Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 7495436, 7 pages
https://doi.org/10.1155/2019/7495436

Research Article

A Novel Deep Learning Method for Obtaining Bilingual Corpus from Multilingual Website

ShaoLin Zhu,[1,2,3] Xiao Li,[1,2] YaTing Yang ,[1,2] Lei Wang,[1,2] and ChengGang Mi[1,2]

[1] *The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi, China*
[2] *Key Laboratory of Speech Language Information Processing of Xinjiang, Urumqi, China*
[3] *University of Chinese Academy of Sciences, Beijing, China*

Correspondence should be addressed to YaTing Yang; yangyt@ms.xjb.ac.cn

Received 3 April 2018; Accepted 10 December 2018; Published 10 January 2019

Academic Editor: Emilio Insfran Pelozo

Machine translation needs a large number of parallel sentence pairs to make sure of having a good translation performance. However, the lack of parallel corpus heavily limits machine translation for low-resources language pairs. We propose a novel method that combines the continuous word embeddings with deep learning to obtain parallel sentences. Since parallel sentences are very invaluable for low-resources language pair, we introduce cross-lingual semantic representation to induce bilingual signals. Our experiments show that we can achieve promising results under lacking external resources for low-resource languages. Finally, we construct a state-of-the-art machine translation system in low-resources language pair.

1. Introduction

Parallel corpus is one of the most invaluable resources for many multilingual natural language processing applications. Especially statistical machine translation (SMT) and neural machine translation (NMT) and parallel corpora play a pivotal role in those applications. Thus, many approaches have been proposed to obtain bilingual corpus automatically. Those can be roughly divided into two categories: (i) the first method is to directly crawl web content and use multifeatures to filter parallel webpages content. Those features mainly contain tokens in URLs, anchor around link, image alt, HTML tags [1–5]. The method recognizes parallel pages by computing the similarity of those features (ii) the other select parallel sentences by building classifiers. Those classifiers mainly consist of maximum entropy, Bayesian, SVM, and neural networks [2, 6–10]. The two methods have been demonstrated promising to obtain bilingual corpus in some language pairs. However, the methods only adapt some specific website or no-low resources languages pairs. Then low-resources languages have still been underachieving to harvest bilingual corpus.

We believe that two major challenges limit obtaining bilingual parallel corpus for low-resources languages. First, dynamic websites have a more complicated structure and it is difficult to filter parallel corpus by recognizing multifeatures. For example, previous work obtains parallel main language corpus from Wikipedia and Twitter [4, 11]. However, many news websites have a more complicated structure than Wikipedia and Twitter. Second, although the classifier is a good solution to select parallel corpus from numerous noise data, the number of parallel sentences has an impact on the classifier [6, 12]. There is not enough training parallel corpus to construct a classifier in low-resource language pairs.

Recently, with the surge of continuous vector representation and extensive application of deep learning, an interesting approach is induce bilingual semantic clues from monolingual data with the new neural-network inspired vector representation. Following those methods, we could alleviate the limit of low-resource language pair to obtain parallel bilingual corpus. In this paper, we use this continuous word embedding to induce bilingual representation by establishing cross-lingual mapping. Then, using those bilingual representations find some parallel sentences. This step avoids the effect of HTML structure as the current website is developed into dynamic modules. Finally, we construct a bidirectional recurrent neural network (LSTM-BiRNN) classifier to extract parallel sentences. We use the

