POSTECH at NTCIR-5 Patent Retrieval: Smoothing Experiments in a Language Modeling Approach to Patent Retrieval

In-Su Kang, Seung-Hoon Na, Jun-Ki Kim, Jong-Hyeok Lee

Div. of Electrical and Computer Engineering
Pohang University of Science and Technology (POSTECH)
Advanced Information Technology Research Center (AITrc)
San 31, Hyoja-Dong, Pohang, Republic of Korea, 790-784
{dbaisk, nsh1979, yangpa, jhlee}@postech.ac.kr

Abstract

This report describes the experimental results of our participation at the Document Retrieval Subtask of NTCIR-5 Patent Retrieval Task. Unlike newspaper articles which belong to the main document type handled in previous information retrieval experiments, patent documents have many different characteristics in terms of length, technicality, structureness, etc. Among these, we focus on the length of patent documents. Since patent documents are long and cover a diverse spectrum of topicality, document models estimated from them by the language modeling approach are expected to show different characteristics from those estimated from short document samples such as newspaper articles. Based on such contemplation, this report investigates the effect of smoothing in the language modeling approach on retrieving patent documents.

Keywords: *Patent retrieval, Invalidity Search, Statistical language model, Smoothing*

1 Introduction

The evaluation of retrieving patent documents has started through recent NTCIR workshops [1, 5]. Unlike newspaper articles which have been the main document type of document collections used in previous information retrieval experiments, patent documents have many different characteristics. First, patent documents are structurally well formed. A typical patent document is composed of a title, bibliographic information, a claim, a detailed description, etc. The structure is firmly fixed by national patent offices. Second, patent documents are long and the variation of their size is large. According to [4], the average length of patent documents is about 24 times larger than that of newspaper articles, and the standard variance of the length of patent documents is approximately 20 times larger than that of newspapers. Third, all patents are assigned international patent classification (IPC) codes. IPC is a standard taxonomy for classifying patents, and has currently about 69,000 nodes which are organized into a 5-level hierarchical system.

The above characteristics of patent documents require the information retrieval community to re-evaluate and modify previous retrieval techniques that have been mainly developed for structure-less and short documents. For example, we need to re-confirm in the context of patent retrieval the effectiveness of well-known retrieval techniques such as the application of logarithmic function to term frequencies, document length normalization, pseudo relevance feedback, query expansion, smoothing in statistical language models, etc. In addition, from the viewpoint of cluster-based retrieval, patent documents could provide the appropriate test bed to evaluate and modify cluster-based retrieval techniques, since all patents have manually-assigned cluster information, IPC.

Among these, we are interested in investigating the effect of smoothing in language modeling approaches on retrieving patent documents. The selection of statistical language models was motivated by the fact that language models are created from target documents themselves that are viewed as samples of document models to be estimated, and thus document models could statistically well reflect the characteristics of patent documents. In language modeling approaches, smoothing is importantly handled since it is directly related to the retrieval performance [10]. According to Zhai and Lafferty [10], smoothing in language modeling approaches has two different roles: *estimation role* and *the role of query modeling*. The first role is to estimate document models from their samples that constitute the target document collection. The second is to reflect the importance of query terms. Our intuition is that patent retrieval would not be dominated by the estimation role, since document samples in patent retrieval are long and cover a diverse spectrum of topicality. From the viewpoint of topicality, note that the claim part in patent documents includes highly specific topic words while the detailed description of patent documents provides more general background information.

Based on the above motivation, our retrieval experiments at the NTCIR-5 Patent Retrieval Task were performed based on the language model. We participated in Document Retrieval Subtask that performs the invalidity search for NTCIR-5 Japanese patent document collections. In order to concentrate on the effect of smoothing on retrieval effectiveness of patent documents, we did not employ advanced term extraction schemes such as the morphological analysis of Japanese patent documents. Instead, character bi-gram terms were extracted from both documents and queries.

The remainder of the paper is organized as follows. Section 2 gives our system description. Section 3 describes the experimental results and discussion. Concluding remarks are given in Section 4.

