Bilingual Dictionary Extraction from Wikipedia

Kun Yu
Graduate School of Information Science and Technology
The University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
kunyu@is.s.u-tokyo.ac.jp

Junichi Tsujii
Graduate School of Information Science and Technology
The University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
tsujii@is.s.u-tokyo.ac.jp

Abstract

The way of mining comparable corpora and the strategy of dictionary extraction are two essential elements of bilingual dictionary extraction from comparable corpora. This paper first proposes a method, which uses the inter-language link in Wikipedia, to build comparable corpora. The large scale of Wikipedia ensures the quantity of collected comparable corpora. Besides, because the inter-language link is created by article author, the quality of collected corpora can also be guaranteed. After that, this paper presents an approach, which combines context heterogeneity similarity and dependency heterogeneity similarity, to extract bilingual dictionary from the collected comparable corpora. Experimental results show that because of combining the advantages of context heterogeneity similarity and dependency heterogeneity similarity appropriately, the proposed approach outperforms both the two individual approaches.

1 Introduction

Bilingual dictionary is a crucial part not only for machine translation (Och and Ney, 2003), but also for other natural language processing applications such as cross-language information retrieval (Grefenstette, 1998). At first, researchers constructed bilingual dictionary from parallel corpora. For example, Wu (1994) extracted English-Chinese translation lexicon through statistical training on a large parallel corpus. But for some languages, collecting parallel corpora is not easy. Thus, utilizing comparable corpora, in which texts are not translation of each other but share similar concepts, to extract bilingual dictionary has drawn more and more attention recently (Fung, 2000; Chiao and Zweigenbaum, 2002; Daille and Morin, 2005; Robitaille et al., 2006; Morin et al., 2007; Otero, 2008; Saralegi et al., 2008).

There are two popular strategies for constructing bilingual dictionary from comparable corpora: context-based strategy and syntax-based strategy.

Context-based strategy is based on the observation that a term and its translation appear in similar lexical contexts (Daille and Morin, 2008). This strategy has shown its effectiveness in terminology extraction (Fung, 2000; Chiao and Zweigenbaum, 2002; Daille and Morin, 2005; Robitaille et al., 2006; Morin et al., 2007; Daille and Morin, 2008; Saralegi et al., 2008). But there exists one problem that some words coming from the same domain may appear in similar contexts even if they are not translation of each other (Yu and Tsujii, 2009).

Besides of using window-based contexts, there were also some works utilizing syntax for bilingual dictionary extraction (Tanaka, 2002; Otero, 2007; Otero, 2008; Yu and Tsujii, 2009). In these works, syntactic contexts of words were acquired through hand-made templates or automatic analyzers. This strategy enlarges the lexical information used for word similarity calculation from a restricted window to the entire sentence. In addition, the usage of syntactic contexts brings richer information to dictionary extraction than using window-based contexts. While, this strategy requires larger corpora for correct dictionary extraction compared with the context-based strategy.

Besides, no matter of using which strategy, a large comparable corpus is an indispensable part in bilingual dictionary extraction from comparable corpora. It has been demonstrated that not only the quantity but also the quality of comparable corpora are important for bilingual dictionary construction (Morin, 2007). Mining the web to build comparable corpora was the most popular way for corpus acquisition. Most of them used the web sites that provide more than one language version for comparable corpora collection (Chiao and Zweigenbaum, 2002; Morin et al., 2007; Robitaille et al., 2006; Daille and Morin, 2008; Saralegi et al., 2008). But the quality of collected corpora cannot be guaranteed sometimes. Some researchers acquired comparable corpora from multi-lingual journals (Daille and Morin, 2005). The collected corpora are more reliable but restricted in a specific domain.

Based on above backgrounds, this paper first proposes using Wikipedia as a resource to mine large-scale and robust comparable corpora. Then, through investigating the context-based strategy and the syntax-based strategy, it presents a new approach combining the advantage of the two strategies properly for bilingual dictionary extraction from the collected comparable corpora. We did experiments to validate the effectiveness of the proposed bilingual dictionary extraction approach. Results show that compared with the approaches based on context heterogeneity similarity or dependency heterogeneity similarity alone, the proposed approach improves the performance of dictionary extraction.

