Automatic Extraction of Bilingual Terms From A Chinese-Japanese Parallel Corpus

Xiaorong Fan
Graduate School of III
University of Tokyo
Tokyo, Japan
81-35-841-2738
han@r.dl.itc.u-tokyo.ac.jp

Nobuyuki Shimizu
Information Technology Center
University of Tokyo
Tokyo, Japan
81-35-841-2738
shimizu@r.dl.itc.u-tokyo.ac.jp

Hiroshi Nakagawa
Information Technology Center
University of Tokyo
Tokyo, Japan
81-35-841-2738
n3@dl.itc.u-tokyo.ac.jp

ABSTRACT
This paper proposes a new approach for the automatic extraction of bilingual terms from a domain-specific bilingual parallel corpus. We combine existing monolingual term extractor and a word alignment tool to extract bilingual terms. Our method is different from those past studies as we simply use a word alignment tool to extract multi-words terms, and we use one monolingual term extractor for both of languages to reduce extraction imbalance. We obtained a good precision and an improved BLEU score in our experiment based on a Chinese-Japanese parallel corpus.

Categories and Subject Descriptors
D.3.3 [**Artificial Intelligence**]: Natural Language Processing – *machine translation*

General Terms
Languages

Keywords
Automatic extraction, bilingual corpus, bilingual term, word alignment, segmentation, multi-words term

1. INTRODUCTION
Terms are the lexical units to represent the most fundamental knowledge of a domain. A bilingual term is a specific terminology that forms a translation pair between two languages. Bilingual terms are very crucial resources for Cross-Language IR and Machine Translation. Automatic extraction of bilingual terms is an important task in Natural Language Processing.

In the recent years, many studies have proposed bilingual term extraction from several kinds of resources. Francis Bond et al.[1] extracted bilingual terms from nomolingual data by using explicit in-text cues such as partenthess "()" and character type. Rogelio Nazar et al. [2] used a monolingual corpus to extract the units to be translated and then found the translation for the source unit from web. Hervé Déjean et al. [3] combined comporable corpora and bilingual lexicons to extract bilingual terms.

Most of the previous studies focused on extracting bilingual terms from/to English, and researchers seldom paied attention to extraction bilingual terms between Asian languages.

Tsunakawa et al.[4] built a Chinese-Japanese lexicons through a pivot language, English, by using phrase-based SMT. Tonoike et al.[5] decomposed a multi-word term into several constituents, and its translation is generated by concatenating translation of each constituent. For Chinese and Japanese, even if there is a bilingual lexicon, it is still difficult to decompose a multi-word term and concatenate its translation.

Terms we mentioned here are technical terms of specific-domain. They can be classified into single-word terms and multi-word terms. In this paper, we focus on multi-word terms, but our method can be applied to single word terms, too. Henceforth, "term" denotes multi-words term.

For Chinese and Japanese, we may get the constituents of a multi-word term through a text segmentation process. However, there are three difficulties. First, the segmentation creteria of Chinese and Japanese are different. For example, the Chinese technical term "二叉树(binary tree)" is segmented into two words "二叉(binary)" and "树(tree)". But the corresponding Japanese technique term "二分木" is segmented into two words "二" and "分木". In this case, it is hard to get the correct translation result because of this discrepancy between Chinese and Japanese causes.

Second, the word order of Japanese and Chinese are not always same. For example, the segmentation result of Chinese term "无监督学习"(unsupervised learning) is "无(no) 监督(supervised) 学习(learning)", and the segmentation result of corresponding Japanese term "教師なし学習" is "教師(teacher) なし(without) 学習(learning)" . We can see from the example that we can't get the correct translation by the simple concatenation of the translations of those single words.

At last, even if the translation of every single word is sometimes right, translation through concatenating the translations of single words is absolutely wrong. For example, literal translation of Japanese term "形態素解析(Morphological Analysis)" is "词法分析". But in Chinese, "词法分析" always means "lexical analysis" in computer science. The right translation of "形態素解析" in NLP should be "汉语(Chinese)词法分析" or "中文(Chinese)词法分析".