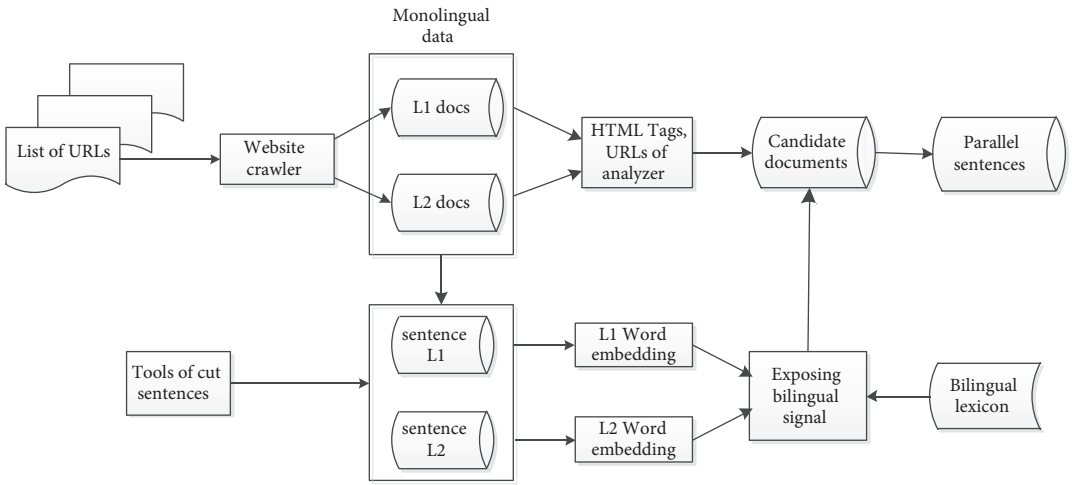

Figure 1: The architecture of obtaining parallel sentences.

parallel corpus obtaining in word-overlap model to train this classifier and perform extracting process. To justify the effectiveness of the proposed approach, we obtain Uyghur-Chinese parallel corpus from multilingual websites to train SMT systems and show improvements in BLEU (BLEU: bilingual evaluation understudy) scores. Our experiments also show that we can achieve promising results by removing the need of any specific feature engineering or external resources.

2. Related Works

The amount of information available on the Internet is expanding rapidly, and many works attempt to construct training corpus for machine translation from websites. A variety of approaches have been proposed to extract parallel sentences from web. Those approaches can be divided into two strategies.

First, many approaches treat collecting parallel sentences as a text classification problem [6, 13, 14], such as SVM classifier and neural network classifier. For example, [6] proposed a latest siamese bidirectional recurrent neural network to construct a state-of-the-art classifier and detect parallel sentences. They remove the need of any domain specific feature engineering or relay on multiples models and only raw parallel sentences. However, parallel sentences also are a very invaluable corpus for some low-resources language pairs. So this excellent method maybe is not suitable for some low-resources applications.

Second, many other works use the HTML structure of the web pages, URLs and image alt et al to detect possible parallel sentences [1, 3, 15] . For instance, [7] use the links between translated articles in Wikipedia to crawl parallel sentences or words. These methods have proven to be useful for specific website, the real challenge is to find strategies that allow to extend them to crawl the Web in an unsupervised fashion.

Esplà-Gomis et al. developed an excellent tool, namely, Bitextor, a free/open-source tool for harvesting parallel data from multilingual websites. It is highly modular and is aimed at allowing users to easily obtain segment-aligned parallel corpora from the Internet. It mainly obtained parallel sentences by comparing HTML structure of the documents and the number of aligned words in bilingual lexicon. The users only provide a bilingual lexicon, and the system can contrast parallel data quickly automatically. The real challenge is that a bilingual lexicon is not easy to obtain for some low-resources language pairs.

3. Methodology

The first step of obtaining parallel corpus is harvesting source of data. We use a web crawler to harvest monolingual data and construct the continuous word representation. Following works of [3], we use multifeatures to get the candidate data. Then, we extend the works of [16, 17] to learn bilingual signal. With the help of bilingual signal, we can induce parallel sentences. The general architecture of obtaining parallel corpus is presented in Figure 1.

3.1. Crawling Web-Data and Candidate Documents. The first step of harvesting bilingual parallel corpus is using web-crawler to download data. However, unlike the pervious works that downloaded a mirror of a webpage, we only download texts that do not contain html tags. As the current website is developed into module, the same theme pages usually have the same HTML structure.

