Unsupervised Parallel Sentence Extraction from Comparable Corpora

Viktor Hangya[1], *Fabienne Braune*[1,2], *Yuliya Kalasouskaya*[1], *Alexander Fraser*[1]

[1]Center for Information and Language Processing
LMU Munich, Germany
[2]Volkswagen Data Lab Munich, Germany
{hangyav, fraser}@cis.lmu.de
fabienne.braune@volkswagen.de

Abstract

Mining parallel sentences from comparable corpora is of great interest for many downstream tasks. In the BUCC 2017 shared task, systems performed well by training on gold standard parallel sentences. However, we often want to mine parallel sentences *without bilingual supervision*. We present a simple approach relying on bilingual word embeddings trained in an unsupervised fashion. We incorporate orthographic similarity in order to handle words with similar surface forms. In addition, we propose a dynamic threshold method to decide if a candidate sentence-pair is parallel which eliminates the need to fine tune a static value for different datasets. Since we do not employ any language specific engineering our approach is highly generic. We show that our approach is effective, on three language-pairs, without the use of any bilingual signal which is important because parallel sentence mining is most useful in low resource scenarios.

1. Introduction

The ability to extract parallel sentences from monolingual corpora is of great interest to the field and many approaches have been proposed [1, 2, 3, 4]. In this paper we explore ways to mine parallel sentences from monolingual data without bilingual supervision.

Our approach is based on bilingual word embeddings (BWEs) which represent words from different languages in the same vector space. While many authors leverage BWEs for parallel sentence extraction, previous work requires a strong bilingual signal to either (i) train the BWEs [5] (ii) train a classifier for sentence-pair extraction [6, 7, 8] or (iii) for feature engineering [9]. The disadvantage of these approaches is that the required bilingual signal is not available for many language pairs which is itself one of the reasons why parallel sentence extraction is important. In contrast to these approaches, our method does not need any bilingual signal. We create BWEs using post-hoc mapping [10] which allows us to leverage large amounts of (cheap) monolingual data to train good monolingual word embeddings (MWEs) which are then mapped into BWEs. We use the method proposed in [11] which combines adversarial training with post-hoc mapping [12] to learn BWEs without any bilingual signal. We show that high performance can be achieved using no parallel sentences nor any bilingual signal.

As a baseline system we produce sentence embeddings by averaging the word embeddings in the source language and target language sentences and compare them using cosine similarity. One difficult aspect of the task is that not all source sentences have a parallel target sentence, thus besides picking the most similar target sentence for a given source sentence it has to be decided if they are actually parallel. We propose a dynamic threshold method which calculates a minimum similarity value in an unsupervised fashion based on the input corpus.

Taking the average of the word embeddings in a sentence tends to give too much weight to irrelevant words [13]. Recently, various word-based sentence similarity metrics were introduced [14, 15]. The disadvantage of these methods is either that they are computationally expensive or that they do not handle irrelevant words. To overcome these issues, we propose a simple method which efficiently pairs source-target words while handling irrelevant words, thus making it feasible to process large datasets. In addition, we consider an important weakness of BWEs that was shown before [16], i.e., that they are poor at capturing the translations of named entities and rare words, showing that this is an important problem for parallel sentence extraction. We alleviate this by combining semantic similarities taken from BWEs with orthographic cues such as Levenshtein distance.

In summary, our contributions are: (i) We evaluate two approaches for parallel sentence extraction utilizing BWEs, based on sentence embeddings and word-by-word similarities respectively, which do not need any bilingual signal, in contrast with previous work. (ii) We introduce a dynamic threshold method for deciding whether a candidate sentence pair is parallel. (iii) We incorporate orthographic similarity to improve performance of parallel sentence extraction. (iv) We show the generality of our method on the German-English, French-English and Russian-English comparable corpora of the BUCC 2017 shared task [17].

2. Building Bilingual Word Embeddings

In this section we present two different scenarios to build BWEs. In particular, we use only monolingual datasets to train MWEs and we map them to the same bilingual space comparing two methods: the first only needs a small seed lexicon while the second does not rely on any bilingual signal.

