A Mutual Iterative Enhancement Model for Simultaneous Comparable Corpora and Bilingual Lexicons Construction

Zede Zhu, Xinhua Zeng, Shouguo Zheng, Xiongwei Sun, Shaoqi Wang, Shizhuang Weng
Institute of Technology Innovation, Hefei Institutes of Physical Science, Chinese Academy of Sciences
Hefei Anhui, 230088, China
zhuzede@126.com, xhzeng@iim.ac.cn, zhshg1985@163.com, xiongweisun@gmail.com, wsq2012@mail.ustc.edu.cn, weng1989@mail.ustc.edu.cn

Abstract

Constructing bilingual lexicons from comparable corpora has been investigated in a two-stage process: building comparable corpora and mining bilingual lexicons, respectively. However, there are two potential challenges remaining, which are out-of-vocabulary words and different comparability degrees of corpora. To solve above problems, a novel iterative enhancement model is proposed for constructing comparable corpora and bilingual lexicons simultaneously under the assumption that both processes can be mutually reinforced. As compared to separate process, it is concluded that both simultaneous processes show better performance on different domain data sets via a small-volume general bilingual seed dictionary.

Keywords: comparable corpora, bilingual lexicons, mutual iterative enhancement, simultaneous construction

1. Introduction

Comparable corpora are selected as pairs of mono-lingual documents based on the criteria of content similarity, non-direct translation and language difference. With respect to parallel corpora, comparable corpora have the advantages in terms of more up-to-date, abundant and accessible (Ji et al., 2009). Furthermore, they are valuable resources for multilingual information processing, from which parallel sentences (Smith et al., 2010), parallel phrases (Munteanu and Marcu, 2006) and bilingual lexicons (Li and Gaussier, 2010; Prochasson and Fung, 2011) can be mined to reduce the sparseness of existing resources (Munteanu and Marcu, 2005; Snover et al., 2008).

Note that previous works of bilingual lexicons construction from comparable corpora consist of two stages separately: building comparable corpora and mining bilingual lexicons (Figure 1(a)). In the first stage, the automatic building of comparable corpora can be completed by focused crawling, cross-language information retrieval or 'inter-wiki' link. However, utilizing the comparability degree to build comparable corpora is still a significant challenging task. The degree of comparability is usually defined as the expectation of finding the translation of source language vocabularies in the target language documents. Therefore, most methods adopt statistical approach to map vocabularies in different languages by a bilingual seed dictionary.

In the second stage, the seminal works of mining bilingual lexicons from comparable corpora are based on the word co-occurrence hypothesis, in which the word and its translation share similar contexts. They assume the corpora are reliably comparable and focus on the improvement of extraction algorithms (Hazem et al., 2012), whereas successful detection of bilingual lexicons is severely influenced by the quality of corpora.

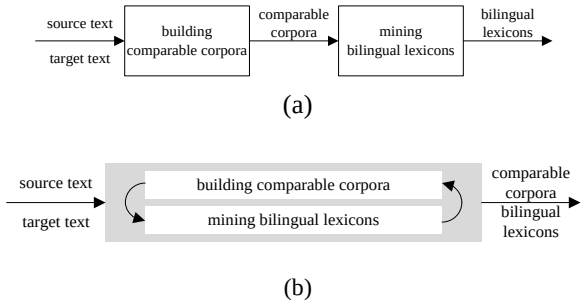


Figure 1: (a) Separate comparable corpora construction and bilingual lexicons construction and (b) joint comparable corpora construction and bilingual lexicons construction.

These two stages respectively suffer from different major challenges: Firstly, if the seed dictionary and the document set are less relevant in domain, out-of-vocabulary words will lower the quality of comparable corpora, especially domain-specific words. Secondly, if the comparable corpora have low comparability degrees, the quality of corpora may limit the performance of the bilingual lexicons construction. To address these potential problems, a novel iterative enhancement model is proposed to construct comparable corpora and bilingual lexicons simultaneously under the assumption that both processes can be mutually boosted (Figure 1(b)). The similar model has success in the domain of cross-domain sentiment classification (Wu et al., 2010).