Extracting bilingual lexicon from parallel corpus is expected to help us extract more accurate results. Since larger and larger bilingual corpora are cretead as training data for SMT system recently, we propose to use bilingual corpora as our resource in our approach. Especially, in this paper, we propose a two-stage model to extract bilingual terms from a Chinese-Japanese parallel corpus without bilingual lexicons.

2. Bilingual Term Extraction

Automatic extraction of bilingual terms generally involves two important steps: (1) extraction of monolingual term candidates, (2) alignment of term candidates.

Different approaches have been proposed for the bilingual term extraction from bilingual corpora. Ha et al. [6] firstly extract terms from corpora of distinct languages and then use a contingency table and log-likelihood to measure how likely a pair of term candidates is to be a correct pair. Ddilie et al. [7] extract terms independently and then find correspondences between candidates across languages using frequency count and bilingual associations of single words.

In this paper, we propose a two-stage extraction process. In the first stage, we extract Chinese and Japanese terms from bilingual corpus independently as other previous works did. In the second stage, we propose a novel word-alignment-based bilingual matching process to produce multi-word term translation pairs.

2.1 Monolingual Term Extraction

For bilingual term extraction, the first step is to extract terms candidates from both the source and target language. In recent years, many software tools that automatically extract terms have been designed and implemented using both statistical and linguistic information.

Most of the previous works select different term extractors for different languages. In this way, extraction imbalance occurs between monolingual extractions due to the different extraction criteria used by term extractors. In our approach, we choose different language versions of one monolingual term extraction tool that can extract both multi-words terms and single word terms using both statistical and linguistic information. They have the same statistical measure and linguistic information, so the imbalance problem is minimized.

2.2 Term alignment

In our approach, we use statistical method of alignment to align term candidates. Unlike the other works described before, we propose a way to align the multi-words term candidates using single word alignment.

We propose a three-step procedure to align term candidates. First, we treat a bilingual corpus through a re-segmentation process and then use a word alignment tool to get the bilingual multi-words term candidates. At last, we use a smoothing process to tune the translations more precisely.

2.2.1 Re-Segmentation

Asian languages, such as Japanese and Chinese, have an important characteristic that they are character-based languages and there are no spaces between characters. Word segmentation must be done prior to most of the natural language processing.

As well known, the main task of word alignment is to identify translation equivalents between lexical items in bilingual parallel texts. In English and other Western Europe languages, the lexical item is a word, but in Japanese and Chinese, the lexical item is a series of continuous characters that is finally recognized by word segmentation process. It can be a multi-words unit or a single word unit. If the lexical items are single word units, we can get the single word translation equivalents using word alignment. On the other hand, multi-words translation equivalents will be extracted when the lexical items are multi-words units using word alignment.

For example, for a bilingual sentence pair S and T,

S(Chinese) 录入了语义信息的词典称为语义词典.

T(Japanese) 意味情報を記載した辞書を意味辞書とよぶ。

(Thesaurus that contains semantic information is called semantic thesaurus.)

We can segment S and T to proper single word unit like below:

S 录入 了 语义 信息 的 词典 称为 语义 词典.

T 意味 情報 を 記載 した 辞書 を 意味 辞書 と よぶ。

After word alignment, we can get translation equivalents with a translation probability between single bilingual words such as 意味/语义/0.9(semantic), 辞書/词典/0.87(dictionary or thesaurus), etc.

If we know that “意味辞書” and “语义词典(semantic thesaurus)” are multi-words terms, we can segment S and T like below:

S 录入 了 语义 信息 的 词典 称为 语义词典.

T 意味 情報 を 記載 した 辞書 を 意味辞書 と よぶ。

The translation equivalents of word alignment on the sentences pairs will be like 意味辞書/语义词典/0.8, 辞書/词典/0.92, etc.

