Overview of CLEF-IP 2013 Lab
Information Retrieval in the Patent Domain

Florina Piroi, Mihai Lupu, and Allan Hanbury

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

Abstract. The first CLEF-IP test collection was made available in 2009 to support research in IR methods in the intellectual property domain; only one type of retrieval task (Prior Art Search) was given to the participants. Since then the test collection has been extended with both more content and varied types of tasks, reflecting various specific parts of patent experts' workflows. In 2013 we organized two tasks – Passage Retrieval Starting from Claims and Structure Recognition – on which we report in this work.

1 Introduction

The patent system is designed to encourage disclosure of new technologies and novel ideas by granting exclusive rights on the use of inventions to their inventors, for a limited period of time [23]. An important requirement for a patent to be granted is that the invention it describes is novel. That is, there is no earlier patent, publication or public communication of a similar idea. To ensure the novelty of an invention, patent offices as well as other Intellectual Property (IP) service providers perform thorough searches called 'prior art searches' or 'validity searches'. Since the number of patents in a company's patent portfolio affects the company market value, well-performed prior art searches that lead to solid, difficult to challenge patents are of high importance.

Patent data has attracted researchers' interest as early as 1977 when, while studying local clustering in full-text searches using local feedback, experiments were done on a database of US patents [5]. In [5], Attar and Fraenkel did an experiment that was a 'technology survey'-like search on a set of 76 US patents. Two decades later an 'invalidity search' was performed on 60000 US patents. Similar to the Prior Art Search task in CLEF-IP 2009–2011, the topics of the invalidity search were patents and citations were used to generate relevance assessments [21].

In the last decades, research in IR methods for the IP domain has intensified. Workshops, conferences and evaluation tracks were organized in an effort to bring IR and IP communities together (see [11,13,27,10]). The National Institute of Informatics (NII), Japan, initiated a series of workshops and evaluations

P. Forner et al. (Eds.): CLEF 2013, LNCS 8138, pp. 232–249, 2013.

using patent data as part of the NTCIR project (the NII Test Collections for IR Systems, currently renamed to the NII Testbeds and Community for Information access Research), focusing on Japanese and Chinese patents, and their translations into English.

In 2009, two further evaluation activities using patent data were launched: TREC–CHEM and CLEF-IP. TREC–CHEM ran from 2009 to 2011 and was organized as a chemical IR track in TREC (Text Retrieval Conference) addressing the challenges in chemical and patent IR [15]. The collection corpus was limited to chemical patent documents and chemical journal articles.

The purpose of the CLEF-IP track, part of the Cross-Language Evaluation Forum (CLEF), is to encourage and facilitate research in the area of multilingual patent retrieval by providing a large, clean data set for experimentation. The data set contains patents in three European languages, patents published by the European Patent Office (EPO), as well as queries and associated relevance judgements.

In 2013, the CLEF-IP lab proposed two tasks: a passage retrieval task where we asked for passages relevant to a given (set of) patent claim(s) and a structure recognition task where we asked to extract the textual representation of flowcharts occuring in patents and represented in black and white images.

2 The 2013 CLEF-IP Benchmark

We begin this section by establishing the patent terminology used throughout this paper and shortly describing the patenting process such that the rationale behind the lab's activities are better understood.

The main phases of obtaining a patent for an invention are[1]:

The Pre-application Phase: a person with a new idea will write down its description as detailed as necessary. Then she or he will usually perform a survey-like search in the domain of the invention. This preliminary search will allow the inventor to avoid unnecessary effort in case a similar invention already exists and will help him to draft the invention claims. The drafted document generally has three parts: an *abstract*, a *description* of the invention with technical drawings, and a *claims* section which states the extent of the protection sought for the described invention.

The Application and Examination Phase: after filing the invention description at a patent office the document (now called a *patent application document*) is given to a patent examiner. He or she will inspect the document and verify that it respects certain criteria, namely: novelty, the existance of a non-obvious

[1] The process described by these phases is typical for EPO patent applications. Though very similar, processes at other patent offices may reveal important differences. For example, the US Patents and Trademark Office,USPTO, makes use of Examiner's Letter or Action to record considered citations and does not publish a distinct search report ([3], chapter 707).

inventive step, and realizability. During the novelty check the patent examiner will search for and create a list of existing patents, a.k.a. prior art, that are relevant to the application document under inspection. At the EPO the list of relevant documents is published as a *search report* document. The search report contains also the relevant documents provided by the inventor as background information to the invention. In the IP vocabulary, the documents listed in the search report are called *patent citations*, the citations provided by the applicant being known, additionally, as *applicant citations*.

In this document, whenever the word 'citation' occurs we mean the patent citations, that is, the documents in the search reports which were considered relevant by patent examiners. This is different from the research community's understanding of 'citation' which refers to later publications citing a research article. A patent citation is more similar to what in the research community is known as a reference at the end of an article[2]. In the IP community, differentiating between the patent citations and later references to patents is done by using the notions of *forward* and *backward citations* [1]. Given a patent application document, the patent citations listed in a search report are known as *backward citations*, while the patent application itself is a *forward citation* for any of the patents listed in the search report.

The patent citations usually have various degrees of relevance to the application document. The main three types of citations are:

- citations that describe prior work but which do not destroy the novelty of the application;
- citations which, taken alone, make a patent application not novel;
- citations that, in combination with other citations, destroy the novelty of an application.[3]

At the end of the examination phase the patent application document and its associated search report are published by the patent office. At the same time, the patent application is given a classification code that assigns the patent to a specific technological area[4].

