Report on the CLEF-IP 2011 Experiments: Exploring Patent Summarization

Parvaz Mahdabi[1] Linda Andersson[2] Allan Hanbury[2] Fabio Crestani[1]
parvaz.mahdabi@usi.ch andersson@ifs.tuwien.ac.at hanbury@ifs.tuwien.ac.at fabio.crestani@usi.ch

[1] University of Lugano, Switzerland
[2] Vienna University of Technology, Austria

Abstract. This technical report presents the work carried out for the Prior Art Candidate Search track of CLEF-IP 2011. In this search scenario, information need is expressed as a patent document (query topic). We compare two methods for estimating query model from the patent document to support summary-based query modeling and description-based query modeling. The former approach utilizes a known text summarization technique, called "TextTiling", and is adopted for patent documents. The latter approach uses the description section of a patent document for estimating the query model. With summary-based query modeling we aspire to capture the main topic of the document as well as the most important subtopics and discard subtopics, which are only marginally discussed in the patent document. We submitted four runs for the Prior Art Candidate Search task. According to recall@1000 our best run was ranked 3rd across 6 participants and 8th, across all 30 submitted runs. In terms of MAP our best run achieved the 3rd rank across participants and 4th rank, across all runs.

keywords: Patent Retrieval, Query Generation, Patent Summarization, CLEF-IP track

1 Introduction

This paper presents the participation of University of Lugano in collaboration with Vienna University of Technology in the Prior Art Candidate Search task of CLEF-IP 2011. This track has been running since 2009 and it is an important platform for comparing the retrieval performance of different patent retrieval systems and testing new ideas. However, compared to other test collections within the IR community, patent retrieval is known to be a difficult search task.

Different term weighting techniques, IR models and ranking functions developed and tested within the CLEF and TREC tracks have been reused on the patent collections, but the expected retrieval effectiveness do not occur [2]. It is shown that even for the best runs of CLEF-IP, the retrieval effectiveness is quite lower compared to other domains in Information Retrieval [4, 5].

The goal of Prior Art Candidate Search is to find all relevant documents for a given patent (considered as query topic) [7]. We submitted four runs and used the topic set of CLEF-IP 2010 as our training data for tweaking the parameters.

In total, 6 participants submitted 30 runs for this task. Our best performing run was ranked 3rd across participants and 4th, across all the runs in terms of MAP. According to recall@1000 our best run was ranked 3rd across participants and 8th, across all the runs.

This paper is organized as follows. In Section 2 we detail our summarization technique and query modeling approach. In Section 3 we describe the details of our experimental setup. In Section 4 we report the evaluation results of our submitted runs. We follow with an analysis in Section 5 and a conclusion in Section 6.

2 Our Approach

An important goal for us is to devise a summary from a patent document. Our intuition is that the patent summary will reflect the main topic as well as the subtopics of a patent document in a concise manner. We focus on generating the summary to improve our query formulation. We were motivated to do so because of the two following reasons:

Our previous work [6] showed that queries generated from the description section outperform generated queries from the claims section. Although, it is known that patent examiners use claims section for query formulation. We tried to merge the results of different sections—to exploit all the available textual information. But this merging did not shown to be helpful. In the present work we address this problem by building a summary from the patent document.

In a recent study [8] on automatic query generation from patent documents, authors experimented with US patents and found that "background summary" performs as the best field for extracting query terms. Since the background summary is not available in the European patents, we decided to create a summary which resembles the background summary.

In this section we describe the details of our approach. We first explain how to build a patent summary. We then discuss our take on query generation. Finally, we explain our citation extraction technique.

2.1 Patent Summarization

Our summarization technique PatTextTiling is a modification of TextTiling—a state of the art text summarization algorithm [3]. Automatic text segmentation and text summarization techniques aspire to capture documents main topic and subtopics by analyzing the pragmatic structure in terms of cohesive markers and text coherence. TextTiling divides the text into sequences with N tokens. The benefit of having a fixed number is that each sequence carries the same amount of information. For each text segment consisting of N number of sentence sequences a depth score will be produced. The depth score indicates the gap cohesion which represents a topic shift in the text.

2.2 Query Generation

In this work, we aim to employ the knowledge embedded in IPC classes, to generate important terms and also to improve the retrieval performance. Patent documents are annotated with IPC classes which represents the different areas of technology to which a patent document pertains. We define the relevance set consisting of documents that have same IPC classes as the query topic. Each relevant document from this sample is considered as evidence towards the estimation of the relevance model. We assume documents in relevance set have equal importance. This set is more specific in contrast to what we used in our previous work [6].

