Improving Bilingual Lexicon Extraction Performance from Comparable Corpora via Optimizing Translation Candidate Lists

Shaoqi Wang
University of Science and Technology of China, Institute of Intelligent Machines
Chinese Academy of Sciences
Hefei, China
wsq2012@mail.ustc.edu.cn

Miao Li, Zede Zhu, Zhenxin Yang, Shizhuang Weng
Institute of Intelligent Machines Chinese Academy of Sciences
Hefei, China
mli@iim.ac.cn,
zhuzede@mail.ustc.edu.cn,
xinzyang@mail.ustc.edu.cn,
weng1989@mail.ustc.edu.cn

Abstract

In this paper, we propose a novel method to optimize translation candidate lists derived from window-based approach for the task of bilingual lexicon extraction. The optimizing process consists of two cross-comparisons between 1th translation candidate of each target word, and between set of all the 1th candidates and that of each word's 2th to *N*th ones. Experiment results demonstrate that the proposed method leads to a significant improvement on *accuracy* over window-based approach in bilingual lexicon extraction from both English-Chinese and Chinese-English comparable corpora.

1 Introduction

Bilingual lexicon is a basic resource in the field of Natural Language Processing such as machine translation and cross-language information retrieval (AbduI-Rauf et al., 2009). Parallel corpora (Och and Ney, 2000) are typically applied to automatically extracting bilingual lexicon with high precision, but they are difficult to obtain in several domains. Due to the high cost of acquiring parallel corpora, comparable corpora, which consist of sets of documents in different languages dealing with a given topic or domain and are much easier to collect from the increasingly rich web data (Xiao and McEnery, 2006), become an alternative resource to the task. Based on comparable corpora, researchers begin to use a variety of approaches to exploit them for bilingual lexicon extraction in recent years (Tanaka and Iwasaki, 1996; Fung and McKeown, 1997; Fung and Yee, 1998; Rapp, 1999; Morin et al., 2007; Saralegui et al., 2008; Kun Yu, Junichi Tsujii, 2009). These approaches mainly share a standard strategy based on the assumption that a word and its translation appear in similar context.

These previous work shows that equivalent extraction from comparable corpora is unstable on all but the most frequent words. An explanation for the phenomenon is that translation candidate lists of target words, coming from matrix of context similarities, are always disturbed by lots of noises introduced by many-to-many mapping between the contexts of words in different languages and only more frequent ones keep comparatively robust (Pekar et al., 2006).

Regardless of the polysemy, in the candidate list of a certain target word, there may be only one correct candidate and the rest ones can be regarded as noises. Moreover, the correct candidate of one target word may become the noise in the candidate list of another target one. Therefore, to retain the correct candidate in one list and remove it (viewed as noise) from others' list when it appears, comparison between candidates in each list need to be done.

In this paper, we propose a novel method to remove these noises via optimizing translation candidate lists. The optimizing process is on the basis of cross-comparison which means comparison object lies on different candidate lists. Firstly, we adopt window-based approach to acquire translation candidate lists (Rapp, 1999; Chiao and Zweigenbaum, 2002). Then, we use the proposed two cross-comparisons of similarity. The first one called identical ranking

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 18–25,
Wuhan, China, 20-21 October 2014

cross-comparison is the comparison between 1^{th} translation candidate of each target word. The second named distinct ranking cross-comparison is the comparison between set of all the 1^{th} candidates and that of each word's 2^{th} to N^{th} ones. Finally, we conduct the experiments to find target words with different frequencies from both Chinese-English and English-Chinese.

The organization of the paper is as follows: Related work is presented in Section 2. Section 3 is devoted to the introduction of window-based approach. In Section 4, we present the proposed optimizing process. In Section 5 we describe the experimental setup and report the results of bilingual lexicon extraction. Section 6 summarizes the paper with a final conclusion.

2 Related work

Previous work about bilingual lexicon extraction from comparable corpora usually focused on utilizing context similarity. Fung (1995) firstly used context heterogeneity in the task. Subsequently, context vectors were modeled and similarities between source-language and target-language contexts were measured with the aid of a general dictionary by many researchers (Fung, 2000; Chiao and Zweigenbaum, 2002; Robitaille et al., 2006; Morin et al., 2007).