When we perform the process of downloading, we use the Scrapy toolkit (https://pypi.python.org/pypi/Scrapy) (it is written in Python). It is an excellent toolkit that allows user set specific content to crawl. The next step is selecting candidate document pair. As we all know, a website contains hundreds of thousand documents, and if we match the whole website, the matching procedure is very low and imprecise. In order to solve this problem, we borrow the idea of [2] that adds a window of time. The main characteristic of news website is time, and every webpage has publication time. The same topic documents often are reported in a period by different

language. Thus, we use a heuristic with which we assume it is more possible that documents with similar content have publication dates that are close to each other. Therefore, each query is fact run only against source documents published within a window of some days around the publication date of the target query document. In this procedure, we set the size of window as three. Then, each query can search only fewer documents and get a higher precision. In next section, we will introduce how to identify two multilingual documents that are parallel.

3.2. Inducing Bilingual Signal. In this paper, we follow works of [16, 17] that induce bilingual lexicon from non-parallel data. In order to learning bilingual lexicon from monolingual corpus, we must construct bilingual semantic representation. However, unlike the usual task that learns a precise bilingual lexicon, our objective is harvesting more bilingual signal from multilingual data. In this step, we care more about recall rather than precision. Our objective function is

$$\mathcal{T}\left(W_{V^s}^i, W_{V^t}^j\right) = \alpha\mathcal{T}_{mono} + \beta\mathcal{T}_{match} \tag{1}$$

where $W_{V^s}^i$ is the one word in vocabulary of V^s, while the reverse direction follows by symmetry for $W_{V^t}^j$. At the same time, in order to normalize $\mathcal{T}(W_{V^s}^i, W_{V^t}^j)$, we set the sum of α and β as 1. Parameters α and β mainly explain the influence of the monolingual and bilingual components.

Unlike the usual monolingual term $\mathcal{T}_{mono}$ explaining regularities in monolingual corpora, we use the term explain the translation probability of two words each other. Since similar words in semantic are closer in distance, we can reveal more translation pairs by measuring distance of two words from seeds. If the two words are closer from one seed in distance, it is more likely translated each other:

$$\mathcal{T}_{mono} = \mathcal{T}_{mono}^s + \mathcal{T}_{mono}^t \tag{2}$$

$$\mathcal{T}_{mono}^s = \min_{\langle ss,tt\rangle\in\mathbf{d}} \left\| W_{V^s}^i - W_{V^s}^{ss} \right\| \tag{3}$$

$$\mathcal{T}_{mono}^t = \min_{\langle ss,tt\rangle\in\mathbf{d}} \left\| W_{V^t}^i - W_{V^t}^{tt} \right\| \tag{4}$$

Our monolingual term $\mathcal{T}_{mono}$ encourages embeddings of word translation pairs from a seed lexicon $\mathbf{d}$ to move closer. $W_{V^s}^i$ and $W_{V^s}^{ss}$ are the two source words in the seed lexicon. $\mathcal{T}_{mono}^s$ computed the semantic similarity between the words $\mathbf{i}$ and ss. For the target, we have the same definition.

Our matching term $\mathcal{T}_{match}$ can expose how source-to-target words translate. The matching term actually can induce

$$\mathcal{T}_{match} = \underset{s\in V^s \text{ and } s\notin\mathbf{d}}{\arg\max} \left[\mathcal{M}_t^s\right] \cos\left(W_{V^s}^s, W_{V^t}^t\right) \tag{5}$$

As we learn bilingual signal from monolingual corpus, it means that source word vector and target word vector are trained independently each other. The two vectors are not in one vector space. In order to solve this problem, we follow the method of [18] to convert the monolingual vector space to a share space. Our objective is optimizing the cross-lingual matching regulizer:

$$\mathcal{M}_t^s = \sum_i \sum_j a_{ij} \left\| w_i^s - w_j^t \right\|^2 \tag{6}$$

$$= \left(\mathcal{R}^S - \mathcal{R}^T\right)^\top \boldsymbol{A} \left(\mathcal{R}^S - \mathcal{R}^T\right) \tag{7}$$

In above formula, we use $\boldsymbol{A}$ as the similarity matrix of two words, where a_{ij} encodes the translation score of word i in source with word j in target. w_i^s is the K-dimensional word embedding which is stacked to form a (V,K)-dimensional matrix $\mathcal{R}$.