Contributions Our contributions are as follows:

① A novel iterative enhancement model is presented to construct two different grained size levels bilingual resources simultaneously.

② A novel method of enriching domain-specific bilingual lexicons directly harvested from the candidate comparable corpora is proposed to enhance the ability of building comparable corpora.

Proceedings of the Ninth Workshop on Building and Using Comparable Corpora, pages 27–33,
LREC 2016, Portorož, Slovenia, May 23, 2016.

③ A novel method of calculating the relativity of cross-language lexicons on the basis of different comparability degrees of comparable corpora is proposed to enhance the ability of mining bilingual lexicons.

④ The model can be effectively applied in various domains, even though it relies on fewer existing resources such as a small-volume general bilingual seed dictionary.

The research hypothesis and motivation are just presented in this section. In the following section, we briefly summarize state-of-the-art approaches of the comparable corpora and bilingual lexicons construction. The iterative enhancement algorithm is described in detail in the "Proposed Model" section. Finally, we present our datasets, experiments and results before concluding the paper.

2. Related Work

2.1 Comparable Corpora Construction

The automatic acquisition of multilingual corpora can be completed by a variety of methods: focused crawling (Talvensaari et al., 2008), cross-language information retrieval (Huang et al., 2010) and 'intewiki' links (Smith, 2010). In fact, the measuring comparability degree of document pairs is still a challenging task to construct comparable corpora.

Recent measuring works mainly adopt statistical approach to map common vocabularies in different languages. To map lexical items, (Li and Gaussier, 2010) made use of a translation table and (Su and Babych, 2012) adopted a bilingual seed dictionary. (Saad et al., 2013) proposed two different comparability measures based on binary and cosines similarity measures using the bilingual dictionary to align words. Given a comparable corpora, (Li and Gaussier, 2010; Su and Babych, 2012) defined the degree of comparability as the expectation of finding the translation of any given source/target words in the target/source corpora vocabulary. In addition (Zhu et al, 2013) utilized the trained bilingual LDA model to calculate the comparability.

These approaches effectively evaluate the metric on the rich-resourced language pairs, thus quality bilingual resources are available. However, this is not the case for all domains in which reliable language resources such as bilingual dictionaries with broad word coverage might be not publicly available. To avoid the limit of existing resources, Tao and Zhai (2005) proposed a purely language-independent method to extract comparable bilingual text without the existing linguistic resources. They assumed that two words with mutual translation should have similar frequency correlation. The association between two documents was then calculated based on this information.

Nevertheless, the performance of the above method may be compromised due to the lack of linguistic knowledge, particularly corpora with low comparability. In this article, the problem can be circumvented by enriching a small general bilingual seed dictionary with a domain-specific bilingual lexicons harvested gradually from candidate comparable corpora to increase the dictionary coverage facing source and target texts.

2.2 Bilingual Lexicons Construction

The seminal works of extracting bilingual lexicons from comparable corpora are based on the word co-occurrence hypothesis, where the term and its translation share similar contexts (Fung, 1998; Rapp, 1999). More recent works usually assume that corpora are reliably comparable and focus on the improvement of extraction algorithms (Hazem et al., 2012). Therefore, less work is focused on the characteristics of comparable corpora (Maia, 2003). In fact, the degree of comparability has the greatly divergence between different corpora. Usually, successful detection of bilingual lexicons from comparable corpora depends on the quality of corpora, especially the degree of their textual equivalence and successful alignment on various text units.

To extract high-quality lexicons, the target and source texts should be highly comparable in a very specific subject domain. If one arbitrarily increases the size of the corpora, he actually takes the risk of decreasing its quality by adding out-of-domain texts. It has been proved that the quality of the corpora is more important than its size. Morin et al. (2007) showed that the discourse categorization of the documents increases the precision of the lexicons despite of the data sparsity. (Li and Gaussier, 2010; Li and Gaussier, 2011) improved the quality of the extracted lexicons when they improved the comparability of the corpora by selecting a smaller–but more comparable corpora from an initial set of documents. (Su and Babych, 2012) presented three different approaches to measure the comparability of cross-lingual comparable documents: a lexical mapping, a keyword and a machine translation approach. The results proved that higher comparability level consistently resulted in more number of parallel phrases extracted from comparable documents. Moreover, (Wang el at., 2014) adopted two step cross-comparisons between translation candidates to improve the quality.

