Passage Retrieval Starting from Patent Claims

A CLEF-IP 2013 Task Overview

Florina Piroi, Mihai Lupu, Allan Hanbury

Vienna University of Technology,
Institute of Software Technology and Interactive Systems,
Favoritenstrasse 9-11, 1040 Vienna, Austria

Abstract. Most of the searches a patent expert at a patent office does are using boolean methods to query large databases of patent data. The CLEF-IP evaluation track is designed to experiment with information retrieval techniques on the patent domain. The data corpus in the CLEF-IP Lab consists of patent documents published by the European Patent Office. One of the main tasks in the Lab has been related to the Prior Art type of search performed by the patent experts at patent offices. The task has went through various changes along the years, from using virtual patent documents as topics (in 2009) to actual patent application documents, and sets of claims from patent application documents (2012 and 2013). Relevance assessments for this task were based on Search Reports published by the European Patent Office.
In this overview we give report on the work we have done in organizing this retrieval task in 2013.

1 The CLEF-IP Passage Retrieval Task

The technological developments in our time are closely coupled with the patent system which encourages inventors to make their ideas public in exchange for a monopoly on the invention for a limited period of time, up to 20 years. A patent can be seen as a contract between a government and the patent owner by which the latter can exclude other parties from manufacturing and exploiting the invention without a permission.

To obtain a patent, one of the main requirements is that the invention is new. To verify this, extensive searches, not only in the patent repositories, but also specialized literature, conference publications, etc., must pe thoroughly done. The amount of data to be searched, as well as the fact that many publications are now digitized, makes it that search operations cannot be done without the help of computers. With the tasks organized in CLEF-IP along the years we investigate how current IR solutions may serve to the needs of patent experts doing novelty searches. This task, in particular, is meant to explore the approaches that IR systems may offer when faced with finding specific pieces of text that are relevant to any given patent claim.

We present here shortly the process of obtaining a patent with focus on the European Patent Office (EPO [2]).

To obtain a patent, a *patent application* must be registered with a patent office. A patent application contains an abstract, a title, a detailed description of the invention, drawings (if necesary) and a set of claims that define the extent of the protection aimed for. An applicant will also cite previously published patents that are considered relevant to the described invention. At the EPO applications can be made in any language. Given that the official languages at the EPO are English, French, and German, whenever another language is used in an application, a translation to one of these three languages must be made. Once the application is registered at the patent office, it will be examined that it is novel, that it has an inventive step, and that it is realizable. During these examinations, at the EPO, a European search report is prepared which lists all the relevant documents found (called *patent citations*).

The EPO publishes patent applications together with their search reports in a time limit of 18 months from the filing date. If the patent applicant, based on the search report, decides to pursue for a patent, a sequence of communications between him and the patent office takes place. Usually, during this process, the claims are adjusted such as not to conflict with existing patents.

The European search report is mainly based on the application claims, and, more often than not, specifies not only the documents relevant to the (various) claims, but also the passages particularly of importance to them. Knowing this, the Passage Retrieval Task Starting from Claims was designed to investigate the effectiveness of Information Retrieva (IR) methods in finding relevant documents and marking passages particularly pertinent to a set of claims.

2 The CLEF-IP Corpus

The CLEF-IP corpus was distributed as a collection of over 3 million XML documents pertaining to over 1.5 million patents published by the EPO and the World Intellectual Property Organization (WIPO) prior to 2002 [8]. The CLEF-IP corpus is an extract of the larger MAREC collection[1] which uses a common normalized XML data format to represent patent documents published by the EPO, WIPO, US Patent and Trademark Office, and Japan Patent Office. We do not describe the collection content here, but we direct the reader to the previous publications that detail it ([7,9]).

3 Task Topics

The Passage Retrieval from Claims Task models closely the novelty search done by patent examiners at the EPO. Topics in this task are sets of claims extracted from actual patent application documents published by the EPO after 2002. Participants had to return passages that are relevant to the topic claims. The

[1] The MAtrixware REsearch Collection. http://ifs.tuwien.ac.at/imp/marec

passages must occur in the documents in the CLEF-IP collection. No other data was allowed to be used in preparing for this task.