We estimate the importance of each term with a weighted log-likelihood based approach as shown in Equation 1 . $H(\theta_Q, \theta_{Coll})$ represents the cross entropy between the query and the collection and $H(\theta_Q, \theta_{Cluster})$ represents the cross entropy between the query and the cluster.

$$H(\theta_Q, \theta_{Coll}) - H(\theta_Q, \theta_{Cluster}) \propto p(w|\theta_Q)\, log\left(\frac{p(w|\theta_{Cluster})}{p(w|\theta_{Coll})}\right) \quad (1)$$

This approach favors terms which have high similarity to the document language model θ_Q and the cluster language model $\theta_{Cluster}$ and low similarity to the collection language model θ_{Coll}. We use maximum likelihood estimates for calculating the language models.

We have two versions of estimating the query model. First, we build a query model for the summary of the patent document. This method for query modeling is referred to as summary-based query modeling (SM). Second, we build a query from the description section of the patent. We refer to this method as description-based query modeling (DM). Full details of the query generation approach can be found in [6].

2.3 Citation Extraction

Making use of distinguishing events (e.g. patent application number) in unrestricted text could be considered as a form of known-item search. The known-item search is applied as a search strategy to facilitate the extraction of key terms and synonyms that later can be used in a non-known item search. Therefore, we chose to extract the citations in the unrestricted text from all language sections and add surrounding text from the English text in the topic queries. We extracted the citations with a two-stage regular expression approach. The first step consists of capturing sentences with at least 4 digit sequences combined with and without hyphen. The next stage aimed to reduce the false positive by a set of regular expression sequences, where letter prefix was checked against a positive stop word list consisting of all accepted country codes.

3 Experiments

We first perform query generation on the patent summary. We refer to this run as SM. Next, we use the same method for query generation but instead of selecting terms from the summary of query topic, we select terms from the description section of the query topic. The output of this method is our second run called DM. We filter the ranked list of both runs by excluding documents which do not have at least one IPC class in common with the query document. After that, we use the list of direct citations extracted for each query topic and we combined this list with our keyword-based run (by performing a linear combination). The output of this combination is two more runs which we refer to as Cit+SM and Cit+DM. The evaluation of our runs is presented in section 4.

In this section we explain our experimental setup and different parameter settings we used for our submitted runs.

3.1 Experimental Set-Up

We index the collection with Terrier[1]. Our preprocessing consists of stop-word removal and stemming using Porter stemmer. In the experiments we use the BM25 implementation of Terrier. We limit our experiments to the English subset of the collection. As explained before, we build two query models: one based on summary and one based on the description. However, for the retrieval we use full text of the documents.

Tables 1 and 2 list some statistical properties of the English subset of the CLEF-IP 2011 collection.

Table 1. Properties of CLEF-IP 2011 collection (EP section)

EP source	Avg. document length	Avg. unique terms	No. Documents
Title	28	23	1,824,499
Abstract	90	57	904,277
Description	5079	718	962,686
Claims	577	123	1,151,609

The average number of unique terms and document length for each section are displayed for both EP and WO subsections of the collection. The last column displays the number of patent documents which were used in the calculations. We considered an additional condition while calculating the statistics. We performed our calculation for documents where the English language meta-tags are consistent with two independent language detection application (one based on stop words and one using n-gram technique). This is due to the fact that there are about 80,000 language meta-tags on section level where the meta-tags show inconsistency with the suggestion of the language detection applications.

[1] http://terrier.org/

Table 2. Properties of CLEF-IP 2011 collection (WO section)

WO source	Avg. document length	Avg. unique terms	No. Documents
Title	20	16	311,755
Abstract	91	57	223,348
Description	5632	916	182,653
Claims	904	147	182,625

3.2 Parameter Settings

We explain the parameter settings used for summarization technique. We used a basic TextTiling Perl module with the following parameters:

1. Number of tokens per sequence which reflects the document length
2. Sequence of window gap (default 2)
3. Smoothing round (default 2)
4. Minimum segment size (default 3)
5. Number of segments

For parameters 1 and 2 we first tried to set these values dynamically according to each section length and the sentence length but the performance decreased. The best performance was obtained when parameter 1 was set to 100 and parameter 2 was set to 2. For the gap sequence and the smoothing round we used the default value. The minimum size was increased to 7 and the number of segments was set according to the number of paragraph meta-tags present in the text.