The left part of this paper is organized as follows: Section 2 shows how to mine comparable corpora from Wikipedia; Section 3 introduces the proposed approach for bilingual dictionary extraction in detail; Experimental results and discussion are listed in Section 4; Section 5 compared the proposed work with related works; finally, Section 6 draws a brief conclusion and gives the direction of future work.

2 Mining Comparable Corpora from Wikipedia

As a rich and free resource, Wikipedia contains very large amount of articles written in different languages and various types of link information showing the relations between articles. It has been used as external resource in many natural language processing tasks successfully (Buscaldi and Rosso, 2006; Mihalcea, 2007; Nakayama et al., 2007).

Among the link information in Wikipedia, the inter-language link, which is created by article authors, connects large amount of articles that describe the same term but are written in different languages. For example, Erdmann et al. (2008) showed that in the English and Japanese Wikipedia database dump data[1] from November/December 2006 with 3,068,118 English articles and 455,524 Japanese articles, there are 103,374 inter-language links from English to Japanese and 108,086 inter-language links from Japanese to English. It has been demonstrated that these inter-language links are useful resources for bilingual dictionary construction (Erdmann et al., 2008). However, only the titles of the linked articles were used as translations of each other to construct bilingual dictionary in previous work (Erdmann et al., 2008). Besides of article titles, there still exists large amount of information that could be used for dictionary construction, such as the text inside the linked articles. After analysis, we find although the linked articles do not always contain the exact contents, they still share large amount of common contents. For example, Figure 1 shows part of the two articles from English and Chinese Wikipedia that describe the same term 'computer'. The listed texts contain the same content about the general introduction and the history of 'computer'.

Based on above analysis, we propose to use inter-language link to collect Chinese-English comparable corpora from Wikipedia. All the articles connected by inter-language links are extracted. Following are the detailed steps:

Step1: downloading Chinese and English Wikipedia database dump data (June/July 2008) from http://download.wikimedia.org.

Step2: extracting English articles that have Chinese inter-language link, then extract the linked Chinese articles.

*Step*3: to ensure the comparability of extracted articles, only keeping the paragraphs in the front part of each article that describes the general information. For example, in both the English and the Chinese articles shown in Figure 1, the last two paragraphs describing 'history of computing' are discarded.

Step4: cleaning extracted articles by removing super-links and unrelated words (e.g. 'Contents (show)' in the English article of Figure 1).

[1]http://download.wikimedia.org

Through these steps, we get Chinese-English comparable corpora with 124,316 article pairs, in which there exist 1,132,492 English sentences and 665,789 Chinese sentences. It is clear that the large scale of Wikipedia ensures the quantity of collected comparable corpora. Besides, because the inter-language links are created by article authors, the quality of collected corpora can also be guaranteed.

Computer

From Wikipedia, the free encyclopedia

This article is about the machine. For other uses, see Computer (disambiguation).
"Computer technology" redirects here. For the company, see Computer Technology Limited.

A **computer** is a machine that manipulates data according to a list of instructions.

Although mechanical examples of computers have existed throughout history, the first resembling a modern computer were developed in the mid-20th century (1940–1945). The first electronic computers were the size of a large room, consuming as much power as several hundred modern personal computers (PC).[1] Modern computers based on tiny integrated circuits are millions to billions of times more capable than the early machines, and occupy a fraction of the space.[2] Simple computers are small enough to fit into a wristwatch, and can be powered by a watch battery. Personal computers in their various forms are icons of the Information Age, what most people think of as a "computer", but the embedded computers found in devices ranging from fighter aircraft to industrial robots, digital cameras, and children's toys are the most numerous.

