# Finding Translation Candidates from Patent Corpus

**Sayori Shimohata**
Corporate Research & Development Center,
Oki Electric Industry Co., Ltd.
2-5-7 Honmachi, Chuo-ku, Osaka,
541-0053, Japan
shimohata245@oki.com

## Abstract

This paper describes a method for retrieving technical terms and finding their translation candidates from patent corpora. The method improves the reliability of bilingual seed words that measure similarity between a target word and its translation candidates. We conducted an experiment with PAJ (Patent Abstracts of Japan), which is a collection of bilingual patent abstracts written in Japanese and English. The experiment result shows that our method achieves a precision of 53.5% and a recall of 75.4%.

## 1 Introduction

Recently, there is an increasing demand for patent translation. Translating a patent is very difficult because it includes various new terms and technical terms. Existing dictionaries are insufficient because such technical terms are highly specialized and continuously increasing.

Our goal is to provide a dictionary-building tool for patent translation. Fortunately, environment for bilingual dictionary compilation is promising in this field. There is a large collection of patent documents with IPC (International Patent Classification) code and a large number of documents have their counterparts. For example, each Japanese patent has its abstract in English and some of them have their whole translation for international application. These bilingual document pairs are not precise translation but approximately same contents; we define them as comparable (loosely-parallel) corpora while we define precise translation pairs as parallel corpora.

Various approaches for automatically retrieving translation pairs from corpora have been proposed. While most of them use parallel corpora (Kupiec 93; Dagan and Church 94; Smadja 94; and Kitamura 04) and achieve high precision, several attempts have been conducted to find translation candidates from comparable corpora (Fung and McKeown 97; Fung and Yee 98; and Rapp 99). However, their precisions are still low.

In this paper, we present an algorithm for finding bilingual technical terms using patent documents.

## 2 Algorithm

### 2.1 Basic idea

Unlike in parallel corpora, the position of a word in a text does not give us information about its translation in the other language. Fung and McKeown 97 assumes that if a term A is closely correlated with another word B in text T, then its counterpart in the other language A' is also closely associated with B', the counterpart of word B, in T'. In their method, word association is measured with bilingual seed words, which are bilingual dictionary entries with certain frequencies in target bilingual corpora. We follow the idea and improve the reliability of seed words. Figure 1 shows an overall process of the proposed method.

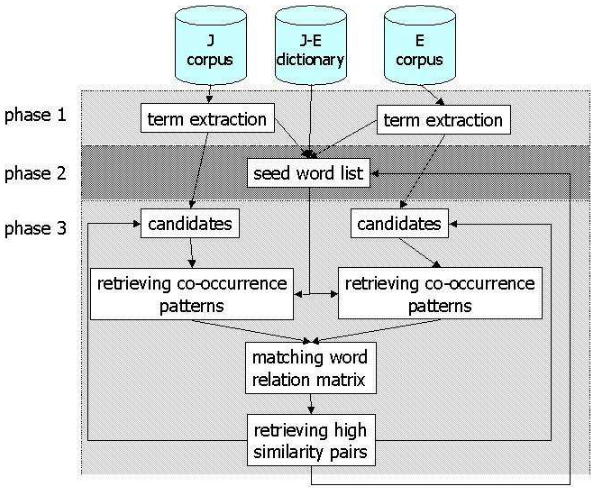
J corpus
J-E dictionary
E corpus
phase 1
phase 2
phase 3
term extraction
term extraction
seed word list
candidates
candidates
retrieving co-occurrence patterns
retrieving co-occurrence patterns
matching word relation matrix
retrieving high similarity pairs

Figure 1 Overview of the Process

The method consists of 3 phases. In the first phase, seed words using a bilingual dictionary are listed. This process is very important in our method. It compares the similarity between each Japanese seed word candidates and its English counterparts, and retrieves only the word pairs with a certain similarity. Then, in the second phase, candidates for technical terms are extracted from each monolingual corpus. And in the third phase, each Japanese technical term candidate is compared with each English candidate, and high similarity word pairs as bilingual technical terms are retrieved.

In the following, each step is described in detail.

