DCU @ CLEF-IP 2009: Exploring Standard IR Techniques on Patent Retrieval

Walid Magdy, Johannes Leveling, Gareth J.F. Jones
Centre for Next Generation Localization
School of Computing
Dublin City University, Dublin 9, Ireland
{wmagdy, jleveling, gjones}@computing.dcu.ie

Abstract

This paper presents the experiments and results for our participation in CLEF-IP 2009, which in newly launched this year. Our work applied standard information retrieval (IR) techniques to patent search. Different experiments tested various methods for the patent retrieval, including query formulation, structured index, weighted fields, filtering, and relevance feedback. Some methods did not show expected good retrieval effectiveness such as blind relevance feedback, other experiments showed acceptable performance. Query formulation was the key task for achieving better retrieval effectiveness, and this was performed through giving some higher weights to the text in certain fields. For the best runs, the retrieval effectiveness is still lower than IR applications for other domains illustrating the fact of the difficulty of patent search. The official results have shown that among fifteen participants we achieved the seventh and the fourth ranks from the mean average precision (MAP) and recall point of view, respectively.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: **H.3.1** Content Analysis and Indexing - Indexing methods; **H.3.3** Information Search and Retrieval;

General Terms

Performance, Experimentation

Keywords

Patent Retrieval; Query Formulation; CLEF-IP track

1 Introduction

This paper presents the participation of Dublin City University (DCU) in the CLEF-IP track 2009. Our participation was in the main task which is retrieving patents prior art. The aim for a given patent (which is considered as the topic) is to retrieve all of its citations automatically [1]. Only three runs were submitted, but more unofficial experiments were performed by us for this task. Fifteen participants have submitted 48 runs; according to MAP scores, our best run achieved the seventh rank across participants and the 22nd across all runs. However, according to recall scores, our best run achieved the fourth rank across all participants and the fourth rank across all 48 submitted runs.

The paper is organized as follows: Section 2 describes the data for the task and an analysis of its nature; Section 3 presents all the experiments for this task; Section 4 shows the results; then Section 5 discusses these results; Finally, Section 6 concludes the paper and provides possible future directions.

2 Data Pre-Processing

More than 1.9M documents were provided representing different versions of 1M patents between the period 1985 and 2000. For our experiments, all different versions for a single patent were merged in one document with fields were updated from its latest versions. Patent structure is very rich, and some fields are presented in three languages (English "EN", German "DE", and French "FR") namely the title and claims. Some fields are for formal use only, hence, candidate relevant fields are extracted from these final merged versions for indexing. Only the patent title, abstract, description, claims, and classifications fields are extracted from the patents. However, many patents lack some of these fields. The only fields that are present in all patents are the title and the classifications; other fields are omitted in some patents (Figure 1). The "description" field is related to "claims" field, as if "claims" field is missing, then "description" is missing too, however, the opposite in not

true, as some documents contain a "claims" field while the "description" field is missing. The "abstract" field is an optional part that is present in some patents.

Figure 1 shows that about 23% of the patents do not contain the claims and description fields, while 73% of them are formed of titles only.

In order to avoid language problems, the English fields only are extracted from these fields. This step will lead to the loss of extra 0.23×(0.23+0.9) = 7.4% of the patents which lack the claims and description fields (these are the German and French patents with claims only in one language). In addition, all non-English patents lack the abstract, description fields. The final outcome resulted in 30% of the collection suffering from missing most of the fields, this portion of the collection is mostly comprises the titles only with a small portion of it containing abstracts too.

In order to maintain the full structure and overcome the lack of some fields in some patents, the abstract (if it exists) is copied to the description and claims fields; otherwise, the title is used instead.