The approaches based on context vectors differ in the way they defined word contexts. Window-based approach uses the window of the compared word to construct context (Rapp, 1999; Chiao and Zweigenbaum, 2002; Dejean et al., 2002; Gamallo, 2007). Apart from that, Syntax-based approach utilizes syntactic information for bilingual dictionary extraction (Otero, 2007).

The above approaches simply yield candidates according to the calculation of vector similarity without any subsequent processing. The proposed method can be viewed as the extension of window-based approach. Different from previous work, we emphasize the optimizing process of translation candidate lists.

3 Window-Based Approach

In window-based approach, some windows of words are firstly considered as forming the context vectors. The approach then translates source words' context vectors by using a general bilingual dictionary, and calculates the similarity between each source and target vector.

3.1 Building Context Vectors

In this step, we first choose a window size β and get β number words from both left and right of every source word w_s in corpora to form the source context information set $I_{w_s} = \{w_{s_1}, \cdots w_{s_{N_s}}\}$. Similarly, we acquire the target context information set $I_{w_t} = \{w_{t_1}, \cdots w_{t_{N_t}}\}$ of target word w_t, where N_s and N_t means the number of words in I_{w_s} and I_{w_t}. The weight $W(w_s, w_{s_k})$ of word $w_{s_k} (1 \le k \le N_s)$, which is represented as follows, is calculated on the basis of mutual information.

$$W(w_s, w_{s_k}) = \ln \frac{count(w_s, w_{s_k})}{count(w_s) \times count(w_{s_k})}. \quad (1)$$

Where $count(w_s, w_{s_k})$ is the number of co-occurrence between w_s and w_{s_k} in all the contexts. $count(w_s)$ and $count(w_{s_k})$ take as values the number of occurrence of w_s and w_{s_k}. We compute weights of every word $w_{s_k} (1 \le k \le N_s)$ in I_{w_s} to form the source context vector $\overrightarrow{V_{w_s}}$. Similar method is adopted to transfer I_{w_t} to the target context vector $\overrightarrow{V_{w_t}}$.

3.2 Vector Similarity

Using a general bilingual dictionary, we map the I_{w_s} into the target language context information $I_{w_s}^{trans}$ whose corresponding context vector is $\overrightarrow{V_{w_s}^{trans}}$: If k^{th} component in I_{w_t} equals to g^{th} component in $I_{w_s}^{trans}$ ($1 \le k \le N_s$, $1 \le g \le N_t$), we assign the value of g^{th} component in $\overrightarrow{V_{w_s}}$ to k^{th} component in $\overrightarrow{V_{w_s}^{trans}}$; if there is no equal word, the value is zero.

By calculating $\overrightarrow{V_{w_s}^{trans}}$ of each w_s and $\overrightarrow{V_{w_t}}$ of each w_t, we create a vector matrix, where rows correspond to $\overrightarrow{V_{w_t}}$, columns to $\overrightarrow{V_{w_s}^{trans}}$ and cells to similarities between each vectors. Finally, we adopt the cosine measure (see equation 2) to calculate the similarities in the matrix and further rank them to generate translation candidate lists.

$$Sim(\overrightarrow{V_{w_t}}, \overrightarrow{V_{w_s}^{trans}}) = \frac{\sum_j v_j v_j^{trans}}{\sqrt{\sum_j (v_j)^2}\sqrt{\sum_j (v_j^{trans})^2}} . \quad (2)$$

where v_j and v_j^{trans} is the component of vector $\overrightarrow{V_{w_t}}$ and $\overrightarrow{V_{w_s}^{trans}}$ respectively.

4 Optimizing Translation Candidate Lists

We take into account top *N* ranking translation candidates in the total *M* lists, where *M* means the number of target words and *N* means the lowest ranking considered in the section, and optimize them with two cross-comparisons of similarity between each candidate. The optimizing process consists of 2 steps: identical ranking cross-comparison between each first 1^{th} candidate; distinct ranking cross-comparison between all the 1^{th} candidates and each word's 2^{th} to N^{th} ones. The architecture of our method is described in Fig.1.