Nevertheless, these methods couldn't effectively make use of comparable corpora of low comparability degree discarded directly. In this article, according to characterize the different comparability, the candidate comparable corpora is awarded different weight to extract good-quality bilingual lexicons from the corpora along with traditional context information.

3. Proposed Model

3.1 Basic Concepts Representation

The model is based on the assumption that the comparability of document pairs can promote the similarity of word pairs, and the similarity of word pairs can enhance the comparability of document pairs, which completes a mutual iterative enhancement model for simultaneous comparable corpora and bilingual lexicons

construction shown in Figure 2.

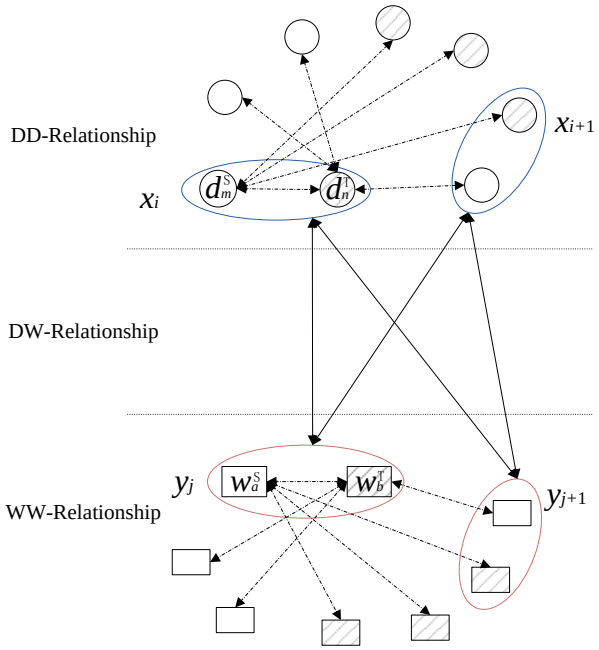


Figure 2: An illustration of the joint model between documents and words, where ○ (d_m^S) means source language document, ⊘ (d_n^T) means target language document, □ (w_a^S) means source language word, ▨ (w_b^T) means target language word, ⬭ (x_i) describes bilingual document pairs consisting of d_m^S and d_n^T, and ⬭ (y_j) describes bilingual word pairs consisting of w_a^S and w_b^T.

3.2 Relationship Formation

To establish the several relationships, the two basic functions have been proposed to measure:

$$\sigma(w^S, w^T) = \begin{cases} 1, if\ w^S\ is\ a\ \text{translation of}\ w^T \\ 0, otherwise \end{cases} \quad (1)$$

Where the function $\sigma(w^S, w^T)$ checks whether the translation of the word w^S in the source language document is equal to another word w^T in its corresponding target language document.

$$\delta(w, w') = \begin{cases} 1, if\ w\ and\ w'\ are\ equivalents \\ 0, otherwise \end{cases} \quad (2)$$

Where the function $\delta(w, w')$ checks whether two words w and w' are equivalence in the same language document.

Step 1: Calculating DD-Relationship

Given the source language document collection $D_S = \{d_m^S | 1 \le m \le \mathrm{M}\}$ (M represents the number of source language documents) and the target language document collection $D_T = \{d_n^T | 1 \le n \le \mathrm{N}\}$ (N represents the number of target language documents), the comparability degree R_{x_i} between multilingual document pairs x_i can be defined by the rate of translation between d_m^S and d_n^T, which is calculated by two bilingual unidirectional seed dictionaries. R_{x_i} is produced by the following formula:

$$R_{x_i} = P_{sim}(d_m^S, d_n^T) \quad (3)$$

$$= \frac{\#_w^{trans}(d_m^S, d_n^T)}{\#_w d_m^S} + \frac{\#_w^{trans}(d_n^T, d_m^S)}{\#_w d_n^T}$$

$$= \frac{1}{\#_w d_m^S} \sum_{w_e^S \in d_m^S} \sum_{w_f^T \in d_n^T} \sigma(w_e^S, w_f^T)$$

$$+ \frac{1}{\#_w d_n^T} \sum_{w_f^T \in d_n^T} \sum_{w_e^S \in d_m^S} \sigma(w_f^T, w_e^S)$$

