Parallel sentences mining with transfer learning in an unsupervised setting

Yu Sun
Zhengzhou University of Light Industry

Shaolin Zhu *
Zhengzhou University of Light Industry
zhushaolin003@163.com

Chenggang Mi
Northwestern Polytechnical University
michenggang@nwpu.edu.cn

Yifan Feng
Zhengzhou University of Light Industry

Abstract

The quality and quantity of parallel sentences are known as very important training data for constructing neural machine translation (NMT) systems. However, these resources are not available for many low-resource language pairs. Many existing methods need strong supervision and hence are not suitable. Although there have been several attempts at developing unsupervised models, they ignore the language-invariant between languages. In this paper, we propose an approach based on transfer learning to mine parallel sentences in an unsupervised setting. With the help of bilingual corpora of rich-resource language pairs, we can mine parallel sentences without bilingual supervision of low-resource language pairs. Experiments show that our approach improves the performance of mined parallel sentences compared with previous methods. In particular, we achieve good results at two real-world low-resource language pairs.

1 Introduction

Parallel sentences are known as very important training data for constructing machine translation (MT) systems (Belinkov and Bisk, 2018). The volumes of quality parallel sentences heavily affect the performance of trained machine translation systems. However, these resources are only available for a handful of language pairs and domains while the others suffer from the scarcity problem (Bouamor and Sajjad, 2018). In this situation, parallel sentences are very crucial for training machine translation systems.

Transfer learning is an effective approach to mine parallel data in low-resource scenarios. (Artetxe and Schwenk, 2019) brought the evidence of cross-lingual transfer to mine parallel data for low-resource language pairs. However, their method is not unsupervised and relies on bilingual supervision (e.g, bilingual lexicon or sentences), which is not available for low-resource language pairs. Although (Kvapilíková et al., 2020) solved the supervised limitation by employing an unsupervised MT, the performance heavily depended on MT's quality.

In this paper, we propose a parallel sentences mining model based on transfer learning in an unsupervised setting[1]. As illustrated in Figure 1, we obtain sentence embeddings by mean-pooling the outputs of multilingual BERT (Lample and Conneau, 2019), which is trained on monolingual corpora. In particular, we use a language discriminator to learn shared and refined language-invariant representations for transfer learning. (Chen et al., 2018; Ziser and Reichart, 2018) pointed out the language-invariant is helpful for transfer learning. Then, we treat detecting parallel sentences as a classification task and generate multi-view semantic representations for the classifier. Generally, data from different views contain complementary information and multi-view learning exploits the consistency from multiple views (Li et al., 2018; Fei and Li, 2020). In our model, we use two views for the classifier: (i) word representations; (ii) sentence representations. In addition to achieving good results on BUCC 2018[2] shared task, we demonstrate the effectiveness of our model using an example of two low-resource language pairs where parallel corpora are almost not available.

In summary, our contributions in this paper are as follows:

(1) We propose an unsupervised method based on transfer learning to mine parallel sentences without any bilingual data for low-resource language

*Corresponding author: Shaolin Zhu, zhushaolin003@163.com

[1] The unsupervised setting means we only have monolingual corpora for a pair of language that bilingual resources are not available, while there are some language pairs have bilingual resources which we use for unsupervised transfer learning in low-resource language pairs.

[2] 11th Workshop on Building and Using Comparable Corpora

Proceedings of NAACL-HLT 2021: Student Research Workshop, pages 136–142
June 6–11, 2021.