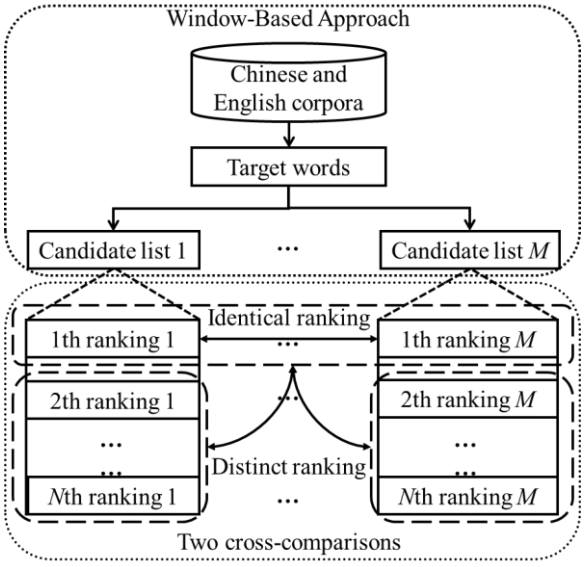

Figure1: Architecture of the proposed method

4.1 Identical Ranking Cross-comparison

Identical ranking cross-comparison relies on the assumption that each target word's 1^{th} candidate is unique. When there are two words having the same 1^{th} candidate, we regard the one with higher similarity as potential correct translation and remove another one defined as noise. This step is presented as follows:

Step1. Choose all the target words' first top ranking candidates $(T^1_{w_{t_1}}, \cdots, T^1_{w_{t_M}})$ and extract their similarities $(Sim^1_{w_{t_1}}, \cdots, Sim^1_{w_{t_M}})$.

Step2. Scan $(T^1_{w_{t_1}}, \cdots, T^1_{w_{t_M}})$. If there exists several equal candidates $(T^1_{w_{t_a}}, T^1_{w_{t_b}}, T^1_{w_{t_c}} \cdots)$ ($1 \le a, b, c \le M$), jump to Step3. If all the candidates are different, go to Step4.

Step3. Compare the corresponding similarities $(Sim^1_{w_{t_a}}, Sim^1_{w_{t_b}}, Sim^1_{w_{t_c}} \cdots)$. Retain candidate with the highest value and remove others. Jump to Step2.

Step4. Complete identical ranking cross-comparison.

4.2 Distinct Ranking Cross-comparison

In light of hypothesis that all target words' 1^{th} candidates are regarded as optimal translations, the main idea of distinct ranking cross-comparison is that these 1^{th} candidates are assumed as noises when they appear in each word's 2^{th} to N^{th} ones with higher similarities. The following describes this step:

Step1. build a noise set $(T^1_{w_{t_1}}, \cdots, T^1_{w_{t_M}})$.

Step2. use the noise set to scan rest candidates $(T^2_{w_{t_n}}, \cdots, T^N_{w_{t_n}})$ of w_{t_n} (*n* ranging from 1 to *M*).

Step3. when $T^j_{w_{t_n}}$ ($2 \le j \le N$) equals to any element $T^1_{w_{t_m}}$ ($2 \le m \le N$) in the noise set, remove $T^j_{w_{t_n}}$ if $Sim^1_{w_{t_m}}$ is higher than $Sim^j_{w_{t_n}}$.