We can see from the examples that if the multi-words terms in both the source and target language are segmented as one lexical unit, we may easily align the multi-words terms.

The question is how to determine the multi-words terms that should be treated as one lexical unit. In our approach, all the terms extracted by term extractor are treated as one lexical unit.

In general, we first run standard word segmentation on corpus. The lexical item of the segmented corpus usually is a single word. Then we extract multi-words term and combine them to one segment unit. For Chinese and Japanese, we just delete the space of multi-words terms and make them look like a single word. This process makes several sing-word lexical units one multi-words lexical unit. We call this process re-segmentation.

2.2.2 Term candidates alignment

We assume that there is only a little imbalance between term candidates extracted in both languages. After re-segmentation, we have a bilingual segmented corpus in which multi-words terms are segmented as one lexical unit. In this step, we used an existed word alignment tool Giza++ (Och et al. [9], Och et al. [10]) to align the bilingual corpus and get a set of multi-words translation equivalents with an alignment probability.

2.2.3 Smoothing

2.2.3.1 Word association score

After the previous step, we get a set of bilingual terms with alignment probability. But the term alignment results still contain noises caused by the alignment error, segmentation error and the term extraction imbalance, etc.

There is an assumption that if two phrases are translation of each other, then one word of source phrase is likely to be the translation of, or at least associated to, one word of target phrase (Dailie 1994[7]).

Based on this assumption, we define a word association score between a pair of bilingual term candidates (S, T) as follows:

$$assoc(S,T) = \frac{\sum_{i,j} assoc(words_i, wordt_j)}{\max(length(S), length(T))} \quad (1)$$

where *assoc* means "association score of", $words_i$ denotes one word composing the candidate term *S* and $wordt_j$ denotes one word composing the term *T*, and *length* stands for the number of words composing a term.

In our approach, the words associations come from word alignment result instead of an existing bilingual lexicon. We use the following steps to compute the word association score of S and T.

1. We firstly align the bilingual corpus, which is not treated by re-segmentation process and its lexical unit is single word, by Giza++ tool. The alignment probability generated by Giza++ between $words_i$ and $wordt_j$ is used as $assoc(words_i, wordt_j)$.

2. Then we sum all of the possible $assoc(words_i, wordt_j)$ between the words composing the term candidates.

3. Finally we normalize the sum by the maximum length between source term and target term to get the associate score *assoc(S,T)*.

In the step2, we use a dynamic process to find the single-word translation candidates. First we define an alignment set Q = $\{p_{ij}: p_{ij} = p(wordt_j|words_i)\}$, where $p(wordt_j|words_i)$ is the alignment probability between $words_i$ and $wordt_j$. The algorithm we used as below.

Algorithm 1.

Input Q: $\{p_{ij}: p_{ij} = p(wordt_j|words_i)\}$

1: set sum to zero

2: **while** Q is not null

3: find the maximum f = max(p_{ij})

4: **if** f is greater than zero

5: add f to sum

6: **else**

7: end while

8: remove $p_{i.}$ and $p_{.j}$ from Q

9: return sum

For example, for a term pair (S,T) in which S is an Chinese term "无(no,un-) 监督(supervised) 学习(learning)" and T is a Japanese term "教師(teacher) なし(no) 学習(learning)", the alignment probabilities of single words generated by Giza++ are listed in Table1.

Table1. Word associations of the example term pair

Alignment probability		Words composing Japanese Term		
		教師	なし	学習
Words composing Chinese Term	无		0.19	
	监督	0.017		
	学习	0.15		0.97

First, we extract the word pair with the highest word alignment probability, (学习/学習) as the most probable associated word pair. Then we extract word pair with the highest alignment probability, 无/なし from the left word pairs. We repeat this process iteratively until we can't find any possible associated word pair. The word association score of example term candidates is (0.97+0.19+0.017)/3 = 0.4.

