Hindawi
Scientific Programming
Volume 2020, Article ID 8879570, 7 pages
https://doi.org/10.1155/2020/8879570

Research Article

Selection of In-Domain Bilingual Sentence Pairs Based on Topic Information

Bin Li[1] **and Jianmin Yao**[2]

[1]*School of Information Engineering, Anhui Open University, Hefei 230041, China*
[2]*Provincial Key Laboratory of Computer Information Processing Technology, Soochow University, Suzhou 215006, China*

Correspondence should be addressed to Bin Li; libin@ahou.edu.cn and Jianmin Yao; jyao@suda.edu.cn

Received 30 September 2020; Accepted 25 November 2020; Published 15 December 2020

Academic Editor: Michele Risi

The performance of a machine translation system (MTS) depends on the quality and size of the training data. How to extend the training dataset for the MTS in specific domains with effective methods to enhance the performance of machine translation needs to be explored. A method for selecting in-domain bilingual sentence pairs based on the topic information is proposed. With the aid of the topic relevance of the bilingual sentence pairs to the target domain, subsets of sentence pairs related to the texts to be translated are selected from a large-scale bilingual corpus to train the translation system in specific domains to improve the translation quality for in-domain texts. Through the test, the bilingual sentence pairs are selected by using the proposed method, and further the MTS is trained. In this way, the translation performance is greatly enhanced.

1. Introduction

At present, the performance of a machine translation system (MTS) is determined by the quality and size of the training data. The larger the size and the higher the quality of training data are, the superior the translation performance is. However, the distribution of existing bilingual resources is quite imbalanced in different domains, and the bilingual resources in some aspects are extremely scarce [1]. When the training corpora and the test texts are subordinated to different domains, a translation system generally presents poor performance. The main reason is that many technical terms are included in those corpora; nevertheless, it fails to obtain effective translation knowledge on the technical terms from the training data. Research on the method for selecting bilingual sentence pairs aims to select subsets of sentence pairs related to in-domain texts from a large-scale bilingual parallel corpus from many different corpora. It is expected to extend the training dataset for an MTS in specific domains to enhance the performance of the machine translation. In addition, the bilingual parallel sentence pairs acquired by using existing methods for mining bilingual resources generally do not contain corresponding labels indicating domains. Thus, determining how to automatically mine bilingual sentence pairs relevant to a specific domain from the bilingual resources becomes an effective approach to improve the performance of machine translation.

2. Related Work

Existing methods for selecting in-domain bilingual sentence pairs can be approximately divided into three types: separately based on information retrieval, a language model, and the combination of the translation model and language model. In terms of the first method, Lu et al. [2] and Huang et al. [3] proposed a selection method for bilingual sentence pairs based on information retrieval. In the method, the sentence pairs related to the corpora in the test set are retrieved from a universal corpus by querying sentences in the test set. The method fails to realize the rapid and real-time translation owing to the test sentences are used as the input of information retrieval to be queried.

As for the second method, Yasuda et al. [4] put forward a method for evaluating and selecting bilingual sentence pairs

based on the language model perplexity in the target domain. In the method, the language models are trained with monolingual corpora within the target domain, and the relevance of the candidate bilingual sentence pairs is evaluated according to the cross entropy of a language model. Duh et al. [5] assessed the quality of the bilingual sentence pairs with the cross-entropy difference of neural network language models. Axelrod et al. [6] screened and explored the in-domain data by means of part-of-speech tagging and N-gram. Additionally, aiming at the selection of bilingual sentence pairs in the universal translation system, Yao et al. [7] proposed a method for selecting bilingual sentence pairs based on the quality and coverage of sentence pairs. The method is able to evaluate the quality of sentence pairs according to the scores of multiple features marked based on an artificially designed weighted fusion. The sentence pairs with a high score are selected, and the redundant sentence pairs are further filtered out based on the coverage of N-gram. Wang et al. [8] proposed a method for selecting parallel corpora based on classification: a classifier with a great difference is constructed based on the features of a small number of sentence pairs to distinguish the quality of bilingual sentence pairs. Shah and Specia [9] evaluated multiple translation methods for sentence pairs through tests. Although the above research methods have achieved a favourable effect, they still show some limitations. At first, these methods only take the domain relevance of sentence pairs into account while ignoring the mutual translation of sentence pairs in the target domain. Thus, the bilingual sentence pairs selected according to these methods possibly present poor mutual translation quality, thus bringing noise to the subsequent training of the translation models. Afterwards, the weights of features for evaluating the qualities of different sentence pairs are adjusted by virtue of manual experience in existing research methods, failing to obtain the optimal performance [10].

