Automatic Extraction of Chinese-Japanese Keyword Pairs via English using Research Paper Abstract

Denny Cahyadi, Toshiaki Nakazawa, and Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
{denny, nakazawa}@nlp.ist.i.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

Abstract

The abstracts of research papers are usually available for free. Some of them contain keywords in two different languages (such as Chinese-English or Japanese-English). We assume these keywords are high quality technical term pairs as they are written by humans. In this paper we present our method of extracting Chinese and Japanese keyword pairs via English collected from research paper abstracts. We first align Chinese-English and Japanese-English keywords within a document, based on their appearance order. We then align Chinese-English and Japanese-English keyword pairs using a simple pivoting technique. Our experiment shows the possibility to collect a large number of keyword pairs, although there is still room for improvement.

1 Introduction

Technical term dictionaries are useful for cross-lingual IR and machine translation tasks in technical domains (e.g. patent translation). Collecting technical terms automatically is harder than collecting general terms because they rarely appear in general documents. We may be able to collect such terms from specific documents such as research papers. However, these documents are usually not available in full for free. On the other hand, we found that some abstracts of research papers are available for free. The abstracts usually contain keywords section. Some of research papers whose original language is not English may contain keywords in two languages (original language and English). Since these sections are written by experts, we assume that keywords in these sections are high quality technical terms. Based on this assumption, we collect these keywords to compile a technical term bilingual dictionary.

Collecting keywords from non-English papers allows us to obtain non-English-English technical term pairs. We would like to extend this possibility into collecting non-English-non-English term pairs (e.g. Chinese-Japanese). This is possible using a pivoting technique. We plan to collect Chinese-English and Japanese-English keywords pairs from Chinese and Japanese research papers first and then use English as a pivot to align Chinese and Japanese keywords.

2 Related Work

An experiment to collect Chinese-English technical term pairs from abstracts of research papers has been done by (Ren et al., 2010). They collected the keyword section (containing both Chinese and English keywords) and domain ID section of research papers from CNKI research portal[1]. They aligned Chinese and English keywords based on their appearance order. The first appearing Chinese keyword is aligned to the first appearing English keyword. From this step, they could collect a large number of keyword pairs. Finally, to ensure the quality of technical terms they collected, they filtered some keywords which are extremely general. It was done based on term frequency and inverse domain frequency score. Their experiment showed promising results.

[1] http://www.cnki.net/

Figure 1: Japanese and English keywords at a CiNii page

We conducted a similar experiment for Japanese research papers. We found that the CiNii research portal[2] provides abstracts and keywords of Japanese research papers for free. We also found that a large number of keywords are written in both Japanese and English. Figure 1 shows a typical page of CiNii which contains keywords in English and Japanese.

3 Proposed Workflow

In order to obtain Chinese-Japanese technical term pairs, we propose a workflow as shown in Figure 2.

3.1 Crawling and Extracting keywords

First, we crawled CNKI research portal to collect Chinese-English keywords. We crawled them by following the links from research paper index page, instead of using seed keywords and BFS method as (Ren et al., 2010) did. We found that papers with English-only title are very likely to contain English-only keywords. Therefore we focus on crawling papers which have at least one Chinese character in the title. We also found that most of papers in the Natural Science domain contain bilingual keywords. For the next step, we cleaned up the HTML and extracted only the Chinese and English keyword part.

For the Japanese part, we did not crawl CiNii, but instead we used a corpus provided by NII[3]. The content of the corpus is similar to the web version of CiNii. We extracted the Japanese and English keywords part of this corpus.

