A System for Japanese/English/Korean Multilingual Patent Retrieval

Mitsuharu Makita[†] , Shigeto Higuchi[†] , Atsushi Fujii[††,†††] and Tetsuya Ishikawa[††]
[†] PATOLIS Corporation
[††] Institute of Library and Information Science, University of Tsukuba
[†††] CREST, Japan Science and Technology Corporation
m_makita@patolis.co.jp

Abstract In response to growing needs for cross-lingual patent retrieval, we propose PRIME (Patent Retrieval In Multilingual Environment system), in which users can retrieve and browse patents in foreign languages only by their native language. PRIME translates a query in the user language into the target language, retrieves patents relevant to the query, and translates retrieved patents into the user language. To update a translation dictionary, PRIME automatically extracts new translations from parallel patent corpora. In the current implementation, trilingual (J/E/K) patent retrieval is available. We describe the system design and its evaluation.

Keywords multilingual, patent retrieval, machine translation, document clustering, translation extraction

1 Introduction

Given the growing number of patents filed in multiple countries, it is feasible that users are interested in retrieving patents across languages. However, many users have difficulty retrieving patents in foreign languages.

To solve this problem, we developed a Japanese/English bilingual patent retrieval system, named PRIME (Patent Retrieval In Multilingual Environment) [4]. In brief, PRIME translates a query in a user language into a document language, retrieves the foreign patent documents relevant to the translated query, and translates retrieved documents into the user language. As a result, users can retrieve and browse foreign patent documents only by their native language.

In this paper, we extend PRIME into a trilingual retrieval system, in which user can utilize Japanese, English, and Korean for both the query and document languages. We describe the architecture of PRIME (Section 2) and its evaluation (Section 3).

2 System Description

2.1 Overview

Figure 1 depicts the overall design of PRIME, in which the online and offline processes are represented by full and dashed lines, respectively.

PRIME consists mainly of five modules: query translation, document retrieval, document clustering, document translation, and translation extraction modules.

In the current implementation, six combinations of query and document languages (EJ, EK, JE, JK, KE, and KJ) can be available. While EJ, JE, JK, and KJ are realized by a direct translation method, in EK and KE Japanese is used as a pivot language.

The JE and EJ retrieval functions are available on a commercial online service[1], in which the query and document translation modules are the same as described in this paper. However, the document retrieval and clustering modules in this paper are used only for research purposes, and are not used in the commercial service.

2.2 Query Translation

The query translation module is based on the method proposed by Fujii and Ishikawa [2], which has been applied to Japanese/English cross-lingual information retrieval. This method translates (compound) words and phrases in a given query, maintaining the word order in the source language. Note that a preliminary study showed that approximately 95% of compound technical terms defined in a bilingual dictionary maintain the same word order in both Japanese and English. This tendency becomes more salient in Japanese and Korean.

[1] http://www.patolis.co.jp/e-index.html

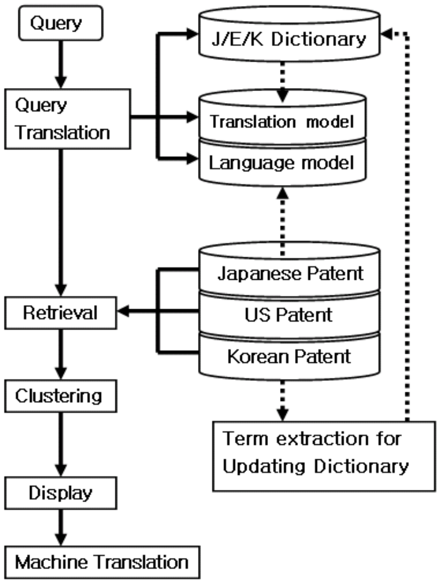


Figure 1: Overview of PRIME (full and dashed lines denote online and offline processes, respectively).

The translation dictionary developed by Cross Language Inc.[2] is used to derive multiple word/phrase translations, and a probabilistic method is used to resolve translation ambiguity.

We represent compound words in user language U and a translation candidate in document language D as follows.

$$\begin{aligned} U &= u_1, u_2, \ldots, u_n \\ D &= d_1, d_2, \ldots, d_n \end{aligned}$$

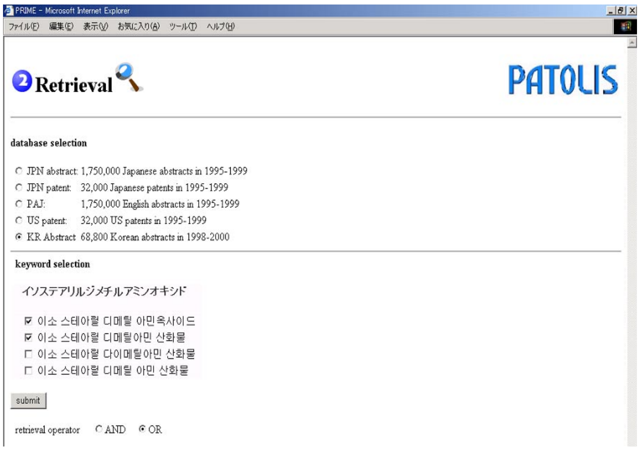
Here, u_i denotes an i-th base word, and d_i denotes a translation candidate of u_i.