As for the third method, the domain relevance and mutual translation of a sentence pair are both taken into account [11]. However, the method for the selection of sentence pairs based on a language model or translation model evaluates the relevance of sentence pairs to the target domain with the aid of the coexistent statistical information of words or N-gram [12]. Limited by the size of monolingual or bilingual resources in a target domain, the method is likely to result in data sparseness; moreover, the topic diversity of in-domain texts is ignored when training a translation model or language model with all dataset [13, 14]. In addition, related work is also conducted from the perspective of the granularity of phrase pairs; phrase pairs of specific types are automatically screened from the phrase table, which can be formed by combining phrases in the phrase table for the target domain. Furthermore, the phrase pairs with high domain relevance are screened and added to the phrase table [15, 16]. Farhath et al. [17] evaluated the impact of different types of data sources in developing a domain-specific SMT for the domain of official government letters.

3. Research Methods

The whole construction scheme is shown in Figure 1. The method for selecting in-domain bilingual sentence pairs based on topic information is used to train the translation systems in specific domains to improve the translation quality. By virtue of the topic relevance between the bilingual sentence pairs and the target domain, the method is capable of selecting subsets of sentence pairs related to the text to be translated from a large-scale bilingual corpus. At first, based on the contextual words of phrase pairs in bilingual corpora, the topic vector of phrase pairs is learnt; afterwards, the topic vectors of the development set in the target domain and candidate bilingual sentence pairs are acquired by using the set of the extracted phrase pairs. Finally, the topic relevance between the candidate bilingual sentence pairs and the texts in the development set in the target domain is calculated. The highly relevant sentence pairs will be preferentially selected as the training data in the target domain. With the aid of the topic relevance of texts, the bilingual sentence pairs relevant to the target domain are selected, which provides a new method for extending the training data for specific MTSs and solves the problem incurred by the lack of training data in specific fields.

3.1. The Training Module for a Topic Model Based on Phrase Pairs. The module learns the topic distribution of bilingual phrase pairs, which shows the occurrence probability of bilingual phrase pairs under different topics. The specific steps are displayed as follows:

(I) Based on the phrase extraction algorithm, phrase pairs are extracted from a word-aligned bilingual parallel corpus and the IDs of the corresponding bilingual parallel sentence pairs containing the phrase pairs are recorded.

(II) Some phrase pairs are stochastically sampled from the extracted phrase pairs, and the contexts of the sentence pairs containing the current phrase pairs are obtained according to the IDs recorded above. Except for the current phrase pairs, the words contained in the contexts of the sentence pairs are combined to form a new document as a pseudo-document for the distribution of phrase pairs. The words frequently occurring in the contexts of different phrase pairs can characterize the semantic meaning of the corresponding phrase pairs. Thus, a pseudo-document of a phrase pair is built with the aid of the contextual words of the phrase pair occurring in a bilingual parallel corpus. Furthermore, based on the topic distribution of the pseudo-document, the topic distribution of the phrase pair is obtained.