2.1. Monolingual Word Embeddings

We train MWEs for all 4 languages in our test set. For this we used monolingual news crawls downloaded between 2011 and 2014 taken from the WMT 2014 shared task [18] containing around 80M, 117M, 31M and 45M sentences for English, German, French and Russian respectively. We used FastText skipgram [19] to train MWEs which computes a distributed representation of words using context and word structure information in the form of character n-grams. Settings used are: Embedding dimension 300; Minimum occurrence frequency 5; Window size 5; Character n-gram sizes between 3 and 6.

2.2. Bilingual Word Embeddings

Our approach to the task of parallel sentence extraction requires BWEs, which is a common vector space for words in two different languages. In previous research BWEs were created either from word-aligned, sentence-aligned or document-aligned parallel data [20, 21] or by using the cross-lingual reference to optimize two monolingual spaces, so called joint training [22, 23, 24]. Similarly to [9] we create BWEs using post-hoc mapping. First, we explain the basic idea of post-hoc mapping in the supervised setup and discuss the way how supervision is eliminated in the unsupervised method which our approach is based on.

Given two MWEs $\mathbb{R}^{d_S}$ and $\mathbb{R}^{d_T}$ post-hoc mapping is performed via a matrix $\mathbf{W} \in \mathbb{R}^{d_S \times d_T}$ which is learned using a bilingual seed lexicon. Each pair of words (s_i, t_i) in the lexicon, with $s_i \in V_s$ and $t_i \in V_t$, is projected into $\vec{x_i} \in \mathbb{R}^{d_S}$ and $\vec{y_i} \in \mathbb{R}^{d_T}$. W can be solved by learning a linear mapping [10]:

$$\mathbf{W}^* = \underset{\mathbf{W} \in \mathbb{R}^{\mathbf{d_S} \times \mathbf{d_T}}}{\arg\min} \; || \mathbf{XW} - \mathbf{Y} ||_{\mathbf{F}} \qquad (1)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are obtained by concatenating all projections $\vec{x_i}$ and $\vec{y_i}$ of words in the seed lexicon. The authors of [12] showed that the mapping can be improved by enforcing an orthogonality constraint on $\mathbf{W}$ which can be achieved by solving the singular value decomposition of $\mathbf{YX}^T$. To achieve good performance a seed lexicon of around 5000 word-pairs is used.

$\mathbf{W}$ can also be solved without any explicitly bilingual signal. The system of [11] uses adversarial training i.e. a generator and discriminator framework to achieve this. The aim of the discriminator is to distinguish mapped source language embeddings $\mathbf{W}\mathcal{X}$ and target language embeddings $\mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are sets of embeddings of words coming from the source and target language. In contrast, the goal of the generator is to learn $\mathbf{W}$ such that it prevents the discriminator from making accurate predictions. After training, $\mathbf{W}$ is used to automatically extract a seed lexicon of best candidate word pairs which is used to perform post-hoc mapping with [12].

We use [11] in our fully unsupervised setup. As a contrastive experiment we report results with [12] using a seed lexicon of 5000 word pairs, which was used as a baseline in [11] as well.

3. Sentence Extraction

We evaluate our model on the shared task data provided by the BUCC Workshop at ACL 2017. We evaluate our system on De-En, Fr-En and Ru-En language pairs. The dataset consists of comparable monolingual corpora (Wikipedia dumps) where the BUCC organizers inserted truly parallel sentences (taken from News Commentary) into the monolingual data for each language pair [17]. Our task is to recover the truly parallel sentences, while minimizing false alarms.