2.2.3.2 Final translation score

Finally we combine alignment score and word association score as follows:

$$sim(S,T) = w_1 align(S,T) + w_2 assoc(S,T) \ (w_1 + w_2 = 1) \quad (2)$$

here *align(S,T)* is the alignment score of term pair *S* and *T*, and *assoc(S,T)* is the word association score of *S* and *T*, and w_1 and w_2 are their corresponding weight. This is the final translation score of term *S* and its translation *T* calculated by our method.

3. Experiment and Evaluation

3.1 Experiment

For experiment, we used a Chinese-Japanese parallel corpus on information science. This corpus contains 378,132 sentences pairs. We selected 200,000 sentences pairs to extract terms, and used the rest of the corpus for evaluation.

We segmented the Japanese corpus using CHASEN[1] and the Chinese Corpus using ICTCLAS[2]. As for monolingual term extractor, GenSen-Web[3] (Yoshida et al.[8]) is used to extract Chinese and Japanese term candidates. Giza++ is applied to align the bilingual corpus. The both parameters w_1 and w_2 in (2) are set to 0.5.

The number of bilingual term candidates and monolingual term candidates extracted are given in Table2.

Table 2. Numbers of terms candidates

Language	Number of term candidates
Japanese	89,665
Chinese	63,247
Bilingual Terms pairs	45,193

[1] http://chasen-legacy.sourceforge.jp/

[2] http://www.ictclas.org/

[3] http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html

Figure 1 shows the distribution of translation score and the number of bilingual terms.

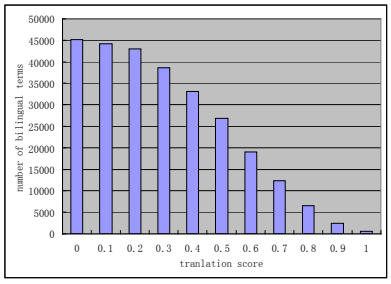

Figure 1. Distribution of bilingual terms

3.2 Evaluation

3.2.1 Evaluation Method

The number of bilingual terms extracted is too large to evaluate manually and it is hard to find a gold standard to evaluate all of the bilingual terms. Instead of sampling a few terms, we use a Statistical Machine Translation (SMT) system to evaluate our result.

Phrases are basic translation unit in phrase-based SMT system. The more accurate phrase table is, the more accurate the translation will be. Thus, in our evaluation, we utilize the bilingual terms extracted by our method as additional phrase table into the phrase-table of SMT system and compare the translation quality with our additional phrase table and without our additional phrase table to evaluate the accuracy of bilingual terms extracted.

We select the Moses toolkit (Koehn et al.[11]) to build the phrase-based SMT system and the tool Giza++ is applied as alignment tool. The SRILM toolkit (Stolcke et al.[12]) is used to build language model. The BLEU is adopted to measure the translation quality.

The training corpus is the same bilingual corpus that we used to extract bilingual terms which has 200,000 sentences pairs. The test data comes from the rest of the same bilingual corpus. We selected 100 sentences pairs from the remaining 178,132 sentences pairs as the test data. The corpus of language model comes from a Chinese information corpus that contains 760,000 sentences.

3.2.2 Results

We first compare the precision of top 20 term pairs before the smoothing and after smoothing manually. We can see from Table3 that the precision is improved from 25% to 90% after smoothing.

Table 3. Precision before and after smoothing

Precision	Before Smoothing	After Smoothing
Top 20 pairs	0.25	0.90

Then we add bilingual terms extracted to phrase table of Moses. To find the influences of bilingual terms with different translation score, we introduce the bilingual terms with different translation scores respectively into phrase table of Moses. To make sure that our phrase table is used in translation, the translation probabilities of our phrase table are all set to 1. The baseline BLUE score is obtained by using the original Moses phrase table. The evaluation result is listed in Table4.