The ability to store and execute lists of instructions called programs makes computers extremely versatile, distinguishing them from calculators. The Church–Turing thesis is a mathematical statement of this versatility: any computer with a certain minimum capability is, in principle, capable of performing the same tasks that any other computer can perform. Therefore computers ranging from a personal digital assistant to a supercomputer are all able to perform the same computational tasks, given enough time and storage capacity.

Contents [show]

History of computing

Main article: History of computer hardware

The first use of the word "computer" was recorded in 1613, referring to a person who carried out calculations, or computations, and the word continued to be used in that sense until the middle of the 20th century. From the end of the 19th century onwards though, the word began to take on its more familiar meaning, decribing a machine that carries out computations.[3]

The history of the modern computer begins with two separate technologies—automated calculation and programmability—but no single device can be identified as the earliest computer, partly because of the inconsistent application of that term. Examples of early mechanical calculating devices include the abacus, the slide rule and arguably the astrolabe and the Antikythera mechanism (which dates from about 150–100 BC). Hero of Alexandria (c. 10–70 AD) built a mechanical theater which performed a play lasting 10 minutes and was operated by a complex system of ropes and drums that might be considered to be a means of deciding which parts of the mechanism performed which actions and when.[4] This is the essence of programmability.

计算机

维基百科，自由的百科全书

关于电子计算机的其他意思，详见计算机。

计算机（英语：**electronic computer**）是一种根据一系列指令来对数据进行处理的机器。所相关的技术研究叫计算机科学，由数据为内核的研究称信息技术。通常人们接触最多的是个人计算机（PC）。

计算机种类繁多。实际来看，计算机总体上是处理信息的工具。根据图灵机理论，一部具有最基本功能的计算机，应当能够完成任何其它计算机能做的事情。因此，只要不考虑时间和存储因素，从个人数码助理（PDA）到超级计算机都应该可以完成同样的作业。即是说，即使是设计完全相同的计算机，只要经过相应改装，就应该可以被用于从公司薪金管理到无人驾驶飞船操控在内的各种任务。由于科技的飞速进步，下一代计算机总是在性能上能够显着地超过其前一代，这一现象有时被称作“摩尔定律”。

计算机在组成上形式不一。早期计算机的体积足有一间房屋大小，而今天某些嵌入式计算机可能比一副扑克牌还小。当然，即使在今天，依然有大量体积庞大的巨型计算机为特别的科学计算或面向大型组织的事务处理需求服务。比较小的，为个人应用而设计的计算机称为微型计算机，简称微机。我们今天在日常使用“计算机”一词时通常也是指此。不过，现在计算机最为普遍的应用形式却是嵌入式的。嵌入式计算机通常相对简单，体积小，并被用来控制其它设备—无论是飞机，工业机器人还是数码相机。[1]

上述对于电子计算机的定义包括了许多能计算或是只有有限功能的特定用途的设备。然而当说到现代的电子计算机，其最重要的特征是，只要给予正确的指示，任何一台电子计算机都可以模拟其他任何计算机的行为（只受限于电子计算机本身的存储容量和执行的速度）。据此，现代电子计算机相对于早期的电子计算机也被称为通用型电子计算机。

目录 [显示]

历史 [编辑]

本来，计算机的英文原词“computer”是指从事数据计算的人。而他们往往都需要借助某些机械计算设备或模拟计算机。这些早期计算设备的祖先包括有算盘，以及可以追溯到公元前87年的被古希腊人用于计算行星移动的安提基特拉机制。随着中世纪末期欧洲数学与工程学的再次繁荣，1623年德国博学家Wilhelm Schickard率先研制出了欧洲第一台计算设备，这是一个能进行六位以内数加减法，并能通过铃声输出答案的“计算钟”。使用转动齿轮来进行操作。

1642年法国数学家布莱士·帕斯卡在英国数学家William Oughtred所制作的“计算尺”的基础上，将其加以改进，使能进行八位计算。还卖出了许多制品，成为当时一种时髦的商品。

Figure 1. Part of the articles describing ‘computer’ in both English and Chinese Wikipedia.