(III) Special characters, stop words, and low-frequency words in the pseudo-document for the distribution of phrase pairs are removed, and the document is

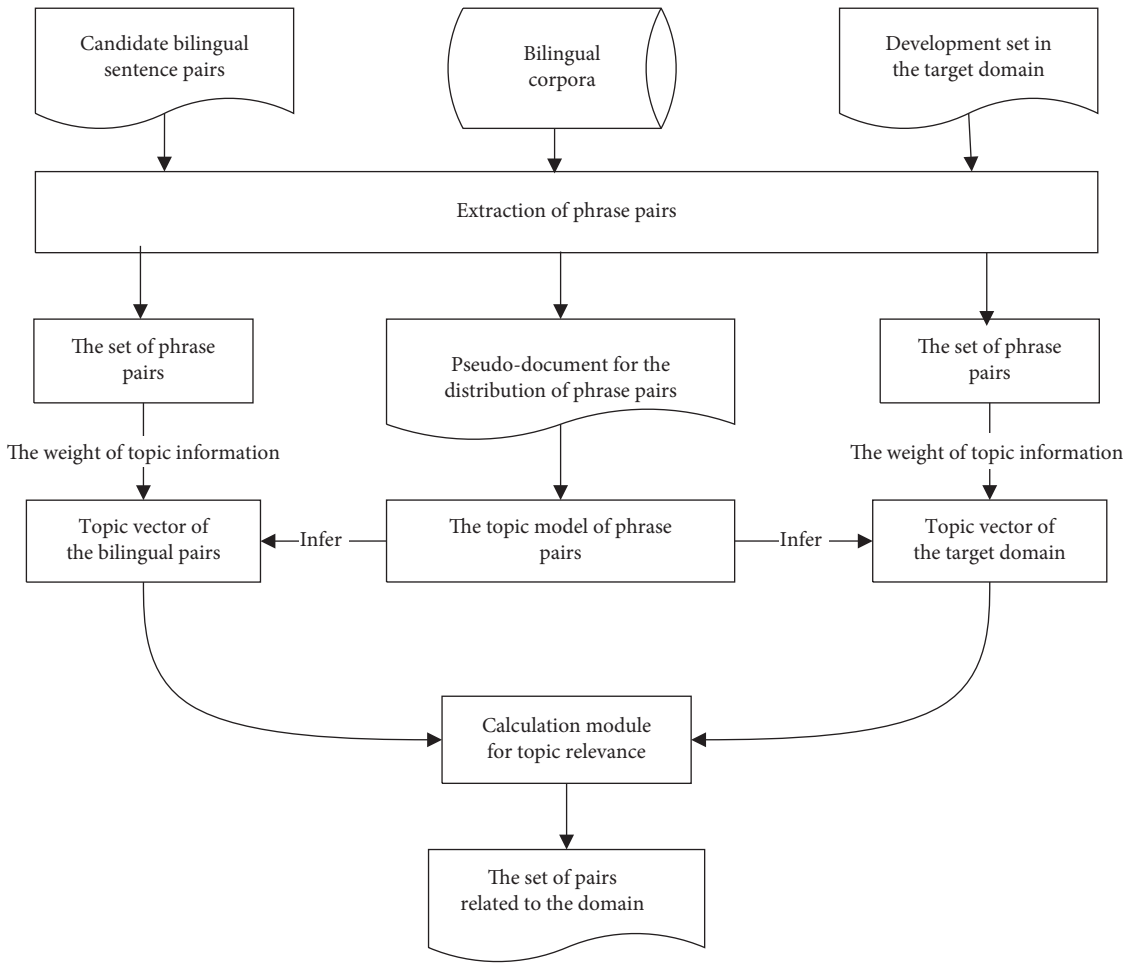

Figure 1: Framework of the method for selecting in-domain bilingual sentence pairs based on topic information.

employed to train the LDA topic model to acquire the topic distribution of the document. Moreover, the topic distribution of the document is taken as that of the corresponding phrase pair.

3.2. Inference Module for Topic Vectors of the Bilingual Sentence Pairs and the Target Domain. The module is used to obtain the topic representation of the bilingual sentence pairs and target domain. The topic vectors of bilingual sentence pairs and the development set in the target domain are calculated by using the set of phrase pairs extracted therefrom. By taking the calculation of the topic vector of the bilingual sentence pairs as an example, the specific steps are summarized as follows:

(I) Extraction of phrase pairs: The phrase pairs satisfying the consistency of word alignment are extracted from the bilingual parallel sentence pairs by virtue of the phrase extraction algorithm.