To select the topics for this task we first had to select the patent application documents out of which we could select, then, various sets of claims. We first selected a pool of candidate application documents from the MAREC collection with a few restrictions:

- the document must be published after 2002 (that is, is not part of the CLEF-IP corpus);
- the document must be published by the EPO (recall that MAREC contains also patents published by the US office, by the WIPO, and the Japanese office);
- the application should contain at least 3 citations and at most 10. This is because the number of patent documents with more than 10 citations in the search report is very small when compared the the number of patents with less than 10 citations. An additional reason for choosing the upper limit is a pragmatic one: patents with more than 10 citations proved to be more difficult and time consuming to process when extracting the relevance judgements;
- the application document does not miss content, that is, it indeed has a description, an abstract and a claims section. We mention here that, according to the Patent Cooperation Treaty [1], for patent applications that are filed first at the WIPO and then at the EPO, EPO does not publish an additional application document, but only a bibliographic entry that points to the original WIPO application. In terms of XML representation, this translates into an XML document that doesn't have a description, an abstract, nor a claims section;
- the document does not count more than 300,000 words. Setting this limit allowed us to avoid selecting patent application documents that are more than 100 pages long. The rationale behind this decision is that, from past experience, task participants sometimes used full patent documents as queries[2], and it has been shown that some retrieval algorithms do not cope well with large queries [5];
- the application document has at least one family member (a patent document published at another patent office) which was filed prior to the document in the pool. This last restriction is an addition to the task organized in 2012. It is, however, an addition that models a widely used practice of the patent examiners, which consists in pulling out everything what was already done at other patent offices with regard to a patent application they have in front of them, before they start their own search.

After applying all these restrictions, we ended up with a pool of over 300,000 patent application documents. The next step, was now, to sample documents from this pool and extract sets of claims to be topics. The sampling was ran-

[2] Although it may have benefits in an IR sense, no patent expert would use such a solution, actually.

domly done, with one restriction, however. Some technological areas are overrepresented in the patent corpus. For example, patents in the farmaceutical domain are more numerous than in other technical domains. Because we intended to have a relatively uniform distribution of the citation numbers the topic documents have, we first grouped the documents in the pool by the number of citations in the search report and in the CLEF-IP collection. We, then, randomly selected 20 patent application documents from each group with the restriction that each document belongs to a different IPC class [3]. We did this three times: once extracting English application documents, once German, and once French application documents. We have now a pool of over 460 patent application documents. Out of this smaller pool we, randomly, inspected over 200 documents, over 60 in each EPO language, to extract claim sets for our topics.

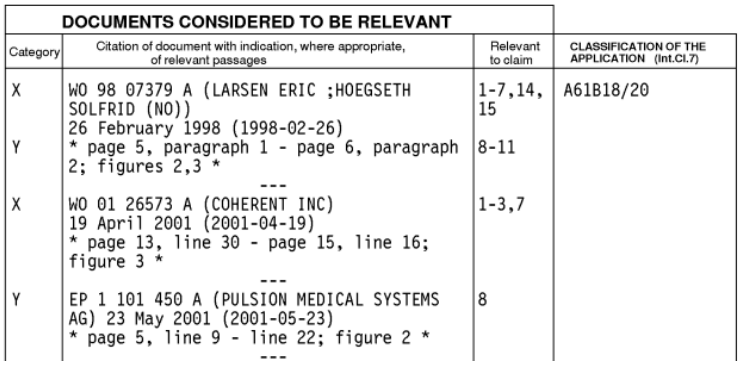
As mentioned in the previous section, a patent application document contains a claims section which define the extent of the legal protection for the described invention. The claims section is a list of sentences (claims) which, for ease of reference, are numbered. Below is an example of the first 8 claims in the application document of patent WO-02058006.

