PRIME: A System for Multi-lingual Patent Retrieval

Shigeto Higuchi[†], Masatoshi Fukui[†], Atsushi Fujii[††,†††], and Tetsuya Ishikawa[††]

[†]PATOLIS Corporation
2-4-29 Shiohama Koto-ku, 135-0043, Japan
[††]University of Library and Information Science
1-2 Kasuga Tsukuba, 305-8550, Japan
[†††]CREST, Japan Science and Technology Corporation
fujii@ulis.ac.jp

Abstract

Given the growing number of patents filed in multiple countries, users are interested in retrieving patents across languages. We propose a multi-lingual patent retrieval system, which translates a user query into the target language, searches a multilingual database for patents relevant to the query, and improves the browsing efficiency by way of machine translation and clustering. Our system also extracts new translations from patent families consisting of comparable patents, to enhance the translation dictionary.

Keywords

multi-lingual patent retrieval, machine translation, document clustering, translation extraction, patent families

1 Introduction

Given the growing number of patents filed in multiple countries, it is feasible that users are interested in retrieving patent information across languages. However, many users find it difficult to perform patent retrieval (i.e., formulating queries, searching databases for relevant patents, and browsing retrieved patents) in foreign languages.

To counter this problem, cross-language information retrieval (CLIR), where queries in one language are submitted to retrieve documents in another language, can be an effective solution. CLIR has of late become one of the major topics within the information retrieval and natural language processing communities. In fact, a number of methods/systems for CLIR have been proposed.

Since by definition queries and documents are in different languages, queries and documents need to be standardized into a common representation, so that monolingual retrieval techniques can be applied. From this point of view, existing CLIR methods are classified into the following three fundamental categories.

The first method translates queries into the document language (Ballesteros and Croft, 1998; Fujii and Ishikawa, To appear; Nie et al., 1999), and the second method translates documents into the query language (McCarley, 1999; Oard, 1998). The third method projects both queries and documents into a language-independent space by way of thesaurus classes (Gonzalo et al., 1998; Salton, 1970) and latent semantic indexing (Carbonell et al., 1997; Littman et al., 1998).

Among those above methods, the first one (i.e., query translation method) is preferable in terms of implementation cost, because this approach can simply be combined with existing monolingual retrieval systems.

Following a query translation method (Fujii and Ishikawa, 1999; Fujii and Ishikawa, To appear), we previously proposed a Japanese/English cross-language patent retrieval system (Fukui et al., 2000), where users submit queries in either Japanese or English to retrieve patents in the other language. In either case, the target database is monolingual.

However, since users are not always sure as to which language database contains patents relevant to their information need, it is effective to retrieve patents in multiple languages *simultaneously*. This process, which we shall call "multi-lingual information retrieval (MLIR)", is an extension of CLIR. In this paper, we propose a Japanese/English multi-lingual patent retrieval system called "PRIME" (Patent Retrieval In Multi-lingual Environment),

The design of our system is based on that for technical documents (Fujii and Ishikawa, 2001), which combines query translation, document retrieval, document translation and clustering modules (Section 2).

Additionally, in this paper we newly introduce a module for enhancing a dictionary used for the query translation module. For this purpose, we propose a method to extract Japanese/English translations from patent families consisting of comparable patents filed in Japan and the United States (Section 3).

2 System Description

2.1 Overview

Figure 1 depicts the overall design of PRIME, which retrieves documents in response to user queries in either Japanese or English. However, unlike the case of CLIR, retrieved documents can potentially be in either

a combination of Japanese and English or either of the languages individually. We briefly explain the entire on-line process based on this figure.

First, a user query is translated into the foreign language (i.e., either Japanese or English) by way of a query translation module.

Second, a document retrieval module uses both the source (user) and translated queries to search a Japanese/English bilingual patent collection for relevant documents.

In real world usage, Japanese and English patents are not comparable in the collection (this is the major reason why cross/multi-lingual retrieval is needed). However, for the purpose of research and development, we currently target a comparable collection.

To put it more precisely, the collection contains approximately 1,750,000 pairs of Japanese abstracts and their English translations, which were provided on PAJ (Patent Abstract of Japan) CD-ROMs in 1995-1999[1].

Third, among retrieved documents, only those that are in the foreign language are translated into the user language through a document translation module.