3.1. Sentence Embeddings

We use a basic sentence embedding approach as a baseline. BWEs are used to embed sentences in both languages into the same space. Each sentence embedding is computed by dimension-wise averaging of the embeddings of words in the given sentence (contained in the BWEs) followed by $l2$ normalization. Once source and target sentences are embedded, their similarity can be efficiently computed via cosine similarity [25]. To overcome the issue of giving too much weight to semantically poor words, which decreases precision and mistakenly selects non-parallel sentences, we remove stopwords [26], digits and punctuation from texts before calculating sentence embeddings. Consider this erroneous example, which shows how weighting stop words like *a, in, by* too highly causes an erroneous match:

> **De:** *Inzwischen sterben mehr Frauen* ***an*** *Gebärmutterhalskrebs – alle zwei Minuten* ***eine*** *– als* ***bei einer*** *Entbindung.*
>
> **Gloss:** *Meanwhile die more women* ***from (literal: in)*** *ovarian cancer – every two minutes* ***one (literal: a)*** *– than* ***at (literal: by) a*** *birth.*
>
> **En:** *For women* ***in*** *the developing world,* ***by*** *contrast, dying* ***in*** *childbirth is simply* ***a*** *fact of life.*

3.2. Dynamic thresholding

To decide whether a candidate sentence pair, i.e., source sentence and its most similar target sentence, is parallel we introduce a method which calculates a minimum similarity value that the candidate has to meet. We calculate this threshold value for each test set with a simple formula:

$$th = \bar{S} + \lambda * std(S) \quad (2)$$

where S is a set containing the similarity values between each source sentence in the test set and its most similar target candidate, $\bar{S}$ and $std(S)$ are its mean and standard deviation. We set $\lambda = 2.0$ based on the De-En development set which worked optimally for the other setups as well. The advantage of this method is that it performed well on all our datasets, while fine tuning a static threshold value on the development sets did not achieve good results (see §4) due to the difference of development and test data. Note also that λ could be quickly and easily adjusted by the user in order to balance between quality and quantity for downstream tasks (in practice inspection of only a few samples is sufficient).

3.3. Bilingual dictionaries

Averaging word embeddings in a sentence tends to give too much weight to irrelevant words. It was shown that hub words, which are similar to a high proportion of other words, have negative effects on performance of embedding based methods [13]. *Word Mover's Distance* was introduced [14], which is based on the minimum distance that the words in one text need to "travel" to reach the words in the other text, to overcome such issues. On the other hand, the approach is computationally intensive which is not desirable in the case of parallel sentence extraction due to the high number of candidate sentence pairs. Furthermore, it was show that WMD performs similarly to maximum alignment based methods on monolingual sentence similarity tasks while the latter is computationally less intensive [15]. We propose an efficient hub word aware maximum alignment approach based on bilingual dictionaries and show that it is more effective than simple sentence embeddings. In this method, we perform bilingual lexicon induction on the trained BWEs to generate large n-best dictionaries, which we then use to mine parallel sentences.

3.3.1. Bilingual lexicon induction

Given a BWE representing two languages V_s and V_t, an n-best list of translations for each word $s \in V_s$ can be induced by taking the n words $t_i \in V_t$ whose representation $\vec{x_t}$ in the BWE is closest to the representation $\vec{x_s}$ according to cosine similarity.

From the source side of the comparable data we compute a list containing the 200,000 most frequent words. For each word in the list, we retrieve the 100-best translations using bilingual lexicon induction on the BWEs. Each translation is given a weight by using cosine similarity computed with the BWE.

3.3.2. Sentence extraction

Given a candidate pair of source and target sentences S and T, the similarity score is calculated by iterating over the words in S from left to right and pair each word s, in a greedy fashion, with the word $t \in T$ that has the highest similarity based on our dictionary. During iteration, we ignore all t which have been already paired to overcome the hubness problem, i.e. by preventing the pairing of multiple source words to the same target word. Then, the averaged word-pair similarity gives the final score. We apply the same stopword filtering as before and use dynamic thresholding for the final decision. Although, we kept our method simple for computational reasons we use pre-filtering as in previous work [6]. For each source sentence we only consider the 100 most similar target sentences as candidates based on sentence embedding similarities. Given a BWE model our method requires around 2.5 hours to process sentences from the De-En test set (164 billion sentence pairs) on a single thread.