What Is Claimed Is:

1. In a paint roller having an inner resilient cylindrical core and an outer annular surface contact material, the outer annular surface contact material forming a paint roll medium that is fixedly attached to the resilient core, the resilient core and paint roll medium rotating about an axis of said cylindrical core; the improvement wherein the paint roll medium is a hydroentangled threedimensional imaged nonwoven fabric.
2. An imaged nonwoven fabric of claim 1, wherein the fabric is formed from a precursor web comprised of staple length fibers.
3. An imaged nonwoven fabric of claim 2, wherein the staple length fibers include surface modification agents.
4. An imaged nonwoven fabric of claim 3, wherein the surface modification agents are selected from the group consisting of hydrophobic modifiers and hydrophilic modifiers.
5. An imaged nonwoven fabric of claim 2, wherein the staple length fibers include the incorporation of melt additives.
6. An imaged nonwoven fabric of claim 5, wherein the melt additives are selected from the group consisting of hydrophobic modifiers and hydrophilic modifiers.
7. An imaged nonwoven fabric of claim 2, wherein the staple length fibers are selected from the group consisting of thermoplastic polymers, thermoset polymers, natural fibers, and blends thereof.

[3] IPC (International Patent Classification System) is a classification system that groups patents by their technological area. IPC is hierarchially organized in sections, classes, subclasses, groups and subgroups. There are 8 sections, 121 classes, and over 630 subclasses in this classification system. A patent may belong to several technological subareas.

Because the relevance judgements for this task are based on European search reports, when selecting the topics we had to inspect, for each application document in the pool, its search report (an example of a search report is shown in Figure 1). A European search report usually has 4 columns. The second column lists the relevant documents (patent citations) together with relevant passages, images, etc. The first column marks the relevance category of the citation, with X and Y being citations that destroy the novelty in the patent application, A being citations that offer background information on the invention but do not destroy its novelty or inventive step. The third column in a European search report writes down the claim numbers to which the patent citations pertain.



DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	WO 98 07379 A (LARSEN ERIC ;HOEGSETH SOLFRID (NO)) 26 February 1998 (1998-02-26)	1-7,14, 15	A61B18/20
Y	* page 5, paragraph 1 - page 6, paragraph 2; figures 2,3 *	8-11	

X	WO 01 26573 A (COHERENT INC) 19 April 2001 (2001-04-19) * page 13, line 30 - page 15, line 16; figure 3 *	1-3,7	

Y	EP 1 101 450 A (PULSION MEDICAL SYSTEMS AG) 23 May 2001 (2001-05-23) * page 5, line 9 - line 22; figure 2 *	8	

Fig. 1. Extract from a search report.

For a patent application document we inspected each patent citation that occured in our corpus[4]. We noted the claim numbers it refered to and the relevant passage information. When the relevant passage information was acceptable, that is, it refered to lines of text and not to figures or whole documents, we retained the set of claims to be a topic in our task. We also took care that the search reports were complete, in the sense that the patent examiner did his search for all the claims in the patent application. When this was not the case, the search reports contain a notice on this fact and we could eliminate these cases from our pool.

Using this procedure, we could extract more topics from one patent application documents. It was often the case that each topic extracted from one patent application document had its own set of relevant documents and passages, and that the sets of relevant documents didn't allways overlap. From the over 200 patent application documents inspected we were able to extract 149 topics from 69 patent documents. From the 149 topics distributed to the participants, we later removed topics 78 and 101 for being erroneous.

[4] Not all patent citations in a European search report occur in the CLEF-IP corpus.

The structure of a CLEF-IP topic is as follows:

```
<tid> topic_id </tid>
<tfile> patent_ucid.xml </tfile>
<tfam-docs> patent_ucid.xml </tfam-docs>
<tclaims> xpaths_to_claims </tclaims>
```

where

- `tid` is the topic identifier;
- `tfile` is the XML file which stores the patent application out of which the topic claims were extracted;
- `tclaims` is the list of XPaths to the claims selected as topic from the source patent document;
- `tfam-docs` contains the XML files that are part of the source patent's family and published prior to the source patent document.

Below is an example of a topic in the CLEF-IP 2013 Passage Retrieval Task:

```
<tid>PSG-22</tid>
<tfile>EP-1267498-A1.xml</tfile>
<tfam-docs>FI-111300-B1.xml,FI-20011095-D0.xml,FI-20011095-A.xml</tfam-docs>
<tclaims>/patent-document/claims/claim[1] /patent-document/claims/claim[2]
/patent-document/claims/claim[3] /patent-document/claims/claim[4]
/patent-document/claims/claim[5] /patent-document/claims/claim[6]
/patent-document/claims/claim[7] /patent-document/claims/claim[8]
/patent-document/claims/claim[9] /patent-document/claims/claim[10]
/patent-document/claims/claim[11]</tclaims>
```

In the topic set distributed to the participants the patent application documents from which the claims were extracted, and the previously published family member documents were also available, such that participants could use them to extend the original queries extracted from the claims.