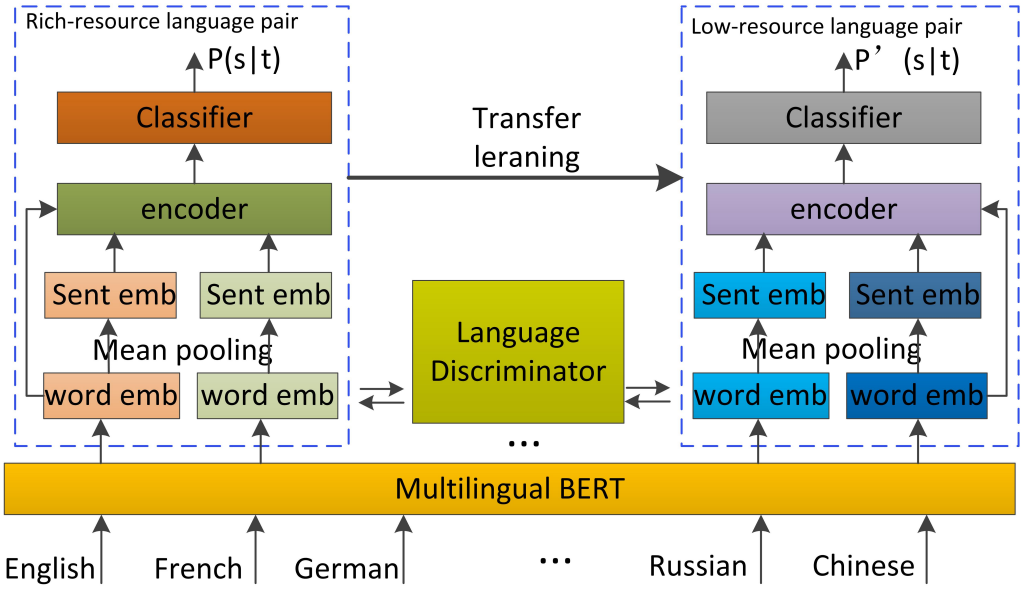

Figure 1: Our proposed method that based on multi-view transfer training for parallel phrase detection on a non-parallel sentence pair.

pairs. By designing a multi-view model, we encode the representations on word-level and sentence-level to obtain high-quality parallel data.

(2) We extensively consider the language-invariant by constructing a language discriminator to well capture the semantic similarity among languages. This makes the robustness of our model for transfer learning.

2 Related Work

Many works mine parallel corpora from monolingual data which contain potential mutual translations. Previous methods depended on engineering features. (Shi et al., 2006; Esplà-Gomis et al., 2016) used metadata information from web crawls to mine parallel data. Recent methods used cross-lingual word embeddings to obtain parallel corpora (Guo et al., 2018; Schwenk, 2018; Bouamor and Sajjad, 2018; Schwenk et al., 2019b,a). (Artetxe and Schwenk, 2019) encoded the universal language embeddings that are agnostic to languages. They used transfer learning to mine parallel sentences of low-resource language pairs. This transfer learning method inspired our work and the main difference is that they required bilingual supervision (e.g, bilingual lexicon, parallel sentences), which is not available for many low-resource language pairs.

Recently, several works developed unsupervised method to mine parallel data (Hangya et al., 2018; Hangya and Fraser, 2019; Kvapilíková et al., 2020; Keung et al., 2020). These approaches mainly rely on unsupervised cross-lingual embeddings (Artetxe et al., 2018; Lample and Conneau, 2019) that be trained on monolingual corpora. However, several researchers question that these methods may not well capture the semantic similarity among languages (Karthikeyan et al., 2019; Pires et al., 2019). Some researchers proposed to use transfer learning to solve cross-lingual applications for low-resource language pairs (Lakew et al., 2018; Kocmi, 2020). (Eriguchi et al., 2018) used a multilingual neural machine translation system to learn the word representations of rich-resource language pairs. Then, they used transfer learning to identify parallel sentences for low-resource language pairs. However, it has an implicit dependency on multilingual NMT that requires pre-training on large parallel sentences. Our transfer learning is inspired by (Fei and Li, 2020). The difference is that they mainly solve cross-lingual unsupervised sentiment classification.

3 Proposed Method

The overview of the model architecture is as shown in Figure 1. Our proposed approach based on transfer learning to mine parallel data is composed of three components: an unsupervised multilingual BERT, a language discriminator, and a multi-view classifier. Motivated by the success of unsupervised cross-lingual word embeddings (Artetxe et al., 2018; Lample and Conneau, 2019) and its

application in mining parallel data (Hangya and Fraser, 2019; Keung et al., 2020), we use multilingual BERT to initialize word and sentence embeddings. Although previous methods are effective, they may ignore sentential context on using multilingual word embeddings, which could harm the performance of mining parallel corpora. In our work, we use multi-view representations to mine parallel data. We can get good performance on rich-resource language pairs. However, our aim is to obtain parallel data for low-resource language pairs. For this purpose, we use transfer learning to mine parallel data of the low-resource scenarios using rich-resource language pairs. Note that our method doesn't rely on any bilingual data of low-resource language pairs. Therefore, we can call that our method is unsupervised for low-resource language pairs.