Where $\#_w *$ is the number of words in the document *; $\#_w^{trans}(d_m^S, d_n^T)$ is the number of translation glossaries from d_m^S to d_n^T in different languages. If the comparability degree R_{x_i} exceeds the predefined threshold R_0, the cross-lingual document pair x_i forms an initial candidate multilingual comparable document pair whose weight is recorded as R_{x_i}.

Step 2: Calculating WW-Relationship

Given the source language word collection $W_S = \{w_a^S | 1 \le a \le \mathrm{A}\}$ (A represents the number of source language words) and the target language word collection $W_T = \{w_b^T | 1 \le b \le \mathrm{B}\}$ (B represents the number of target language words), the statistical relationship L_{y_j} between two words w_a^S and w_b^T can be calculated by the mutual information on the basis of the co-occurrence information. L_{y_j} between w_a^S and w_b^T is calculated as follows:

$$L_{y_j} = P_{co}(w_a^S, w_b^T) = \frac{2 \times \#(w_a^S, w_b^T)}{\#(w_a^S) \times \#(w_b^T)} \quad (4)$$

$$= \frac{2}{\left[\sum_{w_g^S \in D_S} \delta(w_a^S, w_g^S)\right] \times \left[\sum_{w_p^T \in D_T} \delta(w_b^T, w_p^T)\right]}$$

$$\times \sum_{x_i \in \mathrm{X}} min\left\{ \sum_{w_g^S \in d_m^S} \delta(w_a^S, w_g^S), \sum_{w_p^T \in d_n^T} \delta(w_b^T, w_p^T) \right\}$$

Which indicates the degree of statistical dependence between w_a^S and w_b^T. Here, $\#(w_a^S, w_b^T)$ is the number of w_a^S and w_b^T co-occurrence in all candidate comparable document pairs; $\#(w_a^S)$ and $\#(w_b^T)$ are respectively the frequencies of w_a^S and w_b^T in the document collection. If L_{y_j} exceeds the predefined threshold L_0, a cross-lingual word pair y_j is considered as an initial candidate bilingual lexicons pair whose weight is marked as L_{y_j}.

Step 3: Calculating DW-Relationship

Given the candidate bilingual document pairs collection $X = \{x_i | 1 \le i \le \mathrm{I}\}$ (I represents the number of the candidate bilingual document pairs) and the candidate bilingual word pairs collection $Y = \{y_j | 1 \le j \le \mathrm{J}\}$ (J represents the number of the candidate bilingual word pairs), a weighted bipartite relationship $H_{x_i y_j}$ between x_i and y_j can be calculated by the following formula when the word pair y_j appears in the document pair x_i.

$$H_{x_i y_j} = P_{rel}\left(R_{x_i}, L_{y_j}\right) = \frac{2 \times \#_{2w}^{co}\left(x_i, y_j\right)}{\#_w d_m^S + \#_w d_n^T} \quad (5)$$

$$= \frac{2}{\#_w d_m^S + \#_w d_n^T} \times min\left\{ \sum_{w_h^S \in d_m^S} \delta\left(w_a^S, w_h^S\right), \sum_{w_r^T \in d_n^T} \delta\left(w_b^T, w_r^T\right) \right\}$$

Where $\#_{2w}^{co}(x_i, y_j)$ is the number of the times that w_a^S and w_b^T co-occur in the document pair x_i. $H_{x_i y_j}$ can indicate the degree of statistical dependence between x_i and y_j.

