Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity

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

Abstract

This paper proposes an approach for bilingual dictionary extraction from comparable corpora. The proposed approach is based on the observation that a word and its translation share similar dependency relations. Experimental results using 250 randomly selected translation pairs prove that the proposed approach significantly outperforms the traditional context-based approach that uses bag-of-words around translation candidates.

1 Introduction

Bilingual dictionary plays an important role in many natural language processing tasks. For example, machine translation uses bilingual dictionary to reinforce word and phrase alignment (Och and Ney, 2003), cross-language information retrieval uses bilingual dictionary for query translation (Grefenstette, 1998). The direct way of bilingual dictionary acquisition is aligning translation candidates using parallel corpora (Wu, 1994). But for some languages, collecting parallel corpora is not easy. Therefore, many researchers paid attention to bilingual dictionary extraction from comparable corpora (Fung, 2000; Chiao and Zweigenbaum, 2002; Daille and Morin, 2008; Robitaille et al., 2006; Morin et al., 2007; Otero, 2008), in which texts are not exact translation of each other but share common features.

Context-based approach, which is based on the observation that a term and its translation appear in similar lexical contexts (Daille and Morin, 2008), is the most popular approach for extracting bilingual dictionary from comparable corpora and has shown its effectiveness in terminology extraction (Fung, 2000; Chiao and Zweigenbaum, 2002; Robitaille et al., 2006; Morin et al., 2007). But it only concerns about the lexical context around translation candidates in a restricted window. Besides, in comparable corpora, some words may appear in similar context even if they are not translation of each other. For example, using a Chinese-English comparable corpus from Wikipedia and following the definition in (Fung, 1995), we get context heterogeneity vector of three words (see Table 1). The Euclidean distance between the vector of '经济学(economics)' and 'econom-

ics' is 0.084. But the Euclidean distance between the vector of '经济学' and 'medicine' is 0.075. In such case, the incorrect dictionary entry '经济学/medicine' will be extracted by context-based approach.

Table 1. Context heterogeneity vector of words.

Word	Context Heterogeneity Vector
经济学(economics)	(0.185, 0.006)
economics	(0.101, 0.013)
medicine	(0.113,0.028)

To solve this problem, we investigate a comparable corpora from Wikipedia and find the following phenomenon: *if we preprocessed the corpora with a dependency syntactic analyzer, a word in source language shares similar head and modifiers with its translation in target language, no matter whether they occur in similar context or not*. We call this phenomenon as ***dependency heterogeneity***. Based on this observation, we propose an approach to extract bilingual dictionary from comparable corpora. Not like only using bag-of-words around translation candidates in context-based approach, the proposed approach utilizes the syntactic analysis of comparable corpora to recognize the meaning of translation candidates. Besides, the lexical information used in the proposed approach does not restrict in a small window, but comes from the entire sentence.

We did experiments with 250 randomly selected translation pairs. Results show that compared with the approach based on context heterogeneity, the proposed approach improves the accuracy of dictionary extraction significantly.

2 Related Work

In previous work about dictionary extraction from comparable corpora, using context similarity is the most popular one.

At first, Fung (1995) utilized context heterogeneity for bilingual dictionary extraction. Our proposed approach borrows Fung's idea but extends context heterogeneity to dependency heterogeneity, in order to utilize rich syntactic information other than bag-of-words.

After that, researchers extended context heterogeneity vector to context vector with the aid of an existing bilingual dictionary (Fung, 2000; Chiao and Zweigenbaum, 2002; Robitaille et al., 2006; Morin et al., 2007; Daille and Morin, 2008). In these works, dictionary extraction

is fulfilled by comparing the similarity between the context vectors of words in target language and the context vectors of words in source language using an external dictionary. The main difference between these works and our approach is still our usage of syntactic dependency other than bag-of-words. In addition, except for a morphological analyzer and a dependency parser, our approach does not need other external resources, such as the external dictionary. Because of the well-developed morphological and syntactic analysis research in recent years, the requirement of analyzers will not bring too much burden to the proposed approach.