(II) Inference of the topic distribution of phrase pairs: All contexts of each phrase pair in a bilingual parallel corpus are obtained to construct a pseudo-document for the topic distribution.

(III) According to the occurrence frequency of a phrase pair in bilingual sentence pairs, the topic vector of the phrase pair is weighted to calculate the topic vector of the bilingual sentence pairs. The specific mathematical expressions are displayed as follows:

$$w_k(f,e) = \sum_{i=0}^{M}\sum_{j=0}^{N} c_s\left(\widetilde{f}_i,\widetilde{e}_j\right) t_k\left(\widetilde{f}_i,\widetilde{e}_j\right),$$
$$V_s = \langle w_1(f,e),\ldots,w_k(f,e)\rangle, \tag{1}$$

where M and N separately refer to the numbers of the source and target language phrases in bilingual sentence pairs; $c_s(\widetilde{f}_i,\widetilde{e}_j)$, $t_k(\widetilde{f}_i,\widetilde{e}_j)$, K, and V_s denote the occurrence times of the phrase pair $(\widetilde{f}_i,\widetilde{e}_j)$ in bilingual sentence pairs, the value in the kth dimension in the topic vector of the phrase pair $(\widetilde{f}_i,\widetilde{e}_j)$, the number of topics, and the topic vector of the bilingual sentence pairs, respectively.

(IV) To guarantee that the sum of the distribution probabilities of phrase pairs in various topics is

equal to 1, it is necessary to further normalize the topic vectors calculated above.

$$p_i(f,e) = \frac{w_i(f,e)}{\sum_{j=1}^{K} w_j(f,e)}, \quad (2)$$

where $p_i(f,e)$ stands for the probability distribution of bilingual sentence pairs in the ith topic. The topic vector of the development set of the target domain is calculated in the same way.

3.3. Calculation Module for Topic Relevance. The module is applied to calculate the topic relevance between bilingual sentence pairs and the target domain and finally screen the subset of sentence pairs related to the target domain from a large-scale bilingual parallel corpus. The specific steps are described as follows:

(I) The similarities between the topic distributions of all candidate bilingual sentence pairs and the development set of the target domain are calculated by taking Jensen–Shannon divergence as the evaluation index.

(II) According to the score of Jensen–Shannon divergence, all candidate bilingual sentence pairs are listed in an ascending order. A parallel sentence pair with a lower score shows a higher relevance to the target domain. As an index in statistics for calculating the similarity of two probability distributions, Jensen–Shannon divergence is mathematically defined as follows:

$$\mathrm{JSD}(\mathrm{dev}; f,e) = \frac{1}{2}\left[\sum_{i=1}^{i=K} p_i(\mathrm{dev})\log\frac{2p_i(\mathrm{dev})}{p_i(\mathrm{dev}) + p_i(f,e)} + \sum_{i=1}^{i=K} p_i(f,e)\log\frac{2p_i(f,e)}{p_i(\mathrm{dev}) + p_i(f,e)}\right], \quad (3)$$

where $p_i(f,e)$ refers to the probability distribution of bilingual sentence pairs in the ith topic and $p_i(\mathrm{dev})$ represents the probability distribution of the development set of the target domain in the ith topic.

(III) The bilingual sentence pairs sorted as TopN = {100 K, 200 K, 400 K, 600 K, 800 K, 1,000 K} are successively selected as the training data for the target domain to build a specific MTS. The optimal value of N is determined according to the translation performance of the MTS in the development set.

