Harvest Uyghur-Chinese Aligned-Sentences Bitexts from Multilingual Sites Based on Word Embedding

ShaoLin Zhu[3[1-2]], 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
yangyt@ms.xjb.ac.cn

Abstract. Obtaining bilingual parallel data from the multilingual websites is a long-standing research problem, which is very benefit for resource-scarce languages. In this paper, we present an approach for obtaining parallel data based on word embedding, and our model only rely on a small scale of bilingual lexicon. Our approach benefit from the recent advances of continuous word representations, which can reveal more context information compared with traditional methods. Our experiments show that high-precision and sizable parallel Uyghur-Chinese data can be obtained for lacking bilingual lexicon.

Keywords: bilingual parallel data, word embedding, resource-scarce languages

1 introduction

Parallel data is one of the most important linguistic resources for cross-lingual natural language processing (Melamed et al, 2001), especially for statistical machine translation (SMT) and neural machine translation (NMT). Nowadays, the Internet may be seen as a large multilingual corpus as there a large number of websites in which different pages can be found containing the same content written in different languages. In our case, our approach is focused on using the web as a source of bitexts (parallel texts).

Many approaches have been presented for trying to exploit the multilingual sites as bitexts. There are several tools that can be used for automatically crawling parallel data from multilingual websites (Bitextor [1] , 2013; PaCo, 2012;ILSP-FC[2];2012;WPDE,2006). However, all of them share the same limitations:(1) they require the user to provide the URLs of the multilingual websites to be crawled. The crawler downloads the all web pages texts, but the web pages contain a lot of noise such as advertising links, hot recommended et al. To deal with the limitations, we implement the tool called Scrapy[3] to crawl the specific plain text part of web page. (2)

[1] Bitextor: https://sourceforge.net/projects/bitextor
[2] ILSP_FC: http://nlp.ilsp.gr/redmine/projects/ilsp-fc
[3] Scrapy: https://pypi.python.org/pypi/Scrapy/1.4.0

those heavily depend on lexical information, past experiments indicate that the performance improves as the translation lexicon becomes larger. However, it is difficult to obtain a large translation lexicon for scarce resources language such as Uyghur-Chinese.

Unlike the other pair of languages, the Chinese sentences require word segmentation in order to word alignment. Most segmentaion tools are based on semantics can give rise to the data sparse. For example, the same semantic word does not appear in the bilingual lexicon. We can illustrate the problem in Figure 1.

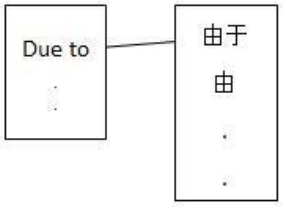

Fig. 1. One source word in the lexicon (left) is matched to a target word (right). However, we can find the other words also can be matched and they don't appear in the lexicon.

Thanks for the emerge of continuous vector representation of words, commonly known as Word Embedding (Mikolov, Le, and Sutskever 2013), which is supposed to carry semantic clues. The biggest contribution of the Word Embedding is to make the relevant or similar semantic words closer in the distance. For example, a Uyghur word can be translated more than one Chinese words, but only one of them in our lexicon. Others can be established connection by using the semantically related word embedding.

The same content of bilingual pages is a most important feature in identifying parallel pages, we encode our intuition into a novel computing term by combining the word embedding to identify the bilingual content. Somewhat surprisingly, even in a small bilingual lexicon(in our experiment, 12,000 bilingual entries), a sizable and high-precision parallel Uyghur-Chinese corpus can be obtained from the multilingual web sites.

The mainly contribution of this paper is as follows:(i) this paper combine word embedding to obtain aligning sentences, to deal with the limitation of the scarce resources. (ii) this approach allows to obtain parallel data in a totally automatic fashion, i.e. without having to provide a large seed lexicon.