Resolving translation ambiguity is equivalent to selecting D that maximizes $P(D|U)$, which can be formalized as in Equation (1) through the Bayesian theorem.

$$\arg\max_D P(D|U) = \arg\max_D P(U|D) \cdot P(D) \qquad (1)$$

At the right side of Equation (1), $P(U|D)$ and $P(D)$ corresponds to translation and language models, respectively (see Figure 1). These models can be decomposed as in Equation (2).

$$\begin{aligned} P(U|D) &\approx \prod_{i=1}^{n} P(u_i|d_i) \\ P(D) &\approx \prod_{i=1}^{n-1} P(d_{i+1}|d_i) \end{aligned} \qquad (2)$$


PRIME - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
PATOLIS

2 Retrieval

database selection

○ JPN abstract: 1,750,000 Japanese abstracts in 1995-1999
○ JPN patent: 32,000 Japanese patents in 1995-1999
○ PAJ: 1,750,000 English abstracts in 1995-1999
○ US patent: 32,000 US patents in 1995-1999
● KR Abstract 68,800 Korean abstracts in 1998-2000

keyword selection

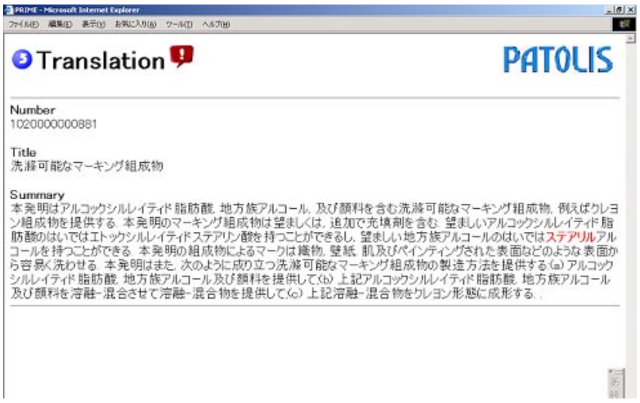
イソステアリルジメチルアミンオキシド

☑ 이소 스테아릴 디메틸 아민옥사이드
☑ 이소 스테아릴 디메틸아민 산화물
□ 이소 스테아릴 다이메틸아민 산화물
□ 이소 스테아릴 디메틸 아민 산화물

submit

retrieval operator ○ AND ● OR

Figure 2: Example of JK query translation.


PRIME - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
PATOLIS

5 Translation

Number
1020000000881

Title
洗滌可能なマーキング組成物

Summary
本発明はアルコックシルレイティド脂肪酸, 地方族アルコール, 及び顔料を含む洗滌可能なマーキング組成物, 例えばクレヨン組成物を提供する. 本発明のマーキング組成物は望ましくは, 追加で充填剤を含む. 望ましいアルコックシルレイティド脂肪酸のはいではエトックシルレイティドステアリン酸を持つことができるし, 望ましい地方族アルコールのはいではステアリルアルコールを持つことができる. 本発明の組成物によるマークは織物, 壁紙, 肌及びペインティングされた表面などのような表面から容易く洗わせる. 本発明はまた, 次のように成り立つ洗滌可能なマーキング組成物の製造方法を提供する:(a) アルコックシルレイティド脂肪酸, 地方族アルコール及び顔料を提供して(b) 上記アルコックシルレイティド脂肪酸, 地方族アルコール及び顔料を溶融-混合させて溶融-混合物を提供して(c) 上記溶融-混合物をクレヨン形態に成形する. .

Figure 3: Example of KJ document translation.

In addition, for words unlisted in the Cross Language's dictionary, transliteration is performed to identify phonetic equivalents in the target language. It is highly effective specifically in processing loanwords spelled out by Japanese phonetic alphabets (i.e., *katakana*).

Figure 2 shows an example of Japanese-to-Korean query translation. In this figure, the top four Korean translation candidates corresponding to a Japanese keyword ("isostearyl dimethyl amine oxide") are displayed. If users can understand Korean to a certain extent, they can select translations used for the subsequent retrieval process. Otherwise, a specific number of top translation candidates are automatically used for retrieval purposes. Figure 3 shows a Korean document retrieved by the Korean keywords in Figure 2 and machine translated into Japanese.

[2] http://www.crosslanguage.co.jp

2.3 Document Retrieval

In the retrieval module, an existing retrieval method [7] is used to compute the relevance score between the translated query and each document in the collection. The relevance score for document d is computed by Equation (3).

$$\sum_t f_{t,q} \cdot \frac{(K+1) \cdot f_{t,d}}{K \cdot \{(1-b) + \frac{dl_d}{b \cdot avgdl}\} + f_{t,d}} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5} \quad (3)$$

Here, $f_{t,q}$ and $f_{t,d}$ denote the frequency that term t appears in query q and document d, respectively. N and n_t denote the total number of documents in the collection and the number of documents containing term t, respectively. dl_d denotes the length of document d, and $avgdl$ denotes the average length of documents in the collection. We empirically set $K = 2.0$ and $b = 0.8$, respectively.