4. Analysis of Experimental Data and Results

4.1. Corpora and Arrangements during the Test

4.1.1. Test Corpora. Based on the English-Chinese translation task in the spoken language domain, the feasibility of the aforementioned methods for the selection of sentence pairs is separately validated through tests. The training corpora for a bilingual topic model involve two parts: (1) training corpora in the spoken language domain, which are taken from the official oral tourism parallel corpora (size: 50 K sentence pairs, 356 K English words, and 317 K Chinese words) offered by the Chinese Language Translation Task, China Workshop on Machine Translation (CWMT09) [18] and (2) training corpora in the universal domain, which correspond to bilingual sentence pairs (size: 16 M sentence pairs, 396 M English words, and 358 M Chinese words) automatically acquired from Web according to the method by Liu et al. [19].

The corpora of phrase-based MTS mainly contain that (1) the training corpora for a translation model correspond to the subset of TopN sentence pairs selected from universal bilingual corpora based on the proposed method for the selection of bilingual sentence pairs; (2) the training corpora for a language model are originated from the local Chinese monolingual corpora (size: 200 M); (3) the development set for a translation system is sourced from the development set (size: 456 English sentences, each of which corresponds to four Chinese translation results) in oral translation evaluation of the National High Technology Research and Development Program of China (HTRDP, also known as 863 program) in 2005; and (4) the test set for the translation system is taken from the public test set (size: 400 English sentences, each of which corresponds to four translation results) of oral translation evaluation of the 863 program in 2004.

4.1.2. Test Setting. The test was conducted with the aid of an open source machine translation tool NiuTrans, which produced by Northeastern University of China [20]. The system environment is set as follows: GIZA++ is used to realize the word alignment of bilingual sentence pairs; by means of a trigram language model, the parameters of the translation system are optimized using the training method for the minimum error ratio [21]; moreover, the BLEU value is used as the evaluation criterion for the performance of the MTS [22]. Different translation system schemes are designed in the test to verify the translation effect under various schemes.

CWMT: the MTS [18] trained with bilingual corpora (with the size of 50 K) in the spoken language from the CWMT09

GE: the MTS [19] trained with the aid of large-scale universal training corpora (with the size of 16 M)

Duh_2013: the MTS [5] trained with TopN sentence pairs selected from a large-scale universal bilingual corpus by employing the method for selecting sentence pairs based on the neural network language model perplexity proposed by Duh et al

Yao_2016: the MTS [1] trained with TopN sentence pairs selected from a large-scale universal bilingual corpus by utilizing the method for the selection of sentence pairs, combining a translation model with a language model proposed by Liu et al. [19]

TIM: the MTS trained with TopN sentence pairs selected from a universal bilingual corpus using the method for the selection of in-domain bilingual sentence pairs based on the topic information proposed in the study

4.2. Results and Analysis. The performances of CWMT and GE systems in the study are shown in Table 1.

It can be seen from Table 1 that the translation performance of the translation system GE trained with large-scale universal bilingual corpora is superior to that trained with the aid of bilingual corpora in the spoken language domain (CWMT). The BLEU value of the GE system based on the same test set is improved by 13.72%. The reason is that the size of the training corpus of the GE system is larger than that of the CWMT system. Large-scale corpora of the GE system cover more translation knowledge and language phenomena, while small-scale corpora of the CWMT system are likely to result in data sparseness, thus leading to poor translation performance.

However, large-scale universal bilingual corpora contain many bilingual sentence pairs having an insignificant correlation or no correlation with the target domain. The sentence pairs impose an adverse effect on the translation performance. By applying different methods for selecting sentence pairs, the sentence pairs sorted as TopN = {100 K, 200 K, 400 K, 800 K, 1000 K} are extracted from large-scale universal bilingual corpora. Moreover, the sentence pairs are used as the training corpus for a translation system to train a translation model. The specific test results are shown in Figure 2.

