Applying a Hybrid Query Translation Method to Japanese/English Cross-Language Patent Retrieval

Masatoshi Fukui[†] Shigeto Higuchi[†] Youichi Nakatani[†]

Masao Tanaka[†] Atsushi Fujii[‡] Tetsuya Ishikawa[‡]

[†]Japan Patent Information Organization
Satoh Daiya Bldg., 1-7 Toyo 4-Chome Koto-ku 135-0016, JAPAN

[‡]University of Library and Information Science
1-2 Kasuga Tsukuba 305-8550, JAPAN

E-mail: fujii@ulis.ac.jp

Abstract

This paper applies an existing query translation method to cross-language patent retrieval. In our method, multiple dictionaries are used to derive all possible translations for an input query, and collocational statistics are used to resolve translation ambiguity. We used Japanese/English parallel patent abstracts to perform comparative experiments, where our method outperformed a simple dictionary-based query translation method, and achieved 76% of monolingual retrieval in terms of average precision.

1 Introduction

Since 1978, JAPIO (Japan Patent Information Organization) has operated PATOLIS, which is one of the first on-line patent retrieval services in Japan, and currently provides clients (i.e., 8,000 Japanese companies) with patent information from 62 countries and 5 international organizations. At the same time, since a patent obtained in a single country can be protected in multiple countries simultaneously, it is feasible that users are interested in retrieving patent information across languages. Motivated by this background, JAPIO manually summarizes each patent document submitted in Japan into approximately 400 characters, and translates the summarized documents into English, which are provided on PAJ (Patent Abstract of Japan) CD-ROMs[1].

In this paper, we target cross-language information retrieval (CLIR) in the context of patent retrieval, and evaluate its effectiveness using Japanese/English patent abstracts on PAJ CD-ROMs.

In brief, existing CLIR systems are classified into three approaches: (a) translating queries into the document language [1, 3], (b) translating documents into the query language [13, 14], and (c) representing both queries and documents in a language-independent space [2, 7, 11, 15]. However, since developing a CLIR system is expensive, we used the CLIR system proposed by Fujii and Ishikawa [5, 6], which follows the first approach.

This system has partially been developed for the NACSIS test collection [10], which consists of 39 Japanese queries and approximately 330,000 technical abstracts in Japanese and English. However, since patent information usually includes technical terms, it is expected that this system also will perform reasonably for patent abstracts.

---

[1]Copyright by Japan Patent Office.

2   System Description

Figure 1 depicts the overall design of our CLIR system, in which we combine a query translation module and an IR engine for monolingual retrieval. Unlike the original system proposed by Fujii and Ishikawa [5, 6] targeting the NACSIS collection, we use the JAPIO collection for the target documents. Here, the JAPIO collection is a subset of PAJ CD-ROMs. We will elaborate on this collection in Section 3. In this section, we briefly explain the retrieval process based on Figure 1.

First, query translation is performed for the source language query to output the translation. For this purpose, a hybrid method integrating multiple resources is used. To put it more precisely, the EDR technical/general dictionaries [9] are used to derive all possible translation candidates for words and phrases included in the source query. In addition, for words unlisted in dictionaries, transliteration is performed to identify phonetic equivalents in the target language.

Then, bi-gram statistics extracted from NACSIS documents in the target language are used to resolve the translation ambiguity. Ideally, bi-gram statistics should be extracted from the JAPIO collection. However, since the number of documents in this collection is relatively small, when compared with the NACSIS collection (see Section 3), we avoided the data sparseness problem.

Since our system is bidirectional between Japanese and English, we tokenize documents with different methods, depending on their language. For English documents, the tokenization involves eliminating stopwords and identifying root forms for inflected content words. For this purpose, we use WordNet [4], which contains a stopword list and correspondences between inflected words and their root form.