3.3.3. Orthographic similarity

As it was shown in previous work the performance of bilingual lexicon induction can be significantly improved by using orthographic cues, especially for rare words. We extend this idea to the sentence level by using a dictionary containing orthographically similar source-target language word pairs and their similarity[1]. We define orthographic similarity as one minus normalized Levenshtein distance. We use this orthographic dictionary with BWE based dictionary when mining parallel sentences by using the bigger value from the two dictionaries. If the given word pair is not in a dictionary we consider their similarity as 0.0.

4. Results

As we mentioned earlier we evaluate our system on the De-En, Fr-En and Ru-En data of the BUCC 2017 shared task [17]. We show results based on BWEs created fully unsupervised with the method of [11] (**unsup**) and the lightly-supervised system of [12] (**lisup**) on the released training sets as in [8]. Our systems only rely on news crawl monolingual data and a small seed lexicon in case of the latter thus we did not use the training set in earlier steps. We will show the performance of our final system and results of previous supervised systems on the official test set at the end[2]. As baseline we use the sentence embedding system with stopword filtering and dynamic thresholding. We report precision, recall and F1-scores.

From table 1 it can be seen that the dictionary based approaches significantly outperform the baseline system for each language pair. Our systems perform best on Fr-En and lowest on Ru-En which strongly correlates with the performance of the mapping approaches, that was shown in [11], on bilingual lexicon induction. Even though the baseline performs the weakest it is competitive on French-English with similar systems [6]. We also ran our baseline system on De-En with *lisup* and static threshold value instead of dynamic.

[1] For speedup we only consider word pairs that have at least 0.8 similarity

[2] We evaluated on the test set by sending our predictions to the shared task organizers.

		De-En			Fr-En			Ru-En		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
lisup	sentence-embedding	27.86	18.01	21.88	29.22	14.38	19.28	6.92	4.42	5.39
	BWE dict.	23.05	42.29	29.83	38.16	52.19	44.08	16.80	24.77	20.02
	BWE+ORT dict.	24.19	45.11	**31.49**	39.00	52.64	44.80	16.32	24.05	19.45
unsup	sentence-embedding	26.53	16.40	20.27	28.99	14.07	18.94	6.54	3.84	4.84
	BWE dict.	22.67	41.90	29.42	37.97	52.30	44.00	17.31	24.97	**20.44**
	BWE+ORT dict.	23.71	44.57	30.96	39.02	52.61	**44.81**	16.75	24.20	19.80

Table 1: Results of our proposed systems on the BUCC 2017 shared task's training set for the 3 language-pairs. Baseline is the sentence embedding based model with stopword filtering and dynamic tresholding. We underline the best F1 scores for a language-pair and BWE method and use **bolding** for the best overall F1 score for a given language-pair.

By fine tuning the value on the development set (achieving high score) we got only 2.69% F1-score on the training set showing the importance of dynamic thresholding. Furthermore, in our preliminary experiments we used a shuffled parallel dataset of parliament proceedings and news articles for BWEs as in previous work [6]. It showed that having strongly comparable data could give 5% performance gain for our setups in average. On the other hand, having access to such data is unrealistic in real life scenarios, so we don't use this data further in our work.

Comparing the dictionary based approaches with and without the orthographic dictionary it can be seen that the orthographic information helped the most for De-En and also increased performance for Fr-En. In the following example the incorrect En sentence is about the same topic but orthography was needed to extract the sentence with correct entities:

De: ***Microsoft*** *hat* ***Nokia Milliarden*** *von Dollar versprochen, wenn es seine* ***Smartphones*** *exklusiv mit* ***Windows Phone*** *ausstattet.*

En-: *In Q1 2008* ***Samsung*** *shipped 46.3* ***million mobile handsets*** *1Q 2008.*

En+: ***Microsoft*** *promised to pay* ***billions*** *of dollars for* ***Nokia*** *to use* ***Windows Phone*** *exclusively.*

On the other hand, it did not help for Ru-En because of their different character sets. We manually analyzed the results and saw that the use of orthographic information gave higher similarity scores to sentence-pairs that contained named entities with the same orthography. These pairs were correctly mined without orthography thus no performance increase was caused. On the other hand, higher similarity scores caused higher dynamic threshold value thus losing some correctly mined pairs. This phenomenon can be fixed by better fine tuning λ for this setup.