4.2 Algorithm Description and Illustration

This part detailedly introduces the proposed method by means of algorithm description. After the description, we illustrate our method with a specific example. Algorithm 1 depicts the identical ranking cross-comparison as follows:

Algorithm 1

Input:

Target words' number *M*, Lowest ranking *N*

Unranked Candidate lists from L_1 to L_M

Unranked similarity lists from S_1 to S_M

Output:

New-ranking candidate lists from L_1^{rank} to L_M^{rank}

1: **for** *i*=1 to *M* **do**

2: rank Candidate list *i*:

3: $L_i \rightarrow L_i^{rank} : (T^1_{w_{t_i}}, \cdots, T^N_{w_{t_i}}, \cdots)$

4: $S_i \rightarrow S_i^{rank} : (Sim^1_{w_{t_i}}, \cdots, Sim^N_{w_{t_i}}, \cdots)$

5: end **for**

6: scan $(T^1_{w_{t_1}}, \cdots, T^1_{w_{t_M}})$

7: **while** equal candidates exist **do**
8: build Set^{i}_{equ}, several sets consist of equal candidates: $(T^{1}_{w_{t_a}}, T^{1}_{w_{t_c}} \cdots), (T^{1}_{w_{t_p}}, T^{1}_{w_{t_q}} \cdots) \cdots$ $1 \le a, c, p, q \le M$
9: build $SimSet^{i}_{equ}$: corresponding similarity sets
10: *Max* = sum of Set^{i}_{equ}, *i* ranging from 1 to *Max*
11: **for** *i*=1 to *Max* **do**
12: scan Set^{i}_{equ} and $SimSet^{i}_{equ}$
13: find the highest similarity: $Sim^{1}_{w_{t_h}}$
14: other $Sim^{1}_{w_{t_x}}$ =0; $1 \le x \le M$, $x \neq h$
15: **end for**
16: re-rank lists, scan $(T^{1}_{w_{t_1}}, T^{1}_{w_{t_2}}, \cdots, T^{1}_{w_{t_M}})$
17: **end while**
18: **return** all the candidate lists

The following Algorithm 2 realizes the distinct ranking cross-comparison.

Algorithm 2

Input:
Target words' number *M*
Lowest ranking *N*
Ranked Candidate lists from L^{rank}_{1} to L^{rank}_{M}
Ranked similarity lists from S^{rank}_{1} to S^{rank}_{M}

Output:
New-ranking candidate lists from L^{rank}_{1} to L^{rank}_{M}

1: **for** *i*=1 to *M* **do**
2: **for** *j*=1 to *M* **do**
3: **for** *k*=2 to *N* **do**
4: **if** $T^{k}_{w_{t_i}} = T^{1}_{w_{t_j}}$ & $Sim^{k}_{w_{t_i}} < Sim^{1}_{w_{t_j}}$ **then**
5: $Sim^{k}_{w_{t_i}}$ =0;
6: **end if**
7: **end for**
8: **end for**
9: re-rank candidate list L^{rank}_{i}
10: **end for**
11: **return** all the candidate lists

For example, following the above algorithm, we get sorted candidate lists (see Tab.1). In identical ranking cross-comparison, we scan all the 1th candidates in each list (see red square in Tab.1) and find two sets of equal candidates: ('market/0.6162', 'market/0.6097') and ('economics/0.5627', 'economics/0.6492') (see black square in Tab.1) . Through the comparison of similarity, the 'market/0.6097' and 'economics/0.5627' become 'market/0' and 'economics/0'. Then we re-rank the lists and scan again, finding that each 1th candidate is unique. So Algorithm 1 is finished. Tab. 2 shows the re-ranking lists after identical ranking cross-comparison.

In distinct ranking cross-comparison, we build a noise set ('market/0.6162', 'theory/0.6012', 'art/0.4982', 'economics/6492', 'human/0.5627') (see red square in Tab.2) to scan each list's 2th to Nth candidates. Taking the list of word '教育' as example, we first use the noise set to scan the remaining candidates ('economics/0.5220', 'theory/0.5136','education/0.5112','art/0.5078', ⋯) (see black square in Tab.2) , and then find that 'economics', 'art' and 'theory' exist in the noise set. So we compare the similarity between 'economics/0.6492' and 'economics/0.5220', 'theory/0.6012' and 'theory/0.5136', and 'art/0.4982' and 'art/0.5078'. Thus, 'economics/0.5220' and 'theory/0.5136' with lower value are turned into 'economics/0' and 'theory/0'. Afterwards, we re-rank this list. Tab. 3 presents the finally optimized lists. Correct translations in Tab.1 to Tab.3 are highlighted in bold.