3.2 Keywords alignment within a document

For the next step, we aligned original Chinese keywords to English keyword for every single Chinese document. We did the similar step for the Japanese

[2]http://ci.nii.ac.jp/

[3]http://www.nii.ac.jp/

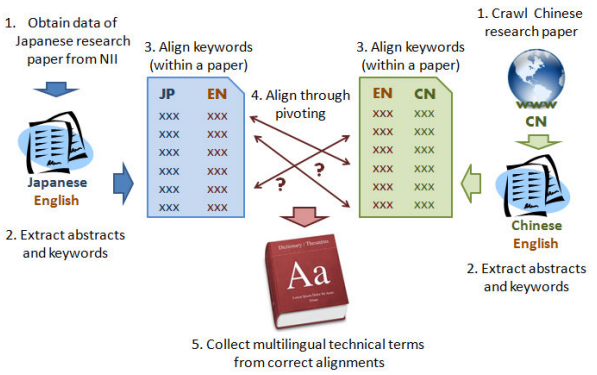

Figure 2: Workflow

documents. We did three experiments to align these keywords: 1) using alignment method proposed by (Ren et al., 2010); 2) using Giza++ (Och and Ney, 2003); 3) using Log Likelihood formula proposed by (Rapp, 1999).

First, we implemented the order-based alignment method proposed by (Ren et al., 2010). This method is based on the assumption that keywords in English are usually written in the same order as keywords in their original language. For this experiment we only used the sentences that have the same number of keywords in English and its original language. We then aligned the keywords based on their position in the keyword list. The first appearing keyword in original language is aligned to the first appearing keyword in English and so on.

We are not sure whether this ordering assumption is true for most of the papers. If many keywords are written not in the same order, we may need better alignment method which can handle different ordering between source language and English keyword. Thus, in the second experiment, we used Giza++ (Och and Ney, 2003) to align the keywords. We treated the list of keywords in a document as a sentence pair and each keyword as a single word. Most keywords consist of several words. We replaced the space between single word with an underscore.

Giza++ is an alignment tool which is designed to align words between languages which have a certain grammar. Our case is different. There is no certain rule of the order of the keywords. Each keyword can be written in different order from document to document. In order to handle this, we ran a third experiment. In this experiment we aligned the key-

	Japanese	Chinese
Total documents	4,200,000	187,237
Documents with bilingual keywords	752,945	75,398

Table 1: Statistics of the Japanese and Chinese documents.

words based on their likelihood score. We used the formula introduced in (Rapp, 1999) to calculate the score.

3.3 Chinese-Japanese Alignment using Pivoting Technique

After we aligned the keywords within each document, we aligned Chinese and Japanese keywords via English using pivoting technique. Our method is very simple; we just align keywords which have a similar English translation. We found there were minor variations in English keywords with the same meaning (e.g. *broad-band noise* and *broadband noise*). If we use an exact match, we cannot tolerate any variations (which maybe correct) and may affect the final alignment. Therefore, we use normalized edit distance score instead of exact matching. Normalized edit distance is defined as the number of operation required to convert a string into another (to make them similar) divied by its length. We allowed small variations of English keywords and treated them as the same keywords if their normalized edit distance is lower than a threshold.

4 Experiments

4.1 Setup

Table 1 shows the total number of documents we crawled and used for the experiments. Each document contains about 2 to 10 keywords in its original language and often the same number of keywords in English. For all experiments we used these 752k Japanese documents and 75k Chinese documents.

We used manually constructed technical term dictionaries to evaluate keyword pairs. We compared our keyword pairs with entries in the dictionaries and labeled our pairs as *correct*, *incorrect*, and *not found* according to the dictionaries.

4.2 Results

Table 2 shows some of our results of Chinese-English-Japanese alignment using order-based alignment method for Chinese-English and Japanese-English alignment and normalized edit distance similarity-based alignment method for English-English alignment at pivoting stage.

Some keywords are correctly aligned and some are not. Based on our analysis, incorrect alignment usually occurs in the original language-English alignment stage. As an example, in the 5th line of Table 2, on the Chinese side, 草莓 is correctly aligned to *strawberry*. However, *strawberry* is misaligned to 収穫ロボット(*harvesting robot*) on the Japanese side. As a result, the final alignment is incorrect. Further analysis shows that some of keyword lists in the documents do not follow the ordering rule. English keywords are written in a different order to their original language.