Using a simple example to explain this procedure, assume that we have an English lexicon {perform, believe, talk} and a Chinese lexicon {zhixing, shixing, jiaotan}, an English-Chinese lexicon {conduct: jinxing}. Assuming that it have already conducting the step (1)(2), we can calculate that {perform} is closer with{conduct} in source and {zhixing, shixing} are closer with {jinxing} in target. We can add {perform: zhixing, perform: shixing} into original lexicon.

3.3. Parallel Sentences Identification. For our particular situation that is seriously low-resource language pairs, although classifier is good method to identify parallel sentence, we have not enough parallel sentences to train this classifier. So in the initial stage, we use a word-overlap model filter to select parallel sentences. The word-overlap model must borrow a bilingual lexicon, and the parallel sentences can be identified by the number of co-occurring word pairs. This process can be represented:

$$\text{score}\left(\mathrm{d}_s, \mathrm{d}_t\right) = \frac{\mathrm{d}_s \cap \mathrm{d}_t}{\min\left(\mathrm{d}_s, \mathrm{d}_t\right)} \tag{8}$$

From the above, we could conclude that inducing bilingual signal is an important step. Using bilingual lexicon, we can quickly calculate the alignment word for two sentences. In order to ensure getting massive parallel sentences, we must get a high coverage bilingual lexicon. However we may not get a large coverage bilingual lexicon, and the result is that we can only get a little parallel corpus using word-overlap model. We can also watch this in the experiments by **Section 4**. In order to get a large parallel sentences and high accuracy, we use classifier to get more parallel data in next. The classifier already has proved that it is an excellent method to extract parallel sentences. We follow the work of [6] to train a BiRNN (bidirectional recurrent neural network) classifier. Our neural network architecture is illustrated as Figure 2.

Like the most previous approaches that train neural network classifier using parallel sentences, our method also convert the sentences into vectors. However, unlike using word vector as input, we use fixed-size sentences vectors as input. For the ReLu layer, we can define as

$$\mathrm{x}_i = \text{sigm}\left(\mathrm{w}_i \mathrm{s}_i + \mathrm{b}\right) \tag{9}$$

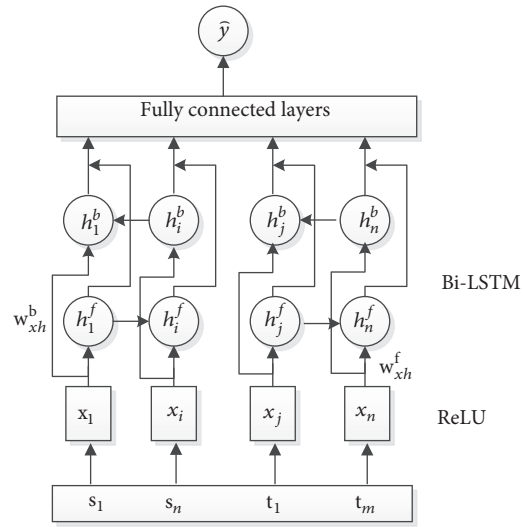

Figure 2: Architecture for bidirectional recurrent neural networks. The fully connected layers predict the probability of parallel sentence pair.

For the BiRNN layer, it contains feed-forward and feed-backward neural networks layer. This can be described by

$$h_i^f = \emptyset\left(w_{xh}^f x_i + w_{hh}^f h_{i-1} + b^f\right) \tag{10}$$

$$h_i^b = \emptyset\left(w_{xh}^b x_i + w_{hh}^b h_{i-1} + b^b\right) \tag{11}$$

$$\mathrm{h} = h_i^f + h_i^b \tag{12}$$

For prediction, a sentence pair can be identified as parallel if the probability exceeds a threshold. We can compute as follows:

$$\hat{y} = \begin{cases} 1 & if\ \ p(\mathrm{y} = 1 \mid \mathrm{h}) > \sigma \\ 0 & otherwise \end{cases} \tag{13}$$

4. Experiments

To assess the effectiveness of our method, we compare it in different setting against the baseline. As we mainly pay attention to solve low-resource problem and construct low-resource language pair translation system. As an instantiation of this goal, we conduct a detailed study on Uyghur-Chinese.