3.3 Iterative Enhancement Algorithm

The core of the algorithm is to calculate the reasonable values of variables R_{x_i} and L_{y_j}. When the algorithm is carried out in the t^{th} iteration, the R_{x_i} and L_{y_j} are denoted as the $R_{x_i}^t$ and $L_{y_j}^t$ respectively. In order to calculate the values of $R_{x_i}^t$ and $L_{y_j}^t$, the iterative enhancement algorithm is mainly proposed on the basis of two basic assumptions as follows:

① If each document pair x_i in different languages contains more bilingual translation vocabularies, x_i should have a greater likelihood to construct comparable corpus;

② If each word pair y_j in different languages appears in the comparable corpora with high comparability degree, y_j should have a greater likelihood to construct bilingual lexicon.

According to the above assumptions, the change of $R_{x_i}^t$ is mainly dependent on $L_{y_j}^{t-1}$, and the change of $L_{y_j}^t$ is mainly dependent on $R_{x_i}^{t-1}$, where the initial values $R_{x_i}^0$ and $L_{y_j}^0$ respectively are calculating with formulas (3) and (4). When $R_{x_i}^t$ is greater than a predefined threshold R, x_i is a candidate comparable corpus. When $L_{y_j}^t$ is greater than a predefined threshold L, y_j is a candidate bilingual word pair. Finally, we can establish the following iterative forms:

$$R_{x_i}^t = \alpha R_{x_i}^{t-1} + \beta \sum_{j=1, L_{y_j}^{t-1} > L}^{J} H_{x_i y_j} L_{y_j}^{t-1} \quad (6)$$

$$L_{y_j}^t = \alpha L_{y_j}^{t-1} + \beta \sum_{i=1, R_{x_i}^{t-1} > R}^{I} (H_{y_j x_i} + \cos < \vec{C}_{w_a^S}, \vec{C}_{w_b^T} >) R_{x_i}^{t-1} \quad (7)$$

Where $\alpha + \beta = 1$, α and β specify the relative contributions to the final scores; The value of $H_{y_j x_i}$ is equal to the value of $H_{x_i y_j}$, which remains unchanged in the iterative process. $\vec{C}_{w_a^S}$ and $\vec{C}_{w_b^T}$ are respectively the context vectors of w_a^S and w_b^T. $\cos < \vec{C}_{w_a^S}, \vec{C}_{w_b^T} >$ is calculated by the standard approach (Fung, 1998; Rapp, 1999).

Finally, the convergence of the iteration algorithm is achieved when the difference of every document pair and word pair falls below a predefined threshold θ, which is formally expressed by the following two formulas: $\left|R_{x_i}^t - R_{x_i}^{t-1}\right| < \theta$ and $\left|L_{y_j}^t - L_{y_j}^{t-1}\right| < \theta$.

4. Experiments and analysis

In this section, several experiments are conducted to verify the effectiveness of this model. The initial thresholds are set as follows: $L_0 = 0.15$, $R_0 = 0.3$, $L = 0.1$, $R = 0.1$ and $\theta = 0.0001$, which are identified by the previous works.

Questions We try to answer the following questions:

① Does the joint model outperform conventional methods of building comparable corpora? (Section 4.1)

② How about the quality of lexicons by the joint model of mining bilingual lexicons? (Section 4.2)

4.1 Comparable Corpora Evaluation

4.1.1. Evaluation Measures

As there is no commonly available data set to evaluate the comparability degree of comparable corpora and then mine bilingual lexicons, we collect our own gold standard comparable corpora as test datasets. They specialize on three different domains on *culture*, *economy* and *sport*, which include 50 English-Chinese bilingual document pairs respectively. The datasets are normalized through the following linguistic preprocessing steps: tokenization, part-of-speech tagging, lemmatization and function word removal. In addition, a small-volume general bilingual seed dictionary is applied which contains 42,373 distinct common entries.

The datasets are acquired by two main steps. Firstly, the initial data are acquired by adopting the focused crawling for automatic acquisition of topic-specific source language web and utilizing interlinks between pages to collect target language web. This method can quickly locate a relative specific domain including 500 page pairs. Secondly, we manually annotate the document pairs on the basis of five comparability levels as gold standard to assess the alignments. Five levels proposed by (Fung P. 1998) are refined the alignments as follows: Same Story, Related Story, Shared Aspect, Common Terminology and Unrelated. Finally, we select 50 document pairs in every domain with Same Story and Related Story as comparable corpora.