### 2.2 Finding bilingual seed words

In Fung 97, seed words are dictionary entries which occur at midrange frequency in the both corpora and have a unique translation in both directions (or at least in one direction). Our method relaxes the restrictions but introduces the word association similarity in both corpora instead. That means word pairs are extracted as seed words in case that they are used in a fixed meaning even if the word has multiple translation candidates.

The seed word extraction is conducted in the following process:

1. Extract bilingual dictionary entries $(D(S_1,T_1),D(S_2,T_2),...D(S_n,T_n))$ , whose frequencies in both corpora are over a given threshold.
2. For each dictionary entry $D(S_i,T_i)$, create a co-occurrence matrix with other dictionary entries. The co-occurrence matrix is obtained from the word association between the target word and given seed words in a specific segment, for example, in a sentence. The co-occurrence matrix for $S_i$ and other dictionary entries $(S_1,S_2,...S_n)$ is represented by $(W(S_i,S_1),W(S_i,S_2),...,W(S_i,S_n))$ . $W(w_i,w_j)$ , here, is the weighted mutual information derived from:
$$\Pr(w_i,w_j)\cdot\log_2\frac{\Pr(w_i,w_j)}{\Pr(w_i)\cdot\Pr(w_j)} \quad (1).$$
3. Compare the co-occurrence matrixes of each entry and its translations and retrieve them if their co-occurrence matrixes are similar over a given threshold. Similarity of dictionary entry $D(S_i,T_i)$ is calculated with the following formula:
$$sim(S_i,T_i)=\sqrt{\sum_{1<j<n}(W(S_i,S_j)-W(T_i,T_j))^2} \quad (2).$$

Thus, only the word pairs whose co-occurrence matrixes are similar in both languages are retrieved and used as bilingual seed words.

### 2.3 Finding translation candidates

1. To find a translation equivalent, our method employs the same process as finding bilingual seed words.
2. Extract technical terms $(s_1,s_2,...s_m)$ from a corpus written in language *S* and $(t_1,t_2,...t_n)$ from a corpus written in language *T*. In this step, every possible n-grams are retrieved from a corpus and filtered with frequency, dispersion of adjacent words (Shimohata 97), and some IR techniques (TF/IDF).
3. For each technical term, create a co-occurrence matrix with the seed words retrieved in the previous phase.
4. Matching a co-occurrence matrix for $s_i$ with co-occurrence matrixes for $t_k(1 \le k \le n)$, retrieve $t_k$ whose co-occurrence matrix is similar with $s_i$'s over a given threshold as a translation equivalent. For the similarity calculation, we used formula (2).

## 3 Experiment

### 3.1 Data profile

We conducted an experiment under the following condition.

- Test corpora PAJ (C12N)
    - 11781 abstracts
    - 38481 Japanese sentences
    - 35343 English sentences
- Reference dictionary
    - JAPIO dictionary (same domain)
    - 4789 entries
- Dictionary
    - EDICT 173 thousand entries

Test corpora are PAJ(Patent Abstracts of Japan). PAJ corpus is a collection of Japanese patents and their English translations, which are loosely correspondent to the original texts. Each text is composed of "title of invention" and "summary". "Title of invention" is completely parallel but "summary" is not necessarily parallel. The corpus domain is biochemistry whose IPC code is C12N.

As a reference, the experiment used JAPIO dictionary, which was a manually compiled bilingual dictionary for machine translation. For 57 entries of JAPIO dictionary whose frequencies were more than 100 in test corpus, we evaluated whether the method presented appropriate translation candidates or not.

### 3.2 Seed word selection

In the seed word extraction process, we used the Japanese-English online dictionary EDICT, thereby extracting 129 bilingual seed words whose frequencies in both corpora were more than 100. As our method did not restrict seed word pairs to one-to-one corresponding translations, more than half of them had alternative translations. Table 1 shows an example of seed words which had more than one translations.

| selected seed word | | dictionary entries |
|---|---|---|
| 昆虫 | insect | insect, bug |
| 分子 | molecule | numerator, molecule |
| 大腸菌 | Escherichia coli | Escherichia coli, colon bacterium |
| コード | code | code, cord, chord |
| 腫瘍 | tumour | neoplasm, tumour |
| 代謝 | metabolism | renewal, regeneration, metabolism |
| 宿主 | host | host, landlord, innkeeper |
| 粒子 | particle | particle, grain |
| 感染 | infection | infection, contagion |
| 投与 | administer | prescribe medicine, administer |