4 Relevance Judgements

Using patent data in evaluation campaigns has one disadvantage when compared to other campaigns: to obtain relevance assessments as in the real life patent search examples experts in the various technological domains are needed. The budget of a research project cannot afford employing them to provide judgements, voluntary participation in creating assessments being for most of the patent experts not an option.

Despite this disadvantage, we are in the very happy situation that relevance judgements of a kind already exist in the form of patent search reports [5]. All CLEF-IP campaigns used, in one form or another, the search reports to extract

[5] Experiments using citation information to design retrieval experiments have been done also in other areas than the patent domain. See for example [11].

relevance assessments. We did the same this year. The difficulty in getting the qrels for our topics in 2013 (and in 2012), is that, although patent citations can be easily obtained in some machine-processable form, relevant passages cannot. Therefore, the relevant passage information extraction was done by manual inspection of the search reports, of the cited documents and by matching them with the textual content of the relevant documents in the CLEF-IP collection.

This proved to be a tedious process, so we developed a system to assist us with selecting the relevant pieces of text from the XML documents in our collection. The system has been used also in 2012 and is described in [9] and [8]. We very shortly present here the main features of the system. We see in Figure 2 that the qrel generating system has three main areas:

- a topic description area where, after typing in the patent application document identifier, we can assign the topic an identifier (unique in the system), we define the set of claims in the topic, save it, navigate among its relevant documents with the 'Prev' and 'Next' buttons.
- a qrel display area where we see the currently selected relevant passages and can save them. Also in this area we give a direct link to the application document on the EPO Patent Register server, which, in turns, gives us a quick link to the document's search report.
- a qrel definition area where individual passages (corresponding to XPaths in the XML documents) are displayed. Clicking on them will select them to be part of the topic's qrels. For convenience, we provide three buttons by which we can select with one click all of the abstract's, description's or claims' passages. When clicking on the 'Save QREL' button the selected passages are saved in the database as relevant passages for the topic in work.

The relevance judgements created contained both relevant documents and relevant passages in them. Though the documents could be differentiated by degrees of relevance, due to their categories in the search reports (X, Y, A), the passages were considered all equally relevant.

Below is an excerpt from the qrel files obtained with the help of our system:

```
PSG-5 EP-1078736-A1 /patent-document/description/p[20]
PSG-5 EP-1078736-A1 /patent-document/description/p[21]
PSG-5 EP-1078736-A1 /patent-document/description/p[18]
PSG-5 EP-1078736-A1 /patent-document/description/p[15]
PSG-5 EP-1078736-A1 /patent-document/claims/claim[1]
PSG-5 EP-1078736-A1 /patent-document/abstract/p
PSG-5 EP-1078736-A1 /patent-document/claims/claim[2]
...
```

5 Submissions and Evaluations

5.1 Submissions to the Task

The submission format for the passage retrieval task required participants to submit text files with retrieval results similar to the qrel format shown above. The number