3 Extracting Bilingual Dictionary with Context Heterogeneity and Dependency Heterogeneity

3.1 Comparison between Context Heterogeneity and Dependency Heterogeneity

Fung (1995) proposed using context heterogeneity similarity for bilingual dictionary extraction from comparable corpora in a specific domain. It is based on the assumption that the context heterogeneity of a given domain-specific word is more similar to that of its translation in another language than that of an unrelated word in the other language (Fung, 1995). The author demonstrated that this feature is more salient than the feature that concerned the occurrence frequencies of words (Fung, 1995).

Through the investigation that some words from the same domain may appear in similar context even if they are not translation of each other, we presented a new feature called as dependency heterogeneity similarity (Yu and Tsujii, 2009). This feature assumes that a word and its translation share similar modifiers and head in comparable corpora. By using dependency heterogeneity similarity, bilingual dictionary from any domains could be extracted successfully.

Table 1. Results of bilingual dictionary extraction

	Accuracy	*MMR*
Context Heterogeneity	0.168	0.064
Dependency Heterogeneity	0.252 (↑50%)	0.119 (↑86%)

Both context heterogeneity similarity and dependency heterogeneity similarity have their own strong points. We did some experiments to do detailed comparison between the two features. We randomly selected 250 word translation pairs from the title of Wikipedia pages collected in Section 2, and used them as test data to evaluate both Fung (1995)'s work and our previous work (Yu and Tsujii, 2009). Two metrics were evaluated, which are *accuracy* (see equation 1) and *MMR* (Voorhees, 1999) (see equation 2). *Accuracy* shows the ability of selecting correct translation candidates. *MMR* shows the ability of precisely ranking the selected translation candidates. Table 1 lists the result of Top5 ranking. It shows the approach of using dependency heterogeneity similarity outperformed the approach of using context heterogeneity similarity. But the increase of *MMR* was 86% and the increase of *accuracy* was 50%. From this result, we could draw a conclusion that compared with context heterogeneity similarity dependency heterogeneity similarity has more potential to successfully rank the selected translation.

$$Accuracy = \sum_{i=1}^{N} t_i \Big/ N \tag{1}$$

$$t_i = \begin{cases} 1, \text{ if there exists correct translation in top } n \text{ ranking} \\ 0, \text{ otherwise} \end{cases}$$

N means the total number of words for evaluation

$$MMR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}, \quad rank_i = \begin{cases} r_i, \text{ if } r_i < n \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

n means top n evaluation

r_i means the rank of correct translation in top n ranking

N means the total number of words for evaluation

3.2 Combining Context Heterogeneity and Dependency Heterogeneity for Bilingual Dictionary Extraction

Based on above analysis, we propose a new approach that combines the merits of context heterogeneity similarity and dependency heterogeneity similarity properly. We utilize context heterogeneity similarity to select translation candidates and apply dependency heterogeneity similarity in candidate ranking. The proposed approach is fulfilled in the following steps:

Step1 (context heterogeneity vector learning): learning context heterogeneity vectors of word *W* in source language and all the words in target language;

Step2 (candidate selection): selecting *m* translation candidates for *W* from the words in target language by calculating the similarity of context heterogeneity vectors learned in *Step1*;

Step3 (dependency heterogeneity vector learning): learning dependency heterogeneity vectors of *W* and the *m* selected translation candidates;

Step4 (candidate ranking): ranking the *m* translation candidates for *W* using the similarity of dependency heterogeneity vectors learned in *Step3*.

The context heterogeneity vector of a word *W* is defined in equation 3. It contains two elements, which represent the heterogeneity of the word appearing in the left or right of *W*. The dependency heterogeneity vector of word *W* is defined in equation 4. It includes four elements. Each of them shows the heterogeneity of a type of dependency relation related with *W*. The types of dependency relations '*NMOD*' (noun modifier), '*SUB*' (subject), and '*OBJ*' (object) are acquired from a syntactic analyzer.