In principle, we need only above three modules to realize multi-lingual patent retrieval in the sense that users can retrieve/browse foreign documents through their native language. However, to improve the browsing efficiency, a clustering module finally divides retrieved documents into a specific number of groups.

Additionally, in the off-line process, a translation extraction module identifies Japanese/English translations in the database, to enhance the query translation module.

2.2 Query Translation

The query translation module is based on the method proposed by Fujii and Ishikawa (1999; To appear), which has been applied to Japanese/English CLIR for the NTCIR collection consisting of technical abstracts (Kando et al., 1999).

This method translates words and phrases (compound words) in a given query, maintaining the word order in the source language. A preliminary study showed that approximately 95% of compound technical terms defined in a bilingual dictionary (Ferber, 1989) maintain the same word order in both Japanese and English.

Then, the Nova dictionary[2] is used to derive possible word/phrase translations, and a probabilistic method is used to resolve translation ambiguity.

The Nova dictionary includes approximately one million Japanese-English translations related to 19 technical fields as listed below:

> aeronautics, biotechnology, business, chemistry, computers, construction, defense, ecology, electricity, energy, finance, law, mathematics, mechanics, medicine, metals, oceanography, plants, trade.

[1] Copyright by Japan Patent Office.

[2] Developed by NOVA, Inc. http://www.nova.co.jp/

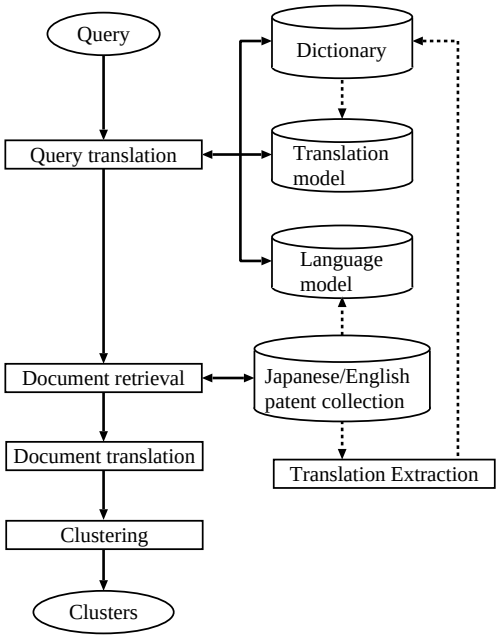

Figure 1: The design of PRIME: our multi-lingual patent retrieval system (dashed arrows denote the off-line process).

In addition, for words unlisted in the Nova dictionary, transliteration is performed to identify phonetic equivalents in the target language. Since Japanese often represents loanwords (i.e., technical terms and proper nouns imported from foreign languages) using its special phonetic alphabet (or phonogram) called "*katakana*", with which new words can be spelled out, transliteration is effective to improve the translation quality.

We represent the user query and one translation candidate in the document language by U and D, respectively. From the viewpoint of probability theory, our task here is to select D's with greater probability, $P(D|U)$, which can be transformed as in Equation (1) through the Bayesian theorem.

$$P(D|U) = \frac{P(U|D) \cdot P(D)}{P(U)} \quad (1)$$

In practice, $P(U)$ can be omitted because this factor is a constant with respect to the given query, and thus does not affect the relative probability for different translation candidates.

$P(D)$ is estimated by a word-based bi-gram language model produced from the target collection. $P(U|D)$ is estimated based on the word frequency obtained from the Nova dictionary. Those two factors are commonly termed language and translation models, respectively (see Figure 1).

2.3 Document Retrieval

The retrieval module is based on an existing probabilistic retrieval method (Robertson and Walker, 1994), which computes the relevance score between the translated query and each document in the collection. The relevance score for document i is computed based on Equation (2).

$$\sum_t \left(\frac{TF_{t,i}}{\frac{DL_i}{avglen} + TF_{t,i}} \cdot \log \frac{N}{DF_t} \right) \quad (2)$$

Here, $TF_{t,i}$ denotes the frequency that term t appears in document i. DF_t and N denote the number of documents containing term t and the total number of documents in the collection. DL_i denotes the length of document i (i.e., the number of characters contained in i), and $avglen$ denotes the average length of documents in the collection.