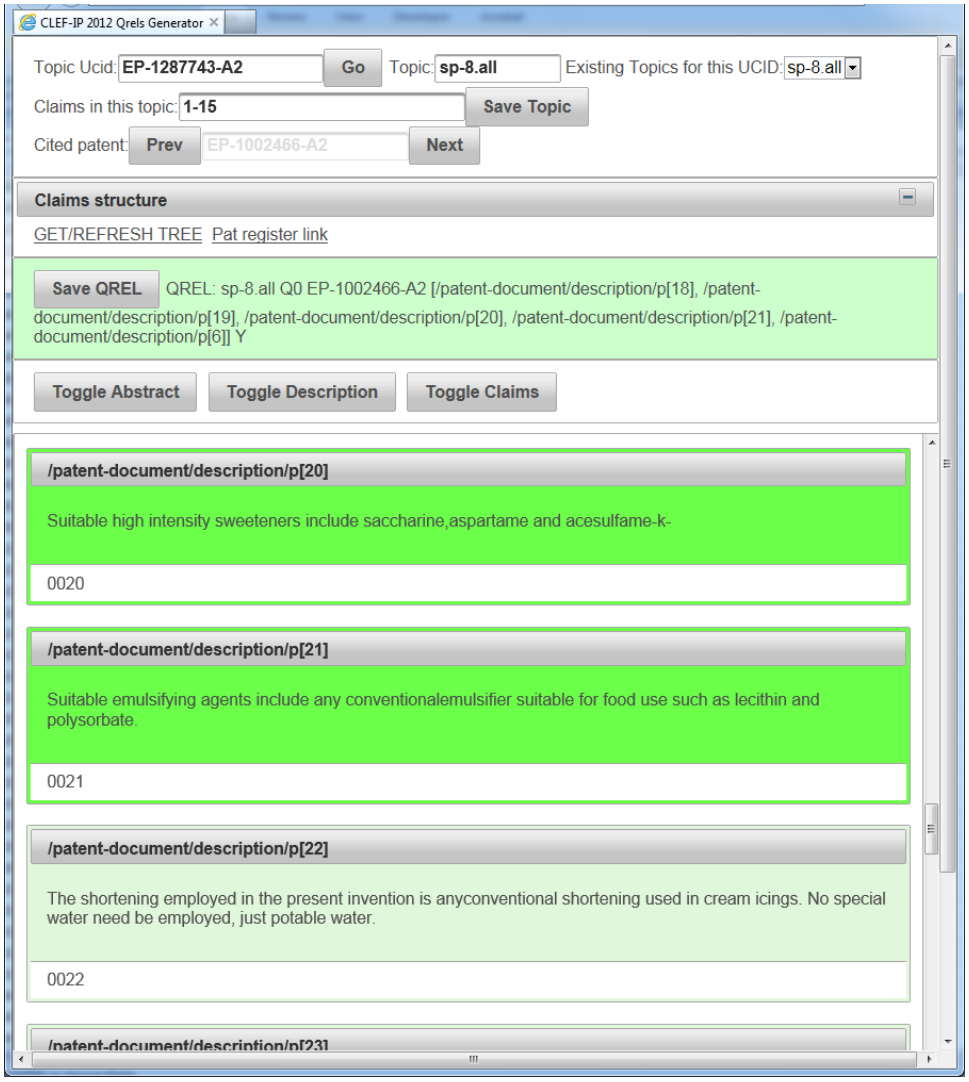


Fig. 2. A system for finding and saving relevant passages.

of documents considered relevant per topic had to be limited at 100, the number of relevant passages in a document was not limited. In addition, to the qrel format, the participant submissions had two more columns, one to specify the order of the results, and another one to specify the retrieval score of a passage/document.

Three participants submitted experiments to the Passage Retrieval task, two of them also included relevant passages in their task. In their experiments a two step approach was used. In the first one, relevant documents were retrieved using various retrieval solutions including Okapi BM25, Language Models and TF-IDF, and Vector Space Models. The participant from the Georgetown University (USA) experienced with various sources for query terms by extracting words from claims and titles, using hyphenating-phrases, Part of Speech tagging and weighted filtering [4]. The team from Innovandio S.A. (Chile) also experienced with a CL-ESA Wikipedia-based multilingual retrieval model ([10], [8], section 3).

The third participant to the task, a team of researchers from Vienna University of Technology and the University of Macedonia, Thessaloniki, used a distributed IR system that queried a split CLEF-IP collection. The split is done by exploiting the hyerarchical structure of the International Patent Classification System (IPC). By dividing the collection into several sub-collections (by IPC class, subclass, and subgroup) the patents are organized according to their technological topic. Then the Lemur indexer was used to index the title, abstract, description, claims, inventor, applicant and IPC class information [3]. The CORI and a multilayer method were used for selecting the sources (sub-collections) on which the retrieval should be performed as well as for joining the results.

In the figures below, the submission files prefixed by 'In' belong to the participant from Chile, the submission files prefixed by 'GU' belong to the participant from Georgetown Unviersity, and the ones prefixed by 'TM' were sent in by the team from Vienna and Thessaloniki.

5.2 Evaluating the Retrieval Results

Three participants submitted a total of 19 runs. Out of these, 8 runs did not provide retrieved passages.