In PatTextTiling additional binary weight was given to the abstract and specific paragraphs with citation or heading (e.g. Prior-Art, Background) in the description section—if present they were included in the final summery. For the description and claims sections the lexical cohesive gap distribution were first computed independently of each other; and once again on the selected text segments. The description section was given a more granular threshold meanwhile the claims section had a reduced granular threshold due to the fact of its stylistic repetitive writing. The threshold for description and claims were twofold: one based upon average difference in the cohesive gap and one fixed to a threshold value (claims 30 and description 20). The fixed values were added due to the fact of the diversity in the gap scores found among topic set documents. The information found in Lists and Tables were not included in the final summery.

4 Results

Organizers used different evaluation scores for evaluating the submitted runs. We used MAP, ndcg, P@100, P@500, recall@100, recall@500 and recall@1000 to report our retrieval performance on this task. Table 3 shows the results of our submitted runs on the English subset of queries which is composed of 1351 queries.

We mainly focused on the textual information for which we submitted SM and DM run. We also detected the direct citation information present in the query topic and we combined it with our first two runs. The output of this combination is two more runs denoted as Cit+SM and Cit+DM.

Table 3. Comparison of two query estimation methods (SM and DM) and the combination with the direct citations (Cit+SM and Cit+DM)

Method	MAP	ndcg	P@100	P@500	recall@100	recall@500	recall@1000
SM	0.0871	0.2305	0.0206	0.0064	0.2789	0.423	0.5254
DM	0.088	0.2318	0.0209	0.0064	0.2822	0.4287	0.5261
Cit+SM	0.0887	0.2331	0.0207	0.0064	0.2808	0.4245	0.5283
Cit+DM	0.0896	0.2344	0.021	0.0065	0.2842	0.4303	0.529

We zoom in to our best performing run Cit+DM to see the effect of extracted citations. We only managed to extract citations for 102 query topics out of 1351 English query topics and the average number of found citation for each topic is 1. Most of the identified citations in the unrestricted text did not have EP and WO numbers. Since we ignored the patent numbers which did not exist in the collection, our citation extraction runs performed just slightly better than our runs without citation. An interesting observation was that several extracted citations were cited by more than one query topic. On average each extracted citation was cited 1.24 times.

In order to fully explore the citation extraction mechanism it has to be used in combination with an online service (e.g. Open Patent Services[2]) to map identified references to a valid patent application number.

Table 4 shows the evaluation scores for Cit run. This run has a comparable MAP to submitted runs but it has a poor recall. This explains why the text and citation combination is not improving their matching runs without citations, as expected.

Table 4. Evaluation scores for Cit run

Method	MAP	ndcg	P@100	P@500	recall@100	recall@500
Cit	0.07	0.1329	0.005	0.001	0.0784	0.0784

5 Analysis

In this section we performed some analysis with the aim to identify the low retrieval effectiveness of the SM run. In order to analyze this we looked into some

[2] http://www.epo.org/searching/free/ops.html

features characterizing both the topics and the qrels. In the following analysis we used the 1348 English topics belonging to the topic set of the CLEF-IP 2010. We used the topic set of last year for performing our analysis. We considered per topic analysis and we first looked into the number of topics in SM run which have a higher Average Precision (AP) value compared to the DM run. The output of this analysis shows that SM run outperforms DM run for 618 topics. While, the DM run outperforms SM run for 628 topics. Figure 1 shows the AP differences between SM and DM. For some topics SM works best while for others DM works best and it is mostly a balanced picture. Therefore, it is not easy to favor one approach against the other.

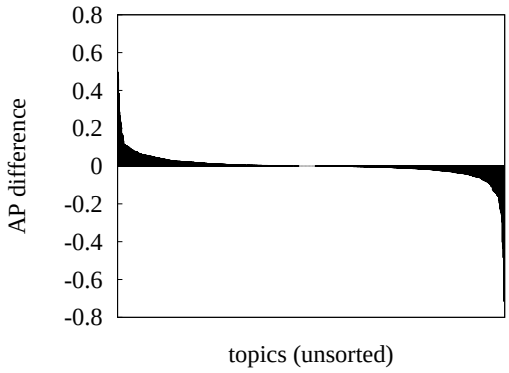

Fig. 1. AP differences between SM and DM

We zoom into one example topic where SM performs better than DM. This example concerns the topic 1038 where the title of the document is *Damping arrangements for Y25 bogies*. Table 5 reports MAP and recall scores and Table 6 shows the top 10 terms for the query models constructed for topic 1038 with SM performing much better than DM.