4.1. Experiment Setup

4.1.1. Data. In our experiments, the actual systems obtain bilingual parallel corpus from three multilingual website: TianShan (http://www.ts.cn/), RenMin (http://www.people.com.cn/) and KunLun (http://www.xjkunlun.cn/) website. We only retain those webpages of documents having more than 20 words. The statistics of the preprocessed corpora is given in Table 1. Pay attention that we only select the length of sentence that exceeds 10 words. The data of our experiment is available at https://pan.baidu.com/s/1EePrHOjhuN-jTb-vNiSgTA.

Table 1: Experiment set statistics.

Websites	languages	#webpages	#sentences
TianShan	Chinese	249,238	3,839,000
	Uyghur	48,907	427,000
RenMin	Chinese	451,972	5,500,000
	Uyghur	99,578	590,000
KunLun	Chinese	44,046	641,000
	Uyghur	27,419	324,000

4.1.2. Evaluations and Ground Truth. In order to carry out an objective evaluation for obtained parallel sentences, we perform two methods to evaluate. The first is translation accuracy, which is the proportion of truly parallel sentence pairs among all obtained sentences pairs. As we obtain data from open data platform, we can't get a standard translation language pairs to compute the translation accuracy of obtained parallel sentences. So we use manually evaluate the accuracy of a random sample of the obtained parallel sentences. In experiments, we randomly select 500 obtained parallel sentences to conduct manual evaluation. Another is that use obtained parallel sentences to construct machine translation system and the BLEU score as an evaluation metric.

4.1.3. Baseline. For comparison, we use a parallel sentences extraction system Bitextor. The system is a free/open-source tool for harvesting parallel data from multilingual websites. User is required to provide one or more URLs of websites to be processed, the two languages for which the parallel corpus will be produced, and a bilingual lexicon in these two languages. This system can automatically analyze the structure of webpages and obtain parallel data by bilingual lexicon. Thus, we alter various size bilingual lexicons to test how it affects the obtaining of parallel sentences.

Another problem is evaluating classifier. We all know that we must use parallel sentences to train classifier and using the classifier predicts parallel sentences. The size of parallel sentences affects the classifier performance. So we, respectively, select different number of parallel sentences to train classifier and test it.

4.2. Effect of Bilingual Lexicon Size. In order to investigate the effectiveness of our system for obtaining parallel sentences in low-resource language pairs, we respectively run ours and Bitextor with different bilingual sice. As the Bitextor does not use the time window as a feature to select parallel data, we use this feature in our system. So we use the time window to filter the results of Bitextor in order to keep experimental consistency. We record the performance by varying the lexicon size to conduct training process, shown in Table 2. The table is for Uyghur-Chinese.

In experiments, we use $\{600; 1,500; 5,000; 10,000\}$ entries of lexicon to conduct the obtaining parallel sentences process. From Figures 3 and 4 we can immediately see the important role the bilingual lexicon plays in obtaining parallel sentences process. We observe that Bitextor does not

Table 2: The size and accuracy of obtaining parallel sentences in different number of training corpus.

Model		The number of training parallel sentences				
		2,000	5,000	10,000	20,000	40,000
LSTM	size	13,000	33,000	65,000	92,000	126,000
	accuracy	0.6	0.71	0.78	0.81	0.82
C-BiRNN	size	14,000	28,000	58,000	86,000	121,000
	accuracy	0.58	0.63	0.68	0.70	0.72

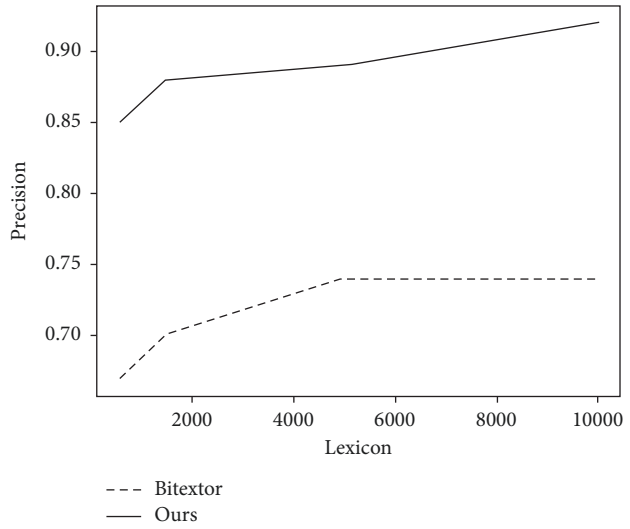

Figure 3: Precision of result as the entries of bilingual lexicon.

