Applying the KISS Principle for the CLEF- IP 2010 Prior Art Candidate Patent Search Task

Walid Magdy, Gareth J.F. Jones

Centre for Next Generation Localisation
School of Computing
Dublin City University, Dublin 9, Ireland
{wmagdy, gjones}@computing.dcu.ie

Abstract. We present our experiments and results for the DCU CNGL participation in the CLEF-IP 2010 Candidate Patent Search Task. Our work applied standard information retrieval (IR) techniques to patent search. In addition, a very simple citation extraction method was applied to improve the results. This was our second consecutive participation in the CLEF-IP tasks. Our experiments in 2009 showed that many sophisticated approach to IR do not improve the retrieval effectiveness for this task. For this reason of we decided to apply only simple methods in 2010. These were demonstrated to be highly competitive with other participants. DCU submitted three runs for the Prior Art Candidate Search Task, two of these runs achieved the second and third ranks among the 25 runs submitted by nine different participants. Our best run achieved MAP of 0.203, recall of 0.618, and PRES of 0.523.

Keywords: Patent Retrieval; Query Formulation; CLEF-IP track

1 Introduction

The Centre for Next Generation Localisation (CNGL) at Dublin City University (DCU) participated in the CLEF-IP track 2010 Candidate Patent Search Task using the KISS principle. KISS stands for "*keep it simple and straightforward*" which describes the methods adopted in our submissions for CLEF-IP 2010. Our participation used standard IR approaches with a very simple information extraction technique. The aim of the task is to automatically retrieve all of citations for a given patent (which is considered as the topic) [3], [7]. We submitted three runs for the task: one is based in citation extraction from patent application descriptions, another one uses simple information retrieval techniques to search the patent documents collection using the patent topic, and the final one is a combination of the first two methods. Nine participants submitted 25 runs in total for this task. These were evaluated using three metrics: mean average precision (MAP), recall, and the patent retrieval evaluation scores (PRES). Our best run (which is the third one) achieved the second rank using all the scores among all runs submitted by participants.

The paper is organized as follows: Section 2 overviews the patent data collection provided by the track organizers, Section 3 gives full details of the experimental setup for our participation, Section 4 reports the results with some analysis to these result, and finally Section 5 concludes the paper and provides possible future directions.

2 Data Collection

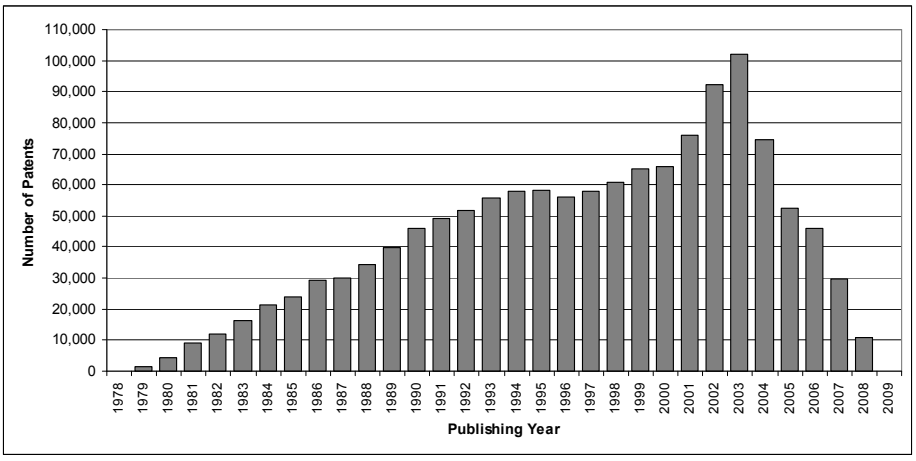

Fig.1. Distribution of patents collection by publishing year

For this task the organizers provided more than 2.68M XML documents representing different versions of 1.35M patents filed between 1978 and 2009 (see Figure 1). For our experiments, all different document versions for a single patent were merged into a single document with fields updated from its latest versions. The patent structure is very rich comprising the 'title' and 'claims', some fields are present in three languages (English "EN", German "DE", and French "FR"). In addition, the patent abstract of non-English patents has an English translation of the abstract included. Only the patent 'title', 'abstract', 'description', 'claims', and 'classifications' fields are extracted from the patents. Some patents lack some of these fields. The only fields that are present in all patents are the 'title' and the 'classifications'. The 'description' field is related to the 'claims' field, thus if the 'claims' field is missing, then the 'description' is missing too. However, the opposite in not true, some documents contain the 'claims' field while the 'description' field is missing. The 'abstract' field is an optional part that is only present in some patents.