The test results show that, with the aid of the method for selecting in-domain bilingual sentence pairs based on topic information, the TopN sentence pairs are extracted from a large-scale bilingual corpus to train a translation model of the MTSs. In this way, the training effect of the translation model can be effectively improved while reducing the cost of model training. When extracting Top400 K sentence pairs (BLEU = 37.25%) related to the target domain from a universal bilingual corpus, the translation effect of the system has exceeded that with all corpora (BLEU = 35.62%). The reason is that the universal bilingual corpus contains training data from various domains, including many sentence pairs less related to the target domain. As a result, many noises are found in the translation rules extracted therefrom, thus influencing the final translation performance.

Table 1: Translation performance of CWMT and GE systems.

System	Source and size of corpora	Test set (BLEU %)
CWMT	Spoken language (50 K)	21.90
GE	Universal (16 M)	35.62

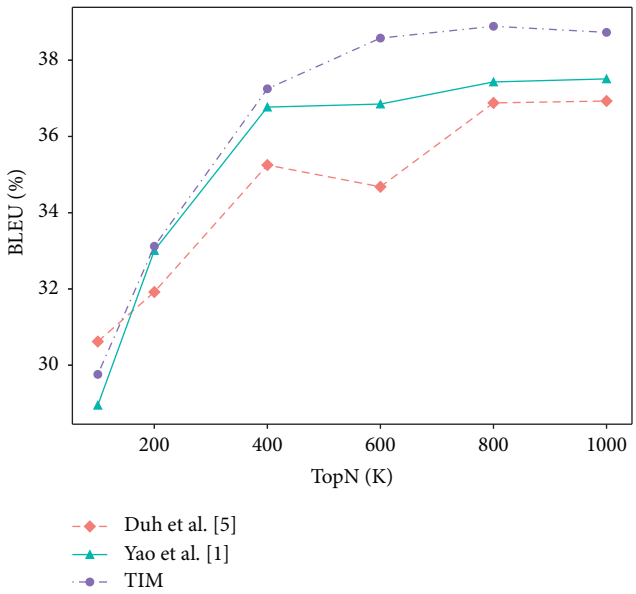

Figure 2: Comparison of translation performances of different MTSs.

As shown in Figure 2, the TIM method for selecting sentence pairs relevant to the target domain based on the topic information proposed in the study is superior to the other test methods. When the extracted sentence pairs satisfy TopN = 800 K, the translation performance of the TIM shows the optimal BLEU value of 38.89%, separately increasing by 2.01% and 1.46% compared with those of Duh et al. [5] and Yao et al. [1]. On the condition of TopN = 1000 K, the BLEU value for measuring the translation performance of the TIM is separately 1.8% and 1.22% higher than those of Duh et al. [5] and Yao et al. [1]. The other several models mainly consider some statistical information (such as text coexistence and language model) to estimate the domain relevance of sentence pairs while ignoring the information of latent semantic meaning in sentence pairs. By contrast, TIM selects bilingual sentence pairs based on the topic information, which generalizes the in-domain translation knowledge at the level of topics. Thus, the information of translation domains is more favourably matched, thus attaining a better translation effect.

5. Conclusions

From the perspective of the domain relevance of bilingual sentence pairs, the bilingual resources for specific translation tasks are selected and extended from a large-scale universal bilingual parallel corpus. On this basis, it is expected to improve the performance of the specific MTSs. The bilingual sentence pairs relevant to the target domain are selected

based on the topic relevance of texts; a domain is depicted from the perspective of topic and the bilingual sentence pairs, and in-domain texts are characterized as the probability distribution of topics; furthermore, the sentence pairs relevant to the target domain are selected with the aid of the topic relevance to train the specific translation system, thus improving the translation quality of in-domain texts. The study provides a new method for extending the training data for the MTS in specific domains and solves the problem incurred by the lack of training data in some specific domains. The test result reveals that the translation performance is greatly enhanced in the case of selecting the bilingual sentence pairs and training the translation system based on the proposed method. In the future work, it is supposed to employ more effective domain features to select the bilingual sentence pairs in specific domains and extend the size of bilingual corpora and the number of domains involved in the corpora.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Key Program of the Natural Science Foundation of Anhui Education Committee under grant no. KJ2019A0969, the Key Program of the Foundation for Young Talents in the Colleges of Anhui Province under grant no. 2013SQL097ZD, the Program of Quality Engineering in the Colleges of Anhui Province under grant nos. 2017SJJD101 and 2018XFSYXM010, and the Key Program of the Teaching Reform Research under grant nos. 2019ZDJG08, 2019ZDJG09, and 2019ZDJG10.