On the other hand, we segment Japanese documents into lexical units using the ChaSen morphological analyzer [12], which has commonly been used for much Japanese NLP research, and extract content words based on their part-of-speech information.
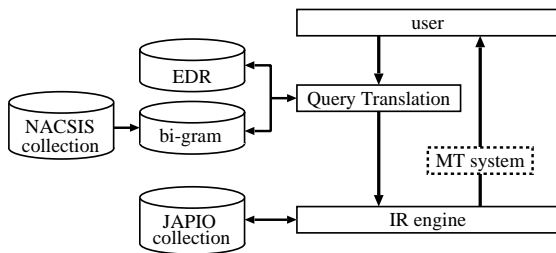

Figure 1: The overall design of our cross-language patent retrieval system.

Second, the IR engine searches the JAPIO collection for documents relevant to the translated query, and sorts them according to the degree of relevance, in descending order. Our IR engine is based on the vector space model, in which the similarity between the query and each document (i.e., the degree of relevance of each document) is computed as the cosine of the angle between their associated vectors. We use the notion of TF·IDF for term weighting. Among a number of variations of term weighting methods [16, 18], we tentatively use the formulae as shown in Equation (1).

$$
\begin{aligned}
TF &= 1 + \log(f_{t,d}) \\
IDF &= \log(\frac{N}{n_t})
\end{aligned}
\tag{1}
$$

Here, $f_{t,d}$ denotes the frequency that term t appears in document d, and n_t denotes the number of documents containing term t. N is the total number of documents in the collection.

For the indexing process, we first tokenize documents as explained above (i.e., we use WordNet and ChaSen for English and Japanese documents, respectively), and then conduct the word-based indexing. That is, we use each content word as a single indexing term.

Finally, since retrieved documents are not in the user's native language, we optionally use a machine translation system to enhance readability of retrieved documents.

3 Experimentation

Since no test collection for Japanese/English patent retrieval is available to the public, we produced our test collection (i.e., the JAPIO collection), which consists of three Japanese queries and Japanese/English comparable abstracts.

Each query, which was manually produced, consists of the description and narrative, and corresponds to different domains, i.e., electrical engineering, mechanical engineering and chemistry. Figure 2 shows the three query descriptions in the second column.

In conventional test collections, relevance assessment is usually performed based on the pooling method [17], which first pools candidates for relevant documents using multiple retrieval systems. However, since in our case only one system described in Section 2 is currently available, a different production method was needed.

To put it more precisely, for each query (domain), target documents were first collected based on the IPC classification number, from PAJ CD-ROMs in 1993-1998. Then, for each query, three professional human searchers, who were allowed to enhance queries based on thesauri and their introspection, searched the target documents for relevant documents.

Thus, in practice, the JAPIO collection consists of three different document collections corresponding to each query. In Figure 2, the third and fourth columns denote the number of relevant documents and the total number of target documents for each query.

We compared the following methods:

- Japanese-English CLIR, where all possible translations derived from EDR dictionaries and the transliteration method were used as query terms (JEALL),

- Japanese-English CLIR, where disambiguation based on bi-gram statistics were performed, and k-best translations were used as query terms (JEDIS),

- Japanese-Japanese monolingual IR (JJ).

Here, we empirically set $k = 1$. Although the performance of JEDIS did not significantly differ as long as we set a small value of k (e.g., $k = 5$), we achieved the best performance when we set $k = 1$.

Figure 3 shows recall-precision curves for the above three methods, where JEDIS generally outperformed JEALL, and JJ generally outperformed both JEALL and JEDIS, regardless of the recall. The difference between JEALL and JEDIS is attributed to the fact that JEDIS resolved translation ambiguity based on bi-gram statistics extracted from the NACSIS collection. Thus, we can conclude that the use of bi-gram statistics (even extracted from a collection other than the JAPIO collection) was effective for the query translation.

Table 1 shows the non-interpolated average precision values, averaged over the three queries, for each method. This table shows that JJ outperformed JEALL and JEDIS, JEDIS outperformed JEALL, and the average precision value for JEDIS was 76% of that obtained with JJ.