Table 1 Example of Seed Words with Multiple Translations

| *Candidates J* | *Freq.* | *Candidates E* | *Freq.* |
|---|---|---|---|
| 遺伝子 | 14203 | acid | 16929 |
| 配列 | 13592 | amino acid | 8347 |
| 細胞 | 11143 | gene | 16333 |
| DNA | 8899 | sequence | 16918 |
| 酵素 | 6794 | protein | 11459 |
| 製造 | 6379 | DNA | 12749 |
| 培養 | 6328 | cell | 13193 |
| をコードする | 5942 | solution | 11134 |
| 活性 | 5719 | amino acid sequence | 6049 |
| タンパク質 | 5405 | microorganism | 4809 |
| アミノ酸 | 5277 | culture | 7892 |
| 新規 | 5166 | encode | 4531 |
| 微生物 | 5144 | polypeptide | 4422 |
| 発現 | 4756 | formula | 4359 |
| ポリペプチド | 4105 | comprise | 4217 |
| 蛋白 質 | 3998 | nucleic acid | 3960 |
| 本 発明 | 3787 | vector | 3921 |
| ヒト | 3555 | enzyme | 5649 |
| 核酸 | 3535 | bacterium | 3952 |

Table 2Result of Technical Term Extraction

### 3.3 Technical term extraction

In the term extraction process, 1038 Japanese word sequences and 1034 English word sequences were extracted. Table 2 shows an output of the term extraction process. Among 57 JAPIO dictionary entries, whose frequencies were more than 100 in the test corpora, 43 entries were retrieved in both languages. That means recall rate for entries over 100 frequencies is 75.4%(43/57).

### 3.4 Translation candidate matching

The translation matching was performed between 1038 Japanese terms and 1034 English terms including 43 JAPIO entries. For 43 Japanese entries, 23 English correspondents (53.5%) were ranked top translation candidates and 35 entries (81.4%) were ranked within top 10. Table 3 shows a result of the translation matching process. The terms indicated in boldface type are translations in JAPIO dictionary.

| Japanese Entries | English Candidates |
|---|---|
| **培養** | **culture** |
| | cultured |
| | objective |
| | culture medium |
| | medium |
| **タンパク質** | **protein** |
| | express |
| | expression |
| | DNA |
| | sequence |
| **アミノ酸** | protein |
| | **amino acid** |
| | formula |
| | DNA |
| | DNA encode |
| **発現** | gene |
| | express |
| | **expression** |
| | protein |
| | transducing |
| **領域** | **region** |
| | domain |
| | construct |
| | link |
| | gene |
| **変異** | variant |
| | mutation |
| | sequence of formula |
| | mutant |
| | gene |
| **分子** | bond |
| | acid |
| | solution |
| | terminal |
| | **molecule** |

Table 3 Result of Candidate Matching

Incorrect matching is classified broadly into 3 categories.

- Either English or Japanese JAPIO entries are common terms.
  Ex.) “application”, “template”
- Some terms have two or more possible translations.

Ex.) “変異”
“variation” (JAPIO dic. entry)
“variant” (top in our method)

- Variation in Japanese spelling

Ex.) “recombinant”
“組み換え”(JAPIO dic. entry)
“組換え” (our method)

In table 3, for example, a desired translation of “変異”, “variation”, is not ranked in the top 5 candidates. This is because “変異” is frequently translated into “variant” in the corpora as well as “variation”. In addition, “変異” appears as a part of “突然変異”(“mutation”). Thus, there is a difficulty in identifying translations for the terms which have broad meaning and variations in expression.

### 3.5 Comparison to previous work

Under the threshold described in the previous section, we obtaind 38 seed words with Fung’s method. We set the number of seed words to 38 and made comparison between our method and Fung’s method. In our method, 38 seed words were selected in descending order of similarity.

Comparing the two, overlapping words are only 8. The difference comes from the constraint on the multiple translations. Fung’s method tends to extract compound words because they have less translation ambiguity. For example, our method took “cancer”-“癌”[1] whereas Fung’s method took “cancer cell”-“癌 細胞”.