The Granting and Opposition Phase: Based on the search report a dialogue between the patent office and the patent applicant is initiated. There are various outcomes to this dialogue: an application may be retracted, rejected, or modified in order not to infringe existing patents. If the patent office reaches the decision to grant a patent, after various fee payments made by the applicant, a *patent document* is published. From this point on, for a certain amount of time (9 months at the EPO) oppositions to a granted patent may be filed to the patent

[2] This observation is critical in understanding how we have selected the topics and how the relevance judgement were created.

[3] The last two patent citation types are referred to as *highly relevant citations* in the CLEF-IP Labs.

[4] We do not expand here on the subject of patent classification codes. See [2] for a description of the classification system we mention later in this work.

office. Note that opposition procedures at a patent office are different from the legal actions to invalidate patents which are taken in justice courts.

The rest of this section describes the main connection between the CLEF-IP tasks and a patent expert's work, and the CLEF-IP test collection: document corpus, topic sets, and judgements.

2.1 The Retrieval Tasks

There are many aspects of the *search for innovation* use case domain that previous evaluation campaigns, including CLEF-IP, have focused on in their retrieval tasks. Creating technical surveys on various chemical subjects (TREC–CHEM [18]) or creating patent translations to be used by non-speakers of certain languages [8] are two such examples.

This year in CLEF-IP we proposed two tasks. The first one models the type of searches examiners do to establish the non-obivousness of an invention, where they closely inspect the claims in the patent application against other existing patent documents. At the EPO, search reports generally show not only the prior art documents, but also the claims in the patent application to which the patent citation pertains and which passages in the citation are particularly of interest (see Figure 1). The retrieval task was designed to investigate the degree of support an IR system offers patent experts in finding relevant documents and text passages to a set of claims in a patent application.

The second task in the lab is not one that models part of an expert's work, but it is designed to support his or her work during patent examination. Technical drawings are often crucial not only in illustrating the embodiments of an invention, but also to quickly filter out non-relevant patents by rapid glances to images in them. The aim of the structure recognition task is to make the content of the images textually searchable and comparable. Out of the many types of images that may occur in patents we limited this retrieval task to images representing flow-charts.

2.2 The Collection Corpus

One of our aims when embarking on the CLEF-IP endeavor was to create a test collection fit for experimenting with patent data, a collection that faithfully mirrors the features and challenges of the data used in the actual working cycles of a patent professional. For this we use actual patent documents published by the EPO and WIPO (World Intellectual Property Organization). These documents contain most of the information that is actively used by patent practitioners in their daily work with patent data.

The bulk of the collection's corpus is made up of patent documents stored as XML files. Since its first release in 2009, consecutive additions were made to the CLEF-IP test collection, so that it currently contains almost 1.5 million patents published before 2002, stored into approximately 3.5 million XML documents.

These patents are an extract from the larger MAREC[5] collection which contains documents representing over 19 million patents published at the EPO, USPTO, WIPO and JPO (Japan Patent Office) stored in a common normalized XML format. The main elements of the XML representations are the ones shown in the simplified listing below:

```
<patent-document>
  <bibliographic-data> ... </bibliographic-data>
  <abstract> ... </abstract>
  <description> ... </description>
  <claims> ... </claims>
</patent-document>
```

The `<abstract>`, `<description>`, and `<claims>` elements store the textual content of the disclosed invention. These fields may occur more than once when, for example, both the English and the German versions of the abstract are stored in a patent document. The abstract, description and claim fields are the parts of the patent file mostly used by the textual retrieval methods. The `<bibliographic-data>` element contains the administrative data related to a patent. In this XML element we will find the application and publication dates and references, family identifiers, the classification symbols, inventors, assignees, postal addresses of the inventors and/or assignees, the invention's title (in three languages), and the citations relevant to the invention in this document.

In the corpus of European patent documents with application date prior to 2002, a high percentage of the patent documents refer to applications internationally filed under the Patent Cooperation Treaty [22], also known as 'EuroPCTs', in which case, the EPO does not republish the whole patent application, but only a bibliographic entry linking to the original application published by the WIPO. Using text-based methods to retrieve such documents is problematic, and therefore, for these patent documents the current CLEF-IP collection contains their WIPO equivalent. Determining that the EuroPCT patent documents refer to a certain invention disclosed in a document published by WIPO is done by the family identifier which for the two documents must be the same.

One of the most important features of the CLEF-IP corpus is its multilingualism. Patent applications to the EPO are written in one of the three official EPO languages (German, English, French), with the additional requirement that, once the decision to grant a patent is made, the claims section of the patent document must be submitted in all these three languages. Although the English language is overrepresented[6] in the CLEF-IP collection, not least due to the EuroPCT applications written in their large majority in English, the collection entails large amounts of content that is in German and French, making the collection suitable for carrying out multilingual retrieval experiments.

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

[6] Almost 70% of the documents in the collection are written in English, about 23% have German as the document language, and about 7% are in French.

2.3 Passage Retrieval Starting From Claims

The topics of this retrieval task are sets of claims occuring in patent application documents. Participants were asked to return documents from the CLEF-IP corpus which were considered relevant and, within these documents, mark the most relevant passages to the set of claims.