We did evaluations at two levels. One at the passage level and one at the patent document level. The evaluation at patent document level was done, as in the previous years, by computing the Recall, MAP, and PRES ([6]) at cutoff 100. At the passage level we computed, first, for each relevant document retrieved the precision and average precision w.r.t. the passage retrieved, then averaged it over the number of relevant documents per topic. Finally, averaging these scores over all topics we obtain the precision and mean average precision scores at the passage level. The evaluation script is available for download on the CLEF-IP project website[6].

Several simple file clean-up operations had to be done in order to ensure that the document encodings matched the expected format by the evaluation script. These operations included duplicate removal, re-grouping the retrieval results such that results belonging to one topic were in a contiguous portion of the files, removing the XPaths refering to headings in the patent document XML files. This last operation was done because headings are not consistently marked as such in the CLEF-IP collection's documents, being left out of the relevance judgements as well.

[6] http://www.ifs.tuwien.ac.at/~clef-ip

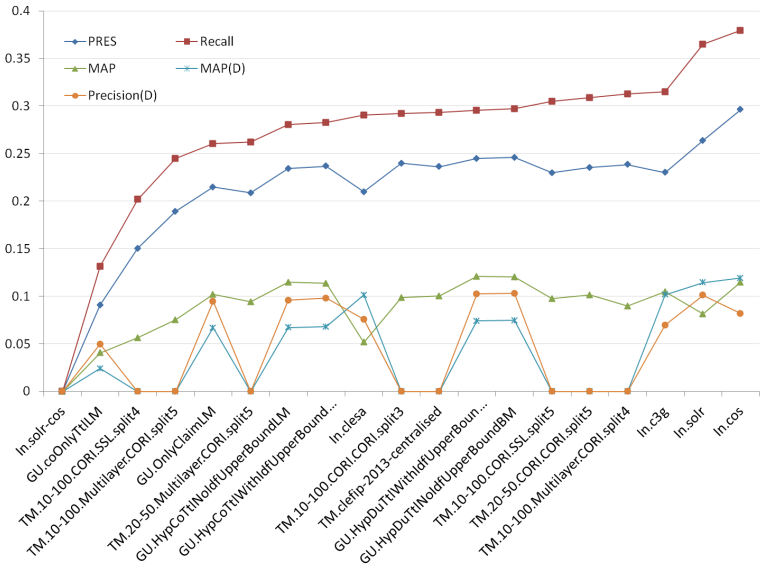


Fig. 3. Evaluation results, ordered by Recall.

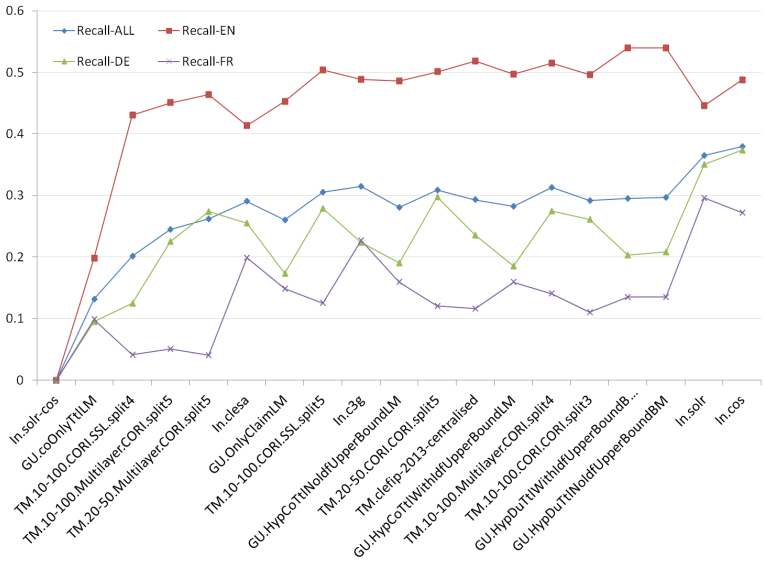


Fig. 4. Evaluation results, document level Recall per language.

