Word Alignment Based Bilingual Terminology Extraction from a Chinese-Japanese Parallel Corpus

Xiaorong Fan Nobuyuki Shimizu Hiroshi Nakagawa
Graduate School of Interdisciplinary Information Studies, University of Tokyo
Information Technology Center, University of Tokyo

ABSTRACT

The automatic extraction of bilingual single-word terms (SWTs) has been very successful, but for multi-word terms (MWTs), the precision is far from enough. This paper proposes a new approach for the automatic extraction of bilingual MWTs from a bilingual parallel corpus. We combine existing monolingual term extractor and a word alignment tool to extract bilingual terms. We introduced a re-segmentation to process a MWT as a single lexical unit so that it can be treated as a single unit by word alignment. We also improved the extraction rules of MWTs for existing term extractor and the experiment shows that our improvement is valid. We obtained a good precision and an improved BLEU score in our experiment based on a Chinese-Japanese parallel corpus.

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 recent years, the automatic extraction of bilingual single-word terms (SWTs) has been very successful, but remains disappointing for multi-word terms (MWTs). In our approach, we propose a new approach to extract bilingual MWTs from a bilingual Chinese-Japanese parallel corpus.

2. Background

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 MWTs extraction from bilingual parallel corpora. Ha et al. [6] firstly extract MWTs from corpora of distinct languages and then use a contingency table and log-likelihood to measure how likely a pair of MWT candidate is to be a correct pair. Ddilie et al. [7] extract MWTs independently and then find correspondences between candidates across languages using frequency count and bilingual associations of single words.

3. Our solution

In this paper, we propose a two-stage extraction process. In the first stage, Chinese and Japanese MWTs are extracted respectively as other previous works did. In the second stage, we propose a novel word-alignment-based bilingual matching process to produce bilingual MWTs.

3.1 Monolingual Term Extraction

First we need to extract term candidates from both the source and target languages. In our approach, we choose an existing monolingual term extraction tool that can extract MWTs from both Chinese and Japanese. We did some improvement to GenSen-Web Chinese version.

3.2 Term alignment

Unlike the other works described before, we treat a MWT as a single lexical unit but not the combination of several lexical units, and then extract the bilingual MWT candidates directly from the bilingual parallel corpus.

We proposed a three-step procedure to align term candidates.

3.2.1 Re-Segmentation

Asian languages, such as Japanese and Chinese, have an important characteristic that there are no spaces between characters.

As well known, the main task of word to word alignment is to identify translation equivalents between lexical items in bilingual parallel texts. In English, the lexical item is a word, but in Japanese and Chinese, the lexical item is a series of continuous characters that is recognized by word segmentation process. It can be a multi-words unit or a single word unit.

If the lexical items in the corpus are single word units, the result of word alignment on such kind of corpus is a set of single word translation equivalents. On the other hand, multi-word 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 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 word alignment result is a set of translation equivalents with 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 treat MWTs “意味辞書” and “语义词典(semantic thesaurus)” like a single lexical 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。

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 MWTs of both source and target language are treated as lexical unit, the MWTs.

In our approach, we combine all the MWTs that are extracted by term extractor as one lexical unit. For Chinese and Japanese, we just deleted the spaces of multi-words terms and make them look like single word. This process looks like a word segmentation process so we call this process re-segmentation.

3.2.2 Term candidates alignment

In previous step, re-segmentation, we treated the bilingual corpus in which MWTs are segmented as one lexical unit. In this step, we used an existed word alignment tool Giza++ (Och et al. 7]) to align the bilingual corpus and get a set of bilingual MWT candidates of alignment probability.

3.2.3 Word association score

The bilingual MWT candidates from the previous step still contain noise because of the alignment error and segmentation error, etc. There is an assumption that if two MWTs are translation of one another, then one word of source MWT is likely to be the translation of or at least associated to one word of target MWT (Dailie 1994[5]).

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

$$assoc(S,T) = \frac{\Delta}{\max(length(S), length(T))}$$

the algorithm we used to calculate Δ as below

Algorithm 1.

```
Input   Q:{$p_{ij}$: $p_{ij}$ =$p(wordt_j|words_i)$}
set sum to zero
while Q  is not null
    find the maximum f = max($p_{ij}$)
     if f is greater than zero
          add f to sum
     else
          end while
     remove $p_{i.}$ and $p_{.j}$ from Q
return sum
```

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

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

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 MWT *S* and its translation *T* calculated by our method.

4. Experiment and Evaluation

Experiments are carried out on a Chinese-Japanese parallel corpus consisting of 378,132 sentence pairs. GenSen-Web (Yoshida et al.[6]), a Term extractor, was used to extract Chinese and Japanese terms candidates. Giza++ is selected as word alignment tool. We extracted 129,127 bilingual term pairs.

We first compare the precision of top 20 term pairs before and after using word association scores manually. We can see from Table1 that the precision is improved from 25% to 90% after using word association score.

Then we selected first 3000 best pairs and evaluate their precision manually. We can see from Figure.1 that the average precision is about 0%. It is acceptable. At last, we utilize extracted bilingual terms as additional phrase table into the phrase-table of SMT system and compare the translation quality with and without our additional phrase table to evaluate the accuracy of bilingual terms by our method. We select the Moses toolkit (Koehn et al.[8]) to build the phrase-based SMT system. The BLEU is adopted to measure the translation quality. 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 in which the final score of each pair is larger than 0.5.

Table 1. Precision before and after using word association score

Precision	Before	After
Top 20 pairs	0.25	0.90

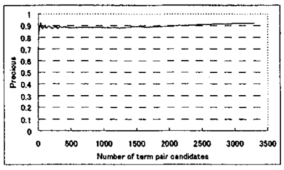

Figure 1. Precision of the top 3000 pairs

Table4. BLEU scores when adding the extracted terms with 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

5. Conclusion

We present a new approach to extract bilingual MWTs from a bilingual corpus. Through an existing monolingual term extractor and a re-segmentation process, we obtain bilingual MWTs candidates directly by a word alignment process. After that we use word association scores to obtain a higher precision.

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

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

[5] 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.

[6] Yoshida, M and Nakagawa, H. Automatic Term Extraction based on Perplexity of Compound Words, In proceedings of IJCNLP 2005. LNAI 3651, 269-27

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

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