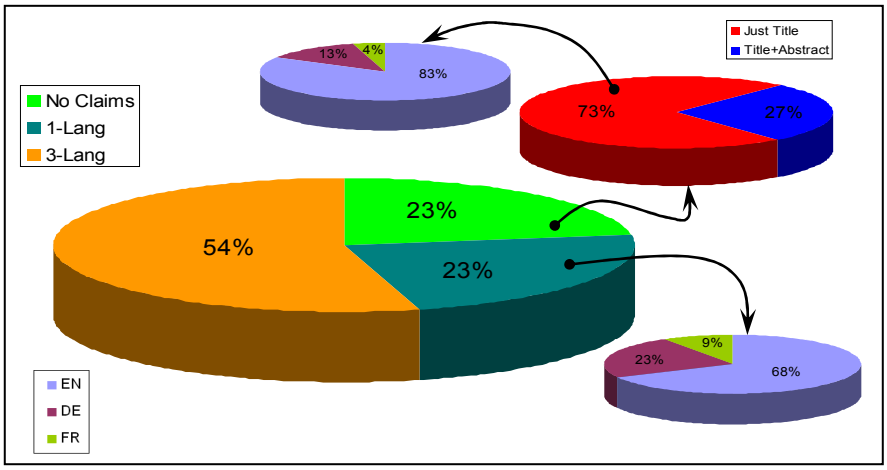

Figure 1: Percentage of patents with claims in EN\FR\DE

3 Experimentations

In this Section, different experiments for indexing and searching the data are discussed. After merging different versions of patents and extracting the relevant fields, some pre-processing is performed for the patent text in order to prepare it for indexing. Different methods were used for query formulation to search the collection.

Many experiments were performed on the training topics provided by the task organizers, however, a small number were submitted on the test data for the official runs. The training set contains 500 patent topics, which was sufficient to compare different methods and select the best for the official submissions.

3.1 Text Pre-Processing

Patent text contains many formulas, numeric references, chemical symbols, and patent-specific words (such as method, system, or device) that can cause a negative effect on the retrieval process. Some filtering of the text is done by removing:

1. Predefined stop words[1],
2. Digits, and
3. Field-specific stop words

To get the fields stop words, the field frequency for terms is calculated separately for each field. The field frequency for a term "T" in field "X" is the number of fields of type "X" containing the term "T" across all documents. For each field, all terms with field frequency higher than 5% of the term with highest field frequency for this field are considered as stop words. For example, for "title" field, the following words have been identified as stop words {method, device, apparatus, process, etc}; for another field such as "claims", the following words have been identified as stop words {claim, according, wherein, said, etc}.

3.2 Structured Indexing

For indexing, Indri [5] was used to create a structured index for patents. A structured index keeps the fields information in the index. The index was structured as shown in Figure 2.

[1] http://members.unine.ch/jacques.savoy/clef/index.html

This structured index allows searching specific fields instead of searching in the full document. It also allows giving different weights for each field while searching. As shown in Figure 1, "desc1" and "claim1" are sub-fields for the description and claims fields respectively. "desc1" is the first paragraph in the description field; typically it carries useful information about the field of the invention and what the invention is about. "claim1" is the first claim in the claims sections, and it describes the main idea of the invention in the patent. The field "class" carries the IPC classification [6] information of the patent, the three top classification levels are used, the rest levels are discarded (example: B01J, C01G, C22B).

As mentioned earlier, for patents that lacks of some fields, the empty fields are filled with the abstract if exists or with the title otherwise. Pre-processing includes stemming using the Porter stemmer [4].

```
<DOC>
      <DOCNO>patent number</DOCNO>
      <TEXT>
            <TITLE>title</TITLE>
            <CLASS>3rd level classification</CLASS>
            <ABSTRACT>abstract</ABSTRACT>
            <DESC>
            <DESC1>1st sentence in description</DESC1>
            Rest of patent description
      </DESC>
      <CLAIMS>
            <CLAIM1>1st claim</CLAIM1>
            Rest of patent claims
      <CLAIMS>
      </TEXT>
</DOC>
```

Figure 2: Structured text for a patent in TREC format

3.3 Query Formulation

Query formulation can be seen as one major task in patent retrieval [3]. As a full patent is considered as the topic, extracting the best representative text with the proper weights is the enabling key for good retrieval results. Using the full patent as a query is not practical due to the huge amount of text in one patent. Hence, text from certain fields was extracted and tested to search the structured index with different weights to different fields. Various combinations of fields were tested with different weights with enabling/disabling filtering using third level classification and enabling/disabling blind relevance feedback **Error! Reference source not found.**.

The patent topic text was pre-processed in the same way as in the indexing phase by removing stop words and digits, in addition to removing special characters, symbols and all words of small length (one or two letters).

Similar to the indexed documents, only English parts are used, which means all non-English patent topics will miss the abstract and description fields to be used in search. However, the amount of text present in claims and titles should be sufficient to create a representative query. In patent topics, there were no missing fields, and all claims and titles are present in the three languages. The following fields were used with different combinations to construct the queries {"title", "abstract", "desc1" (first line in description), "claim_main" (first sentence in first claim), "claim1" (first claim), and "claims"}.