We adopt the *Precision* as evaluation metric:

$$Precision = \left|C_p \cap C_l\right| / \left|C_p\right| \quad (8)$$

Where C_p represent the comparable corpora in the automatic building results; C_l represent the comparable corpora in the labeled results; $|*|$ means the number of document pairs in the corpora $*$.

4.1.2. Results and Analysis

We set two parameters α=0.5 and β=0.5 according to the conclusion of the 'Group 1' in the 4.2.2 subsection. Then

we compare the performance of the joint model with the current representative approach (shown in Table 1).

domain		*culture*	*economy*	*sport*
This paper	**No-iterative**	45	57	49
	Iterative	64	83	69
	Value of improvement	↑ 19	↑ 26	↑ 20
Zhu et al. (2013)		58	77	67

Table 1: Performance (%) of the *Precision* for different domains and existing method.

Overall, the results indicate the robustness and effectiveness of the model. It is concluded that the model can be effectively applied to different domains even through external resources is under adverse conditions that the seed dictionary is a small-volume general bilingual dictionary. In every specific domain, the results reliably depend on the correlation of cross-language document pairs in the datasets. Simultaneously, with respect to the no-iterative process, the performance of the iterative enhancement significantly improves up to 26%. In addition, the scores of this paper outperform the algorithm implemented by (Zhu et al, 2013), which adopts the trained bilingual LDA model to predict the topical structures and calculates the similarity of the documents in different languages. The high quality results of the joint model are due to the fact that out-of-vocabulary words are sufficiently solved in this paper.

4.2 Bilingual Lexicons Evaluation

4.2.1. Evaluation Measures

Automatic evaluation of bilingual lexicons extraction is performed against a gold standard lexicons *G*, which is obtained from the top-ranking nouns or verbs in the gold standard comparable corpora. These lexical items should only appear in a domain bilingual dictionary and be not included in the seed dictionary that is a small-volume general bilingual dictionary. *G* contains 100 Chinese single-word terms with their corresponding English translations. When more than one translation variant are possible for a single English term, each proposed by the model is considered as correct result.

We adopt the *Accuracy* as evaluation metric in bilingual lexicons extraction, which reflects precision among first *K* translation candidates. And the *Accuracy* is calculated in the following equation:

$$Accuray = count_{top\,K}/H \quad (9)$$

Where *H* means the number of the gold standard entries in *G*; $count_{top\,K}$ means all the number of correct translation in top *K* ranking. In this paper, *K* ranges from 1^{th} to 20^{th} ranking.

4.2.2. Results and Analysis

Group 1: Parameter β

In order to better grip the relative contributions from the document x_i and the word y_j, table 2 shows the score with respect to the parameter β in the entire corpora collection and β ranges from 0 to 1 by 0.1 as a step length.

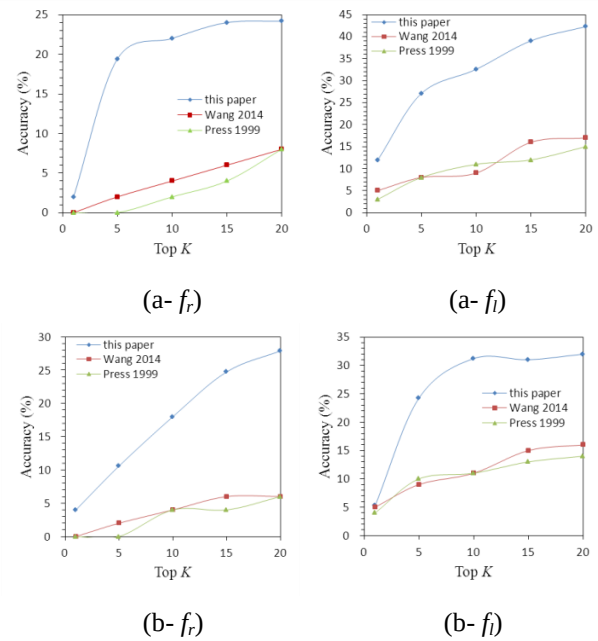
K\β	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
1	0	2	4	6	7	8	7	8	6	3	3
5	1	7	9	16	17	18	19	17	15	11	8
10	2	9	15	17	19	20	21	21	19	14	10
15	2	11	16	19	23	21	20	23	18	15	12
20	4	12	17	20	23	22	22	23	21	15	13
Total	9	41	61	78	89	**91**	90	90	79	58	46

Table 2: Performance (%) of the *Accuray* with value of varied *K* from 1 to 20 by 5 as a step length.