Word	**Candidate/Similarity lists**					
	1	2	3	4	5	…
市场	**market 0.6162**	theory 0.5953	art 0.5837	education 0.5716	human 0.5330	…
理论	market 0.6097	**theory 0.6012**	human 0.5930	family 0.5527	education 0.5326	…
艺术	economics 0.5627	**art 0.4982**	economy 0.4817	job 0.4721	human 0.4330	…
经济学	**economics 0.6492**	market 0.5198	art/ 0.5038	education/ 0.4786	state 0.4687	…
教育	human 0.5407	economics 0.5220	theory 0.5136	**education 0.5112**	art 0.5078	…

Table 1: Ranked lists from window-based approach

Word	Candidate/Similarity lists					
	1	2	3	4	5	…
市场	**market** **0.6162**	theory 0.5953	art 0.5837	education 0.5716	human 0.5330	…
理论	**theory** **0.6012**	human 0.5930	family 0.5527	education 0.5326	nature 0.5008	…
艺术	**art** **0.4982**	economy 0.4817	job 0.4721	human 0.4330	market 0.4291	…
经济学	**economics** **0.6492**	market 0.5198	art 0.5038	education 0.4786	state 0.4687	…
教育	human 0.5407	economics 0.5220	theory 0.5136	**education** **0.5112**	art 0.5078	…

Table 2: Lists after identical ranking cross-comparison

Word	Candidate/Similarity lists					
	1	2	3	4	5	…
市场	**market** **0.6162**	art 0.5837	education 0.5716	job 0.5116	book 0.4930	…
理论	**theory** **0.6012**	human 0.5930	family 0.5527	education 0.5326	nature 0.5008	…
艺术	**art** **0.4982**	economy 0.4817	job 0.4721	book 0.4121	physics 0.4052	…
经济学	**economics** **0.6492**	art 0.5038	education 0.4786	state 0.4687	application 0.4528	…
教育	human 0.5407	**education** **0.5112**	art 0.5078	job 0.4992	state 0.4791	…

Table 3: Final optimized lists

5 Experiments and Analysis

5.1 Experiment Datasets and Setup

We conduct experiments on a Chinese-English corpora derived from the data used in bilingual Wikipedia with 3254 comparable document pairs. The general bilingual dictionary is constructed from an online dictionary which contains 42,373 distinct entries. In addition, we perform the following linguistic preprocessing steps on the comparable corpora: tokenization, lemmatization and removing stop words. After these steps the corpora contain ca. 925,000 Chinese words, and ca. 785,000 English words. The windows size β in building the context vectors is defined as 5, and different sizes are assessed and the above setting turns out to have the best performance in window-based method.

Two experiments are performed on target words with random frequency distribution and certain frequency in order to evaluate the proposed method. During each experiment we also absorb in the extraction performance from both English-Chinese and Chinese-English. The baseline in our experiments is the window-based approach without any optimizing, and we successively use two cross-comparisons in the proposed method and focus on performance respectively.

5.2 Evaluation Metric

We adopt the *accuracy* as evaluation metric. *Accuracy*, which means precision among the top *n* ranking, is a common metric in bilingual lexicon extraction. In this paper, translation candidates in lists from 1^{th} to 20^{th} ranking are kept for automatic and manual evaluation of *accuracy*, and score of *accuracy* is calculated in the following equation:

$$Accuracy = \frac{count_{top_n}}{M}. \qquad (3)$$

Where *n* means top *n* evaluation (*n* ranging from 1 to 20), *M* means the number of target words and $count_{top_n}$ means the number of correct translation in top *n* ranking.

5.3 Results and analysis

Experiment 1: target words with random frequency distribution

When we extract bilingual lexicon from English-Chinese, 1000 (*M*=1000) target words from the Chinese documents are randomly chose. We calculate the vector similarities between these Chinese words and all the English words to generate translation candidate lists, and then optimize them via the proposed method.