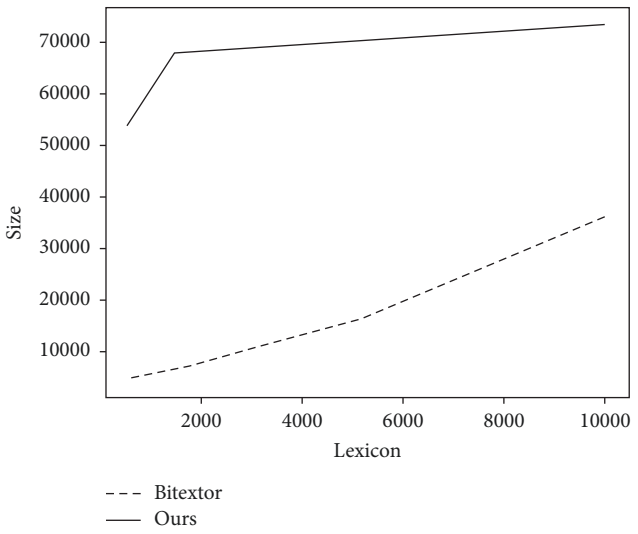

Figure 4: Size of result as the entries of bilingual lexicon.

obtain parallel sentences when the size of bilingual lexicon is very small. However, we see that our system can get a very objective result under low resources. We can easily find that Bitextor has a very unstable performance with different lexicon size. However, ours can keep relatively stable results no matter the number and accuracy of obtaining parallel sentences. When the lexicon entry only is a little such as 600, ours can get a very objective number and accuracy of parallel sentences. This result, combined with the inadequate performance of the baseline, conforms to our expectation that obtain parallel sentences for low-resources language pairs.

From Figures 3 and 4, we can obtain a lot of parallel sentences with a high accuracy. However, we still find that we can't obtain enough parallel sentences for actual natural language processing such as SMT. We analyze two factors affecting the number of obtaining parallel sentences: (1) despite constructing word embeddings and using methods in Section 3, we can find more bilingual signals. However, we only retain words that occur at least 1,000 times (Lower times threshold make the accuracy very low and low time word cannot get a fine word embedding.), and it seriously limits obtaining the size of bilingual signal. (2) We set a time window to filter noisy, it makes large candidate not be obtained in final parallel sentences.

4.3. Effect of Parallel Sentences for Classifier Experiments. Using bilingual signal, we can obtain a certain number of parallel. From Section 4.2, we can see that the number of obtaining parallel sentences is not large enough. Constructing bilingual classifier is a good method to filter bilingual parallel sentences from monolingual corpus. In this section, we will discuss the classifier how affect the obtaining parallel sentences and which factor affects the processing of obtaining parallel sentences.

In this experiment, we construct based on LSTM and classical bidirectional recurrent neural network (C-BiRNN) classifiers to filter parallel sentences from monolingual corpus. At the same time, we use {2,000; 5,000; 10,000; 20,000; 40,000} number of parallel sentences to train classifier. Table 2 shows the results of the tested systems for Uyghur-Chinese. We can observe that the two neural work to have a different performance for size and accuracy of obtaining parallel sentences. From the table, the LSTM have a better result than the other no matter the size and accuracy. We attribute it to the fact that LSTM have a better neural network structure to remember more information than the classical bidirectional recurrent neural network (C-BiRNN). Another interesting finding is that the size of training corpus plays a big role to filter bilingual parallel sentences. When the number of training parallel sentences is only 2,000, the two testing systems only obtain a few results and the most unacceptable is that the result has a very low accuracy so that the result cannot use any natural language process tasks. However, as the training parallel sentences increase, the size and accuracy have a

Table 3: Statistics of the size and precision of parallel sentences extracted from multilingual websites.

