Prior Art Search Using International Patent Classification Codes and All-Claims-Queries

Benjamin Herbert, György Szarvas*, and Iryna Gurevych

Ubiquitous Knowledge Processing (UKP) Lab
Computer Science Department
Technische Universität Darmstadt
Hochschulstr. 10, D-64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract. In this paper, we describe the system we developed for the Intellectual Property track of the 2009 Cross-Language Evaluation Forum. The track addressed prior art search for patent applications. We used the Lucene library to conduct experiments with the traditional TF-IDF-based ranking approach, indexing both the textual content and the IPC codes assigned to each document. We formulated our queries by using the *title* and *claims* of a patent application in order to measure the (weighted) lexical overlap between topics and prior art candidates. We also formulated a language-independent query using the IPC codes of a document to improve the coverage and to obtain a more accurate ranking of candidates. Using a simple model, our system remained efficient and had a reasonably good performance score: it achieved the 6th best Mean Average Precision score out of 14 participating systems on 500 topics, and the 4th best score out of 9 participants on 10,000 topics.

1 Introduction

The CLEF-IP 2009 track was organized by Matrixware and the Information Retrieval Facility. The goal of the track was to investigate the application of IR methods to patent retrieval. The task was to perform prior art search, which is a special type of search with the goal of verifying the originality of a patent. If a prior patent or document is found that already covers a very similar invention and no sufficient originality can be proven for a patent, it is no longer valid. In the case of a patent application, this would prevent it from being granted. If a patent has already been accepted, an opposition procedure can invalidate a patent by providing references to prior art. Therefore, finding even a single prior art document can be crucial in the process, as it may have an adverse effect on the decision about patentability, or withdrawal of an application.

Prior art search is usually performed at patent offices by experts, examining millions of documents. The process often takes several days and requires

* On leave from the Research Group on Artificial Intelligence of the Hungarian Academy of Sciences.

C. Peters et al. (Eds.): CLEF 2009 Workshop, Part I, LNCS 6241, pp. 452–459, 2010.

strict documentation and experienced professionals. It would be beneficial if IR methods could ease this task or improve the speed and accuracy of the search.

Major challenges associated with finding prior art include: the usage of vocabulary and grammar is not enforced and depends on the authors; in order to cover a wide field of applications, often generic formulations and vague language are used; the authors might even try to disguise the information contained in a patent and take action against people that infringe a patent later; the description of inventions frequently uses new vocabulary; information constituting prior art might be described in a different language than the patent under investigation.

Dataset & Task. For the challenge, a collection of 1.9 million patent documents taken from the European Patent Office (EPO) was used. The documents in this collection correspond to approximately 1 million individual patents filed between 1985 and 2000 (thus one patent can have several files, with different versions/types of information). The patents are in the English, German, or French language. The language distribution is not uniform as 70% of the patents are English, 23% are German, and 7% are French. The patents are given in an XML format and supply detailed information such as the title, description, abstract, claims, inventors and classification.

The focus of the challenge was to find prior art for the given topic documents. Several tasks were defined: the *Main* task, where topics corresponded to full patent documents, and the multilingual tasks, where only the *title* and *claims* fields were given in a single language (*English*, *German*, or *French*) and prior art documents were expected to be retrieved in any of these three languages.

Relevance assessments were compiled automatically using the citations pointing to prior art documents found in the EPO files of the topic patent applications. The training data for the challenge consisted of 500 topics and relevant prior art. The evaluation was carried out on document sets having 500 (Small), 1,000 (Medium) and 10,000 (XLarge evaluation) topics, respectively.

For a more detailed description of the task, participating groups, the dataset and overall results, please see the challenge paper [2] and the track Web page[1].

2 Our Approach

For most patents, several files were available, corresponding to different versions of the patent (an application text is subject to change during the evaluation process). We decided not to use all the different versions available for the patent, but only the most up-to-date version. We epxected the latest version to contain the most authoritative information. If a certain field used by our system was missing from that version, we extracted the relevant information from the latest source that included this field. In our system, we used the information provided under the *claims*, *abstract*, *description*, *title* and *IPC codes* fields only. We did not use other, potentially useful sections of patent applications such as authors or date.