References

[1] L. Yao, Y. Hong, and H. Liu, "Combining translation and language models for bilingual data selection," *Journal of Chinese Information Processing*, vol. 30, no. 5, pp. 145–152, 2016.

[2] Y. Lu, J. Huang, and Q. Liu, "Improving statistical machine translation performance by training data selection and optimization," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, June 2007.

[3] J. Huang, Y. Lv, and Q. Liu, "Corpus selection and optimization for statistical machine translation system based on information retrieval method," *Journal of Chinese Information Processing*, vol. 22, no. 2, pp. 40–46, 2008.

[4] K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumita, "Method of selecting training data to build a compact and efficient translation model," in *Proceedings of the Third International Joint Conference on Natural Language Processing*, vol. 2, pp. 655–660, Hyderabad, India, January 2008.

[5] K. Duh, G. Neubig, K. Sudohet, and H. Tsukada, "Adaptation data selection using neural language models: experiments in machine translation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 678–683, Sofia, Bulgaria, August 2013.

[6] A. Axelrod, Y. Vyas, M. Martindale et al., "Class-based N-gram language difference models for data selection," in *Proceedings of the 12th International Workshop on Spoken Language Translation*, Da Nang, Vietnam, December 2015.

[7] S. Yao, T. Xiao, and J. Zhu, "Selection of SMT training data based on sentence pair quality and coverage," *Journal of Chinese Information Processing*, vol. 25, no. 2, pp. 72–78, 2011.

[8] X. Wang, Z. Tu, and J. Xie, "Selection of parallel corpus based on classification," *Journal of Chinese Information Processing*, vol. 27, no. 6, pp. 144–151, 2013.

[9] K. Shah and L. Specia, "Quality estimation for translation selection," in *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, Dubrovnik, Croatia, June 2014.

[10] H. Khayrallah, H. Xu, and P. Koehn, "The JHU parallel corpus filtering systems for WMT 2018," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels, October 2018.

[11] S. Mansour, J. Wuebker, and H. Ney, "Combining translation and language model scoring for domain-specific data filtering," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2011*, San Francisco, CA, USA, December 2011.

[12] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, December 2019.

[13] I. Skorokhodov, A. Rykachevskiy, D. Emelyanenko et al., "Semi-supervised neural machine translation with language models," in *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages*, Boston, MA, USA, March 2018.

[14] H. Xu and P. Koehn, "Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 2017.

[15] X. Zhang, P. Shapiro, G. Kumar et al., "Curriculum learning for domain adaptation in neural machine translation," 2019, http://arxiv.org/abs/1905.05816.

[16] P. Koehn, H. Khayrallah, K. Heafield, and M. L. Forcada, "Findings of the WMT 2018 shared task on parallel corpus filtering," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels, October 2018.

[17] F. Farhath, P. Theivendiram, S. Ranathunga, S. Jayasena, and G. Dias, "Improving domain-specific SMT for low-resourced languages using data from different domains," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May 2018.

[18] C. Zhang, H. Jiang, S. Liu et al., "Technical report of HIT_LTRC for CWMT 2009 evaluation," in *Proceedings of the 5th Conference on China Workshop on Machine Translation*, Nanjing, China, October 2009.

[19] L. Liu, Y. Hong, J. Lu et al., "An iterative link-based method for parallel web page mining," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, October 2014.

[20] T. Xiao, J. Zhu, H. Zhang, and Q. Li, "NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation," in *Proceedings of the ACL 2012 System Demonstrations*, Jeju Island, Korea, July 2012.

[21] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 2003.

[22] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA, July 2002.