We have provided over 150 training topics and the test set contained 149 topics. A third of both the test and the training sets contained topics in English, another third contained topics in German, and yet another third had the topic language French. We did not provide translations of topics from one language into any of the other two. 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 source patent application;
- `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[7] and published prior to the source patent document.

Providing previously published patent documents that are family members of the source patent application is motivated by the patenting process rules and by the practices of the patent examiners at patent offices. More concretely, when an applicant files for a patent grant at, let's say, EPO he is required to provide information on whether he has already applied for a patent grant, *for the same invention*, at other patent offices in the world. Later, when the patent application is examined, the patent examiner pulls whatever search reports are available in the patent databases related to the previous publications of the inventions in order to re-use that information.

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

```
<tid>PSG-2</tid>
<tfile>EP-1445439-A1.xml</tfile>
<tfam-docs>FI-116479-B1.xml,FI-20030196-A.xml,FI-20030196-D0.xml</tfam-docs>
<tclaims>/patent-document/claims/claim[1] /patent-document/claims/claim[2]
/patent-document/claims/claim[3] /patent-document/claims/claim[4]</tclaims>
```

[7] A *patent family* denotes the collection of patent documents that refer to the same invention and are published by different patent offices around the world.

Topic Selection. We created a pool of patent application documents out of which we extracted the set of test topics for this task. The pool of patent applications was extracted from the MAREC collection with the requirements that:

- it was not part of the CLEF-IP corpus (i.e. published after 2002);
- it was a patent published by the EPO;
- there is at least one previously published document in the patent's family;
- it has content for all document parts (claims, abstract, description), and the document word count is lower than 300,000 (we included the XML tags and attributes in this number)[8];
- there are at least two and at most 10 citations in the corresponding search report, and the cited documents occur in the CLEF-IP corpus.

Some technological areas are overly represented in the patent domain. For example, the number of patents filed in US, last year, in the technological area of Electrical Engineering (including Computer Technologies) outnumbered the number of patents filed in any of the other technological areas [29]. To avoid overrepresentation of patents in certain technological classes in the topic set we restricted the sampling process in the following way: we grouped the documents in the pool by the number of citations the documents have, and we randomly selected 20 documents from each group, with the restriction that each selected patent belongs to a different IPC class (there are 121 IPC classes in the topic pool). At this point we have a pool of 462 patent application documents out of which to extract topics.

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

The next step in the topic selection process is, for each patent application document in the pool, to manually retrieve its European search report

[8] We chose this limit in order to avoid pooling documents of excessive length which make some retrieval algorithms fail [19]. Some patent applications are more than 100 pages long which we wanted to avoid being part of the topic test set.

(that is, the search report published by the EPO, see Figure 1) and inspect each citation document with respect to the claims it is relevant to (the third column in Figure 1) and the relevant passage recorded in the report (second column in the same figure). For each citation document in the search report *and* in our data corpus, we extracted the claim numbers the citation referred to[9]. These formed the sets of claims for a candidate topic. Looking, now, at the passages noted as relevant, further decisions had to be made whether a candidate topic is retained. Rejecting candidate topics was done when:

- the relevant documents referred to figures only;
- there was no mention of relevant passages, or only 'whole document' mentions were recorded;
- the search report had the mention 'Incomplete search' which generally means that the patent expert, for various reasons, did not perform the prior art search for all the claims in the patent application.

From one patent application document it was possible to extract several sets of claims as topics, often with completely different sets of relevance judgments. The process just shown has been first used in the CLEF-IP 2012 Lab and is also described in [24].

This has been a lengthy process – being done manually – and we managed to inspect over 200 application documents. The final set of topics (148) was extracted out of 69 patent applications. The citation distribution for the topics and application documents is shown in Table 1, where the topic source documents belong to 66 different IPC classes.

Table 1. Citation distribution in the topic set

Number of citations	3	4	5	6	7	8	9	10	Total
Number of topics	24	25	30	23	16	19	8	3	148
Number of documents	13	11	13	10	10	8	3	1	69

Obtaining the Relevance Judgements. When a topic candidate made it into the final topic set, the next phase was to create its relevance judgements. Judging the retrieved results cannot be confidently done by non-patent experts, therefore, pooling the results and judging them post-submission is not a solution that can be used in CLEF-IP, primarily because engaging patent experts is very costly for a research project. The solution chosen by us, as well as most other evaluation tracks using patent data, is to make use of the patent search reports, which constitute a very reliable source of relevance judgements. More reasons to support this decision can be found in [9].

Patent citation information can be rather easily obtained in a machine processable format (e.g. relational tables). For our task, however, we need relevant passage information which we had to extract manually by matching the passage

[9] In patent documents, claims are numbered for ease of reference.