[1] `http://www.ir-facility.org/the_irf/clef-ip09-track`

2.1 Preprocessing

We performed the following preprocessing steps:

- *Sentence splitting* based on the Java BreakIterator[2] implementation.
- *Tokenization* based on the Java BreakIterator (for the French documents we also used apostrophes as token boundaries: e.g. *d'un* was split to *d* and *un*). We converted all the tokens to lowercase.
- *Stopword removal* using manually crafted stopword lists. We started with general purpose stopword lists containing determiners, pronouns, etc. for each language, and appended them with highly frequent terms. We considered each frequent word (appearing in several hundred thousand documents) a potential stopword and included it in the list if we judged it a generic term or a domain specific stopword; that is, not representative of the patent content. For example, many patents contain words like *figure* (used in figure captions and also to refer to the pictures in the text), or *invention* (it usually occurred in the 1st sentence of the documents).
- for the German language, we applied dictionary-based *compound splitting*[3].
- *Stemming* using the Porter algorithm[4].

The preprocessing pipeline was set up using the *Unstructured Information Management Architecture (UIMA)*, a framework for the development of component based *Natural Language Processing (NLP)* applications. We employed the DKPro Information Retrieval framework [1], which provides efficient and configurable UIMA components for common NLP and Information Retrieval tasks.

2.2 Retrieval

The basis of our system is the extended boolean vector space model as implemented by Lucene. We queried the indices described below and combined the results in a post-processing step in order to incorporate information gathered from both the text and the IPC codes.

Indices. In order to employ Lucene for patent retrieval, we created a separate index for each language just using fields for the relevant language. For example, to create the German index, only fields with a language attribute set to *DE* were used.

For each patent, we extracted the text of a selection of fields (*title* only, *title & claims*, *claims & abstract & description* - limited to a fixed number of words). The concatenated fields were preprocessed in the way described above. For each patent, a single document was added to the Lucene index, and the patentNumber field was added to identify the patent.

Topic documents were preprocessed in the same manner as the document collection. All the text in the *title* and *claims* fields was used to formulate the

[2] `http://java.sun.com/j2se/1.5.0/docs/api/java/text/BreakIterator.html`
[3] `http://www.drni.de/niels/s9y/pages/bananasplit.html`
[4] `http://snowball.tartarus.org`

queries, without any further filtering. This way our system ranked documents according to their lexical overlap with the topic patent. A separate query was constructed for each of the languages.

To exploit the IPC codes assigned to the patents, a separate index was created containing only the IPC categories of the documents. We performed retrieval based on this IPC index, and since single IPC codes usually indentify particular scientific and technological domains, this provided a language independent ranking based on the domain overlap between the query and documents.

Queries. For the main task, sample topic documents were selected that had their *title* and *claims* fields available in all three languages. Moreover, since these documents were full patent applications they contained other fields, possibly in one or more languages, but we did not use any of these additional fields.

We created a separate query for each language and ran it against the document collection index of the corresponding language. Each query contained the whole content of the two above-mentioned fields, with each query term separated by the *OR* query operator.

For the language specific tasks, only the *title* and *claims* fields of the corresponding language were made available. We performed the same retrieval step as we did for the main task, but restricted the search to the respective language index. For instance in the French subtask, we just used the French *title* and *claims* fields to formulate our query and performed retrieval only on the French document index.

To measure the weighted overlap of the IPC codes, a separate query was formulated that included all IPC codes assigned to the topic document (again, each query term *OR*-ed together).

Result Fusion. Language specific result lists were filtered in such a way that documents which did not share an IPC code with the topic were filtered out. The language specific result lists were normalized in order to make the scores comparable to each other. The result list from the IPC code index was normalized in the same way. To prepare our system output for the language specific subtasks, we added the relevance scores returned by the IPC and the textual query and ranked the results according to the resulting relevance score. For the *Main* task submission, the three language-specific lists were combined into a single list by taking the highest score from each language specific result list for each document. For further details about the result list combination, see [3].

3 Experiments and Results

In this section we present the performance statistics of the system submitted to the CLEF-IP challenge and report on some additional experiments performed after the submission deadline. We apply Mean Average Precision (MAP) as the main evaluation metric, in accordance with the official CLEF-IP evaluation. Since precision at top rank positions is extremely important for systems that are

supposed to assist manual work like a prior art search, for comparison we always give the precision scores at 1 and 10 retrieved documents (P@1 and P@10)[5].

3.1 Challenge Submission

We utilized the processing pipeline outlined above to extract text from different fields of patent applications. We experimented with indexing single fields, and some combinations thereof. In particular, we used only titles, only claims, only description or a combination of *title* and *claims* for indexing.