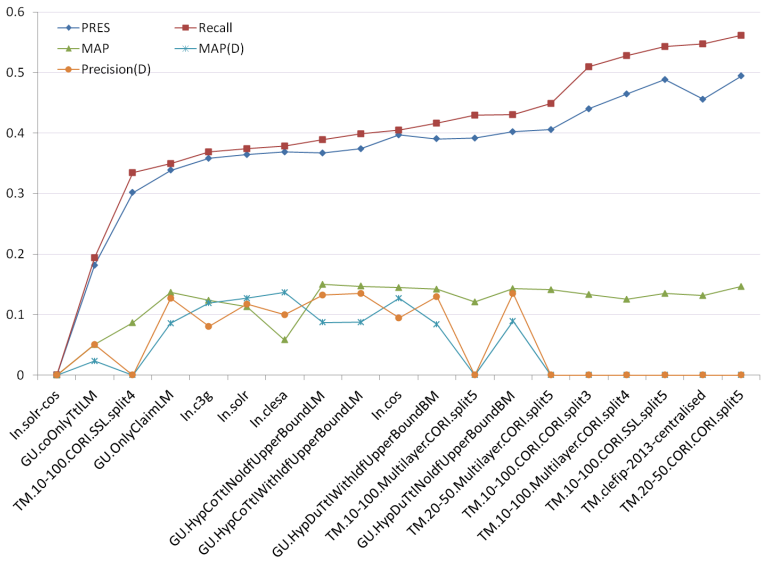


Fig. 5. Second evaluation round, results ordered by Recall.

We ran, then, several evaluations depending on the degree of relevance assigned to the citation documents in the search reports. In each round we computed all of the measures mentioned above, we will not, however, present all of them.

The first evaluation round considered all documents in the relevance judgements as equally relevant and did evaluations on four sets of topics: the set of all 147 topics, on the subset of 50 English topics (1-50), on the subset of 49 German topics (51-100), and the subset of 48 French topics (102-149). The results of these evaluations are shown in Figures 3 and 4. The zero values on the figures belong to the runs that did not contain relevant passages.

Next we were interested in the metric scores when only the highly relevant citation documents were considered, ignoring the applicant citations. From the 147 topics only 116 have highly relevant citations in the CLEF-IP corpus, so the new evaluation round is done for this smaller set. Figures 5 and 6 show plots for the metrics for this smaller topic set, for the 38 English topics, for the 42 German topics, and for the 22 French topics in it.

To compare how the different retrieval strategies perform with respect to the different relevant documents required (highly relevant only, or both highly relevant and relevant) we computed a third round of evaluations, where we restricted the set of qrels used in the first round of evaluation to the 116 topics evaluated in the second round. Although we computed all the mentioned metrics for all three languages, we present only the results for the whole Recall and MAP(D) for the 116 topics, in Figure 7.

6 Final Words

This paper presented the activities we have done to organize the Passage Retrieval Starting from Patent Claims Task in CLEF-IP 2013. We started with selecting patent

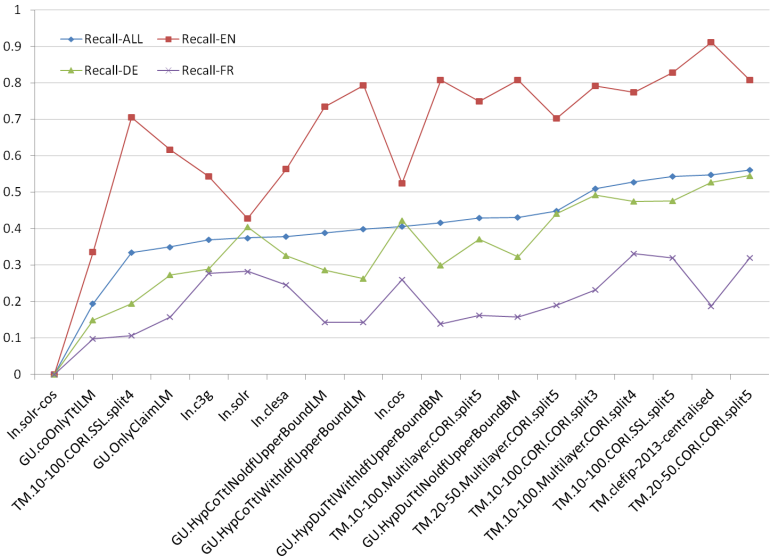


Fig. 6. Second evaluation round, document level Recall per language.