2 System Description

2.1 Query and Document Processing

Invalidity search uses a patent application itself as a query in order to retrieve published or registered patents that contain some conflicting claim parts that are enough to invalidate (or reject) the patent application. Thus, the whole parts of a patent application could be used to extract query terms. However, we only use the claim part to create the query as described in the task description of NTCIR-5 patent retrieval. The claim part is preprocessed to eliminate punctuation marks and special characters, and to normalize two-byte numbers and two-byte English characters into one-byte characters. Next, a sequence of numbers or a string of English characters is extracted as a query term from the claim part. English words are not stemmed. Then, we simply generate Japanese character-based bi-grams as query terms from the claim part.

The above term extraction procedure is equally applied to documents to produce index terms. The only exception is that index terms are extracted from the whole parts of the patent documents.

2.2 Retrieval Model

The language modeling approaches to information retrieval assume individual models for documents and views a query as a random sample from each document model [8]. At retrieval, documents are then generally ranked by the query likelihood that a document model D_M will generate a given query Q. The simple and common approach of calculating the query likelihood views queries as a sequence of independent terms as shown in Formula (1), where *freq*(q) is the count of query term q in Q. This multinomial view of document models was chosen by Miller et al. [7], Song and Croft [9], and Hiemstra [3].

$$P(Q \mid D_M) = \prod_{q \in Q} P(q \mid D_M)^{freq(q)} \qquad (1)$$

Then, the retrieval problem is reduced to estimating a unigram language model for each document and each term. However, the simple maximum likelihood estimation for unigram language models assign zero probabilities to unseen document terms. To avoid this data sparseness problem, the language modeling approach normally employs smoothing techniques among which the simple and popular one is Jelinek-Mercer smoothing [10] as shown in Formula (2), where λ is a smoothing parameter, *mle* indicates maximum likelihood estimation, and *Coll* means the collection model. Our participation system at the NTCIR-5 patent retrieval was based on the Formulas (1) and (2).

$$P(q \mid D_M) = (1-\lambda)P_{mle}(q \mid D_M) + \lambda P_{mle}(q \mid Coll) \qquad (2)$$

3 Experiment

3.1 Experimental Setup

As preliminary experiments, we first evaluated the retrieval effectiveness of the language model on invalidity search of patent retrieval using the NTCIR-4 patent test set [1]. The document collection consists of 1,707,185 patent documents of unexamined Japanese patent applications published in 1993 through 1997. The test set has 101 search topics that correspond to Japanese patent applications rejected by the Japanese Patent Office. A search topic is composed of a title, a claim, the date of filing, and other parts. We used claim parts and the date of filing as queries. Relevant documents of invalidity search task should be the prior art which had been open to the public before the topic patent was filed. Thus, we used the date of filing to filter out retrieved documents of which filing date is more than the filing date of the topic patent.

Using the optimal value of a smoothing parameter determined from the preliminary experiments, we applied the language model to NTCIR-5 patent documents which includes the NTCIR-4 patent documents and consists of 10-year unexamined Japanese patent applications published in 1993 through 2002. See the overview paper [2] of NTCIR-5 patent retrieval for the detailed description of NTCIR-5 patent test set.

There are two types of relevance judgments for the evaluation of NTCIR patent documents: *type A* and *type B*. The documents that can invalidate the demands of all essential components in a target claim are considered as type A, and the documents that can invalidate the demands of most of the essential components in a target claim are considered as type B. All retrieval results in this paper are reported using non-interpolated mean average precision which is computed by executing the TREC_EVAL program.

3.2 Experimental Results

In language modeling approaches, it is generally known that document models should be smoothed a lot with the collection model [6]. In other words, better performance is observed when λ in Formula (2) is more than 0.5. To see whether this observation applies also to patent documents, we have performed invalidity search over different values of λ using the Jelinek-Mercer language model which corresponds to Formulas (1) and (2). Figure 1 shows the retrieval result for the NTCIR-4 patent test set using *type A* relevance judgment.

As shown in Figure 1, smoothing is helpful to estimating better document language models. As λ increase from 0.01 to 0.1, the performance improves, demonstrating the estimation role of smoothing. In terms of the role of its query modeling, however, the result of Figure 1 violates the general behavior of query-modeling role that Jelinek-Mercer smoothing shows, considering that topic claims are long. According to Zhai and Lafferty [10], long and verbose queries need more smoothing, since heavy smoothing could discriminate the common and non-informative terms from informative terms by emphasizing IDF factor in terms of term weighting.