Besides of using window-based contexts, there were also some works utilizing syntactic information for bilingual dictionary extraction. Otero (2007) extracted lexico-syntactic templates from parallel corpora first, and then used them as seeds to calculate similarity between translation candidates. Otero (2008) defined syntactic rules to get lexico-syntactic contexts of words, and then used an external bilingual dictionary to fulfill similarity calculation between the lexico-syntactic context vectors of translation candidates. Our approach differs from these works in two ways: (1) both the above works defined syntactic rules or templates by hand to get syntactic information. Our approach uses data-driven syntactic analyzers for acquiring dependency relations automatically. Therefore, it is easier to adapt our approach to other language pairs. (2) the types of dependencies used for similarity calculation in our approach are different from Otero's work. Otero (2007; 2008) only considered about the modification dependency among nouns, prepositions and verbs, such as the adjective modifier of nouns and the object of verbs. But our approach not only uses modifiers of translation candidates, but also considers about their heads.

3 Dependency Heterogeneity of Words in Comparable Corpora

Dependency heterogeneity means *a word and its translation share similar modifiers and head in comparable corpora*. Namely, the modifiers and head of unrelated words are different even if they occur in similar context.

Table 2. Frequently used modifiers (words are not ranked).

经济学(economics)	economics	medicine
微观/micro	keynesian	physiology
宏观/macro	*new*	Chinese
计量/computation	institutional	traditional
新/new	positive	biology
政治/politics	*classical*	internal
大学/university	labor	science
古典派/classicists	*development*	clinical
发展/development	engineering	veterinary
理论/theory	finance	western
实证/demonstration	international	agriculture

For example, Table 2 collects the most frequently used 10 modifiers of the words listed in Table 1. It shows there are 3 similar modifiers (italic words) between '经济学(economics)' and 'economics'. But there is no similar word between the modifiers of '经济学' and that of 'medicine'. Table 3 lists the most frequently used 10 heads (when a candidate word acts as subject) of the three words. If excluding copula, '经济学' and 'economics' share one similar head (italic words). But '经济学' and 'medicine' shares no similar head.

Table 3. Frequently used heads (the predicate of subject, words are not ranked).

经济学(economics)	economics	medicine
是/is	is	is
均衡/average	has	tends
毕业/graduate	was	include
承认/admit	*emphasizes*	moved
能/can	non-rivaled	means
分化/split	became	requires
剩下/leave	assume	includes
比/compare	relies	were
成为/become	can	has
偏重/emphasize	replaces	may

4 Bilingual Dictionary Extraction with Dependency Heterogeneity

Based on the observation of dependency heterogeneity in comparable corpora, we propose an approach to extract bilingual dictionary using dependency heterogeneity similarity.

4.1 Comparable Corpora Preprocessing

Before calculating dependency heterogeneity similarity, we need to preprocess the comparable corpora. In this work, we focus on Chinese-English bilingual dictionary extraction for single-nouns. Therefore, we first use a Chinese morphological analyzer (Nakagawa and Uchimoto, 2007) and an English pos-tagger (Tsuruoka et al., 2005) to analyze the raw corpora. Then we use MaltParser (Nivre et al., 2007) to get syntactic dependency of both the Chinese corpus and the English corpus. The dependency labels produced by MaltParser (e.g. SUB) are used to decide the type of heads and modifiers.

After that, the analyzed corpora are refined through following steps: (1) we use a stemmer[1] to do stemming for the English corpus. Considering that only nouns are treated as translation candidates, we use stems for translation candidate but keep the original form of their heads and modifiers in order to avoid excessive stemming. (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 dependencies including punctuations and remove the sentences with

---

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

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.