For both Japanese and English collections, we use content words extracted from documents as terms, and perform a word-based indexing. For the Japanese collection, we use the ChaSen morphological analyzer (Matsumoto et al., 1999) to extract content words. However, for the English collection, we extract content words based on parts-of-speech as defined in WordNet (Fellbaum, 1998).

2.4 Document Translation

The document translation module consists of the the Transer Japanese/English MT system, which uses the same dictionary used for the query translation module.

In practice, since machine translation is computationally expensive and degrades the time efficiency, we perform machine translation on a phrase-by-phrase basis. In brief, phrases are sequences of content words in documents, for which we developed rules to generate phrases based on the part-of-speech information. This method is practical because even a word/phrase-based translation can potentially improve on the efficiency for users to find relevant foreign documents from the whole retrieval result (Oard and Resnik, 1999).

2.5 Clustering

For the purpose of clustering retrieved documents, we use the Hierarchical Bayesian Clustering (HBC) method (Iwayama and Tokunaga, 1995), which merges similar items (i.e., documents in our case) in a bottom-up manner, until all the items are merged into a single cluster. Thus, a specific number of clusters can be obtained by splitting the resultant hierarchy at a predetermined level.

The HBC method also determines the most representative item (centroid) for each cluster. Thus, we can enhance the browsing efficiency by presenting only those centroids to users.

The similarity between documents is computed based on feature vectors that characterize each document. In our case, vectors for each document consist of frequencies of content words appearing in the document. We extract content words from documents as performed in word-based indexing (see Section 2.3).

Given the clustering module, the system can facilitate an interactive retrieval. To put it more precisely, through the interface, users can discard irrelevant clusters determined by browsing representative documents, and re-cluster the remaining documents. By performing this process recursively, relevant documents are eventually remained.

3 Extracting Translations Using Patent Families

3.1 Overview

Since patents are usually associated with new words, it is crucial to translate out-of-dictionary words. The transliteration method used in the query translation module is one solution for this problem (see Section 2.2).

On the other hand, it is also effective to update the translation dictionary. For this purpose, a number of methods to extract translations from bilingual (parallel/comparable) corpora (Smadja et al., 1996; Yamamoto and Matsumoto, 2000) are applicable. However, it is considerably expensive to obtain bilingual corpora with sufficient volume of alignment information.

To resolve this problem, we use patent families, which are patent sets filed for the same/related contents in multiple countries, as comparable corpora. Thus, patents contained in the same family are not necessarily parallel, but quite comparable.

Among a number of ways to apply for patents in multiple countries, we focus solely on patents claiming priority under the Paris Convention, because we can easily identify patent families by the identification number assigned to each patent.

In addition, the number of patent families is still increasing. Thus, we can easily update a large-scale bilingual comparable corpus based on patent families. To the best of our knowledge no research has utilized patent families for extracting translations.

3.2 Methodology

Since patents are structured with a number of fields (e.g., titles, abstracts, and claims), our method first identifies corresponding fragments based on the document structure, to improve the extraction accuracy.

However, structures of paired patents are not always the same. For example, the number of fields claimed in a single patent family often varies depending on the language. Thus, we use only the title and abstract fields, which usually parallel in Japanese and English patents. In other words, unlike the case of

most existing extraction methods, our method does not need sentence-aligned corpora.

We use the ChaSen morphological analyzer (Matsumoto et al., 1999) and Brill tagger (Brill, 1995) to extract content words from Japanese and English fragments, respectively. In addition, we combine more than one word into phrases, for which we developed rules to generate phrases based on the part-of-speech information.

We then compute the association score for all the possible combinations of Japanese/English phrases co-occurring in the same fragment, and select those with greater score as the final translations. For this purpose, we use the weighted Dice coefficient (Yamamoto and Matsumoto, 2000) as shown in Equation (3).

$$score(W_j, W_e) = \log F_{je} \cdot \frac{2F_{je}}{F_j + F_e} \qquad (3)$$

Here, W_j and W_e are Japanese and English phrases, respectively. F_j and F_e denote the frequency that W_j and W_e appear in the entire corpus, respectively. F_{je} denotes the frequency that W_j and W_e co-occur in the same fragment. The logarithm factor is effective to discard infrequent co-occurrences, which usually decrease the extraction accuracy.

