.3 (12.1)
newsRare	ridge	23.2 (37.7)	6.8 (13.5)
	max-marg	**25.3 (39.5)**	15.1 (24.1)
medRare	ridge	12.2 (19.0)	8.2 (14.2)
	max-marg	**12.5 (20.0)**	8.8 (15.3)
genFreq	ridge	33.8 (51.4)	16.2 (32.1)
	max-marg	**38.7 (56.5)**	28.3 (45.3)
medFreq	ridge	17.8 (33.6)	14.9 (26.7)
	max-marg	**29.3 (42.7)**	19.9 (33.2)

Table 3: Bilingual lexicon induction using MWEs trained with FASTTEXT (FTT).

in very few contexts only. Through post-hoc mapping, these (poor) embeddings get projected randomly into the bilingual space which results in very poor performance on BLI especially for the medical domain.

4.1 Using Subword Models

A first way to create BWEs that are better adapted to rare words is to generate MWEs that provide better vector representations for the words. One simple idea is to try to add subword information. We show empirically this helps BLI of rare words, which has not been shown before, to our knowledge. FASTTEXT (Bojanowski et al., 2017) extends W2V by adding subword information $s(w, c)$ to the context-based objective as follows:

$$s(w, c) = \sum_{g \in G_w} z_g^\top v_c \quad (3)$$

where $G_w \subset \{1, ..., G\}$ is a set of character n-gram indices corresponding to the n-grams that appear in the word w, z_g is the vector representation of the n-gram and v_c is the vector of the context words. Subword information helps for rare words (by using n-gram information shared between words) and generates more accurate MWEs especially for morphologically rich languages like German. We create 300 dimensional MWEs using FASTTEXT skip-gram and cbow models with default parameters and with the same exception as before, i.e., we lowered the minimum word

Domain	mapping	ensemble	ensemble + edit	edit only	% orth. close
crawlRare	ridge	10.3 (14.6)	19.5 (22.1)	19.0 (21.7)	60.3
	max-marg	13.2 (17.6)	**19.5 (22.3)**		
newsRare	ridge	24.3 (39.9)	32.0 (42.8)	21.8 (29.5)	39.8
	max-marg	27.2 (40.0)	**32.8 (43.5)**		
medRare	ridge	15.1 (20.6)	25.5 (26.8)	26.0 (28.2)	76.7
	max-marg	12.5 (21.2)	**26.3 (28.2)**		
genFreq	ridge	42.9 (60.7)	44.8 (62.0)	16.0 (27.1)	26.5
	max-marg	45.4 (63.2)	**47.2 (63.6)**		
medFreq	ridge	31.6 (37.0)	35.5 (44.2)	20.7 (32.9)	55.7
	max-marg	37.7 (46.7)	**38.6 (47.4)**		

Table 4: Bilingual lexicon induction of **low-frequency** and **frequent** word-pairs in different domains. *Ensemble* denotes the results for ensembling all BWE models, *ensemble + edit* shows results by adding orthographic similarity, *edit only* denotes the results obtained by using only orthographic distance (all other weights set to 0) and *% orth. close* shows the percentage of orthographically similar gold standard word pairs (whose normalized Levenshtein distance is at most 0.3).

No.	source	model1	model2	corpus	glossary
		W2V skip	FTT skip		
1.	snowstorms	skinhead	**schneestürme**	crawlRare	skinhead
2.	fire-extinguishers	goldzertifikate	**feuerlöschern**	crawlRare	gold certificates
3.	tissue-specificity	basismilieu	**gewebespezifizität**	medRare	base environment
4.	university	**universität**	harvard-universität	crawlRare	Harvard University
5.	cabin	**kabine**	flugzeugkabine	newsRare	airplane cabin
		FTT skip	ensemble		
6.	rubbish	mülltonnen	**müll**	newsRare	trashcan
7.	bathtub	badezimmer	**badewanne**	newsRare	bathroom
8.	parenthood	vaterschaft	**elternschaft**	newsRare	fatherhood
9.	cognitively	neurokognitiven	**kognitiv**	medRare	neuro cognitive
10.	nanojoules	**nanojoule**	mikrotröpfchen	medRare	microdroplet
		ensemble	ensemble + edit		
11.	sleddogs	pferdeschlittenfahrten	**schlittenhunde**	crawlRare	sled rides
12.	gnome-applets	gtkhtml	**gnome-anwendungen**	crawlRare	layout engine used by Gnome
13.	glutenins	getreideproteinen	**glutenine**	medRare	grain proteins
14.	esterify	**verestern**	esteröl	medRare	ester oil

Table 5: Examples comparing the predictions of the indicated models using ridge for the mapping where *model1* and *model2* shows the induced words for the given source. Bolding indicates the correct prediction and we give glosses for the incorrect predictions.