4.2 Dependency Heterogeneity Vector Calculation

Equation 1 shows the definition of dependency heterogeneity vector of a word *W*. It includes four elements. Each element represents the heterogeneity of a dependency relation. 'NMOD' (noun modifier), 'SUB' (subject) and 'OBJ' (object) are the dependency labels produced by MaltParser.

$$(H_{NMODHead}, H_{SUBHead}, H_{OBJHead}, H_{NMODMod}) \quad (1)$$

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

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

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

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

4.3 Bilingual Dictionary Extraction

After calculating dependency heterogeneity vector of translation candidates, bilingual dictionary entries are extracted according to the distance between the vector of W_s in source language and the vector of W_t in target language. We use Euclidean distance (see equation 2) for distance computation. The smaller distance between the dependency heterogeneity vectors of W_s and W_t, the more likely they are translations of each other.

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

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

For example, following above definitions, we get dependency heterogeneity vector of the words analyzed before (see Table 4). The distances between these vectors are D_H(经济学, economics) = 0.222, D_H(经济学, medicine) = 0.496. It is clear that the distance between the vector of '经济学(economics)' and 'economics' is much smaller than that between '经济学' and 'medicine'. Thus, the pair '经济学/economics' is extracted successfully.

Table 4. Dependency heterogeneity vector of words.

Word	Dependency Heterogeneity Vector
经济学(economics)	(0.398, 0.677, 0.733, 0.471)
economics	(0.466, 0.500, 0.625, 0.432)
medicine	(0.748, 0.524, 0.542, 0.220)

5 Results and Discussion
5.1 Experimental Setting

We collect Chinese and English pages from Wikipedia[2] with inter-language link and use them as comparable corpora. After corpora preprocessing, we get 1,132,492

English sentences and 665,789 Chinese sentences for dependency heterogeneity vector learning. To evaluate the proposed approach, we randomly select 250 Chinese/English single-noun pairs from the aligned titles of the collected pages as testing data, and divide them into 5 folders. *Accuracy* (see equation 3) and *MMR* (Voorhees, 1999) (see equation 4) are used as evaluation metrics. The average scores of both *accuracy* and *MMR* among 5 folders are also calculated.

$$Accuracy = \sum_{i=1}^{N} t_i \Big/ N \quad (3)$$

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

$$MMR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}, \quad rank_i = \begin{cases} r_i, \text{ if } r_i < n \\ 0, \text{ otherwise} \end{cases} \quad (4)$$

n means top *n* evaluation,
r_i means the rank of the correct translation in top *n* ranking
N means the total number of words for evaluation

5.2 Results of Bilingual Dictionary Extraction

Two approaches were evaluated in this experiment. One is the context heterogeneity approach proposed in (Fung, 1995) (*context* for short). The other is our proposed approach (*dependency* for short).

The average results of dictionary extraction are listed in Table 5. It shows both the average *accuracy* and average *MMR* of extracted dictionary entries were improved significantly (McNemar's test, p<0.05) by the proposed approach. Besides, the increase of top5 evaluation was much higher than that of top10 evaluation, which means the proposed approach has more potential to extract precise bilingual dictionary entries.

Table 5. Average results of dictionary extraction.

	context		dependency	
	ave.accu	*ave.MMR*	*ave.accu*	*ave.MMR*
Top5	0.132	0.064	0.208(↑57.58%)	0.104(↑62.50%)
Top10	0.296	0.086	0.380(↑28.38%)	0.128(↑48.84%)

5.3 Effect of Dependency Heterogeneity Vector Definition

In the proposed approach, a dependency heterogeneity vector is defined as the combination of head and modifier heterogeneities. To see the effects of different dependency heterogeneity on dictionary extraction, we evaluated the proposed approach with different vector definitions, which are

only-head: $(H_{NMODHead}, H_{SUBHead}, H_{OBJHead})$
only-mod: $(H_{NMODMod})$
only-NMOD: $(H_{NMODHead}, H_{NMODMod})$

Table 6. Average results with different vector definitions.