As the *claims* field is the legally important field, we decided to include the whole *claims* field in the indices for the submitted system. We employed an arbitrarily chosen threshold of 800 words for the indexed document size. That is, for patents with a short *claims* field, we added some text taken from their abstract or description respectively, to have at least 800 words in the index for each patent. When the claims field itself was longer than 800 words, we used the whole field. This way, we tried to provide a more or less uniform-length representation of each document to make the retrieval results less sensitive to document length. We did not have time during the challenge timeline to tune the text size threshold parameter of our system, so this 800 words limit was chosen arbitrarily – motivated by the average size of *claims* sections.

Table 1 shows the MAP, P@1 and P@10 values and average recall (over topics, for top 1000 hits) of the system configurations we tested during the CLEF-IP challenge development period, for the Main task, on the 500 training topics. These were: **1)** the system using the IPC-code index only; **2)** the system using a text-based index only; **3)** the system using a text-based index only, the result list filtered for matching IPC code; **4)** a combination of result lists of 1) and 2); **5)** a combination of result lists of 1) and 3).

Table 1. Performance on Main task, 500 train topics

Nr.	Method	MAP	P@1	P@10	avg. recall
(1)	IPC only	0.0685	0.1140	0.0548	0.6966
(2)	Text only	0.0719	0.1720	0.0556	0.4626
(3)	Text only - filtered	0.0997	0.1960	0.0784	0.6490
(4)	IPC and text	0.1113	0.2140	0.0856	0.6960
(5)	**IPC and text - filtered**	**0.1212**	**0.2160**	**0.0896**	**0.7319**

The bold line in Table 1 represents our submitted system. This configuration gave the best scores on the training topic set for each individual language. Table 2 shows the scores of our submission for each language specific subtask and the Main task on the 500 training and on the 10,000 evaluation topics.

[5] During system development we always treated every citation as an equally relevant document, so we only present such an evaluation here. For more details and analysis of the performance on highly relevant items (e.g. those provided by the opposition), please see the task description paper [2].

Table 2. Performance scores for different subtasks on training and test topic sets

	Train 500				Evaluation 10k			
Task	**MAP**	**P@1**	**P@10**	**avg. recall**	**MAP**	**P@1**	**P@10**	**avg. recall**
English	0.1157	0.2160	0.0876	0.7265	0.1163	0.2025	0.0876	0.7382
German	0.1067	0.2140	0.0818	0.7092	0.1086	0.1991	0.0813	0.7194
French	0.1034	0.1940	0.0798	0.7073	0.1005	0.1770	0.0774	0.7141
Main	0.1212	0.2160	0.0896	0.7319	0.1186	0.2025	0.0897	0.7372

3.2 Post Submission Experiments

After the submission, we ran several additional experiments to gain a better insight into the performance limitations of our system. We only experimented with the English subtask, for the sake of simplicity and for time constraints. First, we experimented with different weightings for accumulating evidence from the text- and IPC-based indices.

We found that slightly higher weight to text-based results would have been beneficial to performance in general. Using 0.6/0.4 weights, which was the best performing weighting on the training set, would have given a 0.1202 MAP score for English, on the 10k evaluation set – which is a 0.4% point improvement.

We also examined retrieval performance using different document length thresholds. Hence, we extracted the first 400, 800, 1600 or 3200 words of the concatenated claims, abstract and description fields to see whether more text could improve the results. Only a slight improvement could be attained by using more text for indexing documents. On the training set, the best score was achieved using 1600 words as the document size threshold. This would have given 0.1170 MAP score for English, on the 10k evaluation set – which is only a marginal improvement over the submitted configuration.

Previously we discarded all resulting documents that did not share an IPC code with the topic. This way, retrieval was actually constrained to the cluster of documents that had overlapping IPC codes. A natural idea was to evaluate whether creating a separate index for these clusters (and thus having in-cluster term weighting schemes and ranking) is beneficial to performance. We found that using such local term weights improved the performance of our system for each configuration. The best parameter settings of our system on the training topics were: 1600 words threshold; 0.6/0.4 weights for text/IPC indices; indexing the cluster of documents with a matching IPC code for each topic. This provided a MAP score of 0.1243, P@1 of 0.2223 and p@10 of 0.0937 on the 10,000 document evaluation set, for English. This is a 0.8% point improvement compared to our submitted system.