Our *lisup* and *unsup* systems are on par with each other. Regarding F1 scores, the seed lexicon caused higher performance only for De-En while the unsupervised method performed better for the rest of the language pairs. In figure 1 we show precision-recall curves comparing the two systems on the three language pairs. This also shows that their performance is similar. There is a bigger gap between the systems

		P (%)	R (%)	F1 (%)
De-En	[27]	88	80	84
	lisup BWE+ORT dict.	24	45	32
	unsup BWE+ORT dict.	24	45	31
Fr-En	[27]	80	79	79
	lisup BWE+ORT dict.	39	53	45
	unsup BWE+ORT dict.	39	53	45
Ru-En	lisup BWE+ORT dict.	16	24	19
	unsup BWE+ORT dict.	17	24	20

Table 2: Results on the test set. We show the best performing supervised system of the shared task [27].

in the case of Ru-En in higher precision ranges in favor of the unsupervised system. Overall, these results show that good performance can be achieved in a fully unsupervised manner, i.e., using only monolingual data for training BWEs and using only these for mining parallel sentence-pairs.

We show negative examples which our unsupervised system with orthographic information made on the De-En set in table 3. Examples 1-3 are incorrectly mined sentence pairs. In the first case the meaning of the mined pair is very similar although it is not parallel while high named entity content causes the error in the next two. Although names in example 2 are not orthographically similar they are close in embedding space which causes the error. Similarly, cardinal directions are different in example 3 but in general they appear in similar contexts thus get represented similarly in the word embedding space. In contrast, examples 4-6 have not been mined by our system. The first two pairs have extra information on the source side, although they are parallel, which caused error for our system. Example 6 contains the compound noun *Entwicklungsländern* (developing countries) which is not handled by our system.

Finally, we show results on the official shared task test sets in table 2[3]. For comparison we also include the results of the best performing supervised system on De-En and Fr-En [27]. There were no submissions for Ru-En. It can be seen that the performance of our systems on the test set are very close to the performance on the training set which we presented in table 1. This shows that our dynamic threshold approach, with λ tuned on the De-En development set, is gen-

[3]Results are rounded for consistency with the shared task paper.

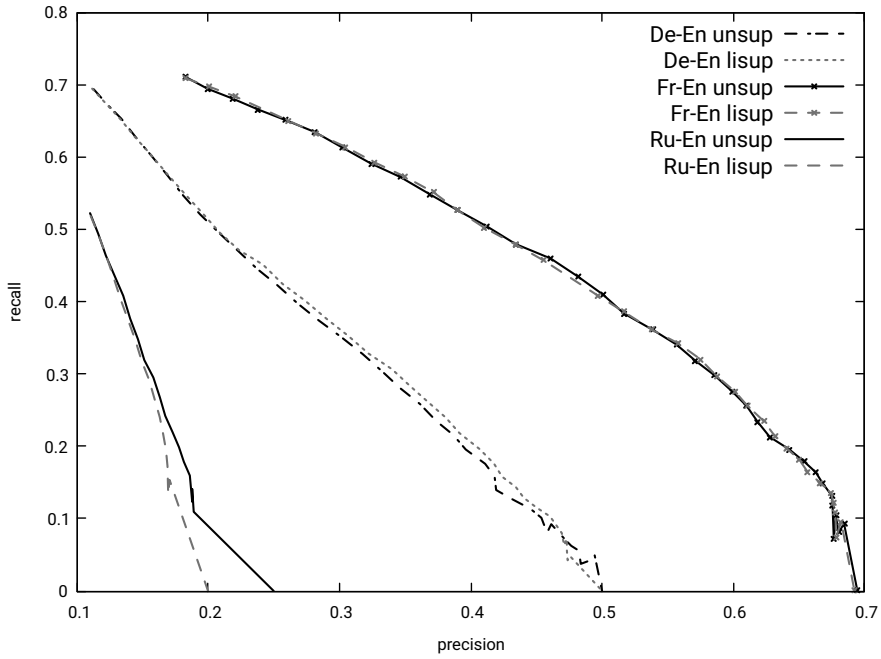

Figure 1: Precision-recall curves comparing unsup and lisup systems on the three language pairs.