3.1 Language Discriminator

Previous works (Chen et al., 2018; Fei and Li, 2020) indicate that cross-lingual transfer learning work well when their representations are language-invariant. We use the unsupervised multilingual BERT to map the word representations into a shared space. Although we can generate shared word representations for different languages by using the unsupervised multilingual BERT, there is still a semantic gap between languages. Following (Chen et al., 2018; Lample et al., 2018), we employ a language discriminator for getting fine-tuned word representations, which is necessary to preserve language-invariant on language transfer. In detail, the language discriminator is trained to distinguish between the mapped source and target embeddings. Then, we refine-turn the two language embeddings with a cross-lingual Procrustes method according to (Lample et al., 2018). The language discriminator contains a feed-forward neural network with two hidden layers as an encoder and one softmax layer. The objective of the discriminator is to maximize its ability to identify the source and target embeddings. The discriminator loss can be written as follows:

$$L(\theta_D|W) = -\log P_{\theta_D}(source = 1|Wx) + \log P_{\theta_D}(target = 1|y) \quad (1)$$

Where Θ_D denotes parameters of the discriminator, (x, y) corresponds to source and target language. $P_{\theta_D}(source = 1|z)$ is a probability that a vector z is the mapping W of a source embedding, $P_{\theta_D}(target = 1|z)$ is similar. In parallel, we use the Procrustes analysis to fine-tune the mapping W as follows (Lample et al., 2018). We can obtain universal language-agnostic embeddings when the discriminator is not able to identify the origin of an embedding.

3.2 Transfer Learning for Mining Parallel Data

In this paper, we propose to use transfer learning to mine parallel data of the low-resource scenarios by rich-resource language pairs. In this paper, we first consider two views of input for classifier in rich-resource language pairs:(i) the word-level representations from languages; (ii) the sentence-level representations from languages. The multi-view classifier has been demonstrated useful as data from different views contains complementary information (Chen and Qian, 2019; Fei and Li, 2020). In this paper, we use a feed-forward neural network based on LSTM with two hidden layers as an encoder to balance two view representations. Then, we train a classifier to match predicted labels with ground truth from the parallel sentences in rich-resource language pairs as follows:

$$P(s|t) = \frac{e^{enc(\theta)}}{1 + e^{enc(\theta)}}\epsilon(0, 1) \quad (2)$$

Where $enc(\theta)$ denotes parameters of the encoder. Then, we use transfer learning to mine parallel data for low-resource language pairs. The detail process is as follows: We firstly train a classifier on rich-resource language pairs (such as English-Chinese or English-French). In parallel, we use the language discriminator to fine-tune the different language representations into a shared space to keep language-invariant between languages. After that, we transfer the pre-trained classifier to detect parallel sentences for low-resource pairs. Finally, we use detected parallel data to train the classifier again in low-resource language pairs for better performance.

4 Experimental Setting

In this section, we mainly present our experimental settings and describe the datasets used.

Dataset: We test our proposed method on four language pairs of BUCC sample data (English-French, English-German, English-Russian, English-Chinese). The shared task of

	En-Fr			En-De			En-Ru			En-Zh		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
supervised methods												
Bouamor and Sajjad, 2018	87.5	65.8	75.1	-	-	-	-	-	-	-	-	-
Schwenk, 2018	84.8	68.6	75.8	84.1	70.7	76.9	81.1	67.6	73.8	77.7	66.4	71.6
Artetxe and Schwenk, 2019	91.5	93.3	92.3	95.6	95.1	95.4	90.6	94.0	92.2	91.9	91.3	91.6
unsupervised methods												
Hangya and Fraser, 2019	50.5	38.1	43.4	48.5	39.1	43.3	37.4	18.7	24.9	-	-	-
Keung et al., 2020	-	-	73.0	-	-	74.9	-	-	69.6	-	-	60.1
Hangya et al., 2018	39.0	52.6	44.8	23.7	44.5	30.9	17.3	24.9	20.4	-	-	-
Kvapilíková et al., 2020	-	-	78.7	-	-	80.1	-	-	77.1	-	-	67.0
Proposed method	81.6	79.5	80.6	88.5	85.5	86.9	80.4	78.1	80.6	78.4	76.3	77.3