3.3 Significance Analysis

We used the *paired t-test* for significance tests. Due to the huge number of topics (we presented results for 10,000 topics) even small differences in MAP values tend to be statistically significant using very low significance thresholds.

First, the difference in performance between our submitted system and that of our post-submission experiments was statistically significant ($P < 10^{-4}$) even though the latter system only utilized English texts, not all three languages. This means that for our approach it was more important to set the corresponding weights for the combination, indexed text size and to use accurate term weights (in-cluster) than to exploit results for less frequently used languages. The performance of our submitted system (0.1186 MAP, placed 4th in the XL evaluation) is significantly different from both the one placed third (0.1237 MAP, $P < 10^{-3}$) and fifth (0.1074 MAP, $P < 10^{-4}$). The performance of our post-submission system (0.1243 MAP) is significantly different from the one placed second in the challenge (0.1287 MAP, $P < 10^{-2}$), but the difference compared to the third place system is not significant (0.1237 MAP, $P > 0.5$).

4 Discussion

In the previous section we introduced the results we obtained during the challenge timeline, together with some follow-up experiments. We think our relatively simple approach gave fair results, our submission came 6th out of 14 participating systems on the evaluation set of 500 topics[6] and 4th out of 9 systems on the larger evaluation set of 10,000 topics. Taking into account the fact that just one participating system achieved remarkably higher MAP scores and the simplicity of our system, we find these results promising.

We should mention here that during the challenge development period, we made several arbitrary choices regarding system parameter settings, and that even though we chose reasonably well performing parameter values tuning these parameters could have improved the accuracy of the system to some extent.

The limitations of our approach are obvious though. First, as our approach mainly measures lexical overlap between the topic patent and prior art candidates, prior art items that use substantially different vocabulary to describe their innovations are most probably missed by the system. Second, without any sophisticated keyword / terminology extraction from the topic claims, our queries are long and probably contain irrelevant terms that place a burden on the system's accuracy. Third, the patent documents provided by the organizers were quite comprehensive, containing detailed information on inventors, assignees, priority dates, etc. Out of these information types we only used the IPC codes and some of the textual description of patents. Last, since we made a compromise and searched among documents with a matching IPC code (and only extended the search to documents with a matching first three digits of IPC when we had an insufficient number of retrieved documents in the first step), we missed those prior art items that have a different IPC classification from the patent being investigated. We think these patents are the most challenging and important items to identify and they are rather difficult to discover for humans as well.

[6] Since the larger evaluation set included the small one, we consistently reported results on the largest set possible. For more details about performance statistics on the smaller sets, please see [2].

5 Conclusions and Future Work

In this study, we demonstrated that even a simple Information Retrieval system measuring the IPC-based and lexical overlap between a topic and prior art candidates works reasonably well: our system gives a True Positive (prior art) top ranked for little more than 20% of the topics. We think that a simple visualization approach like displaying content in a parallel view highlighting textual/IPC overlaps could be an efficient assistant tool for a manual prior art search (performed at Patent Offices).

In the future we plan to extend our system in several different ways. We already know that local and global term weightings behave differently in retrieving prior art documents. A straightforward extension would be therefore to incorporate both weightings in order to improve our results even further. Similarly, experimenting with other weighting schemes than the one implemented in Lucene is another straightforward way of extending our current system.

Acknowledgements

This work was supported by the German Ministry of Education and Research (BMBF) under grant 'Semantics- and Emotion-Based Conversation Management in Customer Support (SIGMUND)', No. 01ISO8042D, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.

References

1. Müller, C., Zesch, T., Müller, M.C., Bernhard, D., Ignatova, K., Gurevych, I., Mühlhäuser, M.: Flexible UIMA Components for Information Retrieval Research. In: Proceedings of the LREC 2008 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP', Marrakech, Morocco, pp. 24–27 (May 2008)
2. Roda, G., Tait, J., Piroi, F., Zenz, V.: CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In: Working Notes of the 10th Workshop of the Cross Language Evaluation Forum (CLEF), Corfu, Greece (2009)
3. Szarvas, G., Herbert, B., Gurevych, I.: Prior Art Search using International Patent Classification Codes and All-Claims-Queries. In: Working Notes of the 10th Workshop of the Cross Language Evaluation Forum (CLEF) (August 2009)