$$(H_{Left}, H_{Right}) \tag{3}$$

$$H_{Left}(W) = \frac{\text{number of different words appearing in the left of } W}{\text{total number of words appearing in the left of } W}$$

$$H_{Right}(W) = \frac{\text{number of different words appearing in the right of } W}{\text{total number of words appearing in the right of } W}$$

$$(H_{NMODHead}, H_{SUBHead}, H_{OBJHead}, H_{NMODMod}) \quad (4)$$

$$H_{NMODHead}(W) = \frac{\text{number of different heads of } W \text{ with } NMOD \text{ label}}{\text{total number of heads of } W \text{ with } NMOD \text{ label}}$$

$$H_{SUBHead}(W) = \frac{\text{number of different heads of } W \text{ with } SUB \text{ label}}{\text{total number of heads of } W \text{ with } SUB \text{ label}}$$

$$H_{OBJHead}(W) = \frac{\text{number of different heads of } W \text{ with } OBJ \text{ label}}{\text{total number of heads of } W \text{ with } OBJ \text{ label}}$$

$$H_{NMODMod}(W) = \frac{\text{number of different modifiers of } W \text{ with } NMOD \text{ label}}{\text{total number of modifiers of } W \text{ with } NMOD \text{ label}}$$

Euclidean distance is used to calculate both the similarity between context heterogeneity vectors of W_s in source language and W_t in target language (see equation 5) and the similarity between dependency heterogeneity vectors of W_s and W_t (see equation 5).

$$D_{Context}(W_s, W_t) = \sqrt{D_{Left}^2 + D_{Right}^2} \quad (5)$$

$$D_{Left} = H_{Left}(W_s) - H_{Left}(W_t)$$

$$D_{Right} = H_{Right}(W_s) - H_{Right}(W_t)$$

$$D_{Dependency}(W_s, W_t) = \sqrt{D_{NMODHead}^2 + D_{SUBHead}^2 + D_{OBJHead}^2 + D_{NMODMod}^2} \quad (6)$$

$$D_{NMODHead} = H_{NMODHead}(W_s) - H_{NMODHead}(W_t)$$

$$D_{SUBHead} = H_{SUBHead}(W_s) - H_{SUBHead}(W_t)$$

$$D_{OBJHead} = H_{OBJHead}(W_s) - H_{OBJHead}(W_t)$$

$$D_{NMODMod} = H_{NMODMod}(W_s) - H_{NMODMod}(W_t)$$

Before extracting bilingual dictionary by the proposed approach, we need to preprocess the collected comparable corpora, which includes: (1) a stemmer[2] is used to do stemming for the English corpus. To avoid excessive stemming, we use stems only for translation candidates because we only consider about dictionary extraction for nouns currently. (2) stop words are removed. For English, we use the stop word list from (Fung, 1995). For Chinese, we remove '的 (of)' as stop word. (3) we remove the sentences with more than k (set as 30 empirically) words from both English corpus and Chinese corpus, in order to reduce the effect of parsing error on dictionary extraction. After preprocessing, we use a Chinese morphological analyzer (Nakagawa and Uchimoto, 2007) and an English pos-tagger (Tsuruoka et al., 2005) to analyze the raw corpora. Then, a syntactic analyzer MaltParser (Nivre et al., 2007) is applied to get dependency relations.

4 Results and Discussion

4.1 Experimental Setting

The comparable corpora mined in Section 2 are used for context heterogeneity vector and dependency heterogeneity vector learning. We did two experiments using different data sets.

- *exp1*: this experiment uses 500 Chinese-English single-noun pairs that are randomly selected from the aligned titles of the collected pages. We divide them into 10 folders. 5 folders are for testing and the other 5 folders are for development.
- *exp2*: because the data from Wikipedia page titles used in *exp1* could be more likely to have a translation in the corresponding article than the non-title words, only using *exp1* may not prove the effectiveness of the *proposed* approach in real bilingual dictionary extraction. Therefore we did another experiment, which uses 150 Chinese-English single-noun pairs that are randomly selected from the Chinese-English translation lexicon from LDC[3] as testing data. We also divide them into 3 folders, with each folder containing 50 translation pairs.