Meanwhile, we conduct the experiment of finding translations of 1000 target words from English documents. N in this experiment is assign as 1020. Fig. 2 and Fig. 3 demonstrate the resulting *accuracy* of different methods from two directions.

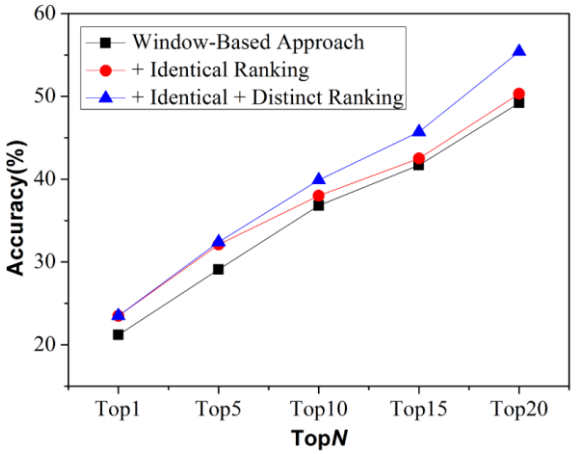

Figure 2: Extraction Results of different methods from English-Chinese

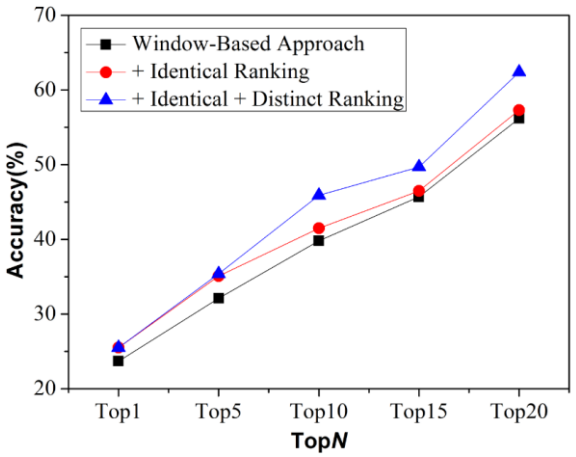

Figure 3: Extraction Results from Chinese-English

The results show that *accuracy* is improved significantly from both English-Chinese and Chinese-English, thereby indicate the robustness and effectiveness of our method. In particular, two steps in the proposed method can gradually improve the *accuracy*. Improvements of *accuracy* in top1 and top5 are mainly attributed to identical ranking cross-comparison as it processes candidate lists' top-ranking area. Distinct ranking cross-comparison can markedly boost *accuracy* in top10, top15 and top20, since it removes noises in larger area of the lists.

Experiment 2: target words with certain frequency

Previous work showed that frequent words' correct translations are easier to be found than infrequent ones (Pekar et al., 2006). Allowing for this fact, we distinguish different frequency ranges to assess the validity of the proposed approach. Target words with frequency more than 400 are defined as high-frequency words (W_H), whereas words with frequency less than 100 are low-frequency words (W_L). The number of target words from either Chinese or English documents is 1000 (M=1000) and N equals to 1020. Extraction performance on *accuracy* beyond W_H and W_L are showed in Fig. 4, Fig. 5, Fig. 6 and Fig. 7.