The table 2 shows that the parameter β has a remarkable impact on the performance of the model. When the value of β is set as 0.4 or 0.6, the *Accuracy* mostly achieves the peak with each value of *K*. If β becomes large enough (near to 1) or very small (near to 0), the *Accuracy* sharply falls into decline. These results demonstrate that both documents and words are very important contributions to rank comparable corpora. The loss of each element will greatly deteriorate the final performance. The total of *Accuracy*, which shows the overall performance of the algorithm with all values of *K*, arrives the best performance under the condition of β =0.5. So the optimal β is set to 0.5 in the subsequent experiments according to the analysis of influence.

Group 2: Existing Methods Comparison

In order to verify the excellence of the model in the paper, we make use of all the document pairs as test dataset. Then we compare the performance of our model with the other two existing representative approaches: one is proposed by (Press, 1999) which reflects a baseline level, the other one is proposed by (Wang el at., 2014) which represents the current state of the art (Shown in Figure 3).



(a- f_r) (a- f_l)

(b- f_r) (b- f_l)

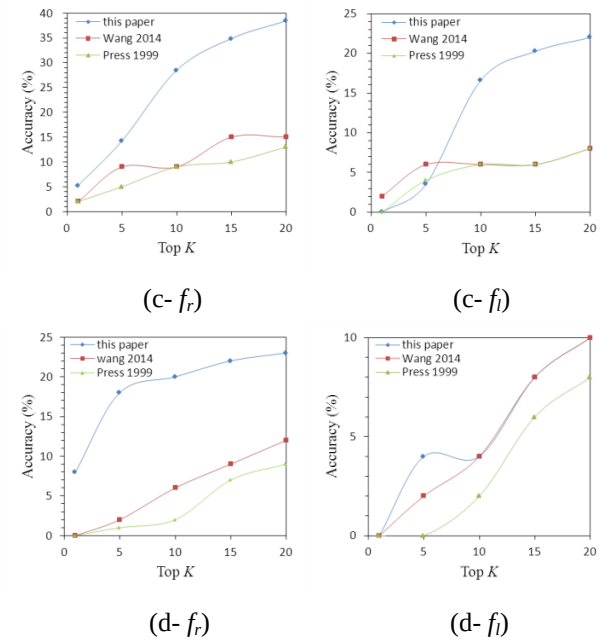


Figure 3: Performance of bilingual lexicons construction from different the methods with varied *K* values from 1 to 20. (a) culture, (b) economy, (c) sport, (d) mixed: containing three domains. (f_r) random frequency words, (f_l) low frequency words.

The figure 3 shows that the score obtained by this paper practically outperforms the other two approaches in three different domains regardless of word frequency, which indicates that iterative enhancement model is valid to construct bilingual lexicons. (Press, 1998) extracts bilingual lexicons in the view of the context information. (Wang el at., 2014) adopts two step cross-comparisons between translation candidates of each target word to improve the quality of bilingual lexicons. But the correlation between vocabularies completely depends on the coverage of seed dictionary and the comparability of the document pairs are utilized as equalization. When the dictionary cannot cover the most of glossaries in the corpora due to different domains, the method will loss the advantage.

The model proposed in the paper not only can distinguish the comparability degree of different document pairs to mine bilingual lexicons, but also utilize domain-specific bilingual lexicons producing in this process to calculate the comparability degree, which are continuous iteration and mutually reinforced. Only when low frequency bilingual lexicons are extracted from the mixed corpora, does the model proposed by this paper have almost equivalent performance with the method put forward by (Wang el at., 2014) shown in figure 3 (d- f_l). The main reason is that the mixed corpora have great differences of the domain knowledge, which lead to a very small promotion in the iterative process, especially when the target bilingual lexicons are the low frequency vocabularies.