4 Results

Some of the tested experiments seemed to be ineffective with retrieval. Blind relevance feedback and structured search have negative impact on the results. All experiments with blind relevance feedback led to a degradation in the MAP to around 60% of the original runs without feedback, and this can stem from the low quality of the highly ranked results. On the other hand, structured retrieval was tested by searching each field in the patent topic to its corresponding field in the index. Different weights for fields were tested; however, all experiments led to lower MAP and recall than searching in the full index as a whole without directing each field to its correspondent. Since patent documents were treated as a full document neglecting its structure, patent topics which were used for formulating the queries were tested by giving different weights to the text in each field. Giving higher weight to text in "title", "desc1", and "claim_main" has been proven to produce the best results across all runs.

According to the tested experiments on the training data, three runs were submitted on the official topics with the same setup as the best results in training. The three runs are nearly the same but with minor modifications. The common setup for the three runs was as follows:

1. The patent document is treated as a full document, neglecting its structure.
2. English text only is indexed with stemming (Porter stemmer).
3. Stop words are removed, in addition to digits and words of length less than two letters.
4. A query is formulated from the following fields with the following weights:
    a. 5×Title

b. 1×Abstract (English topics only)
c. 3×Desc1 (English topics only)
d. 2×Claim_main
e. 1×Claims

5. Additional bi-grams words with a frequency in the text higher than one were used in query. The text of the fields: "title", "abstract", "desc1", and "claim_main" was used for extracting the bi-grams words.

The difference between the three runs is as follows:

1. 1st Run: No filtering is performed
2. 2nd Run: Filtering is performed for all results that do not match up to the third level classification of the patent topic (at least one common classification should be present)
3. 3rd Run: The same as 2nd run, but with removing all words in the query of length less than three

Runs were submitted on the extra large collection that contains 10,000 patent topics. The average time for running this amount of topics was around 30 hours (about ten seconds on average for retrieving results of one topic).

Table 1 shows the results of the three submitted runs on three different collection sizes S: small (500 topics), M: medium (1,000 topics), and XL: extra large (10,000 topics) [2]. In Table 1, it is shown that the 3rd Run always gets the best results from the precision and recall perspective. The 1st run always has the worst results, which shows that applying the filtering over the results based on the patent classification is always useful. For all runs (for official and training ones), the retrieval effectiveness is relatively low when compared to other IR tasks; this can stem from the nature of patent documents itself in addition to the task of finding cited patents which are relative to the patent topic from the conceptual point of view, not from the word matching. This is discussed in the next section.

Table 1: Three submitted runs results for three different collection sizes (P: Precision, P<n>: Precision at n, R: Recall, and R<n>: Recall at n)

Size	Run #	P	P5	P10	P100	R	R5	R10	R100	MAP	nDCG
S	**Run 1**	0.0032	0.0916	0.0694	0.0184	0.5380	0.0818	0.1222	0.3086	0.0815	0.4142
	Run 2	0.0038	0.1008	0.0774	0.0208	0.6156	0.0908	0.1373	0.3526	0.0896	0.4120
	Run 3	0.0038	0.1036	0.0768	0.0208	**0.6236**	0.0952	0.1363	0.3531	**0.0913**	0.4183
M	**Run 1**	0.0032	0.0984	0.0709	0.0183	0.5289	0.0916	0.1302	0.3135	0.0891	0.4265
	Run 2	0.0037	0.1074	0.0785	0.0203	0.6071	0.1013	0.1440	0.3504	0.0970	0.4254
	Run 3	0.0037	0.1092	0.0785	0.0204	**0.6128**	0.1032	0.1433	0.3516	**0.0979**	0.4264
XL	**Run 1**	0.0033	0.1081	0.0768	0.0185	0.5438	0.1010	0.1405	0.3188	0.0968	0.4281
	Run 2	0.0038	0.1169	0.0839	0.0208	0.6240	0.1102	0.1546	0.3608	**0.1074**	0.4291
	Run 3	0.0038	0.1173	0.0840	0.0208	**0.6267**	0.1107	0.1546	0.3612	**0.1074**	0.4266

5 Discussion

In this section some analysis is done trying to identify the reason behind the low retrieval effectiveness for the patent retrieval task. In order to analyze this problem, the overlap between each topic in the training data and its relevant cited patents is computed; in addition, the overlap between the topics and the top five ranked non-relevant documents is calculated (the reason behind selecting the number "five" is that the average number of relevant documents for all topics is between five and six). The overlap is measured using two measures: 1) cosine measure between each two corresponding fields of the two compared patents. 2) Percentage of zero overlap (no shared words) between two corresponding fields of the two compared patents. The same pre-processing is done for all patents and topics, where stop words are removed including digits, and the comparison is based on the stemmed version of words. From Figure 3 and 4, it seems that relying on common words between topics and relevant documents for patent retrieval is not the best approach. Figure 3 shows that the cosine measure between the top ranked non-relevant documents to the topic is nearly double that of the relevant documents for all fields. The same is shown in Figure 4, where surprisingly, 12% of the relevant documents for topics have not any shared words in any field to the topics. This outcome has proven the importance of introducing different approaches for query formulation instead of relying on word matching in the patent topics only.