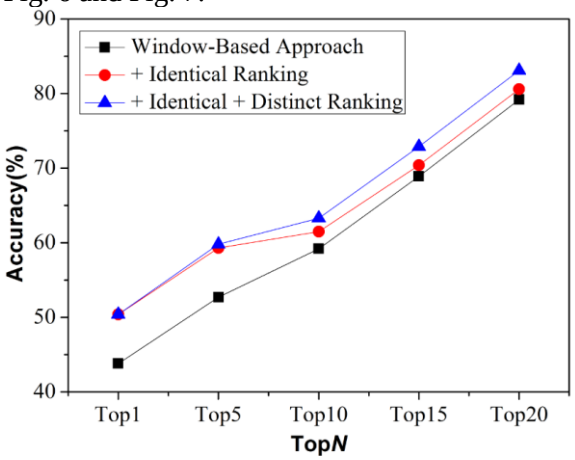

Figure 4: Extraction Results of W_H from English-Chinese

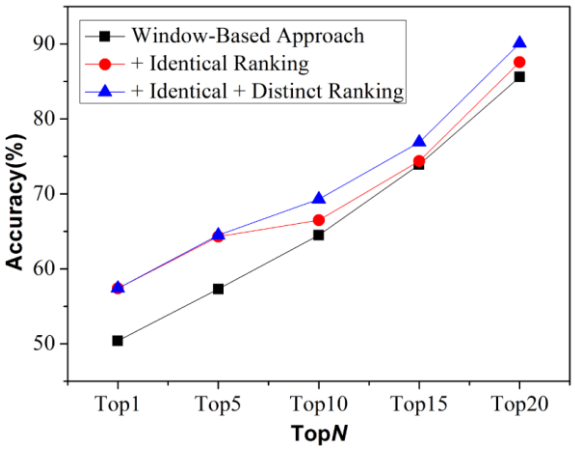

Figure 5: Extraction Results of W_H from Chinese-English

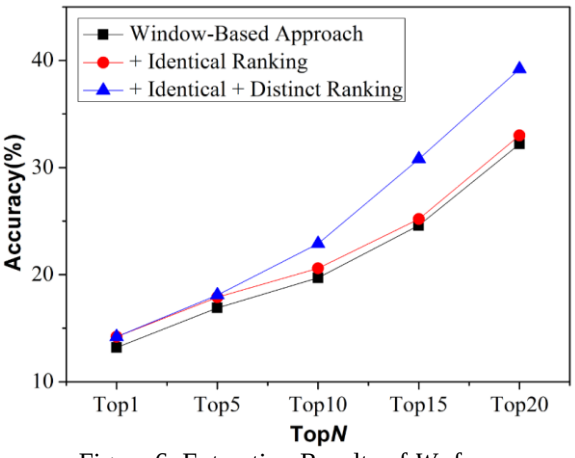

Figure 6: Extraction Results of W_L from English-Chinese

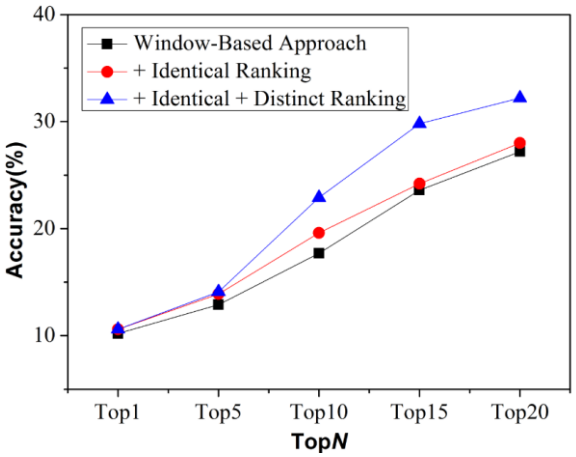

Figure 7: Extraction Results of W_L from Chinese-English

From Fig. 4 and Fig. 5 we have the following observation: *accuracy* improvement effect of identical ranking cross-comparison in top1 and top5 becomes more obvious in the process on W_H. In addition, Fig. 6 and Fig. 7 indicate that in processing W_L distinct ranking cross-comparison promotes *accuracy* in top10, top15 and top20 to a larger extent. The main reason is that for W_H each word's correct translation, which is also high-frequency source word, happens to be noise existing in top-ranking area of other words' lists. This situation leads to increasing number of identical ranking cross-comparison which can eliminate noises more effectively. Meanwhile, for W_L noises in each target word's translation candidate lists are all high-frequency source words, leading high repetition rate between the noises set and top N candidates in the lists. Therefore, distinct ranking cross-comparison can boost most optimal translations which locate in lower ranking before to concentrate in the area between 5^{th} and 20^{th} ranking.