Model	Training corpus	#sentences	#precision
Bitextor&LSTM	30,000	117,900	0.70
	40,000	124,200	0.70
Ours&LSTM	30,000	120,200	0.81
	40,000	127,900	0.82

Table 4: BLEU scores on Uyghur-Chinese SMT using different training corpus.

Model	BLEU	#sentences
Bitextor&LSTM &SMT(baseline)	5.6	100,000
Ours &LSTM &SMT	15.81	100,000

great improvement no matter C-BiRNN and LSTM. We can conclude that the number of training parallel sentences has a big influence on the performance of classifier. This conclusion can present the importance of our inducing bilingual signal. Only by the methods detailed in Section 3.2 and experiment in Section 4.2 can we obtain enough parallel sentences to train an state-of-the-art classifier.

4.4. Machine Translation Evaluation. Our final objective of obtaining parallel sentences is training a machine translation system to perform translation task for low-resource language pair. In order to justify the effectiveness of our methods, we obtain parallel sentences to construct a machine translation system in low-resources Uyghur-Chinese language pair and evaluate its quality by measuring the BLEU score on SMT system. We use an state-of-the-art free/open source Moses [19] to train phrase-based translation system.

In our experiment, we use the Bitextor and ours method to obtain training parallel sentences, and the classifiers all use LSTM neural network. The reason of using Bitextor is that we need a baseline system to measure. For two methods, we select {30,000; 40,000} sentence pairs as training corpus to construct classifier and obtain enough training corpus to train machine translation system. The first experiment is selecting enough number of parallel sentences (see Table 3). We can see that ours exceed Bitextor under using same classifier. Although the two can get many of candidate parallel sentences, the results of Bitextor have a low precision. We attribute the reason that the Bitextor needs an enough bilingual lexicon and ours does not.

In next section, we will use the collocated extraction procedure described in Section 3 to train some machine Uyghur-Chinese translation systems. As the baseline SMT system, we use parallel sentences obtained by Bitextor to train a classifier to obtain final training corpus for SMT system. Table 4 shows the BLEU scores for the different SMT systems.

We can see that our approach can get a higher BLEU score than the baseline. In the experiment, we all use 30,000 sentence pairs to train classifier. Combining Table 3 with Table 4, we can believe that the baseline cannot get a very high accuracy of parallel sentences and makes SMT system have a low performance. As we all know, the quality of training corpus heavily affects the performance of SMT system. We further analyze the Bitextor need of a bilingual lexicon to guarantee a high accuracy of parallel corpus. Although it is an excellent system to obtain parallel corpus, it will show a poor performance for low-resources language pairs. This experiment clearly indicates the benefit of obtaining parallel sentences using our method. It is important to note that we can construct a machine translation system with low-resources language pairs.

5. Conclusion

In this paper, we present a new minimal supervision method to obtain parallel sentences for solving low-resources problem in natural language processing. Our experiments show that our approach outperforms the traditional system to obtain parallel corpus from multilingual websites for low-resources language pairs.

Our methods mainly contain three steps. First, we use Word2vec to train two monolingual word embeddings. By a small bilingual lexicon about hundreds of words, we can induce more bilingual signals. Then, using a word-overlap model finds some parallel sentences. This step avoids the effect of HTML structure as the current website is developed into dynamic modules. Finally, we construct a LSTM-BiRNN classifier to extract parallel sentences. We use the parallel corpus obtaining in above step to train this classifier and perform extracting process. We use the final obtaining parallel sentences to construct a Uyghur-Chinese SMT system to measure our method. The experiments indicate that our method can get state-of-the-art results in low-resources language pair.

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 work is supported by the Xinjiang Fun (Grant no. 2015KL031), the West Light Foundation of the Chinese Academy of Sciences (Grant no. 2015-XBQN-B-10), the Xinjiang Science and Technology Major Project (Grant no.

