Prior art retrieval using the claims section as a bag of words

Suzan Verberne and Eva D'hondt
(s.verberne|e.dhondt)@let.ru.nl

Information Foraging Lab, Radboud University Nijmegen

Abstract. In this paper we describe our participation in the 2009 CLEF-IP task, which was targeted at prior-art search for topic patent documents. We opted for a baseline approach to get a feeling for the specifics of the task and the documents used. Our system retrieved patent documents based on a standard bag-of-words approach for both the Main Task and the English Task. In both runs, we extracted the claim sections from all English patents in the corpus and saved them in the Lemur index format with the patent IDs as DOCIDs. These claims were then indexed using Lemur's BuildIndex function. In the topic documents we also focused exclusively on the claims sections. These were extracted and converted to queries by removing stopwords and punctuation. We did not perform any term selection or query expansion. We retrieved 100 patents per topic using Lemur's RetEval function, retrieval model TF-IDF. Compared to the other runs submitted to the track, we obtained good results in terms of nDCG (0.46) and moderate results in terms of MAP (0.054).

1 Introduction

In 2009 the first CLEF-IP track was launched by the Information Retrieval Facility (IRF)[1] as part of the CLEF 2009 evaluation campaign.[2] The general aim of the track was to explore patent searching as an IR task and to try to bridge the gap between the IR community and the world of professional patent search.

The goal of the 2009 CLEF-IP track was "to find patent documents that constitute prior art[3] to a given patent" [1]. In this retrieval task each topic query was a (partial) patent document which could be used as one long query or from which smaller queries could be generated. The track featured two kinds of tasks: In the Main Task prior art had to be found in any one (or combination) of the three following languages: English, French and German; three optional subtasks used parallel monolingual topics in one of the three languages. In total 15 European teams participated in the track.

At the Radboud University of Nijmegen we decided to participate in the CLEF-IP track because it is related to the focus of the Text Mining for Intellectual Property (TM4IP) project[4] that we are currently carrying out. In this project we investigate how linguistic knowledge can be used effectively to improve the retrieval process and facilitate interactive search for patent retrieval. Because the task of prior-art retrieval was new to us, we chose to implement a baseline approach to investigate how well traditional IR techniques work for this type of data and where improvements would be most effective. These results will effectively serve as a baseline for further experiments as we explore the influence of using dependency triplets[5] for various IR tasks on the same patent corpus.

[1] See http://www.ir-facility.org/the_irf/clef-ip09-track
[2] See http://www.clef-campaign.org/
[3] Prior art for a patent (application) means any document (mostly legal or scientific) that was published before the filing date of the patent and which describes the same or a similar invention.
[4] http://www.phasar.cs.ru.nl/TM4IP.html
[5] A dependency triplet is a unit that consists of two open category words and a meaningful grammatical relation that binds them.

2 Our methodology

2.1 Data selection

The CLEF-IP corpus consists of EPO documents with publication date between 1985 and 2000, covering English, French, and German patents (1,958,955 patent-documents pertaining to 1,022,388 patents, 75GB) [2]. The XML documents in the corpus do not correspond to one complete patent each but one patent can consist of multiple XML files (representing documents that were produced at different stages of a patent realization).

In the CLEF-IP 2009 track, the participating teams were provided with 4 different sets of topics (S,M,L,XL). We opted to do runs on the smallest set (the S data set) for both the Main and the English task. This set contained 500 topics. There appeared to be a number of differences in the information that is contained in the topics for the Main task and the English task: the topics for the Main Task contained the abstract content as well as the full information of the granted patent except for citation information, while the topic patents for the English Task only contained the title and claims sections of the granted patent [2].

Therefore, we decided to use the field that was available in all topics for both tasks: the (English) claims sections. Moreover, as [3], [4] and [5] suggest, the claims section is the most informative part of a patent, at least for prior-art search. We found that 70% of the CLEF-IP corpus contained English claims, as a result of which a substantial part of the corpus was excluded from our experiments. Of the 30% that could not be retrieved by our system, 7% were documents that only had claims in German or French but not in English, 6% only contained a title and abstract, usually in English and 17% only contained a title.

2.2 Query formulation

At the CLEF-IP meeting there was much interest on term extraction and query formulation, for example from the University of Glasgow [6]. Though this seems to be a promising topic, we choose not to distil any query terms from the claims sections but instead concatenated all words in the claims section in one long query. The reason for this was twofold. First, adding a term selection step in the retrieval process makes the retrieval process more prone to errors because it requires the development of a smart selection process. Second, by weighting the query and document terms using the TF-IDF ranking model, a form of term selection is carried out in the retrieval and ranking process. We did not try to enlarge the set of query words with any query expansion technique but only used the words as they occurred in the texts.