Given a translated query, the retrieval module searches a target patent collection for relevant documents and sorts them according to the score in descending order.

We use the following document collection for the purpose of research and development of PRIME:

- Japanese Patent Application (32,000),
- Japanese Patent Abstract (1,750,000),
- US Patent Application (32,000),
- Patent Abstract of Japan (1,750,000),
- Korean Patent Application (68,800),
- Korean Patent Abstract in Korean (68,800),
- Korean Patent Abstract in English (78,000).

Here, Patent Abstract of Japan (PAJ) is an English translation of the Japanese Patent Abstract.

We extract content words, such as nouns, and perform word-based indexing. We use ChaSen[3], the Brill Tagger [1], and the Cross Language's morphological analyzer, to extract content words from Japanese, US, and Korean patent documents, respectively.

2.4 Document Clustering

In the document clustering module, we use the Hierarchical Bayesian Clustering (HBC) method [5], which merges similar items (i.e., patents in our case) in a bottom-up method, 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 two patents is computed based on the feature vectors that characterize each patent. In our case, vectors for each patent consist of the frequencies of content words appearing in the patent. We extract content words from patents as performed in word-based indexing.

2.5 Document Translation

In the document translation module, we use the PAT-Transer Japanese/English and Japanese/Korean machine translation systems (developed by Cross Language Inc.), which use the same dictionary for the query translation module.

2.6 Translation Extraction

Because 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.

At the same time, it is also effective to update the translation dictionary itself. For this purpose, a number of methods to extract translations from bilingual (parallel and comparable) corpora are applicable. However, it is expensive to obtain bilingual corpora with sufficient volume of alignment information.

To resolve this problem, we use patent families as a comparable corpus. Here, patent family is a set of patents filed for the same or related invention in multiple countries. Thus, the patents in the same family are not necessarily parallel, but quite comparable.

We first extract (compound) words from patents in different languages (e.g., Japanese and English), and use a Dice-like coefficient to determine plausible word combinations.

3 System Evaluation

3.1 Evaluating Japanese/English retrieval

To evaluate the accuracy of the query translation, assessors evaluated the quality of translated queries subjectively. The accuracy of JE was 94.2% and that of EJ was 92.8%. These results show that the query translation module in PRIME is highly practical.

Fukui et al. [3] evaluated a cross-lingual patent retrieval system which corresponds to predecessor of PRIME. Using the NACSIS test collection [6], cross-lingual retrieval and monolingual retrieval methods were compared in per-

[3] http://chasen.aist-nara.ac.jp/

Table 1: Results for KJ document translation.

	# of Words	Proportion
Correct	541	90.5%
Incorrect	57	9.5%
Reasons of errors		
unknown words	20	35.1%
syntactic	9	15.8%
dictionary	8	14.0%
morphological analysis	6	10.5%
others	14	24.5%

Table 2: Results for JK query translation.

Rank of the best translation	Proportion to total # of query
1st	47.5 %
2nd or 3rd	36.3 %
Less than 3rd	15.2 %
Uncertain	1.00 %

formance. While the mean average precision (MAP) of monolingual retrieval was 0.4051, MAP of cross-lingual retrieval was 0.3156.

3.2 Evaluating Japanese/Korean retrieval

We asked an assessor (excluding the authors of this paper) to evaluate the quality of the JK query translation and KJ document translation. The assessor evaluated the Japanese translations of 15 documents in Korean Patent Abstract and the Korean translations of 99 Japanese keywords. In both case, the evaluation was performed on a word-by-word basis. In the latter case, if the correct translation was in the top three candidates, the assessor judged it correct. The accuracy of document translation was 90.5% (Table 1). The accuracy of query translation was 83.9% (Table 2).

As in Section 3.1, we confirmed that the accuracy of query and document translation for Japanese and Korean was also practical. At the same time, The performance of document retrieval for Korean patents remains an open question and needs to be further explored.

4 Conclusion

In this paper, we proposed a multilingual system for Japanese/English/Korean patent retrieval and evaluated its performance from different perspectives. The experimental results showed the utility of PRIME. We are currently extending PRIME to Chinese retrieval.

Acknowledgments

The authors would like to thank the Cross Language, Inc. for their support with the dictionaries and the Pat-Transer system, and Korean Institute of Patent Information (KIPI) for permission to use the Korean patent documents.

Bibliography

[1] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, Vol. 21, No. 4, pp. 543–565, 1995.

[2] Atsushi Fujii and Tetsuya Ishikawa. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, Vol. 35, No. 4, pp. 389–420, 2001.

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

[4] Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. PRIME: A system for multi-lingual patent retrieval. In *Proceedings of MT Summit VIII*, pp. 163–167, 2001.

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

[6] 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.

[7] S.E. Robertson and S. Walker. 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*, pp. 232–241, 1994.