2 Related Works

One of the most common strategies to crawl parallel data from the websites is to focus on multilingual web sites that make it straight-forward to detect parallel documents (Nie et al., 1999;Koehn, 2005; Tiedemann, 2012; Miquel, 2013). Many approaches use content-based metrics (Jiang et al., 2009; Utiyama et al.,2009; Yan et al., 2009; Hong et al., 2010; Sridhar et al., 2011; Antonova and Misyurev, 2011; Barbosa

et al., 2012), such as bag-of-words co-occurrence. The method of content-based metrics is the most important feature in detecting the parallel texts. Although these metrics have proved to be useful for parallel data detection, their main limitation is that they require amount of linguistic resources (such as a bilingual lexicon or a basic machine translation system) which may not be available for some language pairs(such as Uyghur-Chinese). To avoid this problem, other works use the HTML tags of the web pages, which usually remains stable between different translations of the same document (Ma and Liberman, 1999; Nie et al., 1999; Resnik and Smith, 2003; Zhang et al., 2006; Forcada, 2010; San Vicente and Manterola, 2012; Papavassiliou et al., 2013). Another useful strategy is to identify language markers in the URLs (Ma and Liberman, 1999; Nie et al., 1999; Resnik and Smith,2003; Zhang et al., 2006; D esilets et al., 2008; Espla-Gomis and Forcada, 2010; San Vicente and Manterola, 2012) that help detect possible parallel documents. However, those methods only can be used in some specific sites and not be applicable to some news websites. The popularity of the dynamic website makes the application of this method gone away (see figure 2). We cannot learn any parallel information from the structure pages.

Fig. 2. URLs are very similar, but we cannot get any bilingual information.

Munteanu and Marcu(2005) adopt a useful strategy to obtain bitexts, who compare the time of news published in news websites written in different languages by using a publication time stamp window. Ying Zhang and Ke Wu(2006) present multiple features to identify English-Chinese bitexts via a k-nearest-neighbor classifier. To determine the parallelism between potential document pairs, they calculated the similarity of content translation feature by a large English-Chinese lexicon containing 250,000 entries. It indicates selecting alignment text sentences depending heavily on bilingual lexicon.

Even though these methods have proven to be useful for specific web sites, the real challenge is to find a large bilingual lexicon. For some language pairs (such as English-Spanish, English-French or English-Chinese et al), it is easy to obtain. However, it is a challenge to obtain parallel corpora for some low resource language such as Uyghur-Chinese. On the other hand, those method use crawl the whole web pages to find potential parallel texts, but the current pages contains too many noise. Such as the news web sites, it often contains advertising links, hot recommended or directory information et al.

In this paper we propose a novel strategy for building parallel corpora automatically only rely on a small scale of bilingual lexicon. This strategy mainly deal with the limitation of scarce bilingual lexicon, and the experiments indicate a sizable number of high-precision aligning Uyghur-Chinese sentences can be obtained, even in the a small bilingual lexicon(in our experiment, the entry is 12,000).

3 methods

In this paper, we proposed a word embedding based parallel corpora extraction method, which can obtain a large scale of Uyghur-Chinese parallel corpus only rely on a small bilingual lexicon. In this section, we describe the details of our method, which include two main parts: (i) we firstly need to detect align-documents from the multilingual websites. (ii) Our objective is obtaining parallel sentences from the documents.

3.1 Detection of Alignment Document

Previous works extracted the alignment documents are based on multiple features such as the file length, the HTML tags, the co-occurrence of words et al. The co-occurrence is the most important feature to filter alignments. However, the features heavily depend on the bilingual lexicon (see Table 1).

Table 1. the number of words in our corpora is far more than the vocabulary

language	#pages number	#tokens	Vocabulary Size
Bitextor	295,303	323,102	7009
Ours	52,336	91,235	8528

In order to extract more candidate precise pair of documents, we urgently need a large bilingual lexicon to count the number of co-occurrence words in two language web pages. However, the actual situation deeply strike us, the lexicon is far away from us for scarce language resources. We only could obtain a small Uyghur-Chinese lexicon. Fortunately, we can reveal a lot new semantic information utilize word embedding, it convert words into vectors and the similar two vector in semantics is closer in distance(such as Cosine, Euclidean distance and others). For example, the words “提出” and “提议” are closer distance than “表明”. We use Word2vec[4](Mikolov etal. 2013) to convert words into word embedding. We attempt to utilize word embedding to provide further bilingual information about whether the monolingual document should be aligned.