1.	Das Werk wurde außerdem mit vier Academy Awards (Oscars) prämiert, darunter der Trophäe für den besten fremdsprachigen Film. In 2011, it was awarded the Academy Award for Best Documentary Feature at the 83rd Academy Awards. *The work has also received four Academy Awards (Oscars), including the Best Foreign Language Film Trophy.*
2.	Am 7. Juli 1957 wurde Angelas Bruder Marcus, am 19. August 1964 ihre Schwester Irene geboren. In April 1976 a daughter, Josina, was born, and in December 1978 a son, Malengane. *On July 7, 1957 Angela's brother Marcus was born, on August 19, 1964 her sister Irene.*
3.	Im Osten Serbien, im Südosten Montenegro, sowie im Norden, Westen und Südwesten Kroatien. It was in the modern Vojvodina (in northern Serbia), northern Croatia and western Hungary. *In east Serbia, southeast Montenegro, as well as in the north, west and southwest Croatia.*
4.	Aber die meisten Frauen, die Hillary Clinton wählen sollen, sind nicht Unternehmensjuristinnen oder Staatssekretärinnen. But most of the women sought as voters are not corporate attorneys or secretaries of state. *But most women who are to vote for Hillary Clinton are not corporate lawyers or state secretaries.*
5.	Durch diese Kürzungen ging jedoch die Produktion weiter zurück und die wirtschaftliche Misere verschlimmerte sich nur noch mehr. As they cut, output fell further and economic misery deepened. *As a result of these cuts, however, production continued to decline and the economic misery deepened.*
6.	Für Frauen in den Entwicklungsländern dagegen ist es ganz normal, bei der Entbindung sterben zu können. For women in the developing world, by contrast, dying in childbirth is simply a fact of life.

Table 3: Samples from the manual analysis. 1-3 are incorrectly mined examples (translation of De sentences where differing from En pair shown in italic) while 4-6 are the missed parallel sentences.

eral enough to work well on multiple languages and datasets. Interestingly, the supervised system performed better on De-En comparing with Fr-En while our approach reached higher F1 scores on the latter. One reason for this could be the better mapping quality of the word embedding space for Fr-En which was show in [11]. Our results with the fully unsupervised system are lower comparing to the supervised method since the latter has access to a large parallel corpus during training. Using parallel data supervised systems can learn features which help to decide if a sentence pair is parallel, e.g. word order of a source side phrase in the target side. In contrast, in the unsupervised case, we can only rely on word similarity information which can cause errors when syntax is the deciding factor in the case of a sentence-pair with similar words. On the other hand, our approach performed well on this task and will serve as a strong baseline for future unsupervised methods. With our dynamic thresholding method it is also easy to calculate a good initial threshold value which can be changed manually by the user in order to balance between quantity and quality of the mined sentence pairs.

5. Conclusion

In this work we introduced our first steps for the task of unsupervised parallel sentence extraction. We showed the performance of a simple sentence embedding system based on unsupervised BWEs and proposed a novel technique for dynamically setting the decision threshold. We improved upon this baseline system by proposing a simple word pair similarity based method which is efficient for large corpora. Furthermore, we addressed the shortcomings of BWEs when applying them for parallel sentence mining by using orthographic similarity. We showed that our system works well for various language pairs where BWEs could be built by achieving good results on De-En, Fr-En and Ru-En. In addition, we showed that unsupervised BWEs perform as well as BWEs based on a small seed lexicon. The goal of this short work is to provide a strong baseline for the unsupervised parallel sentence extraction task, and we are hoping to encourage more research on this important problem.