68% of the patents in the collection are English, 24% are German, and 8% are French. In keeping with our KISS approach to avoid complications with language processing, only the English fields were used for indexing. Hence, English patents were fully indexed, but for German and French patents the title, abstract and claims were only indexed where a translation into English was found. This simplification meant that 32% of the collection (the non-English portion) the description field was not indexed. Additionally, as mentioned above, since some sections were already missing, some patents had no English field content to be indexed except the title or the title and abstract. Figure 2 shows the content that has been indexed for the patent collections based on the existing fields. It can be seen that 22% of the patents have only the titles indexed; leading them to effectively be very short indexed documents since they consist only of the single item of the title. Additionally 10% of the patents have only the titles and abstract indexed and 16% of the patents have only the title

and the claims indexed, with some of them having the abstract section as well. Only 52% of the patents have nearly the full document indexed, where the title, description, and claims sections are present. This 52% of the collection are the English patents where all the main sections found.

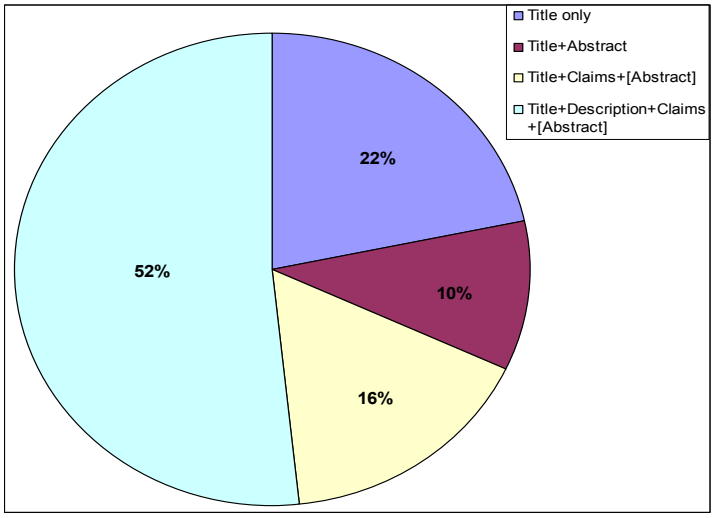

Fig.2. Proportions of patent texts in the collection indexed by field combinations

3 Experimental Setup

In this section we describe details of our experimental setup for our three submitted runs. The first run is a very standard information retrieval method where the patent topics were used to search the indexed collection after translating the non-English topics into English. The second run involves extracting the patent citation disclosed within the description of the patent topic without the use of any kind of IR techniques. Finally, the third run is generated through merging the first two runs.

3.1 IR Experiment

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 information retrieval process. In order to minimise these problems the text was filtered to remove predefined stop words[1], digits, and field-specific stop words.

To obtain the stop words for each field, the field frequency for each term in each field was calculated separately. The field frequency for a term "T" in field "X" is the

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

number of fields of type "X" across all documents containing the term "T". For each field, all terms with field frequency higher than 5% of the highest term field frequency for this field were considered to be stop words [4]. For example, for the 'title' field, the following words were identified as stop words: *method*, *device*, *apparatus*, *process*, etc; for another field such as 'claims', the following words were identified as stop words: *claim*, *according*, *wherein*, *said*, etc.

In addition to stop word removal, Porter stemming was applied to the text in order to normalize different surface forms of a given word [6].

Indexing. The Indri search toolkit [8] was used to index the extracted English parts of the patent collection. In the indexing process, the text of the following sections in the patents, if they existed in English, was included in the index:

1. Title
2. Abstract
3. Description
4. Claims

In addition, the patent IPC classification [9] was included in the index to be used later in filtering the retrieved results in the search process. Only the top three classification levels were retained for the filtering process with the deeper levels being discarded (example: B01J, C01G, C22B).

Further fields in the patent were not used; these included the fields which carry logistic information such as the patent filing date, institute, inventor's name and address, etc. Nevertheless, the 'inventors' field was tested using the training data assess its effectiveness in retrieving relevant documents. Results of these investigations showed it to have a weak effect on the quality of search. Hence, it was discarded from the index.

Translating non-English topics. 2,005 patent topics were provided including 1,351 English patent topics, 520 German topics, and 134 French topics. Out of these 5 topics were later excluded by the track organizers. Since the index was built only in English, German and French topics were translated into English using Google translate[2]. The 'title', 'description', and 'claims' sections were translated into English, while the 'abstract' field already had its English translation.

Query formulation. One major challenges in patent retrieval is query formulation [2], [7]. As a full patent is taken to be the topic, extracting the best representative text with the proper weights is a key to achieving good retrieval results.

Earlier experiments from our participation in CLEF-IP 2009 showed that using the full patent text to search the collection achieves the best results, especially after filtering all results which do not carry any overlap in the 'classification' section (only the first three levels of classification are used) [4]. The same setup was used this year by forming the query as follows:

- Unigram tokens were extracted from the 'description' field after stemming and stop word removal.

[2] http://translate.google.com/

- Bigram tokens of frequency higher than 3 were extracted from all the patent fields combined and added to the query after stemming and stop word removal.

Extracted queries can be seen a bit long, however, this formulation proved to be the best in our experiments using the training set. This first run is called the "IR" run in our submitted runs.