We evaluated three approaches in all the experiments, which are:

- *context*: only using context heterogeneity similarity for both translation candidate selection (step2) and ranking (step4);

- *dep*: only using dependency heterogeneity similarity for both translation candidate selection (step2) and ranking (step4);

- *proposed*: the proposed approach.

Accuracy (see equation 1) and *MMR* (see equation 2) are used as evaluation metrics in both the two experiments. The average scores of both *accuracy* and *MMR* among folders are also calculated.

In all the experiments, the number of selected candidates m (See Section 3.2) in step2 was set as 20 (see Section 4.3 for detailed explanation).

[2]http://search.cpan.org/~snowhare/Lingua-Stem-0.83/

[3]http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L27

4.2 Results of Experiment 1

Table 2 lists the average evaluation results with Top5 ranking on testing data. These results prove that because combining context heterogeneity similarity and dependency heterogeneity similarity appropriately, the *proposed* approach outperformed the *context* approach and the *dep* approach. In addition, compared with the result of the *context* approach, the increase of *Ave.MMR* by the *proposed* approach was much larger than the increase of *Ave.Accuracy*. It demonstrated again that the usage of dependency heterogeneity similarity in the *proposed* approach gave great help to candidate ranking.

Table 2. Average results with Top5 ranking on testing data of *exp1*.

	Ave.Accuracy	*Ave.MMR*
context	0.152	0.053
dep	0.216	0.112
proposed	**0.228** (↑50.0%: *context*) (↑5.6%: *dep*)	**0.125** (↑135.8%: *context*) (↑11.6%: *dep*)

4.3 Results of Experiment 2

The evaluation results of *exp2* are shown in Table 3. It indicates that when testing on the data from a real bilingual dictionary, the *proposed* approach still outperformed the *context* approach and the *dep* approach.

Table 3. Average results with Top5 ranking on testing data of *exp2*.

	Ave.Accuracy	*Ave.MMR*
context	0.167	0.078
dep	0.140	0.079
proposed	**0.193**	**0.097**

While, compared with the *context* approach, the *dep* approach got lower average accuracy but a little higher average *MMR*. One possible reason is the occurrence time of the translation candidates in the comparable corpora. In our previous work (Yu and Tsujii, 2009), we indicated that the dependency heterogeneity similarity was easily affected by the occurrence time of the translation candidates. And our analysis shows that among the 150 Chinese-English translation pairs in the testing data, there were 81 Chinese words that only appeared in the corpora less than 50 times. But this problem was well solved by combining the context heterogeneity similarity with the dependency heterogeneity similarity in the *proposed* approach.

4.4 Discussion

In the proposed approach, the number of translation candidates selected by context heterogeneity similarity (i.e. m) affects the performance of dictionary extraction. This parameter was set as 20 in our experiments. This setting was based on the learning curve on the development data (see Figure 2). In Figure 2, there were two peaks (when m=20 and m=40) in the curve of *AVE.Accu* on development data, but the best *Ave.MMR* was obtained when m was 20. Considering about that better ranking of extracted dictionary entries is more important in real application, the setting m=20 was selected in our experiments.

In addition, for Top5 ranking, m=50 means only using dependency heterogeneity similarity for dictionary extraction and m=5 means only using context heterogeneity similarity. In Figure 2, compared with the results when m was set as 5, the *Ave.Accuracy* was improved greatly when m was set as 50 on testing data. This result demonstrates that dependency heterogeneity similarity not only performs better than context heterogeneity similarity in translation candidate ranking, but also contributed more in translation candidate selection.