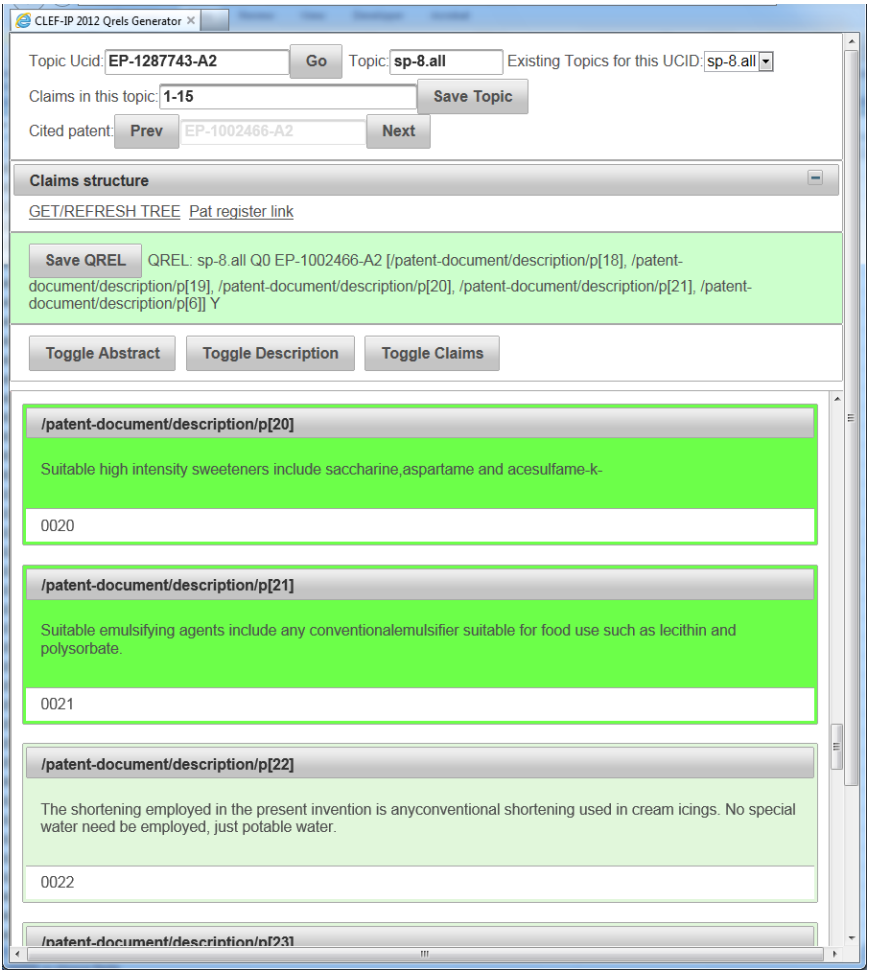


Fig. 2. A system for extracting and storing qrels

indications in the search reports with the textual content of the patent documents in our corpus. When matched, we extracted the XPaths of the identified content and saved them to a database.

To assist us in this tedious process we used an in-house developed system, developed in 2012 (Figure 2), which read the XML patent citation documents and displayed the individual XPath passages. We see in Figure 2 three screen areas: In the upper part, given a topic source document ('Topic UCID'), we define the topic id ('sp-8.all' – an intermediary topic identifier), and the claim numbers that are to be part of this topic. In the middle part of the screen we have buttons for toggling (marking as selected) all passages in the abstract, description, or claim sections at once, and a button ('Save QREL') for storing the currently selected XPaths to the database. In the lower part of the screen each textual content at the end of an XPath in the citation document selected

for the source topic id is displayed and can be selected into the topic's qrels (the green text). Displaying the different citations for the source patent application is done with the navigation buttons 'Prev' and 'Next' of the top screen area.