Table 5. Performance on topic # 1038

run	MAP	recall
SM	0.2599	1
DM	0.022	0.75

Table 6. Query models for topic "Damping arrangements for Y25 bogies"

*S*M	DM
pedestal	bogie
piston	axle
bogie	pedestal
axle	box
box	spring
arrangement	resilient
damp	damp
spring	relative
lenoir	movement
mount	wagon

SM managed to identify all relevant documents for this query through the terms introduced by SM query model. As it is displayed in the Table 6 the terms *piston* and *arrangement* are only selected with the SM and not with the DM.

Next feature we examined is the *non-retrievable* topics of each run, i.e. the number of topics that no relevant document was retrieved by that run. The size of non-retrievable set for SM is 63 and the size of non-retrievable set for DM is 52. We calculated the overlapping between the non-retrievable set of SM and DM and we found that for 42 topics none of the runs managed to retrieve any relevant documents. We first looked into the number of relevant documents for non-retrievable topic set in the qrels. Based on our findings, this feature was not able to distinguish non-retrievable topic set from the other topics.

We then decided to check the document length of the relevant documents for non-retrievable topic set. To our surprise, this investigation showed us that 0.48 of relevant documents do not contain any English text apart from the title. As a contrast, we decided to compare this with the *easy-retrievable*, i.e. topics which were retrieved by both methods with an AP value over 0.9. The corresponding value for this set is 112 topics and 0.18 of retrieved relevant documents for this set contain only title section in English.

These rather contradicting facts indicate that our methods managed to retrieve relevant documents where only the title of relevant documents existed for the easy-retrievable set. One reason for the good performance of our methods despite the lack of the text, could be the extra weight given to the surrounding text of the unrestricted citation.

The lack of the text in the qrels of non-retrievable set is one of the reasons which explains the low retrieval effectiveness of our runs on this set. According to Bashir and Rauber [1] this problem can be considered as a retrievability bias.

Depending on how the similarity between a query and a document is measured, some documents maybe more or less retrievable in certain systems, up to some documents not being retrievable at all within common threshold settings. Retrieval biases are due to different factors such as the popularity of a document (e.g. increasing weight of references), length of documents and structural infor-

mation such as metadata or headings. Therefore, in such scenarios one search strategy alone (e.g. keyword search) does not perform well.

6 Conclusion and Future Work

In this paper, we presented the experiments and results of our participation in CLEF-IP 2011 Prior Art Candidate Search task. We submitted four runs to this track. For our first run we built a summary of the patent document and then we introduced a method for sampling query terms from the patent summary.

In our second run we used the description section of a patent document for sampling query terms. For our third and fourth runs, we combined the extracted citations from the topics with our first two runs. According to the evaluation results our text summarization run performed slightly lower than the run based on the description. One reason for this is that words in Lists and Tables of the query topic were not included in the patent summary. In addition, the parameter setting of the text summarization technique needs to be further explored.

In this work, we used the documents with same IPC classes as query topic to calculate the sampling distribution. In an extension to this, we can also take the citations and use them for estimating the relevancy. Moreover, a document's importance can be approximated by its relevance to the original query and this can be used as a document prior.

For our future work, we need to explore other retrieval mechanisms such as bibliographic data to address the problem of missing text. In terms of query modeling, in addition to unigrams, we need to consider n-grams to capture concepts.

7 Acknowledgements

Authors would like to thank Information Retrieval Facility (IRF) for the support of this work.

References

1. S. Bashir and A. Rauber. Analyzing document retrievability in patent retrieval settings. In *Database and Expert Systems Applications, 20th International Conference, DEXA 2009*, pages 753–760, 2009.
2. S. Bashir and A. Rauber. On the relationship between query characteristics and IR functions retrieval bias. In *Journal of the American Society for Information Science and Technology*, volume 2:8, pages 1515–1532, 2011.
3. M. A. Hearst. Context and structure in automated full-text information access. In *PhD Thesis, UC Berkeley Computer Science Technical Report number UCB/CSD-94/836*, 1994.
4. P. Lopez and L. Romary. Experiments with citation mining and key-term extraction for prior art search. *Workshop of the Cross-Language Evaluation Forum, LABs and Workshops, Notebook Papers*, 2010.

5. W. Magdy and G. J. F. Jones. Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task. *Workshop of the Cross-Language Evaluation Forum, LABs and Workshops, Notebook Papers*, 2010.
6. P. Mahdabi, M. Keikha, S. Gerani, M. Landoni, and F. Crestani. Building queries for prior-art search. In *IRFC*, pages 3–11, 2011.
7. F. Piroi. CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. *Workshop of the Cross-Language Evaluation Forum, LABs and Workshops, Notebook Papers*, 2010.
8. X. Xue and W. B. Croft. Transforming patents into prior-art queries. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 808–809, 2009.