6. Acknowledgments

We thank Helmut Schmid for helpful discussions and comments. 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).

7. References

[1] D. S. Munteanu, A. Fraser, and D. Marcu, "Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora," in *Proc. NAACL-HLT*, 2004.

[2] J. R. Smith, C. Quirk, and K. Toutanova, "Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment," in *Proc. NAACL-HLT*, 2010.

[3] C. Chu, R. Dabre, and S. Kurohashi, "Parallel Sentence Extraction from Comparable Corpora with Neural Network Features," in *Proc. LREC*, 2016.

[4] K. Krstovski and D. A. Smith, "Bootstrapping Translation Detection and Sentence Extraction from Comparable Corpora," in *Proc. NAACL-HLT*, 2016.

[5] J. Grover and P. Mitra, "Bilingual Word Embeddings with Bucketed CNN for Parallel Sentence Extraction," in *Proc. ACL, Student Research Workshop*, 2017.

[6] F. Grégoire and P. Langlais, "BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora," in *Proc. 10th Workshop on Building and Using Comparable Corpora*, 2017.

[7] H. Bouamor and H. Sajjad, "H2@ BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings," in *Proc. Workshop on Building and Using Comparable Corpora*, 2018.

[8] H. Schwenk, "Filtering and Mining Parallel Data in a Joint Multilingual Space," in *Proc. ACL*, 2018.

[9] B. Marie and A. Fujita, "Efficient Extraction of Pseudo-Parallel Sentences from Raw Monolingual Data Using Word Embeddings," in *Proc. ACL*, 2017.

[10] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting Similarities among Languages for Machine Translation," *CoRR*, vol. abs/1309.4168, 2013.

[11] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word Translation Without Parallel Data," *CoRR*, vol. abs/1710.04087, 2017.

[12] C. Xing, D. Wang, C. Liu, and Y. Lin, "Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation," in *Proc. NAACL-HLT*, 2015.

[13] G. Dinu, A. Lazaridou, and M. Baroni, "Improving Zero-Shot Learning by Mitigating the Hubness Problem," in *Proc. workshop track at international conference on learning representation*, 2015.

[14] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From Word Embeddings to Document Distances," in *Proc. ICML*, 2015.

[15] T. Kajiwara and M. Komachi, "Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings," in *Proc. COLING*, 2016.

[16] F. Braune, V. Hangya, T. Eder, and A. Fraser, "Evaluating bilingual word embeddings on the long tail," in *Proc. NACL-HLT*, 2018.

[17] P. Zweigenbaum, S. Sharoff, and R. Rapp, "Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora," in *Proc. BUCC*, 2017.

[18] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, "Findings of the 2014 Workshop on Statistical Machine Translation," in *Proc. 9th Workshop on Statistical Machine Translation*, 2014.

[19] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *CoRR*, vol. abs/1607.04606, 2016.

[20] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning, "Bilingual Word Embeddings for Phrase-Based Machine Translation." in *Proc. EMNLP*, 2013.

[21] K. M. Hermann and P. Blunsom, "Multilingual Distributed Representations without Word Alignment," in *Proc. ICLR*, 2014.

[22] S. Gouws, Y. Bengio, and G. Corrado, "BilBOWA: Fast Bilingual Distributed Representations without Word Alignments," in *Proc. ICML*, 2015.

[23] A. Klementiev, I. Titov, and B. Bhattarai, "Inducing Crosslingual Distributed Representations of Words," in *Proc. COLING*, 2012.

[24] H. Soyer, P. Stenetorp, and A. Aizawa, "Leveraging Monolingual Data for Crosslingual Compositional Word Representations," *CoRR*, vol. abs/1412.6334, 2014.

[25] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *CoRR*, vol. abs/1702.08734, 2017.

[26] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," in *Proc. ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002.

[27] A. Azpeitia, T. Etchegoyhen, and E. Martínez, "Weighted Set-Theoretic Alignment of Comparable Sentences," in *Proc. 10th Workshop on Building and Using Comparable Corpora*, 2017.