Below is an excerpt from the qrel files obtained with the help of the 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]
...
```

2.4 Structure Recognition from Patent Images

From the outset, non-textual patent content, like tables, technical drawings, formulae, was not a part of the CLEF-IP campaigns. But these non-textual items have an important role in taking quick decisions about the relevance of a document to an information need. During a patent search, plenty of documents may be returned as the result of a query. An experienced patent professional will often be able to expeditiously dismiss non-relevant documents by glances at images in the patent documents.

In 2012 we designed a task that aimed at making the patent images searchable and comparable by textual means. Two separate sets of images were given, flow-charts and chemical structures. This year we continued this task only with a set of flow-chart images that contained more complicated graphical structures than in 2012.

The topics of this task are black and white images representing flow-charts, images ocurring in patents. We made available the 2012 training and test topic sets as training data (150 images). The test set we used in 2013 contains 747 images of flow-charts[10]. The retrieval task required the participants to extract the information stored in the image files and store it into a textual form that encoded the graph-like structure of the flow-charts, where the text is seen as node or edge labels.

Topic Selection. By comparison with the topic selection process in the Passage Retrieval Task, shown above, selecting the topics for the structure recognition task was 'a walk in the park': we re-used the set of flow-chart images that were part of the Patent Image Classification task in the CLEF-IP 2011 Lab [23]. We slightly modified the encodings used last year to accomodate for the more complicated flow-charts in this year's topic set.

Relevance Judgements. Before creating the qrels we have to establish a textual encoding of the flow-charts. For the purpose of this task, we decided that a text file encoding a flow-chart is a sequence of text lines, each line being one of the below:

`MT` for 'Meta', refers to meta information in the flow-chart:

- `MT Title "figure's title"`: title of the chart, in double quotes

[10] In CLEF-IP 2012, the set of flow-charts selected for the Structure Recognition Task was filtered to contain less complex flow-charts, w.r.t. type of nodes, edges, and lines enclosing other nodes–meta nodes in 2013.

- `MT NO <number>`: number of nodes in the flow-chart;
- `MT DE <number>`: number of directed edges in the flow-chart;
- `MT UE <number>`: number of undirected edges in the flow-chart

`NO` for 'Node'. Lines starting with `NO` describe the node of the chart. Each node description line must contain an identifier of the node (unique in the chart), a node-type that describes the shape of the node (oval, rectangle, etc.), the text of the node (empty string of no text is present), and a pair of coordinates marking the graphical location of the node's center. The coordinates are intended for later use with graph representation tools to graphically display the encoded graph and visually compare it with the original image.

`MN` for 'Meta-node'. Lines beginning with `MN` describe a meta-node of the chart. Each such node must have a unique identifier (different from the `NO`'s identifiers), a comma separated list of `NO` nodes identifiers enclosed in square brackets, a text attached to the meta-node (or the empty string).

`DE |UE` for 'directed' and 'undirected' edges. The lines starting with one of these identifiers describe the edges connecting the flow-chart nodes. Each such line must contain the identifiers of the start and end nodes of the edge, the type of the edge (plain, wiggly, dotted, etc.), and the label attached to the edge, if any.

`CO` for 'Comment'. These lines are not to be considered by the evaluation scripts.

Figure 3 shows an example of a flow-chart textually encoded using the format given above.

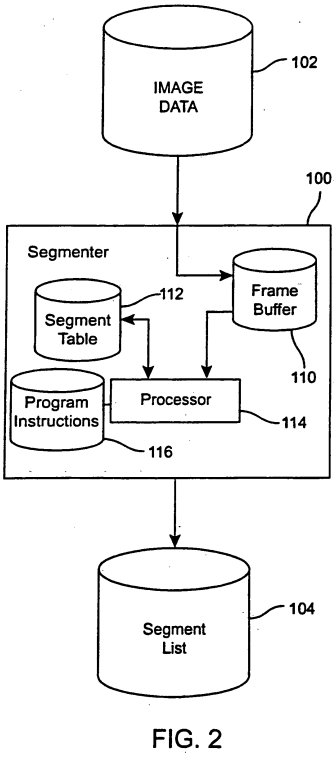
3 Submitted Runs

Three participants submitted a total of 19 retrieval experiments, we shortly describe the main retrieval approaches used.

Gerogetown University, USA. The participants from Georgetown University focused on formulating representative queries using patent metadata (embedded in the collection's XML patent documents). The queries were then submitted to a Lemur search engine [14]. Several indexes were created: one for the stemmed content words in the CLEF-IP collection, and several other for specific patent metadata (title, inventor, application date). The retrieval engines used were TF-IDF based, Language Modelling based, and Okapi BM25.

Six experiments were submitted, each of them using a different approach to obtaining query terms. Extracting the words occuring in claims, titles (experiment with the id `GU.OnlyClaimLM, GU.coOnlyTtlLM`), the hyphenating phrases, Part of Speech tagging (`GU.HypCoTtlNoIdfUpperBoundLM`) and a combination of `idf` filterings on the extracted query terms (`GU.HypCoTtlWithIdfUpperBoundLM`, `GU.HypDuTtlNoIdfUpperBoundBM`, `GU.HypDuTtlWithIdfUpperBoundBM`) are among the tested options. The queries thus generated were used to retrieve relevant documents. The passages in these documents were ranked using a `tf-idf` weighting scheme, returning the top 10 ranked passages.

Innovandio S.A. The participants from Chile submitted five runs to the Passage Retrieval starting from Claims task. The general approach used a two step model in which relevant documents were first retrieved, which were further processed to extract relevant passages. The best placed run was obtained using a Vector Space Model with word 1-grams, and `tf-idf` weighting scheme for the word/dimensions (runID: `In.cos`). Using cosine similarity computations, the first 100 patent documents were retrieved,