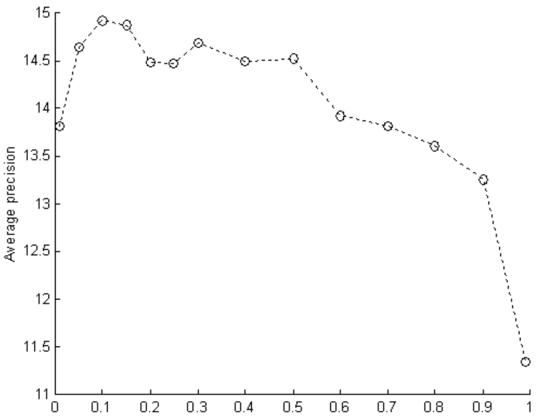

Figure 1. Effect of smoothing in retrieving NTCIR-4 patent documents using *type A* relevance judgment

Currently, we believe that this results from the characteristics of patent documents and invalidity search. The first characteristic is as follows. As mentioned in Section 1, the average length of patent documents is about 24 times larger than that of newspaper articles [4]. Thus, the degree of smoothing of document models for a patent collection is expected to be smaller than that of document models for much shorter newspaper articles. The second characteristic is that although the claim part is long, but it is not verbose. This means that the differentiation of claim terms by heavy smoothing could be harmful to obtaining better query model in the invalidity search of patent retrieval.

Table 1. NTCIR-5 retrieval results

Model	Type A	Type B
Jelinek-Mercer LM	18.23	14.38
The top score at NTCIR-5	21.07	16.84

Table 1 shows the retrieval results of the Jelinek-Mercer language model for NTCIR-5 patent collection using λ=0.1 at which the language model performed best in the case of NTCIR-4 data. Interestingly, the performance of our system was not much less than that of the best system at NTCIR-5 patent retrieval, although our system did not rely on any advanced retrieval techniques such as word-based indexing or pseudo relevance feedback. Unfortunately, however, the result of Table 1 was obtained by our post-experiments. Our official results received by NTCIR-5 patent organizers were 7.86 for type A relevance judgment and 7.57 for type B. After submitting our runs of NTCIR-5 patent retrieval, we have found that our implementation of the retrieval system had some errors.

4 Conclusions

This paper presented the retrieval results of the effect of smoothing in the language modeling approach to patent retrieval. An empirical result from our experiments is as follows. While heavy smoothing is common in language modeling approaches [6], it can be harmful to the invalidity search of patent documents. In other words, when smoothing document models in patent retrieval, it can be disadvantageous to smooth document models heavily with the collection model, because topic claims are long but not verbose. This finding leads us to the further study on the effect of other smoothing techniques for patent documents.

Acknowledgements

This work was supported by the KOSEF through the Advanced Information Technology Research Center (AITrc) and by the BK21 project.

References

[1] Fujii, A., Iwayama, M. and Kando, N. Overview of patent retrieval task at NTCIR-4. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages 225-232, 2004.

[2] Fujii, A., Iwayama, M. and Kando, N. Overview of patent retrieval task at NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop*, 2005.

[3] Hiemstra, D. Using language models for information retrieval. *PhD Thesis*, University of Twente, 2001.

[4] Iwayama, M., Fujii, A., Kando, N. and Marukawa, Y. An empirical study on retrieval models for different document genres: patents and newspaper articles. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 251-258, 2003.

[5] Iwayama, M., Fujii, A., Kando, N. and Takano, A. Overview of patent retrieval task at NTCIR-3. In *Working Notes of the Third NTCIR Workshop Meeting*, pages 1-10, 2002.

[6] Kraaij, W., Nie, J.Y. and Michel S. Embedding Web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29, 1-37, 2003.

[7] Miller, D., Leek, T. and Schwartz, R. A hidden Markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214-221, 1999.

[8] Ponte, J.M. and Croft, W.B. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275-281, 1998.

[9] Song, F. and Croft, W.B. A general language model for information retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 279-280, 1999.

[10] Zhai, C. and Lafferty, J. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334-342, 2001.