5. Conclusions

Previous works on bilingual lexicons construction from comparable corpora are completed by two independent tasks. In this paper, we propose a simultaneous comparable corpora and bilingual dictionary construction method based on a mutual iterative enhancement model. Our evaluation shows the simultaneous construction approach improves the accuracy of the outcome comparable corpora and bilingual dictionary via a small-volume general bilingual seed dictionary. In addition, based on the encouraging results, we are going to explore more other sizes of bilingual resources simultaneously, such as bilingual parallel sentences and bilingual multi-word expressions.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of Anhui No.1608085QF127 and No. 1508085QC65, the Open Projects Program of National Laboratory of Pattern Recognition No.201306320, the National Natural Science Foundation of China No.31401285 and No. 61475163, China Postdoctoral Science Foundation No.2015M570548.

7. References

Ji H. (2009). Mining name translations from comparable cor-pora by creating bilingual information networks. In *Proceedings of BUCC*, pp. 34--37.

Jason Smith, Chris Quirk and Kristina Toutanova. (2010). Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Proceedings of NAACL*, pp. 403--411.

Dragos Munteanu and Daniel Marcu. (2006). Extracting parallel sub-sentential fragments from nonparallel corpora. In *Proceedings of ACL*, pp. 81--88.

Li B., and Gaussier E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 644--652.

Emmanuel Prochasson and Pascale Fung. (2011). Rare Word Translation Extraction from Aligned Comparable Documents. In *Proceedings of ACL-HLT*, pp. 1327--1335.

Dragos Stefan Munteanu and Daniel Marcu. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4), pp. 477--504.

Matthew G. Snover, Bonnie J. Dorr, and Richard M. Schwartz. (2008). Language and translation model adaptation using comparable corpora. In *Proceedings of EMNLP*, pp. 857--866.

Hazem A. and Morin E. (2012). Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 288--292.

Wu Q, Tan S, Cheng X, et al. (2010). MIEA: a mutual iterative enhancement approach for cross-domain sentiment classification. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 1327--1335.

Tuomas Talvensaari, Ari Pirkola, Kalervo Järvelin, Martti Juhola, and Jorma Laurikkala. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5), pp. 427--445.

Huang D. G., Zhao L., Li L. S., et al. (2010). Mining Large-scale Comparable Corpora from Chinese-English News Collections. In *Proceedings of Coling*, pp. 472--480.

Su F., Babych B. (2012). Measuring Comparability of Documents in Non-Parallel Corpora for Efficient Extraction of (Semi-) Parallel Translation Equivalents. In *Proceedings of the EACL*, pp. 10--19.

Motaz Saad, David Langlois, and Kamel Smaïli. (2013). Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia-Social and Behavioral Sciences*, 95(4), pp. 40--47.

Zhu Z., Li M., et al. (2013). Building Comparable Corpora Based on Bilingual LDA Model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 278--282.

Tao Tao, Chengxiang Zhai. (2005). Mining comparable bilingual text corpora for cross-language information integration. In *Proceedings of ACM SIGKDD*, pp. 691--696.

Fung Press. (1998). A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pp. 1--17.

Rapp R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pp. 519--526.

Maia B. (2003). Some languages are more equal than others. Training translators in terminology and information retrieval using comparable and parallel corpora. *Corpora in Translator Education*, pp. 43--53.

Morin E., Daille B., Takeuchi K., et al. (2007). Bilingual terminology mining-using brain, not brawn comparable corpora. In *Proceedings of the annual meeting of the Association for Computational Linguistics*, 45(1): 664--671.

Li B., Gaussier E., Aizawa A. (2011). Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers-Volume 2*, pp. 473--478.

Wang Shaoqi, Li Miao; et al. (2014). Improvement of Bilingual Lexicon Extraction Performance From Comparable Corpora via Optimizing Translation Candidate Lists. In *Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp. 18--25.