2016A03007-3), and Natural Science Foundation of Xinjiang (Grant no. 2015211B034)

References

[1] L. Barbosa, V. Sridhar K, and M. Yarmohammadi, "Harvesting Parallel Text in Multiple Languages with Limited Supervision," *International Conference on Computational Linguistics*, pp. 201–214, 2012.

[2] D. S. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol. 31, no. 4, pp. 477–504, 2005.

[3] M. Esplà-Gomis, M. Forcada, S. Ortiz Rojas, and J. Ferrández-Tordera, "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*, pp. 685–691, Berlin, Germany, August 2016.

[4] W. Ling, L. Marujo, C. Dyer, A. W. Black, and I. Trancoso, "Crowdsourcing High-Quality Parallel Data Extraction from Twitter," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 426–436, Baltimore, Maryland, USA, June 2014.

[5] A. Khwileh, H. Afli, G. Jones, and A. Way, "Identifying Effective Translations for Cross-lingual Arabic-to-English User-generated Speech Search," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, pp. 100–109, Valencia, Spain, April 2017.

[6] F. Grégoire and P. Langlais, "A Deep Neural Network Approach To Parallel Sentence Extraction," 2017, https://arxiv.org/abs/1709.09783.

[7] J. R. Smith, C. Quirk, and K. Toutanova, "Extracting parallel sentences from comparable corpora using document level alignment," in *Proceedings of the 2010 Human Language Technologies Conference ofthe North American Chapter of the Association for Computational Linguistics, NAACL HLT 2010*, pp. 403–411, USA, June 2010.

[8] C. Tillmann and S. Hewavitharana, "An efficient unified extraction algorithm for bilingual data," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*, pp. 2093–2096, Italy, August 2011.

[9] R. G. Hussain, M. A. Ghazanfar, M. A. Azam, U. Naeem, and S. Ur Rehman, "A performance comparison of machine learning classification approaches for robust activity of daily living recognition," *Artificial Intelligence Review*, pp. 1–23, 2018.

[10] M. A. Ghazanfar, S. A. Alahmari, Y. F. Aldhafiri, A. Mustaqeem, M. Maqsood, and M. A. Azam, "Using machine learning classifiers to predict stock exchange index," *International Journal of Machine Learning and Computing*, vol. 7, no. 2, pp. 24–29, 2017.

[11] C. Chu, T. Nakazawa, and S. Kurohashi, "Constructing a Chinese-Japanese parallel corpus from wikipedia," in *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pp. 642–647, Iceland, May 2014.

[12] A. Barrón-Cedeño, C. España-Bonet, J. Boldoba, and L. Màrquez, "A Factory of Comparable Corpora from Wikipedia," in *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pp. 3–13, Beijing, China, July 2015.

[13] V. K. Rangarajan Sridhar, L. Barbosa, and S. Bangalore, "A scalable approach to building a parallel corpus from the Web," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*, pp. 2113–2116, Italy, August 2011.

[14] A. Antonova and A. Misyurev, "Building a web-based parallel corpus and filtering out machine-translated text," *The Workshop on Building Using Comparable Corpora: Comparable Corpora the Web*, pp. 136–144, 2011.

[15] V. Papavassiliou, P. Prokopidis, and G. Thurmair, "A modular open-source focused crawler for mining monolingual and bilingual corpora from the web," *The Workshop on Building & Using Comparable Corpora*, pp. 43–51, 2013.

[16] T. Mikolov, K. Chen, and G. Corrado, "Efficient Estimation of Word Representations in Vector Space," *Computation and Language*, 2013.

[17] M. Zhang, H. Peng, Y. Liu, H. Luan, and M. Sun, "Bilingual lexicon induction from non-parallel data with minimal supervision," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pp. 3379–3385, USA, February 2017.

[18] S. Gouws, Y. Bengio, and G. Corrado, "BilBOWA: Fast bilingual distributed representations without word alignments," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pp. 748–756, France, July 2015.

[19] P. Koehn, R. Zens, C. Dyer et al., "Moses: open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07)*, pp. 177–180, Prague, Czech Republic, June 2007.