3.3 Experimentation

A preliminary study showed that out of approximately 1,750,000 patents filed in Japan (1995-1999), approximately 32,000 patents were paired with those filed in the United States as patent families. Thus, in practice we obtained a bilingual comparable corpus consisting of 32,000 Japanese/English pairs. From this corpus, our method extracted 1,234,347 phrase-based translations, which were judged it correct or incorrect.

However, we selected translations association whose score was above 1.5, and manually judged their correctness, because a) the judgement can be considerably expensive for the entire translations, and b) translations with small association scores are usually incorrect. The total number of selected translations was 37,669.

We then evaluated the accuracy of our extraction method. The accuracy is the ratio between the number of correct translations, and the number of cases where the association score of the translation is above a specific threshold. By raising the value of the threshold, the accuracy also increased, while the number of extracted translations decreased, as shown in Table 1. According to this table, we could achieve a high accuracy by limiting the number of translations extracted.

We spent only four man-days in judging the 37,669 translations and identifying 5,879 correct translations. In other words, our method facilitated to produce bilingual lexicons semi-automatically with a trivial cost.

Table 1: Accuracy for translation extraction.

Threshold for Score	1.5	2.0	3.0	4.0	5.0
# of Translations	37,669	24,869	4,419	962	356
# of Correct Translations	5,879	4,129	1,399	564	240
Accuracy (%)	15.6	16.6	31.7	58.6	67.4

4 Conclusion

In this paper, we proposed a multi-lingual system for Japanese/English patent retrieval. For this purpose, we used a query translation method explored in cross-language information retrieval (CLIR).

However, unlike the case of CLIR, our system retrieves bilingual patents simultaneously in response to a monolingual query. Our system also summarizes retrieved patents by way of machine translation and clustering to improve the browsing efficiency.

In addition, our system includes an extraction module which produces new translations from patent families consisting of comparable patents, and updates the translation dictionary.

Future work would include improving existing modules in our system, and the application of our framework to other languages.

Acknowledgments

The authors would like to thank NOVA, Inc. for their support with the Nova dictionary and Transer system, and Makoto Iwayama for his support with the HBC software.

References

Lisa Ballesteros and W. Bruce Croft. 1998. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. 1997. Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 708–714.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Gene Ferber. 1989. *English-Japanese, Japanese-English Dictionary of Computer and Data-Processing Terms*. MIT Press.

Atsushi Fujii and Tetsuya Ishikawa. 1999. Cross-language information retrieval for technical documents. In *Proceedings of the Joint ACL SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 29–37.

Atsushi Fujii and Tetsuya Ishikawa. 2001. Evaluating multi-lingual information retrieval and clustering at ULIS. In *Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*.

Atsushi Fujii and Tetsuya Ishikawa. (To appear). Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*.

Masatoshi Fukui, Shigeto Higuchi, Youichi Nakatani, Masao Tanaka, Atsushi Fujii, and Tetsuya Ishikawa. 2000. Applying a hybrid query translation method to Japanese/English cross-language patent retrieval. In *ACM SIGIR Workshop on Patent Retrieval*.

Julio Gonzalo, Felisa Verdejo, Carol Peters, and Nicoletta Calzolari. 1998. Applying EuroWordNet to cross-language text retrieval. *Computers and the Humanities*, 32:185–207.

Makoto Iwayama and Takenobu Tokunaga. 1995. Hierarchical Bayesian clustering for automatic text classification. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1322–1327.

Noriko Kando, Kazuko Kuriyama, and Toshihiko Nozue. 1999. NACSIS test collection workshop (NTCIR-1). In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–300.

Michael L. Littman, Susan T. Dumais, and Thomas K. Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In Gregory Grefenstette, editor, *Cross-Language Information Retrieval*, chapter 5, pages 51–62. Kluwer Academic Publishers.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, and Masayuki Asahara. 1999. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. Technical Report NAIST-IS-TR99009, NAIST.

J. Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 208–214.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81.

Douglas W. Oard and Philip Resnik. 1999. Support for interactive document selection in cross-language information retrieval. *Information Processing & Management*, 35(3):363–379.

Douglas W. Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, pages 472–483.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241.

Gerard Salton. 1970. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Kaoru Yamamoto and Yuji Matsumoto. 2000. Acquisition of phrase-level bilingual correspondence using dependency structure. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 933–939.