These results are also observable in existing CLIR experiments using the TREC and NACSIS collections. Thus, we conclude that our cross-language patent retrieval system is relatively comparable with those for newspaper articles and technical abstracts in performance.

However, we could not conduct statistical testing, which investigates whether the difference in average precision is meaningful or simply due to chance [8], because the number of queries is small. We concede that experiments using a larger number of queries need to be further explored.

4 Conclusion

In this paper, we explored Japanese/English cross-language patent retrieval. For this purpose, we used an existing cross-language IR system relying on a hybrid query translation method, and evaluated its effectiveness using Japanese queries and English patent abstracts. The experimental results paralleled existing experiments. That is, we found that re-

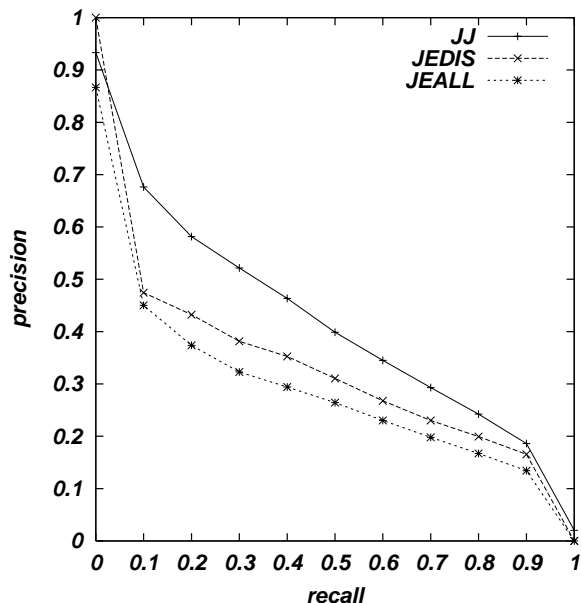| IPC	Description	#Relevant	#Documents
electronics	GPS car navigation system based on VICS	930	7,526
mechanics	eliminating dioxin in burning solid wastes	451	8,214
chemistry	antibacterial plastic combining inorganic materials	473	5,902

Figure 2: Query descriptions in the JAPIO collection.

Figure 3: Recall-precision curves for different methods.

solving translation ambiguity was effective for the query translation, and that the average precision value for cross-language IR was approximately 76% of that obtained with monolingual IR. Future work will include qualitative/quantitative analyses based on a larger number of queries.

References

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

Table 1: Non-interpolated average precision values, averaged over the three queries, for different methods.

Method	Avg. Precision	Ratio to JJ
JJ	0.4151	–
JEDIS	0.3156	0.7603
JEALL	0.2709	0.6526

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

[3] Mark W. Davis and William C. Ogden. QUILT: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 92–98, 1997.

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

[5] Atsushi Fujii and Tetsuya Ishikawa. Cross-language information retrieval at ULIS. In *Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 163–169, 1999.

[6] Atsushi Fujii and Tetsuya Ishikawa. 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*, pp. 29–37, 1999.

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

[8] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329–338, 1993.

[9] Japan Electronic Dictionary Research Institute. EDR electronic dictionary technical guide, 1995. (In Japanese).

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

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

[12] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Osamu Imaichi, and Tomoaki Imamura. Japanese morphological analysis system ChaSen manual. Technical Report NAIST-IS-TR97007, NAIST, 1997. (In Japanese).

[13] J. Scott McCarley. 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*, pp. 208–214, 1999.

[14] Douglas W. Oard. 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*, pp. 472–483, 1998.

[15] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, Vol. 21, No. 3, pp. 187–194, 1970.

[16] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523, 1988.

[17] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 315–323, 1998.

[18] Justin Zobel and Alistair Moffat. Exploring the similarity space. *ACM SIGIR FORUM*, Vol. 32, No. 1, pp. 18–34, 1998.