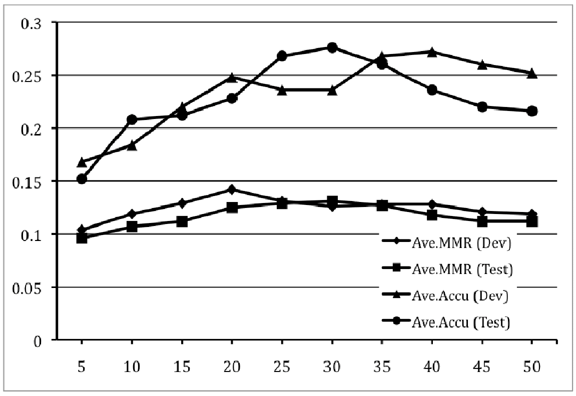

Figure 2. Performance of bilingual dictionary extraction (Top5) with different m (horizontal axis)

We also evaluated the three approaches using translation pairs with different occurrence times, in order to see the effect of word occurrence on context heterogeneity similarity and dependency heterogeneity similarity. Table 4 and Table 5 list the results. These results first show that no matter how many times the translation pairs appear in the comparable corpora, combining context heterogeneity similarity and dependency heterogeneity similarity through the proposed approach achieved the best performance. They also show that when the words appeared frequently (*occur*>50), the improvement of performance (especially *Ave.MMR*) was much larger than the improvement when the occurrence of words was small (*occur*<=50). These phenomena imply that the quantity of comparable corpora has large effect on dependency heterogeneity similarity.

Table 4. *Average accuracy* with Top5 ranking on different testing data of *exp1*.

	occur <= 50	*occur > 50*
context	0.112	0.156
dep	0.124	0.180
proposed	0.148 (↑32.1%)	0.228 (↑46.2%)

Table 5. *Average MMR* with Top5 ranking on different testing data of *exp1*.

	occur <= 50	*occur > 50*
context	0.053	0.061
dependency	0.059	0.098
proposed	0.077 (↑45.3)	0.125 (↑104.9%)

5 Related Work

Previous work about bilingual dictionary extraction from comparable corpora mainly focused on using context similarity. Fung (1995) utilized context heterogeneity similarity to compile English-Chinese dictionary. Other researchers (Fung, 2000; Chiao and Zweigenbaum, 2002; Daille and Morin, 2005; Robitaille et al., 2006; Morin et al., 2007; Daille and Morin, 2008; Saralegi et al., 2008) extracted bilingual dictionaries by comparing the similarity between the context vectors of words in both source and target languages with the aid of an external dictionary. Compared with these works, the proposed approach only used context heterogeneity similarity to select translation candidates, but applied dependency heterogeneity similarity in translation candidate ranking.

Other researchers introduced syntactic similarity to bilingual dictionary extraction from comparable corpora (Tanaka, 2002; Otero, 2007; Otero, 2008; Yu and Tsujii, 2009). Similar to these approaches, the proposed approach utilized rich syntactic information for translation candidate ranking. The main difference between them is the combination of context heterogeneity similarity for candidate selection in the proposed approach.

In addition, this paper presented an effective method to build comparable corpora from Wikipedia by using inter-language links. Previous work about using Wikipedia in bilingual dictionary extraction (Erdmann et al., 2008) only concerned about the title of pages collected by inter-language links. But in the proposed corpus mining method, the content of collected pages are also processed to acquire robust and large-scale comparable corpora.

6 Conclusion and Future Work

Extracting bilingual dictionary from comparable corpora has drawn great attention in recent years, in which how to collect comparable corpora and how to extract bilingual dictionary are two essential problems. In this paper, a new method for mining comparable corpora from Wikipedia by using the inter-language link is introduced first. Through this method, robust and large-scale comparable corpora could be collected. Then, this paper presents an approach combining both context heterogeneity similarity and dependency heterogeneity similarity for bilingual dictionary extraction from the collected comparable corpora. The experimental results show that by combining the advantages of context heterogeneity similarity and dependency heterogeneity similarity properly, the proposed approach outperformed the approaches that use the two features alone.

There are still several future works under consideration. Currently, the proposed bilingual dictionary extraction approach was only tested on single-words. In the future, we will extend it to extracting bilingual dictionary for multi-words. Besides, the experimental results prove that the usage of syntactic information performs better than lexical context in both translation candidate selection and candidate ranking. In our future work, we would like to try richer features, such as semantic information, to see their effects on dictionary extraction. Finally, although the experimental results have proven the effectiveness of the proposed approach, the accuracy of bilingual dictionary extraction is still low. In the current work, we combine the context heterogeneity similarity and dependency heterogeneity similarity simply. In the future, we will apply some machine learning methods in combination to improve the dictionary accuracy further.