We calculate the similarity of bitexts content combining word embedding with k-nearest neighbor, the basic calculation is as the following measure:

$$F(d_1,d_2) = \frac{\sum Matching(s_i,t_j)}{\max len(d_1,d_2)} \qquad (1)$$

Where d_1 and d_2 are the source and target document, $s_i \in \{w_1, w_2 \dots\dots, w_n\}$ and w_i is the word of source document, the target side are similar. $Matching(s_i, t_j)$ is the matching mechanism which reveal the source word are translated into target word. The matching mechanism is calculated as following:

[4] Word2vec: https://code.google.com/p/word2vec/

$$\mathrm{m}(s_i, t_j) = \begin{cases} 1 & if\ s_i, t_j\ in\ the\ bilingual\ lexicon\ and\ document \\ c & if\ s_i, t_j\ in\ the\ bilingual\ lexicon\ but\ one\ of\ them\ in\ document \\ 0 & others \end{cases} \quad (2)$$

$$Matching(s_i, t_j) = \begin{cases} 1 & if\ \mathrm{m}(s_i, t_j) = \mathrm{m}(t_i, s_j) \\ 0 & if\ \mathrm{m}(s_i, t_j) \neq \mathrm{m}(t_i, s_j) \end{cases} \quad (3)$$

Specifically, to calculate the parameter c, we present combining word embedding with k-nearest neighbor. For $\mathrm{m}(s_i, t_j)$, we firstly retrieval source word s_i in lexicon, if the result w is not target word $t_j$and not null. We utilize the word embedding to calculate the k-nearest words z_i close to result w. $z_i \in \{W_1, W_2, \dots, W_k\}$, which the set is k-nearest words from "w". The "k" explains the number of nearby the target side which can be retrieved. If the target document contains one of the "w", we will set the parameter c as $c = 1$ and other situation we set $c = 0$. The corresponding $\mathrm{m}(t_i, s_j)$ is generated analogously using this method. This method cannot conduct an explicit matching process in inference. To deal with the malpractice, we present two kinds of ways:(i) As our task is obtaining parallel corpora from the news websites, we consider it likely that articles with similar content have publication dates that are close to each other. Thus, each query is actually run only against documents published within a window of five days around the publication date of the other side query document. (ii) We detect the parallel documents by calculating alignment sentences model (we will explain in Sec 4), which is a probability generative model based on Word Embedding.

3.2 Alignment Sentences Model

Like most approaches that obtain parallel sentences from websites, our learning objective is obtaining parallel sentences from the documents. However, unlike the previous works that need a large bilingual seed lexicon, ours additionally includes generative model that attempts to maximize translation probability from source side to target.

Our probability generative model is inspired by IBM model 1 (Brown et al. 1993). We begin by an exposition of source-to-target linking, at the same time the reverse direction follows by symmetry. We assume each source word in the source sentence s^s should have synonyms, namely can be formatted multivariate Gaussian (see Equation 4); Then we use the Word2vec transform words into vector $\mathrm{w}_i \epsilon \mathbb{I}^v$, which represent v-dimensional real number space vector. Then we can write out the basic source-to-target probability p_{s2t}:

$$w_i \sim \mathcal{N}(0, i_d) \quad (4)$$

$$P_{S2T} = \prod \log P(w^s | w^t) = \prod \log P(w^s | \{w^{s2t}\}_1^k) \quad (5)$$

w^{s2t} is a target word corresponding translation of the $w^s$and assuming that the vector w^{s2t} has k synonyms, which the set is $\{\mathrm{w}^{s2t}\}_1^k$ and the set is k-nearest words

from w^{s2t} . We leave the discussion on a practical choice to a later section. For the vector w^t in the target sentence, $w^t \in \{\mathrm{w}^{s2t}\}_1^k$ specifies whether target word can be linked. The reverse direction follows by symmetry. We assume each vectors in the set is independent of each other, and it only depends on the premier word w^s. Therefore, we have:

$$P\left(w^s \middle| \{w^{s2t}\}_1^k\right) = \mathrm{P}\left(w^s \middle| w^t \in \{w^{s2t}\}_1^k\right) = P(w^s, w_i^{s2t}) \quad (6)$$

Where $\{w^{s2t}\}_1^k$ can be computing by

$$\{w^{s2t}\}_1^k = \arg\ \min_k \left\| \{w_t^T\}_{t=1}^{V^t} - w^{s2t} \right\|^2 \quad (7)$$

Where $\{w_t^T\}_{t=1}^{V^t}$ is a set that the target corpus contains the number of words are nearest distance from w^{s2t}. Finally, the parameterization of the probability can be expressed by the similarity distance, we can conclude:

$$P\left(w^s, w_i^{s2t}\right) = \begin{cases} 1 & if\ w_i^{s2t} = w^t \\ \dfrac{\max\left\|\{w_t^T\}_{t=1}^k - w^{s2t}\right\|^2 - \left\|w^{s2t} - w_i^{s2t}\right\|^2}{\max\left\|\{w_t^T\}_{t=1}^k - w^{s2t}\right\|^2} & otherwise \end{cases} \quad (8)$$