frequency value to 3. We perform PHM using RIDGE and MAX-MARGIN. The results in Table 3 show that this procedure yields impressive performance improvements. After evaluation[8] we manually looked at the prediction of our models. We present examples in table 5. Examples 1–3 shows that the model improves non-trivial cases as well where the meanings of the incorrect predictions induced by W2V are not close to that of the input. We also show counterexamples 4 and 5 where subword elements cause errors by inducing hyponymies of the correct words. Generating BWEs with MAX-MARGIN on these improved MWEs is particularly effective. By analyzing word similarities we saw that in BWEs acquired with RIDGE rare English words are often mapped near to noise. Because MAX-MARGIN uses negative noisy word pairs as training examples this phenomenon is not as strongly present there.

[8]We added these examples to the camera-ready paper after the results were finalized.

4.2 Model Ensembling

Although BWEs obtained with FASTTEXT and MAX-MARGIN clearly outperform other methods on rare words, a combination of BWEs obtained with different models can further improve performance by integrating several sources of information. We ensemble BWEs obtained using different MWEs as follows: we generate n-best lists ($n = 100$) of translation candidates using each model. For each pair (s, t) of candidate translations, we compute an ensemble weight given by a weighted sum of similarity scores $Sim_i(s, t)$ ob-

tained on each BWE:

$$\sum_{i=1}^{M} \gamma_i Sim_i(s,t) \quad (4)$$

$Sim_i(s,t)$ is computed using cosine similarity. When a candidate pair (s,t) is not in the n-best list generated by a model i then $Sim_i(s,t)$ is set to 0. The weights γ_i for each test set are tuned on validation sets separately (presented in §2) using grid search. The results (Table 4) show that ensembling yields significant gains over subword models alone for all data sets. We again looked for examples after evaluation (Table 5) where ensembling helped compared with the previous best setup (examples 6–9) and saw that the method again improves upon hard cases where the incorrect predictions are very close, in terms of meaning, to the gold annotation. Row 10 shows a counter-example. We note that this idea could be used in a supervised neural network for BLI as well, where information from multiple models could be integrated by concatenating embeddings from them for a given word.

4.3 Adding Orthographic Distance

While subword information captures orthographic properties of words to a certain extent, it cannot precisely represent the orthographic distance of each word pairs in a predefined number of dimensions, especially not that of source and target word pairs when performining post-hoc mapping (MWEs are trained separately thus there is no such cross-lingual information). Thus, it is beneficial to strengthen BWEs by integrating a similarity measure between word strings directly. The BWEs ensemble in Equation 4 can easily be augmented with a weighted term $\gamma_{M+1} OSim(s,t)$ that measures the orthographic similarity (which we define as one minus the normalized Levenshtein distance) between the surface-forms of words s and t. We generate n-best lists of candidate translations using different BWE models as in §4.2. In addition, we generate a list containing the n closest target words according to $OSim(s,t)$ and ensemble all lists together. Results are shown in Table 4. To measure the impact of orthographic information alone, we also report results obtained when using this information only (all other ensemble weights set to 0). For **low frequency** word-pairs, orthographic information leads to massive performance gains. We analyzed the gold standard word pairs in our datasets from the perspective of orthographic similarity. For CRAWLRARE and MEDRARE the ratios of similar words are high which explains the large improvements obtained by adding this measure. Even though the ratio is not high for NEWSRARE and the two frequent datasets, orthographic information still improves performance which shows the advantage of using the technique in all cases. Table 5 shows non-trivial examples (11–13) where orthographic distance improves performance. Example 11 shows the advantage of combining the vector representation with orthographic distance, i.e., our model could find translations of *sleddogs* that have similar meaning, while in examples 12 and 13 orthographic distance helped to pick the correct translation which is the closest in terms of edit distance. On the other hand, in example 14 orthographic distance caused an error because the incorrect prediction is too close to the source word in orthographic distance.

5 Conclusion

We evaluated BWEs on the novel task of rare term mining in different domains. Our experiments show that previous approaches to bilingual lexicon induction fail when mining rare words. We have studied techniques for decreasing the impact of these problems. By ensembling different BWEs and combining those with orthographic cues, we have reached state-of-the-art results. By making our code and datasets publicly available, we hope to encourage other researchers to further enhance BWEs to perform well on this important task. In the future, we would like to work on BLI of multi-word translations and compound words.

Acknowledgments

We would like to thank Helmut Schmid and the anonymous reviewers for their valuable input. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement № 644402 (HimL). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL* .

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proc. EMNLP*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. EACL*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proc. NAACL*.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proc. EACL*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proc. ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR* abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pages 3111–3119.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proc. ACL*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proc. NAACL*.