3.2 Citation Extraction

One of the features of patents is the presence of some of the cited patent numbers within the text of the description of the patents. These patent numbers were not filtered out of the text of the patent topics, which can be considered as the presence of part of the answer within the question. Despite this fact, we have not focused on building extra experiments based on this information, since in real life patent search situation this information is not always presented in the patent application, and hence, creating results on it can be considered as a misleading conclusion in the area of patent retrieval.

However, in the experimental results, adding this information to the tested methods is reported to demonstrate the impact of using this kind of information. Results show that a misleadingly high MAP can be achieved, but with a very low recall. This observation is significant since of course recall is usually the main objective for patent retrieval tasks.

For the large topic collection containing 2,005 patent topics, 2,307 citations were extracted from 771 patent topics and found to be IDs of patents in the indexed collection. Other extracted citations that do not exist in the collection were discarded. The extracted citations were put into the TREC format to be the second run submitted to the CLEF-IP track with the ID "Cit", standing for "citation".

The first run results list was appended to the second run list after removing the duplicates to act as our third run submitted to the track. This run ID is "IR+Cit".

4 Results

Several evaluation scores have been used for evaluating the submitted runs. Here, we focus on five scores: mean average precision (MAP), recall, recall@100, patent retrieval evaluation score (PRES) [5], and PRES@100. MAP and recall have so far been the two main metrics used for evaluating this task. However, our recently introduced PRES metric is designed to be a dedicated score for recall-oriented IR applications, such as patent search. PRES reflects the quality of the system in retrieving a large portion of the relevant documents in a relatively high ranks based on a user specific cut-off (N_{max}) [5]. This is the reason behind using the cut-off of 100, as it has been in shown in [1] that the average number of documents to be checked by a patent examiner is 100. In addition, PRES and recall are also calculated at the cut-off specified by the track organizers (1000).

Table 1 shows results for our three submitted runs for the large topic collection (2000 topics). The table shows two extreme runs, namely the "IR" run and the "Cit"

run. The "IR" run achieved high recall and moderate precision. On the other hand, the "Cit" run achieved a very high precision while a very low recall. Although the MAP of the "IR" run is higher than "Cit" run, the "Cit" run has a very high precision, as was mentioned in section 3.2, that only 771 topics out of the 2000 had citations extracted, which means that the MAP for these topics alone is 0.3. The last run "IR+Cit" achieves the highest recall and precision since it comes from a simple merging of the two previous runs. The PRES and PRES@100 scores reflect both the recall and the quality of ranking of the system. The "IR+Cit" run and the "IR" runs achieved the second and the third best results among the 25 runs submitted to the track according to PRES and PRES@100.

Table 1. MAP, recall, recall@100, PRES, and PRES@100 for the three submitted runs in CLEF-IP 2010.

Run #	MAP	R	R@100	PRES	PRES@100
IR	0.1216	0.57	0.3036	0.4614	0.228
Cit	0.112	0.1187	0.1187	0.1186	0.1176
IR+Cit	0.2029	0.618	0.3846	0.5229	0.3162

5 Conclusion and Future Work

In this paper, we have described our participation in the CLEF-IP 2010 Prior Art Patent Search Task. Three runs were submitted to the track, with one of them achieving the second best run among 25 runs submitted by 9 participants in this task. The three runs represent very simple and straightforward approaches for achieving high effectiveness in this task. Our run using standard IR techniques achieved the third highest performance among all submitted runs by participants according to recall and PRES. Our second run using straightforward citation extraction from patent topics achieved a very high precision performance. The third run with is a very simple merging of the first two runs achieved both high recall and precision (both reflected in PRES) to act as the best second run among the 25 runs submitted by the participants.

For future work, utilizing the information of the automatically extracted citation can be an interesting avenue for investigation. Semi-pseudo relevance feedback can be applied through extracting additional terms from these citations to help in improving the results. In addition, different approaches for translating the non-English patents can be tested, since further investigation showed that retrieval performance for the non-English topics was relatively lower than that of the English ones.

6 Acknowledgment

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

7 References

[1] Azzopardi L., H. Joho and W. Vanderbauwhede. A Survey on Patent Users Search Behavior, Search Functionality and System Requirements. *IRF Report* 2010-00001, (2010)

[2] Fujii A., M. Iwayama, and N. Kando. 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)

[3] Graf E. and L. Azzopardi. A methodology for building a patent test collection for prior art search. In *Proceedings of the Second International Workshop on Evaluating Information Access (EVIA)*, (2008)

[4] Magdy W., J. Leveling, and G. J. F. Jones. Exploring Structured Documents and Query Formulation Techniques for Patent Retrieval. In *CLEF working notes 2009, Corfu, Greece*, (2009)

[5] Magdy W. and G. J. F. Jones. PRES: A Score Metric for Evaluating Recall-Oriented Information Retrieval Applications. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, Geneva, Switzerland, pp. 611-618, (2010)

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

[7] Roda G., J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In *CLEF working notes 2009, Corfu, Greece*, (2009)

[8] Strohman T., 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)

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