Next, we can further elaborate the documents alignment problem which is mentioned. Through calculating the number of alignment sentences, it is easy to decide whether the documents are aligned. For example:

$$\mathrm{m}(s_i, t_j) = \frac{\sum_1^k \prod \log P(w^s | \{w^{s2t}\}_1^k)}{\max\ len(d_s, d_t)} \quad (9)$$

It mainly explains the probability of alignment sentences how affect the parallel documents. If a pair of documents has many non-parallel sentences, we can mark them as non-parallel documents.

4 Experiment

In this section, we will explain the implementation details of our proposed system and verify the performance by experiment.

4.1 Data

In our experiments, the tested systems obtain parallel sentences form multilingual web sites on Uyghur-Chinese language pair by the crawler Scrapy[5]. For we couldn't obtain a large bilingual lexicon, we have a small lexicon about 12,000 entries. For the

[5] Scrapy: https://pypi.python.org/pypi/Scrapy/1.4.0

Chinese side, we first implement OpenCC[6] to normalize characters to be simplified, and perform Chinese word segmentation using Jieba.[7] The preprocessing of Uyghur side involves tokenization, POS tagging, lemmatization, which are carried out by a tool developed by our team. Then we use the Word2vec to convert the words into word embedding. The statistics of the preprocessed data is given in Table 1.

4.2 Baselines

We compare our approach to two existing system:

1. Bitextor (Espla-Gomis, 2013).
2. INSP-FC (Vassilis, 2012)

The first baseline (Bitextor) is 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. The core component of Bitextor to find document and sentence alignments is content-based and URL-based heuristics and algorithms applied to identify and align the parallel web pages in a website.

The second baseline (INSP-FC) is a modular system that includes components and methods for all the tasks required to acquire domain-specific corpora from the Web. The system is available as an open-source Java project and due to its modular architecture, each of its components can be easily substituted by alternatives with the same functionalities. Depending on user-defined configuration, the crawler employs processing workflows for the creation of either monolingual or bilingual collections.

4.3 Results and Discussion

In this section we examine how our system performance contrasting the baseline. Then we will discuss how the parameter “k” and the size of bilingual lexicon affect obtaining parallel data.

4.3.1 Overall Performance

Table 2 shows the performance of our system compare with two baselines. The bilingual lexicon size is 12,000 and the “k” is set as 3 in our experiment. As we downloaded the news data which can be marked release-time easily, the results of all system can be filtered by time. In our experiment, we only save the two documents which are published in three days.

[6] OpenCC: https://pypi.python.org/pypi/opencc-python/
[7] Jieba: https://pypi.python.org/pypi/jieba/

Table 2. the number of obtained data for Uyghur-Chinese corpora

methods	#documents	#sentences
Bitextor	71	828
INSP-FC	83	1047
Ours	316	5,628

Compared to the baseline, our system has a considerable promotion in obtaining parallel data. However, no matter what method the obtained data is too small to application. We analyze two factors affecting the obtaining:(i) the multilingual website contains a few of parallel pages so that we can't get a large data. For this problem, we can download more multilingual website to select bitexts which we need. (ii) another factor perhaps is the bilingual lexicon. In our experiment, Uyghur-Chinese is a resource-scarce language pair with limited parallel data. In fact, although it is very small, ours is still having a significant performance than the others.

Although ours outperform the others, we should analyze the precision of results and if the precision is too small, the system is not good. We use manual criteria to examine our system performance contrasting baseline. We first manually select 20 pairs of documents and 100 pairs of sentences randomly, then examine the accuracy of them. The precision of documents is defined as the number of correctly obtained over the total number of pairs documents obtained. The precision of sentences is defined similarly. The result is given in Table3.

Table 3. Accuracies of random samples

methods	Accuracy of document(%)	Accuracy of sentence((%)
Bitextor	83	93
INSP-FC	88	92
Ours	80	90

Table 3 shows the three system have an almost same precision. Combining Table 2 with Table 3, we can find that attain considerably better performance. We have more parallel data compared with the two baseline. The poor performance of the baseline should be attributed to the harsh condition they have to face, which only 12,000 entries can be used. It is too few for them to reveal bilingual signals and obtain parallel data. Table 1 shows that our monolingual corpus contains 323,102 and 91,235 words separately, but the bilingual lexicon only has 7009 and 8528 words each other. However, the success of our approach certificates that it is actually possible obtaining a considerable language pairs.