Table 1: Results of our proposed systems on the BUCC shared task's training set for the 4 language-pairs. We also report the results of baselines as described in their paper. "-" represents the result are not reported in ther paper, respectively.

the workshop on Building and Using Comparable Corpora (BUCC) is a well-established evaluation framework for mining parallel corpora (Zweigenbaum et al., 2018). The shared task provides a gold standard to assess retrieval systems for precision, recall, and F_1-score. We applied our approach to all language pairs of the BUCC18 shared task. Moreover, we carry out an experiment on real-world low-resource scenarios (English-Esperanto, Chinese-Kazakh). For the monolingual data, we extract corpora from Wikipedia using WikiExtractor[3]. As there is no gold standard to evaluate mining parallel sentences, we use mined parallel sentences to train a machine translation system that can reflect the quality of mined parallel sentences.

Baselines: In our experiments, we consider supervised baselines (Bouamor and Sajjad, 2018; Schwenk, 2018; Artetxe and Schwenk, 2019). We also compare several unsupervised baselines which contains (Hangya and Fraser, 2019; Keung et al., 2020; Hangya et al., 2018; Kvapilíková et al., 2020).

5 Results and Discussions

In this section, we present the results of mining parallel sentences and our comparison to previous work. We also present results on real-world low-resource language pairs and demonstrate our obtained parallel corpora can improve the performance of machine translation.

[3]https://github.com/attardi/wikiextractor

5.1 Results on BUCC

As BUCC provides a gold standard to assess mined parallel data, we test our method on the BUCC dataset. Although the language pairs used for evaluation are all high-resources, we only simulate the low-resource scenario to justify our method here and we will present results on real-world low-resource language pairs in the section 5.3. We show precision (P), recall(R) and F_1 scores in Table 1 for the four language pairs. Noted that, we use English-German as the rich-resource language pair to initialize our model. Then, we transfer this model into other low-resource language pairs. We also test different rich-resource language pairs for transfer learning as Table 2.

Noted that, our method doesn't rely on any bilingual data of low-resource language pairs. Therefore, we can call that our method is unsupervised for low-resource language pairs. This is a fair comparison to other unsupervised methods. From Table 1, we achieve an increase of F_1 compared with unsupervised baselines for all language pairs. It also can be seen that the precision and recall of the proposed method is significantly increased for all language pair than unsupervised methods. (Artetxe and Schwenk, 2019) also used transfer learning to mine parallel sentences. However, their method needs strong supervision which is not available in low-resource language pairs. The proposed method overcomes the limitation and obtains relatively good results against (Artetxe and Schwenk, 2019).

	En-Fr			En-De			En-Ru			En-Zh		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
-language discriminator												
(En-Fr)	87.5	85.8	86.6	78.1	76.8	77.4	63.6	63.4	62.5	63.1	61.8	62
(En-Ru)	66.3	63.1	64.7	64.2	60.7	62.4	86.8	83.6	85.2	63.7	63.4	63.5
(En-Zh)	61.2	63.3	62.2	60.6	62.1	61.3	60.8	64.2	62.5	83.7	81.6	82.6
(En-De)	75.5	74.5	75.0	88.5	85.5	86.9	74.1	74.2	74.2	71.7	70.7	71.3
+language discriminator												
(En-Fr)	87.5	85.8	86.6	82.3	81.2	81.7	79.6	76.6	78.1	77.2	74.6	75.9
(En-Ru)	80.6	82.3	81.4	81.1	80.7	80.9	86.8	83.6	85.2	76.8	75.3	76.1
(En-Zh)	78.2	76.1	77.1	80.7	78.6	79.6	77.6	78.8	78.2	83.7	81.6	82.6
(En-De)	81.6	79.5	80.6	88.5	85.5	86.9	80.4	78.1	80.6	78.4	76.3	77.3

Table 2: Ablation study on the BUCC shared task. Note that, the first column indicates that we use different rich-resource language pairs for transfer learning.