Table4. Comparison of various translation scores

Translation Score	BLEU Score(%)
Baseline	26.75
>=0.3	27.13
>=0.4	27.16
>=0.5	**27.23**
>=0.6	27.19
>=0.7	26.73
>=0.8	26.72

As is illustrated in Table4, we computed the BLEU score with different phrase table. We can see from the table that the BLEU score is improved by up to 0.48 point with the case of additional phrase table where a translation score is larger than 0.5.

When we introduce our phrase tables with translation score lower than 0.5, we can see even though the BLEU scores decreased, they are still higher than baseline of BLEU score. When we use phrase tables with a translation score higher than 0.5, the BLEU scores are increased firstly then decrease. The reason is that for additional phrase table with higher translation score, the accuracy of phrase table is high but the number of translation pairs in effect is fewer. On the contrary, for additional phrase table with lower translation score, the number of useful translation pairs become larger but the accuracy of phrase table decreases.

4. Conclusion

We present a new approach to extract bilingual terms especially multi-words terms from a bilingual corpus without the need of an initial word-to-word bilingual lexicon. Through an existing monolingual term extractor and a re-segmentation process, we obtain bilingual term candidates just by a word alignment process. After that we use a smoothing process that is based on word association scores to obtain a higher precision.

The results of the experiments indicate that the bilingual term pairs extracted by our method have a higher precision than that extracted by Moses.

Our approach is initially based on a Chinese-Japanese parallel corpus, but it can be easily applied to any languages.

5. REFERENCES

[1] Bond, F., Nichols, Chang, Z.Q. and Uchimoto, K. 2008. Extracting Bilingual Terms from Mainly Monolingual Data. In 14th Annual Meeting of the Association for Natural Language Processing, (Tokyo, Japan).

[2] Nazar, R., Wanner, L. and Vivaldi, J. Two Step Flow in Bilingual Lexicon Extraction from Unrelated Corpora, Conference of EAMT 2008, (Hamburg, Germany, September 2008). 140-149.

[3] Déjean, H., Gaussier, E. and Sadat, F. Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In proceedings of COLING 2002, (Taipei, Taiwan, Aug. 24-Sep. 1, 2002), 218-224.

[4] Tsunakawa, T., Naoaki, O. and Tsujii J. Buliding a Bilingual Lexicon Using Phrase-based Statistical Machine Translation via a Pivot Language. In the Proceedings of the 22nd COLING 2008: Companion volume Posters and Demonstrations. 127 – 130.

[5] Tonoike, M., Kida, M., Takagi, T., Sasaki, Y., Utsuro, T. and Sato, S. A Comparative Study on Compositional Translation Estimation using a Domain/Topic-Specific Corpus collected from the Web. Proceedings of the 2nd International Workshop on Web as Corpus, EACL-2006 Workshop, pp.11-18, Trento, Italy, 2006.

[6] Ha, L. A., Fernandez, G., Mitkov, R. and Corpas, G. Mutual bilingual terminology extraction. In proceedings of LREC 2008, Marrakesh, Morocco .

[7] Daille, B., Gaussier and Lange, J. M. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In proceeding of the 15th International Conference on Computational Linguistics (Kyoto, Japan, 1994), 515-521.

[8] Yoshida, M and Nakagawa, H. Automatic Term Extraction based on Perplexity of Compound Words. In proceeding of IJCNLP 2005. LNAI 3651, 269-279.

[9] Och, Franz Josef, 2000. Giza++: Training of statistical translation models. http://www.isi.edu/~och/GIZA++.html.

[10] F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1), March.

[11] Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C.Moran, R. Zens, C. Dyer, O.Bojar, A.Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In proceedings of 45th Annual Meeting of the Association for Computational Linguaistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177 – 180, Prague, Czech Republic.

[12] A. Stolcke. SRILM – an extensible language modeling toolkit. Proceeding of International Conference on Spoken Language Processing, 2002.