4.3.2 Effect of Bilingual Lexicon Size

In this section, we will investigate how the number of bilingual lexicon may affect the performance of obtaining parallel data. We change the bilingual lexicon size

in {5,000; 8,000; 10,000; 1,2000}. Figure 3 shows the accuracies of the tested systems for Uyghur-Chinese. Figure 4 shows the results size varies as the bilingual lexicon size. We observe that although the result precision of ours is not performing out the baseline, the results size of ours far performance than the baseline. From the Figure 3 we could find that obtaining parallel data form websites heavily depend on the bilingual lexicon. However, even in the more difficult cases, ours can obtain a sizable high-precision parallel data. We conjecture this is due to that our method provides a larger search-space to find bilingual signal, and then we can obtain more data.

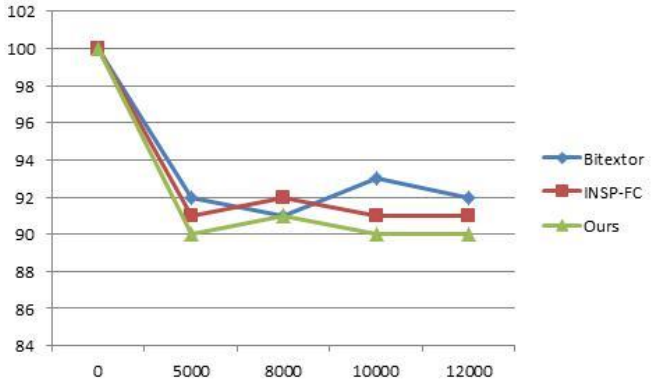

Fig. 3. Accuracies vary with the bilingual lexicon size

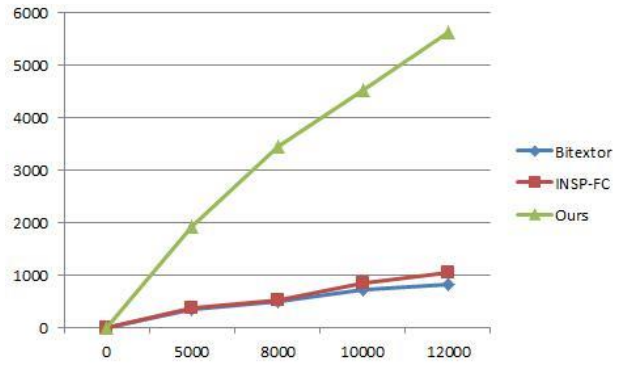

Fig. 4. The number of sentences in results varies with the bilingual lexicon size

4.3.3 Effect of the value "k"

In this section, we discuss the value "k" proposed in section 3. In order to investigate the effectiveness of the value "k", we run a version of our system with different

value. We mainly discuss that the value “k” how affect the results size and accuracy. The bilingual lexicon size is 12,000 in our experiments.

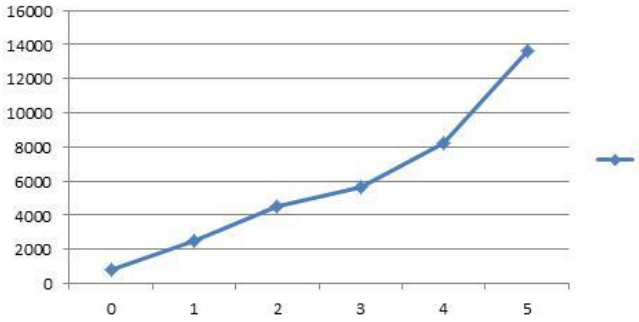

Fig. 5. The number of sentences in results varies with the value “k”

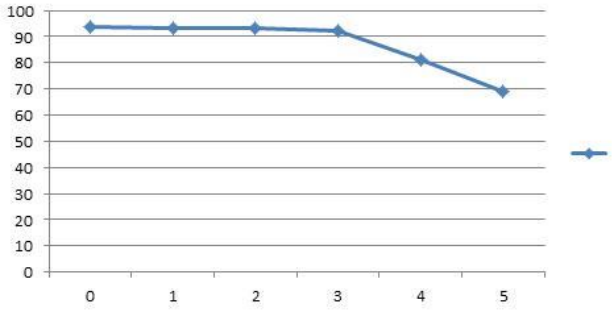

Fig. 6. Accuracies vary with the value “k”