5.2 Ablation Study

To understand the effect of different components in our model on the overall performance, we conduct an ablation study in Table 2 to test the language discriminator whether affects transfer learning or not. "-language discriminator" is not adding the language discriminator and "+language discriminator" is adding the language discriminator. In Table 2, the first column is that we use different rich-source language pairs to implement transfer learning for mining parallel sentences. We firstly can find that different sources have similar results for transfer learning of our model. Then, we can find that when we don't add the language discriminator, the performances of the model are not good for transfer learning. When we add the language discriminator for transfer learning, we can find that our model gets an obvious and stable improvement in all language pairs. So from Table 2, we can conclude that language-invariant is very important for transfer learning.

5.3 Results on Low-resource Language Pair

In the above section, we simulate the low-resource scenario to justify our method on the BUCC dataset. In this section, we evaluate our mined parallel sentences on real-world low-resource language pairs. We apply our method to the English-Esperanto(En-Es) and Chinese-Kazakh(Zh-Kz) language pairs. As there is no gold standard to evaluate mining parallel sentences, we use mined parallel sentences to train a machine translation system that can reflect the quality of mined parallel sentences.

Methods	**En-Es**	**Zh-Kz**
(Hangya and Fraser, 2019)	18.5	21.6
(Keung et al., 2020)	20.2	22.8
(Hangya et al., 2018)	16.3	19.3
(Kvapilíková et al., 2020)	23.6	22.7
Proposed method	24.3	25.8

Table 3: BLEU scores on different language pairs.

We use openNMT[4] to train the machine translation system. The results are as in Table 3. Based on the scores in Table 3 it can be seen that we achieve a significant performance increase compared to the unsupervised baseline. It is well-known that the quality and quantity heavily affect the performance of machine translation. The results of Table 3 demonstrate that the proposed method is effective, especially for low-resource language pairs.

6 Conclusion

In this paper, we propose an unsupervised method that uses multi-view transfer learning to mine parallel sentences. Our method can effectively use the bilingual data of rich-resource language pairs. We transfer the model of rich-resource language pairs into a low-resource situation without any supervision of low-resource language pairs. In particular, we employ a language discriminator to capture language-invariant for benefiting transfer learning. In the experiments, the results show that our method significantly and consistently outperforms the baselines.

[4]https://opennmt.net/

For the future, we would like to apply our model on other low-resource language pairs to test universal applicability in different language pairs.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (61906158), the Project of Science and Technology Research in Henan Province (212102210075).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. 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.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *International Conference on Learning Representations*.

Houda Bouamor and Hassan Sajjad. 2018. H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proc. Workshop on Building and Using Comparable Corpora*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 547–556.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Miquel Esplà-Gomis, Mikel L Forcada, Sergio Ortiz Rojas, and Jorge Ferrández-Tordera. 2016. Bitextor's participation in WMT'16: shared task on document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 685–691.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, et al. 2018. Effective Parallel Corpus Mining using Bilingual Sentence Embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised Parallel Sentence Extraction from Comparable Corpora. In *Proc. IWSLT*.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A Smith. 2020. Unsupervised Bitext Mining and Translation via Self-trained Contextual Embeddings. *arXiv preprint arXiv:2010.07761*.

Tom Kocmi. 2020. Exploring Benefits of Transfer Learning in Neural Machine Translation. *arXiv preprint arXiv:2001.01622*.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262.

SM Lakew, A Erofeeva, M Negri, M Federico, and M Turchi. 2018. Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary. In *15th International Workshop on Spoken Language Translation*, pages 54–62.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Yingming Li, Ming Yang, and Zhongfei Zhang. 2018. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Holger Schwenk. 2018. Filtering and Mining Parallel Data in a Joint Multilingual Space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv*, pages arXiv–1907.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A dom tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 489–496.

Yftah Ziser and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 238–249.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.