```
MT Title "FIG. 2"
MT NO 16
MT DE 5
MT UE 7
CO ------
CO ------ Nodes in the chart ----
NO 1 cylinder "IMAGE DATA" (100,100)
NO 2 cylinger "Frame Buffer" (150, 180)
NO 3 rectangle "Processor" (100, 250)
NO 4 cylinger "Segment Table" (70,200)
NO 5 cylinger "Program Instructions" (30,200)
NO 6 point "" (100,120)
NO 7 no-box "110" (180, 240)
...
CO ---- List of edges ----
DE 1 6 plain ""
DE 6 2 plain ""
DE 2 3 plain ""
UE 2 7 wiggly ""
...
CO --- Meta nodes ----
MN 16 [2,3,4,5] "Segmenter" (180,180)
CO --- Done ----
```

Fig. 3. A flow chart and an excerpt of its textual encoding

then another cosine similarity was computed at the passage level, between the passages of these 100 documents and the topic's passage vector representations. A similar retrieval approach was done using character 3-gram computations (`In.c3g`).

To tackle the multilingual aspect of the topics and collection, a method that tested the CL-ESA Wikipedia-based multilingual retrieval model was applied [26,4]. 10,000 Wikipedia articles with the most amount of available translations were used to create CL-ESA vector representations, which, together with the `tf-idf` weighting scheme, were used in similarity computations (run `In.clesa`).

In another approach, using the open-source Apache/Solr framework, the entire collection corpus was indexed and the topic content was used to generate a sequence of queries per topic which were sent to the framework. The top 100 documents retrieved were indexed at the passage level (including their XPaths) and using the queries formed out of the most frequent words (10 per query) the Solr was taped to retrieve the most relevant passages (`In.solr`).

In the last of the submitted runs, a combination of the Solr index and word 1-gram solution was aimed for. We suspect (as do the participants) that due to a mistake the wrong data was processed, since all computed scores were 0. This run is not shown in the figures displaying the evaluation scores below.

In all retrieval approaches stopwords and diacritics were removed and a stemmer was applied.

Vienna University of Technology - University of Macedonia, Thessaloniki. The `TM` team participated in the Passage Retrieval task and used a distributed IR system that queried a split CLEF-IP collection. The split is done by exploiting the hyerarchical structure of the IPC system. By dividing the collection into several subcollections (by IPC class `TM.split3`, subclass `TM.split4`, and subgroup `TM.split5`) the patents are organized according to their technological topic. Because patents may be assigned several IPC codes, these splits are not disjunct.

The documents in the CLEF-IP collection were preprocessed to remove the stopwords, and to apply the Porter stemmer. Different documents referring to one patent were merged to form a single (virtual) document to represent the patent. Then the Lemur indexer was used to index the title, abstract, description (first 500 words), claims, inventor, applicant and IPC class information [7].

The CORI and a multilayer method were used for selecting the sources (subcollections) on which the retrieval should be performed as well as for joining the results.

We note that the `TM` team did only document level retrieval, therefore the passage specific metric scores in the next section are zero.

4 Evaluation Results

We present in this section the measures and the numeric values of these measures we obtain when evaluating the participant's submissions against the task's relevance judgements.

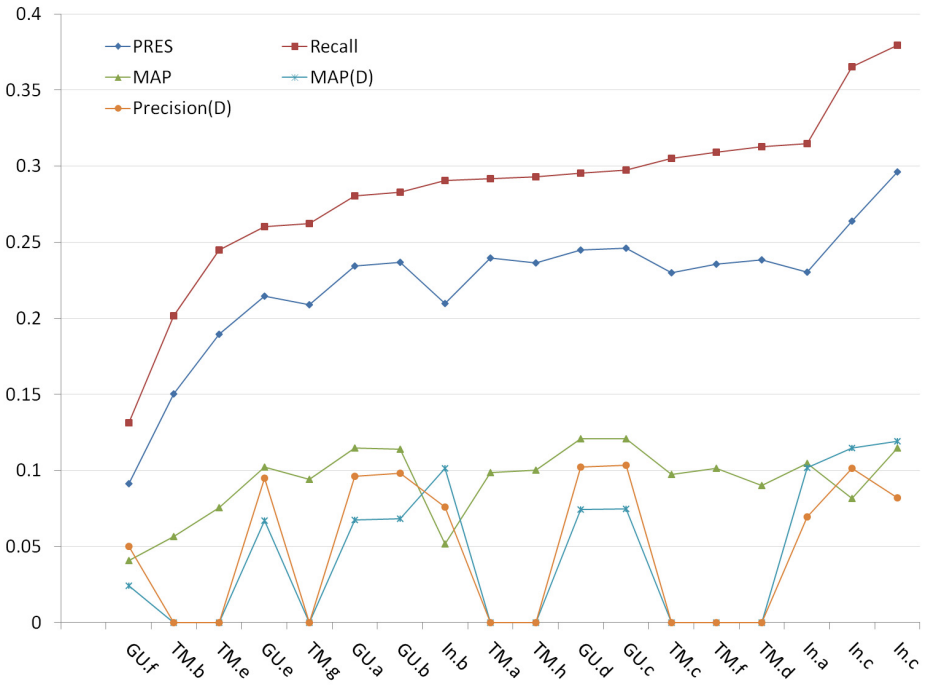


Fig. 4. Evaluation results, ordered by Recall

Passage Retrieval Starting from Claims. Considering the submission requirements, where both the patent document relevant to a topic as well as the most important passages in the document are given in the retrieval experiments, we proceeded to do evaluations at two levels: the document level and the passage level.

The evaluation at the document level ignored the passage information in the submitted runs. The metrics computed were PRES (Patent Retrieval Evaluation Score [20]), Recall, and MAP (Mean Average Precision).

At the passage level we compute, for each relevant document retrieved, the Precision and Average Precision scores of the retrieved passages. We then average over the number of relevant document retrieved to get the passage retrieval scores per topic. Averaging over all topics we obtain then the Precision(D) and Mean Average Precision MAP(D) for the retrieval experiment. For more details on these computations see [24]. The idea behind the solutions chosen to compute Precision(D) and MAP(D) are based on the measures used in the 'Relevant in Context' task of the INEX evaluation track [12].