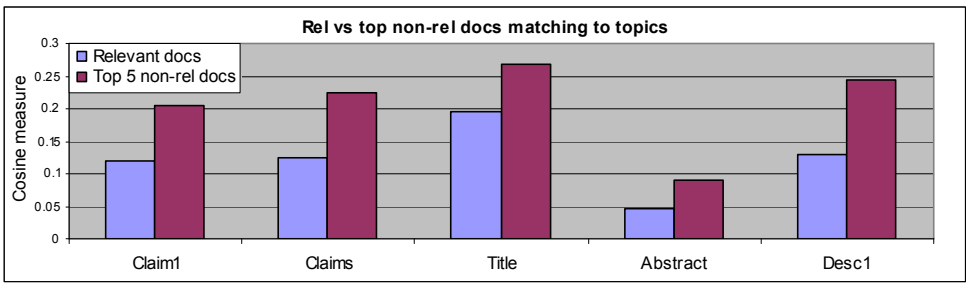

Figure 1: Cosine measure between fields of topics and the corresponding ones in relevant and top retrieved documents

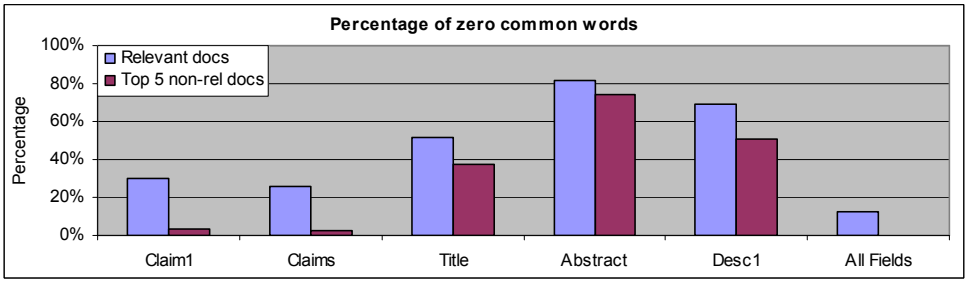

Figure 2: Percentage of fields with zero common (shared) words between that of topics and the corresponding ones in relevant and top retrieved documents

6 Conclusion and Future Work

In this paper, we described our participation in the CLEF-IP track 2009. Standard IR techniques were tested with some focusing on query formulation. Our experiments illustrated the challenge of patent search task, where some analysis showed that depending on word matching is not the best solution as in other IR applications. Our best result was through treating patents as a full document with some pre-processing by removing standard stop words in addition to patent-specific stop words; and on the query phase, text is extracted from patent topic with giving higher weights to some specific fields. Some additional experiments showed the poor effectiveness of using blind relevance feedback or using patent structure in index.

For future work, more investigation is required for checking the best use of patent structure in both index and query phases; machine learning can be a useful approach for identifying the best weights for different fields. Furthermore, query expansion through the conceptual meaning of words is a potential approach to be tested. Finally, machine translation can be a good solution to overcome the problem of multi-lingual documents and queries.

7 Acknowledgment

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project

References

[1] E. Graf and L. Azzopardi. A methodology for building a patent test collection for prior art search. EVIA-2008 Workshop, NTCIR-7, 2008

[2] F. Piroi, G. Roda, and V. Zenz. CLEF-IP 2009 Evaluation Summary. CLEF 2009, 2009

[3] Fujii, A., Iwayama, M., and Kando, N. Overview of patent retrieval task at NTCIR-4. In *Proceedings of the fourth NTCIR workshop on evaluation of information retrieval, automatic text summarization and question answering, June 2–4, Tokyo, Japan*. 2004

[4] M.F. Porter, An Algorithm for Suffix Stripping, *Program* 14 (3) (1980), pp. 130–137

[5] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis,* 2004.

[6] IPC (International Patent Classification): http://www.epo.org/patents/patent-information/ipc-reform.html