	Top5		Top10	
	ave.accu	*ave.MMR*	*ave.accu*	*ave.MMR*
context	0.132	0.064	0.296	0.086
dependency	**0.208**	**0.104**	**0.380**	**0.128**
only-mod	0.156	0.080	0.336	0.103
only-head	0.176	0.077	0.336	0.098
*only-NMOD*s	0.200	0.094	0.364	0.115

The results are listed in Table 6. It shows with any types of vector definitions, the proposed approach outperformed the *context* approach. Besides, if comparing the results of *dependency*, *only-mod*, and *only-head*, a conclusion can be drawn that head dependency heterogeneities and modifier dependency heterogeneities gave similar contribution to the proposed approach. At last, the difference between the results of *dependency* and *only-NMOD* shows the head and modifier with NMOD label contributed more to the proposed approach.

5.4 Discussion

To do detailed analysis, we collect the dictionary entries that are not extracted by *context* approach but extracted by the proposed approach (*good* for short), and the entries that are extracted by *context* approach but not extracted by the proposed approach (*bad* for short) from top10 evaluation results with their occurrence time (see Table 7). If neglecting the entries '护照/passports' and '上海/shanghai', we found that the proposed approach tended to extract correct bilingual dictionary entries if both the two words occurred frequently in the comparable corpora, but failed if one of them seldom appeared.

Table 7. Good and bad dictionary entries.

Good		Bad	
Chinese	English	Chinese	English
犹太人/262	jew/122	十字架/53	crucifixion/19
速度/568	velocity/175	水族箱/6	aquarium/31
历史/2298	history/2376	混合物/47	mixture/179
组织/1775	organizations/2194	砖/17	brick/66
运动/1534	movement/1541	量化/23	quantification/31
护照/76	passports/80	上海/843	shanghai/1247

But there are two exceptions: (1) although '上海 (shanghai)' and 'shanghai' appeared frequently, the proposed approach did not extract them correctly; (2) both '护照(passport)' and 'passports' occurred less than 100 times, but they were recognized successfully by the proposed approach. Analysis shows the cleanliness of the comparable corpora is the most possible reason. In the English corpus we used for evaluation, many words are incorrectly combined with 'shanghai' by '**br**' (i.e. line break), such as 'airport**br**shanghai'. These errors affected the correctness of dependency heterogeneity vector of 'shanghai' greatly. Compared with the dirty resource of 'shanghai', only base form and plural form of 'passport' occur in the English corpus. Therefore, the dependency heterogeneity vectors of '护照' and 'passports' were precise and result in the successful extraction of this dictionary entry. We will clean the corpora to solve this problem in our future work.

6 Conclusion and Future Work

This paper proposes an approach, which not uses the similarity of bag-of-words around translation candidates

but considers about the similarity of syntactic dependencies, to extract bilingual dictionary from comparable corpora. Experimental results show that the proposed approach outperformed the context-based approach significantly. It not only validates the feasibility of the proposed approach, but also shows the effectiveness of applying syntactic analysis in real application.

There are several future works under consideration including corpora cleaning, extending the proposed approach from single-noun dictionary extraction to multiwords, and adapting the proposed approach to other language pairs. Besides, because the proposed approach is based on the syntactic analysis of sentences with no more than *k* words (see Section 4.1), the parsing accuracy and the setting of threshold *k* will affect the correctness of dependency heterogeneity vector learning. We will try other thresholds and syntactic parsers to see their effects on dictionary extraction in the future.

References

Y.Chiao and P.Zweigenbaum. 2002. Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. *Proceedings of LREC 2002*.

B.Daille and E.Morin. 2008. An Effective Compositional Model for Lexical Alignment. *Proceedings of IJCNLP-08*.

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.

E.Morin et al.. 2007. Bilingual Terminology Mining – Using Brain, not Brawn Comparable Corpora. *Proceedings of ACL 2007*. T.Nakagawa and K.Uchimoto. 2007. A Hybrid Approach to Word Segmentation and POS Tagging. *Proceedings of ACL 2007*.

J.Nivre et al.. 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 MT 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 on Comparable Corpora*. pp. 19-26.

X.Robitaille et al.. 2006. Compiling French Japanese Terminologies from the Web. *Proceedings of EACL 2006*.

Y.Tsuruoka et al.. 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*.