6 Conclusion

In this paper, we address the 'noise' problem in extracting translation equivalent from comparable corpora. To solve the problem, we develop a novel method to optimize translation candidate lists. The optimizing process includes two step cross-comparisons between translation candidate of each target word. Experimental results show that the proposed method can boost *accuracy* significantly and outperform window-based approach in bilingual lexicon extraction from both English-Chinese and Chinese-English. Moreover, identical ranking and distinct ranking cross-comparison can improve the *accuracy* respectively in different ranking area, and their improvements depend on the frequency of target words. Future work may focuses on conducting experiment between the proposed method and syntax-based approach, and eliminating our method's impact on synonyms.

Acknowledgements

The work is supported by the Informationization Special Projects of Chinese Academy of Science under No. XXH12504-1-10 and the Open Projects Program of National Laboratory of Pattern Recognition.

Reference

AbduI-Rauf S, Schwenk H. On the use of comparable corpora to improve SMT performance[C] //Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 16-23.

Chiao Y-C, Zweigenbaum P. Looking for candidate translational equivalents in specialized, comparable corpora[C] //Proceedings of COLING. 2002.

Dejean H, Gaussier E, Sadat F. Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora[C] //Proceedings of COLING, Tapei, Taiwan. 2002.

Fung Pascale. Compiling Bilingual Lexicon Entries from a Nonparallel English-Chinese Corpus[C] // Proceedings of the 3rd Annual Workshop on Very Large Corpora. 1995: 173-183.

Fung Pascale, Kathleen McKeown. Finding terminology translation from non-parallel corpora[C] //5th Annual Workshop on Very Large Corpora, Hong Kong. 1997: 192-202.

Fung Pascale, Lo Yuen Yee. An IR approach for translating new words from nonparallel, comparable texts[C] //Proceedings of the 17th international conference on Computational linguistics, Montreal, Quebec, Canada. 1998: 414–420.

Fung Pascale. A Statistical View on Bilingual Lexicon Extraction from Parallel Corpora to Non-parallel Corpora[J]. Parallel Text Processing: Alignment and Use of Translation Corpora, 2000.

Hiroyuki Kaji. 2005. Extracting translation equivalents from bilingual comparable corpora[C] //Proceedings of the LREC-2008 Workshop on Comparable Corpora. 2008: 313–323.

Morin E, Daille B, Takeuchi K, Kageura K. Bilingual terminology mining – using brain, not brawn

comparable corpora[C] //Proceedings of the 45th annual meeting of the Association of Computational Linguistics, Prague, Czech Republic. 2007: 664–671.

Och F J, Ney H. Improved Statistical Alignment Models[C] //Proceedings of ACL. 2000: 440-447.

Otero P. Learning Bilingual Lexicons from Comparable English and Spanish Corpora[C] // Proceedings of MT Summit XI. 2007: 191-198.

Pekar, Viktor, Ruslan Mitkov, Dimitar Blagoev, Andrea Mulloni. Finding translations for low-frequency words in comparable corpora[J]. Machine Translation, 2006, 20(4): 247–266.

Pablo Gamallo. Learning bilingual lexicons from comparable english and spanish corpora[C] // Machine Translation SUMMIT XI, Copenhagen, Denmark. 2007.

Rapp, Reinhard. Automatic identification of word translations from unrelated English and German corpora[C] //Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, USA. 1999: 519–526.

Robitaille X. Compiling French Japanese Terminologies from the Web[C] //Proceedings of EACL. 2006.

Saralegi X, San Vicente I, Gurrutxaga A. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain[C] //Workshop on Building and Using Comparable Corpora in LREC. 2008.

Tanaka K, Iwasaki H. Extraction of lexical translations from non-aligned corpora[C] //Proceedings of COLING-96: The 16th international conference on computational linguistics, Copenhagen, Denmark. 1996: 580–585.

Xiao Z, McEnery A. Collocation, Semantic Prosody and Near Synonymy: A Cross-linguistic Perspective[J]. Applied Linguistics, 2006, 27(1): 103-129.

Yu Kun, Junichi Tsujii. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity[C] //Proceedings of HLTNAACL, Boulder, Colorado, USA. 2009: 121–124.