2.3 Indexing and Retrieval using Lemur

We extracted the claims sections from all English patents in the corpus after removing the XML markup from the texts. Since a patent may consist of multiple XML documents, which correspond to the different stages of the patent realization process, one patent can contain more than one claims section. In the index file, we concatenated the claims sections pertaining to one patent ID into one document. We saved all patent claims in the Lemur index format with the patent IDs as DOCIDs. One entry in the index looks like this:

```
<DOC><DOCNO>EP-0148743</DOCNO>
<TEXT> A thermoplastic resin composition comprising a melt mixed product of
(A) 70% to 98% by weight of at least one thermoplastic resin selected from the
group consi sting of polyamides, polyacetals, polyesters, and polycarbonates
and (B) 30% to 2% by weight of a modified ultra-high molecular weight polyolefin
powder having an average powder particle size of 1 to 80 m and having a particle
size distribution such that substantially all of the powder particles pass through
a sieve having a sieve mesh of 0.147 mm and at least 20% by weight of the total
powder particles pass through a sieve having a sieve mesh of 0.041 mm, said
```

```
polymer being modified by graft copolymerizing unsaturated carboxylic acid derivative
units having at least one polar group selected from the group consisting of
acid groups, acid anhydride group,and ester groups and derived from an unsaturated
carboxylic acid or the acid anhydride, the salt, or the ester thereof to ultra-high
molecular weight polyolefin having an intrinsic viscosity [#] of 10 dl/g or
more, measured in decalin at 135C.
</TEXT>
</DOC>
```

The claims sections were then indexed using Lemur's BuildIndex function with the Indri IndexType and a stop word list for general English. The batch retrieval and ranking was then performed using the TF-IDF ranking model as it has been included in Lemur. We did not compare the different ranking models provided by Lemur to each other since the goal of our research is not to find the optimal ranking model[6] but to explore the possibilities and difficulties of any BOW approach.

3 Results

We performed runs for the Main and English Task with the methodology described above. Since we used the same data for both runs, we obtained the same results. These results are in Table 1. The first row shows the results that are obtained if all relevant assignments are taken into consideration; the second row contains the results for the highly-relevant citations only [8].

Table 1. Results for the clefip-run 'ClaimsBOW' on the small topic set using English claims sections for both the Main Task and the English Task.

	P	P5	P10	P100	R	R5	R10	R100	MAP	nDCG
All	0.0129	0.0668	0.0494	0.0129	0.2201	0.0566	0.0815	0.2201	0.0540	0.4567
Highly-relevant	0.0080	0.0428	0.0314	0.0080	0.2479	0.0777	0.1074	0.2479	0.0646	0.4567

4 Discussion

Although the results we obtained with our ClaimsBOW approach may seem poor on first sight, they are not bad compared to the results obtained by other participants. In terms of nDCG, our run performs well (ranked 6th of 70 runs); in terms of MAP our results are moderate (ranked around 35th of 70 runs). The low performance achieved by almost all runs (except for one submitted by Humboldt University) shows that the task at hand is a difficult one.

There are a number of reasons for these low scores: First of all, some of the documents were 'unfindable': 17% of the patent documents in the collection contained so little information, e.g. only the title which is poorly informative for patent retrieval [9], that they could not be retrieved. Secondly, the relevance assessments were based on search reports and the citations in the original patent only. This means that they were incomplete [1].

Finally, in order to perform retrieval on the patent level, instead of the document level, some of the participating groups created 'virtual patents': for each field in the patent the most recent information was selected from one of the documents with that patentID. These fields were glued together to form one whole 'virtual' patent. It is, however, not necessarily true that the most recent fields are the most informative [9]. This selection may have resulted in a loss of information. However, even without these impediments, it is clear that patent retrieval is a difficult task for standard retrieval methods.

[6] Such experiments where conducted by the BiTeM group who also participated in this track [7].

The discussion at the CLEF-IP meeting showed that merely text-based retrieval is not enough for patent retrieval. Those groups that made use of the metadata in the patent documents (e.g. classification information) scored remarkably better than those relying on standard text-based methods.

5 Conclusion

The CLEF-IP track was very valuable to us as we now have a baseline that is based on standard bag-of-words text retrieval techniques. In future work we are going to focus on improving the ranking of the result list that we produced in the CLEF-IP experiment. We plan to apply an additional reranking step to the result set using syntactic information in the form of dependency triplets [10].

References

1. Roda, G., Tait, J., Piroi, F., Zenz, V.: CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In: CLEF working notes 2009. (2009)
2. Piroi, F., Roda, G., Zenz, V.: CLEF-IP 2009 Track Guidelines. Technical report, Information Retrieval Facility (2009)
3. Graf, E., Azzopardi, L.: A methodology for building a test collection for prior art search. In: Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA). (2008) 60–71
4. Shinmori, A., Okumura, M., Marukawa, Y., Iwayama, M.: Patent claim processing for readability: structure analysis and term explanation. In: Proceedings of the ACL-2003 workshop on Patent corpus processing, Morristown, NJ, USA, Association for Computational Linguistics (2003) 56–65
5. Iwayama, M., Fujii, A., Kando, N., Marukawa, Y.: Evaluating patent retrieval in the third NTCIR workshop. Information Processing Management **42**(1) (2006) 207–221
6. Graf, E., Azzopardi, L., Van Rijsbergen, K.: Automatically Generating Queries for Prior Art Search. In: CLEF working notes 2009. (2009)
7. Gobeill, J., Theodoro, D., Ruch, P.: Exploring a wide Range of simple Pre and Post Processing Strategies for Patent Searching in CLEF IP 2009. In: CLEF working notes 2009. (2009)
8. Piroi, F., Roda, G., Zenz, V.: CLEF-IP 2009 Evaluation Summary. Technical report, Information Retrieval Facility (2009)
9. Tseng, Y., Wu, Y.: A study of search tactics for patentability search: a case study on patent engineers. In: Proceeding of the 1st ACM workshop on Patent information retrieval. (2008) 33–36
10. D'hondt, E., Verberne, S., Oostdijk, N., Boves, L.: Re-ranking based on Syntactic Dependencies in Prior-Art Retrieval. In: Proceedings of the Dutch-Belgium Information Retrieval Workshop 2010. (2010) To appear.