Before running any evaluation scripts, we did a clean-up of the submissions by checking that the data follows the required format, that no duplicates occur, and that the retrieval results for one topic were not scattered in the submission file (this caused the evaluation script to exit with an error code). We also removed all XPaths refering to headings since we deliberately left them out of the relevance judgements as well. On the qrels side we found that out of 149 two topics were erroneous (topic 78 and topic 101) so we removed them from the evaluation data.

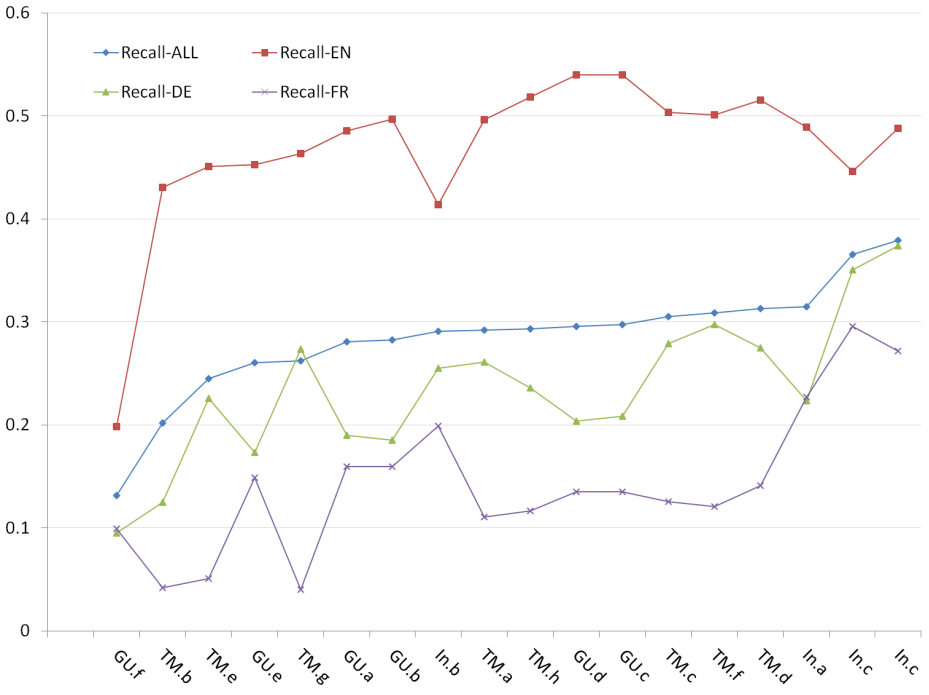


Fig. 5. Evaluation results, document level Recall per language

In the figures below we will use a shortened name for the experiment files. The mapping between the short and the original file name is shown in the appendix.

In our evaluations we considered all documents equally relevant and did evaluations on four sets of topics: the set of all 147 topics, the subset of 50 English topics (1-50), the subset of 49 German topics (51-100, with topic 78 removed), and the subset of 48 French topics (102-149, topic 101 previously removed). The results of the evaluation for the whole topic set are shown in Figure 4, and the document level Recall scores per languages are ploted in Figure 5. One participant submitted retrieval results to the document level, only, which is the reason for the zero Precision(D) and MAP(D) scores.

Further evaluations were done depending on the relevance degree of the patent citation documents, evaluations presented in [25]. A thorough statistical analysis of the retrieval result scores is yet to be done in the near future and will be reported on.

Structure Recognition Evaluations. To our dismay, there were no submissions to this task. Nevertheless, in the eventuality that image information extraction experiments were submitted, we were prepared to do evaluation using a set of measures to assess the effectiveness of flowchart recognition. The first set of measures are based on a graph distance metric using the notion of 'most common subgraph' (see [6,28] for a definition of the metric and [24] for how it was used in the evaluation last year). Using the experimental data participants submitted in 2012 we also investigated a functional view of the flow-chart recognition results (see [17]).

5 Final Words

The CLEF-IP Lab and its tasks have evolved considerably over the last five years, from a rough approximation of a prior art search task in 2009, to, in 2013, a good simulation of the passage-level search carried out by patent searchers. Along the way we have also investigated other important aspects of patent search such as patent classification and patent image search.

The increase in the realism of the tasks over the five years has also raised the bar for participation. In 2009, the CLEF-IP task was similar to a standard ad-hoc retrieval task, and participants could straightforwardly apply general IR solutions and achieve good results. As the tasks have been more closely modelled on actual patent search workflows, participants have been required to invest increasing time in understanding how the patent system works and in developing more granular retrieval solutions.

These factors have likely led to the decline in CLEF-IP submissions in recent years. In 2013, although the number of initial registrations to the tasks was promising, the small number of result submissions is visible in this paper. These factors have made us decide not to pursue the organisation of another round of CLEF-IP evaluations.

The comprehensive, curated test collection containing patent data, with tasks closely related to various activities of a patent expert's daily workflow, created during the five years of running the CLEF-IP Lab, will remain available to the research community. This should give researchers more than the few months available in the CLEF cycle to develop solutions meeting the demanding requirements of professional patent searchers. A conclusion of CLEF-IP is that patent IR is certainly not a solved problem — many challenges [16] in applying IR solutions in the intellectual property domain remain to be overcome.

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. ***. Citations, `http://www.intellogist.com/wiki/Citations` (last retrieved: July 2013)
2. ***. International Patent Classification (IPC), `http://www.wipo.int/classifications/ipc/en/` (last retrieved: March 2013)
3. Manual of Patent Examining Procedure (MPEP), revision 2012 (2012) (last retrieved: June 2013)
4. Anderka, M., Stein, B.: The ESA Retrieval Model Revisited. In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C.X., Zobel, J. (eds.) Proceedings of SIGIR, pp. 670–671. ACM (2009)
5. Attar, R., Fraenkel, A.S.: Local Feedback in Full-text Retrieval Systems. J. ACM 24(3), 397–417 (1977)