Table 3 shows the number of *correct*, *incorrect*, and *not-found* pairs according to the dictionary. From the table, we can see that most of our pairs were not found in the dictionary. We found that our technical term pairs cover a different or larger scope than the dictionary. We found that some keyword pairs that are labeled as *not found* are actually correct pairs (e.g. the bottom 3 lines of Table 2). In the future we may use human evaluation for better results.

For term pairs that can be found in the dictionary, the number of *incorrect* is high for Chinese-English pairs. We found that many English keywords in the Chinese documents are not written in the same order with Chinese keywords. We ran the other two experiments (using Giza++ and Log-likelihood) to see whether statistical method could improve the result. The number of *correct* term pairs obtained from our three experiments is shown in table 4.

For Japanese-English alignment, the results of three different alignment methods are only slightly different. Order-based alignment still gives the best result while Log-likelihood score-based alignment gives the worst. However, for Chinese-English alignment, there is significant improvement by using Giza++. We think that Giza++ may be able to find the correct alignment even though Chinese and English keywords are written in different order. Log-

Chinese	English	Japanese	Result
口蹄疫病毒	foot-and-mouth disease virus	口蹄疫ウイルス	Correct
紫外吸收光谱	ultraviolet absorption spectrum	紫外吸収スペクトル	Correct
差向异构化	epimerization	エピマー化	Correct
肠肌丛	myenteric plexus	筋層間神経叢	Correct
草莓	strawberry	収穫ロボット	Incorrect
振动控制	vibration control	可变减衰器	Incorrect
近红外光谱	near infrared reflectance spectroscopy	黒ボク土壌	Incorrect
挥发性有机化合物	volatile organic compounds	性有機化合物	Not found
双曲型偏微分方程	hyperbolic partial differential	双曲型偏微分方程式	Not found
类金属硫蛋白	metallothionein	メタロチオネイン	Not found

Table 2: The example results of Chinese-English-Japanese alignment

	JP-EN	CN-EN	CN-JP
Documents with bilingual keywords	752,945	75,398	-
Keyword pairs	308,342	105,462	39,699
Correct pairs	59,630	5,228	4,695
Incorrect pairs	20,878	12,041	15,955
Pairs not found	227,834	88,193	19,049

Table 3: Statistics of order-based alignment result according to manually-compiled dictionaries

Alignment	JP-EN	CN-EN
Order-based	59,630	5,228
Giza++	59,258	12,486
Log-likelihood	56,453	3,001

Table 4: Number of correct term pairs obtained from three alignment methods

likelihood score-based method still gives the worst alignment.

While Giza++ improves the result of Chinese-English alignment, we think that the number of correct alignments is relatively small compared to the number of all keyword pairs. Giza++ and log likelihood score-based methods may not be able to find most of pairs effectively due to data sparseness. While we obtained a large number of documents, the variation of keywords is also large. Some keywords only appear in a few documents. Since these methods highly depend on the statistical properties of the corpus, they may not very effective for our case. We need some improvement in our method.

5 Conclusion and Future Improvement

Our experiments show that it is possible to collect a large number of Chinese-Japanese technical term pairs from abstract part Chinese and Japanese research papers.

We found that simple order-based alignment is not efficient enough and statistical-based methods are facing data sparseness problem. To overcome these problems in the future, we are planning to implement multi-tier alignment method: to align each join word within each keywords first, and then to align each keyword afterwards. We are also planning to improve our evaluation method, either by human evaluation or other dictionaries with a larger scope.

References

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Association for Computational Linguistics*, 29(1):19–51.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526, Stroudsburg, PA, USA. Association for Computational Linguistics.

Feiliang Ren, Jingbo Zhu, and Huizhen Wang. 2010. Web-based technical term translation pairs mining for patent document translation. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on*, pages 1 –8, aug.