From the two figures 5 and 6 we immediately see the important role the value “k” plays in our method. Varying the value “k” can result in dramatic accuracy and size gain. We conjecture this is due to that when the value is very small such as “1”, although the 1-nearest words are very similar, the search-space is also too small to obtain more bilingual signal. It will cause that we can’t get a sizable size. With the increase of the value, the accuracy will gradually decrease. We could find it is very obvious when the value is set as “5”. We conjecture this is due to that increasing the value can bring a lot of noise, it arouses the system product many incorrect bilingual signals. Considering the size and accuracy, we set the value as “3”.

5 Conclusions and Outlook

In this paper, we explore harvesting aligned-sentences parallel data from multilingual websites using currently popular Word Embedding. We mainly deal the limitation that the bilingual lexicon is heavily scarce in mining parallel data from the multilingual websites. We train monolingual word embedding in obtained monolingual corpora. In addition, we properly embed more signals cross-lingual by introducing the k-nearest words. We show our method dramatically improve the obtaining size, and allows that it has a high-precision.

Due to the training data used in our experiments is relatively small; therefore, we can't obtain a very large parallel data. We will increase the number of training data in our future work. In addition, we will further test our proposed approach in other language pairs.

Acknowledgments. This work is supported by the Xinjiang Fun under Grant (No.2015KL031), the West Light Foundation of The Chinese Academy of Sciences(No.2015-XBQN-B-10), the Xinjiang Science and Technology Major Project(No.2016A03007-3) and Natural Science Foundation of Xinjiang(No.2015211B034)

References

1. Miquel Espla-Gomis and Mikel L. Forcada.: Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. The Prague Bulletin of Mathematical Linguistics, 77–86 (2010)
2. Ying Zhang, Ke Wu, Jianfeng Gao and Phil Vines.: Automatic acquisition of Chinese–English parallel corpus from the web. In Advances in Information Retrieval, volume 3936. 420–431(2006).
3. Inaki San Vicente and Iker Manterola.: PaCo2: A fully automated tool for gathering parallel corpora from the web. In Proceedings of the 8th International Conference on Language Resources and Evaluation, 1–6(2012).
4. Philip Resnik and Noah A. Smith.: The Web as a parallel corpus. Computational Linguistics, 349–380(2003).
5. Vassilis Papavassiliou, Prokopis Prokopidis, and Gregor Thurmair.: A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, 43–51(2013).
6. Munteanu, D. S. and Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics, 477–504(2005).
7. M. Espla-Gomis.: Bitextor, a free/open-source software to harvest translation memories from multilingual websites. In Beyond Translation Memories Workshop (MT Summit XII). (2009)

8. M. Espla-Gomis and Mikel L. Forcada.: Bitextor's participation in WMT'16: shared task on document alignment. Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, 685–691(2016).
9. Xiaoyi Ma and Mark Y. Liberman.: BITS: A Method for Bilingual Text Search over the Web. Linguistic Data Consortium, 538-542(1999)
10. Miquel Espla-Gomis, Filip Klubicka and Nikola Ljube.: Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites. LREC 2014 Proceedings, 1252-1256(2014)
11. Nie, Jian-Yun, Simard, Michel, Isabelle, Pierre, and Durand, Richard.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In Proceedings of the 22ndAnnual International ACM SIGIR Conference on Research and Development in Information Retrieval, 74–81(1999).
12. Wang Ling,Lu ́ s Marujo, Chris Dyer, Alan Black and Isabel Trancoso.: Crowdsourcing High-Quality Parallel Data Extraction from Twitter. In Proceedings of the Ninth Workshop on Statistical Machine Translation,426-436(2014)
13. Dragos Stefan Munteanu and Daniel Marcu.: Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist, 477–504(2005).
14. Mikolov, T. Chen, K. Corrado, G and Dean, J. 2013a. Efficient Estimation of Word Representations in Vector Space. In ICLR Workshop, 1-12(2013)
15. Mikolov, T, Sutskever, I. Chen, K. Corrado, G. S and Dean.: Distributed Representations ofWords and Phrases and their Compositionality. In NIPS,3111-3119(2013)