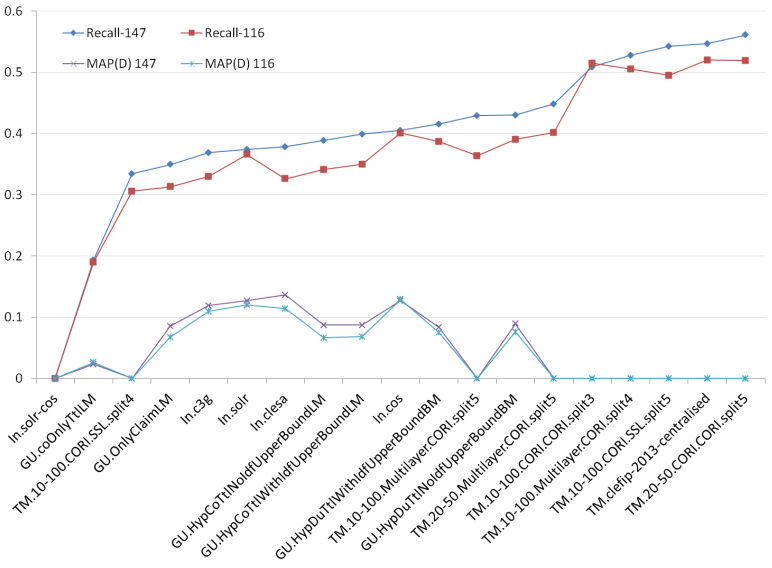


Fig. 7. Third evaluation round, document level Recall and MAP(D).

application documents and sets of claims in these documents that were our final topics. The most time consuming part of these activities has been extracting the XPaths to relevant passages identified by patent experts in their search reports. Participants were not given any specific queries, but were allowed to build them out of the information provided in the topics: claims, patent application document, previously published family member documents.

Over 20 teams registered to submit retrieval experiments to this task, a number similar with the number of registrations in the previous years. We received submissions from three groups, two of them with relevant passage information as well.

Acknowledgements This work was partly supported by the EU Network of Excellence PROMISE(FP7-258191) and the Austrian Research Promotion Agency (FFG) FIT-IT project IMPEx(No. 825846).

References

1. ***. *Patent Cooperation Treaty*. 1970. last retrieved: March, 2013.
2. ***. *Guidelines for Examination in the European Patent Office*, 2012. www.epo.org/law-practice/legal-texts/guidelines.html, latest retrieved in June 2013.
3. Anastasia Giachanou, Michail Salampasis, Maya Satratzemi, and Nikolaos Samaras. Report on the CLEF-IP 2013 Experiments: Multilayer Collection Selection on Topically Organized Patents. In *CLEF (Notebook Papers/LABs/Workshops)*, 2013.
4. Jiyun Luo and Hui Yang. Query formulation for prior art search - georgetown university at CLEF-IP 2013. In *CLEF (Notebook Papers/LABs/Workshops)*, 2013.
5. Yuanhua Lv and ChengXiang Zhai. When documents are very long, BM25 fails! In Wei-Ying Ma, Jian-Yun Nie, Ricardo A. Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceedings of SIGIR*, pages 1103–1104. ACM, 2011.
6. W. Magdy and G. J. F. Jones. PRES: A score metric for evaluating recall-oriented information retrieval applications. In *SIGIR 2010*, 2010.
7. F. Piroi, M Lupu, A. Hanbury, and V. Zenz. CLEF-IP 2011: Retrieval in the intellectual property domain, September 2011.
8. Florina Piroi, Mihai Lupu, and Allan Hanbury. Overview of CLEF-IP 2013 Lab: Information Retrieval in the Patent Domain. In *Proceedings of CLEF 2013*, Lecture Notes for Computer Science, 2013. to appear.
9. Florina Piroi, Mihai Lupu, Allan Hanbury, Alan P. Sexton, Walid Magdy, and Igor V. Filippov. Clef-ip 2012: Retrieval experiments in the intellectual property domain. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.
10. Martin Potthast, Benno Stein, and Maik Anderka. A wikipedia-based multilingual retrieval model. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Proceedings of ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 522–530. Springer, 2008.
11. A. Ritchie, S. Teufel, and S. Robertson. Creating a Test Collection for Citation–based IR Experiments. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006.