Table 4 is a summary of experiments. While Fung’s method achieved 34.9% of precision for top candidates and 58.1% for top 10 candidates, our method achieved 39.5% and 76.7% respectively. This suggests that adapting the seed word to the corpus is effective to boost term matching quality.

## 4 Related work

Attempts at using non-parallel corpora for terminology translation are very few (Rapp, 1995; Fung and McKeown, 1997; Fung and Yee, 1998). Among these, (Rapp, 1995) proposes that the association between a word and its close collocates is similar in any language. (Fung and McKeown, 1997) and (Fung and Yee, 1998) suggest that the associations between a word and many seed words are also similar in another language. These attempts are highly suggestive, but they haven't achieved in practical use.

| | Fung97 | experiment1 | experiment2 |
|---|---|---|---|
| # of seed word | 38 | 129 | 38 |
| # of top | 15 | 23 | 17 |
| # of top 10 | 25 | 35 | 33 |

Table 4 Summary of Experiments

In accordance with precedent studies, we focused on the seed word refining process and improved the quality of candidate ranking. The method supports whole process for dictionary building, ranging from technical term identification to bilingual term matching.

## 5 Conclusion

The handling of technical terms is considered to be a very important aspect when translating highly specialized documents. In this paper, a method for extracting technical terms and finding their translations from comparable (loosely parallel) bilingual corpora has been described.

The method can retrieve technical terms from each monolingual corpus with IR techniques and align them by comparing a similarity of co-occurrence patterns between retrieved terms and bilingual seed words. After having applied the method for patent documents, we were able to achieve high accuracy compared to the former approaches. By using the method, time and effort for dictionary building will be reduced and the quality of the dictionary will be improved.

As future work, we will develop a user-friendly dictionary-building tool with the proposed method. More specifically, we will try to find a methodology for measuring output certainty and add the certainty information to the output.

## 6 Acknowledgements

Our thanks go to Japan Patent Information Organization (JAPIO) for permitting to use the PAJ corpus in the experiment and the members of AAMT-JAPIO Special Interest Group on Patent Translation for their valuable comments and discussions. This work was supported by National Institute of Information and Communications Technology (NICT) of Japan.

[1] Both “cancer” and “癌” have multiple translations. “Cancer” can be translated into “癌” (“malignant tumor”)and “蟹” (“crab”), while “癌” can be translated into “cancer” and “curse”.

## References

Jim Breen. 2003. *The EDICT Project*, http://www.csse.monash.edu.au/~jwb/edict.html.

Ido Dagan and Kenneth W. Church. 1994. *Termight: Identifying and translating technical terminology.* In Proceedings of the 4th Conference on Applied Natural Language Processing, pp.34-40.

Pascale Fung and Kathleen McKeown. 1997. *Finding terminology translations from non-parallel corpora.* In Proceedings of 5th Annual Workshop on Very Large Corpora, pages 192—202.

Pascale Fung and Lo Yuen Yee. 1998. *An IR approach for translating new words from nonparallel, comparable texts.* In Proceedings of the 36th conference on Association for Computational Linguistics, pp.414-420.

Mihoko Kitamura and Yuji Matsumoto. 2004. *Practical Translation Pattern Acquisition from Combined Language Resources.* In Proceedings of The First International Joint Conference on Natural Language Processing, pp.652--659.

Julian Kupiec. 1993. *An algorithm for finding noun phrase correspondences in bilingual corpora.* In proceedings of the 31st Annual Conference of the Association for Computational Linguistics, pp. 17-22.

Reinhard Rapp. 1999. *Automatic Identification of Word Translations from Unrelated English and German Corpora.* In Proceedings of 37th Annual Meeting of the Association for Computational Linguistics. pp.5190--526.

Sayori Shimohata, et al. 1997. *Retrieving Collocations by Co-occurrences and Word Order Constraints.* In Proceedings of 35th Annual Meeting of the Association for Computational Linguistics, pp.476--481.

Frank Smadja and Kathleen R. McKeown. 1996. *Translating Collocations for Bilingual Lexicons: A Statistical Approach*. In Computational Linguistics, Vol.22, No.1, pp.1—38.