6. Bunke, H., Shearer, K.: A Graph Distance Metric Based on the Maximal Common Subgraph. Pattern Recognition Letters 19(3-4), 255–259 (1998)
7. Giachanou, A., Salampasis, M., Satratzemi, M., Samaras, N.: Report on the CLEF-IP 2013 Experiments: Multilayer Collection Selection on Topically Organized Patents. In: CLEF (Notebook Papers/LABs/Workshops) (2013)
8. Goto, I., Lu, B., Chow, K.P., Sumita, E., Tsou, B.K.: Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In: Proceedings of NTCIR, vol. 9, pp. 559–578 (2011)
9. Graf, E., Azzopardi, L.: 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)
10. Hanbury, A., Zenz, V., Berger, H.: 1st international workshop on advances in patent information retrieval (AsPIRe 2010). SIGIR Forum 44(1), 19–22 (2010)
11. Iwayama, M., Fujii, A., Kando, N., Marukawa, Y.: An Empirical Study on Retrieval Models for Different Document Genres: Patents and Newspaper Articles. In: Proc. 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003, pp. 251–258. ACM (2003)
12. Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., Robertson, S.: INEX 2007 Evaluation Measures. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) INEX 2007. LNCS, vol. 4862, pp. 24–33. Springer, Heidelberg (2008)
13. Kando, N., Leong, M.-K.: Workshop on Patent Retrieval (SIGIR 2000 Workshop Report). SIGIR Forum 34(1), 28–30 (2000)
14. Luo, J., Yang, H.: Query Formulation for Prior Art Search - Georgetown University at CLEF-IP 2013. In: CLEF (Notebook Papers/LABs/Workshops) (2013)
15. Lupu, M., Huang, J., Zhu, J., Tait, J.: TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at TREC. SIGIR Forum 43(2) (December 2009)
16. Lupu, M., Hanbury, A.: Patent Retrieval. FnTIR. NOW Publishers (2012)
17. Lupu, M., Piroi, F., Hanbury, A.: Evaluating Flowchart Recognition for Patent Retrieval. In: Song, R., Webber, W., Kando, N., Kishida, K. (eds.) The Fifth International Workshop on Evaluating Information Access (EVIA), pp. 37–44 (2013)

18. Lupu, M., Piroi, F., Huang, X., Zhu, J., Tait, J.: Overview of the TREC 2009 Chemical IR Track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, 2009, Gaithersburg, Maryland, USA, November 17-20, Special Publication 500–278. National Institute of Standards and Technology, NIST (2009)
19. Lv, Y., Zhai, C.: When Documents are Very Long, BM25 Fails! In: Ma, W.-Y., Nie, J.-Y., Baeza-Yates, R.A., Chua, T.-S., Croft, W.B. (eds.) Proceedings of SIGIR, pp. 1103–1104. ACM (2011)
20. Magdy, W., Jones, G.J.F.: PRES: A Score Metric for Evaluating Recall-oriented Information Retrieval Applications. In: SIGIR 2010 (2010)
21. Osborn, M., Strzalkowski, T., Marinescu, M.: Evaluating Document Retrieval in Patent Database: A Preliminary Report. In: Proceedings of the Sixth International Conference on Information and Knowledge Management, CIKM 1997, pp. 216–221. ACM, New York (1997)
22. PCT. Patent Cooperation Treaty (1970), `http://www.wipo.int/pct/en/treaty/about.html` (last retrieved: March 2013)
23. Piroi, F., Lupu, M., Hanbury, A., Zenz, V.: CLEF-IP 2011: Retrieval in the Intellectual Property Domain (September 2011)
24. Piroi, F., Lupu, M., Hanbury, A., Sexton, A.P., Magdy, W., Filippov, I.V.: Clef-IP 2012: Retrieval Experiments in the Intellectual Property Domain. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
25. Piroi, F., Lupu, M., Hanbury, A.: Passage Retrieval Starting from Patent Claims. A Clef-Ip2013 Task Overview. In: CLEF (Online Working Notes/Labs/Workshop) (2013)
26. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-Based Multilingual Retrieval Model. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer, Heidelberg (2008)
27. Tait, J., Harris, C., Lupu, M.: The 3rd International Workshop on Patent Information Retrieval (PaIR 2010) (October 2010)
28. Wallis, W.D., Shoubridge, P., Kraetzl, M., Ray, D.: Graph Distances Using Graph Union. Pattern Recognition Letters 22(6/7), 701–704 (2001)
29. WIPO Econimics & Statistics Series. 2013: PCT Yearly Review. WIPO Publication No. 901E/2013 (2013), `http://www.wipo.int/export/sites/www/freepublications/en/patents/901/wipo_pub_901_2013.pdf`

Appendix

Table 2. Original and short experiment file names

Original file ID	Short ID	Comment
GU.HypCoTtlNoIdfUpperBoundLM	GU.a	
GU.HypCoTtlWithIdfUpperBoundLM	GU.b	
GU.HypDuTtlNoIdfUpperBoundBM	GU.c	
GU.HypDuTtlWithIdfUpperBoundBM	GU.d	
GU.OnlyClaimLM	GU.e	
GU.coOnlyTtlLM	GU.f	
In.c3g	In.a	
In.clesa	In.b	
In.cos	In.c	
In.solr-cos	In.d	Probably an error in generating this experiment file.
In.solr	In.e	
TM.10-100.CORI.CORI.split3	TM.a	no relevant passages returned
TM.10-100.CORI.SSL.split4	TM.b	-"-
TM.10-100.CORI.SSL.split5	TM.c	-"-
TM.10-100.Multilayer.CORI.split4	TM.d	-"-
TM.10-100.Multilayer.CORI.split5	TM.e	-"-
TM.20-50.CORI.CORI.split5	TM.f	-"-
TM.20-50.Multilayer.CORI.split5	TM.g	-"-
TM.clefip-2013-centralised	TM.h	-"-