Acknowledgments

This research is sponsored by Microsoft Research Asia Web-scale Natural Language Processing Theme.

References

D.Buscaldi and P.Rosso. 2006. Mining Knowledge from Wikipedia for the Question Answering Task. *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Y.Chiao and P.Zweigenbaum. 2002. Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

B.Daille and E.Morin. 2005. French-English Terminology Extraction from Comparable Corpora. *Proceedings of the 2nd International Joint Conference on Natural Language Processing*.

B.Daille and E.Morin. 2008. An Effective Compositional Model for Lexical Alignment. *Proceedings of the 3rd International Joint Conference on Natural Language Processing*.

M.Erdmann, K.Nakayama, T.Hara and S.Nishio. 2008. An Approach for Extracting Bilingual Terminology from Wikipedia. *Proceedings of the 13th International Conference on Database Systems for Advanced Applications*.

P.Fung. 1995. Compiling Bilingual Lexicon Entries from a Non-parallel English-Chinese Corpus. *Proceedings of the 3rd Annual Workshop on Very Large Corpora*. pp. 173-183.

P.Fung. 2000. A Statistical View on Bilingual Lexicon Extraction from Parallel Corpora to Non-parallel Corpora. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer Academic Publishers.

G.Grefenstette. 1998. The Problem of Cross-language Information Retrieval. *Cross-language Information Retrieval*. Kluwer Academic Publishers.

R.Mihalcea. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT 2007)*.

E.Morin, B.Daille, K.Takeuchi and K.Kageura. 2007. Bilingual Terminology Mining – Using Brain, not Brawn Comparable Corpora. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. pp. 664-671.

T.Nakagawa and K.Uchimoto. 2007. A Hybrid Approach to Word Segmentation and POS Tagging. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

K.Nakayama, T.Hara and S.Nishio. 2007. Wikipedia Mining for An Association Web Thesaurus Construction. *Proceedings of the 8th International Conference on Web Information Systems Engineering*.

J.Nivre, J.Hall, J.Nilsson, A.Chanev, G.Eryigit, S.Kubler, S.Marinov and E.Marsi. 2007. MaltParser: A Language-independent System for Data-driven Dependency Parsing. *Natural Language Engineering*. 13(2): 95-135.

F.Och and H.Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1): 19-51.

P.Otero. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. *Proceedings of Machine Translation Summit XI*. pp. 191-198.

P.Otero. 2008. Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora. *Proceedings of LREC 2008 Workshop of Building and Using Comparable Corpora*. pp. 19-26.

X.Robitaille, Y.Sasaki, M.Tonoike, S.Sato and T.Utsuro. 2006. Compiling French Japanese Terminologies from the Web. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

X.Saralegi, I.S.Vicente and A.Gurrutxaga. 2008. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. *Proceedings of LREC 2008 Workshop of Building and Using Comparable Corpora*.

T.Tanaka. 2002. Measuring the Similarity between Compound Nouns in Different Languages Using Non-parallel Corpora. *Proceedings of the 19th International Conference on Computational Linguistics*.

Y.Tsuruoka, Y.Tateishi, J.Kim, T.Ohta, J.McNaught, S.Ananiadou and J.Tsujii. 2005. Developing a Robust Part-of-speech Tagger for Biomedical Text. *Advances in Informatics – 10th Panhellenic Conference on Informationcs*. LNCS 3746. pp. 382-392.

E.M.Voorhees. 1999. The TREC-8 Question Answering Track Report. *Proceedings of the 8th Text Retrieval Conference*.

D.Wu. 1994. Learning an English-Chinese Lexicon from a Parallel Corpus. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*.

K.Yu and J.Tsujii. 2009. Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT 2009)*.