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

György Szarvas*, Benjamin Herbert, Iryna Gurevych
UKP Lab, Technische Universität Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

**Abstract**

In this study, we describe our system at the Intellectual Property track of the 2009 Cross-Language Evaluation Forum campaign (CLEF-IP). The CLEF-IP track addressed prior art search for patent applications. We used the Apache Lucene IR library to conduct experiments with the traditional TF-IDF-based ranking approach, indexing both the textual content of each patent and the IPC codes assigned to each document. We formulated our queries by using all claims and the title of a patent application in order to measure the (weighted) lexical overlap between topics and prior art candidates. We also formulated a language-independent query using the IPC codes of a document to improve the coverage and to obtain a more accurate ranking of candidates. Additionally, we used the IPC taxonomy (the categories and their short descriptive texts) to create a Concept Based Query Expansion [14] model for measuring the semantic overlap between topics and prior art candidates and tried to incorporate this information to our system's ranking process. Probably due to an insufficient length of definition texts in the IPC taxonomy (used to define the concept mapping of our model), incorporating the concept based similarity measure did not improve our performance and was thus excluded from the final submission. Using the extended boolean vector space model of Lucene, our system remained efficient and still yielded fair performance: it achieved the 6th best Mean Average Precision score out of 14 participating systems on 500 topics, and the 4th best score out of 9 participants on 10.000 topics.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

## General Terms

Measurement, Performance, Experimentation

## Keywords

Patent Information Retrieval, Invalidity Search

# 1 Introduction

The CLEF-IP 2009 track was organized by Matrixware and the Information Retrieval Facility. The goal of this track was to investigate the application of IR methods for patent retrieval. The task was to perform prior art search, which is a special type of search with the goal of verifying the originality of a patent. If another patent or document is found that already covers a very similar

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

invention and no sufficient originality is given in a patent, it would no longer be valid. In the case of a patent application, this would prohibit the acceptance. If a patent is already accepted, opposition can render a patent invalid with citations of prior art they found. Therefore, finding even a single prior art document can be crucial in the process, as it can have direct effect on the decision about patentability, or withdrawal of the patent application.

Prior art search is usually performed manually at patent offices by experts over millions of documents. The process often takes several days and requires strict documentation and experienced professionals. It would be beneficial if IR methods could ease this task or improve the efficacy of search.

Major challenges associated with finding prior art are the following:

- The usage of vocabulary and grammar is not enforced and depends on the authors.
- In order to cover a wide field of applications, many times very general formulations and vague language are used.
- Some authors might try to disguise the information contained in a patent for taking actions against people that infringe a patent later.
- The description of inventions frequently uses new vocabulary, as probably no such thing existed before.
- Since patents can be submitted in three different languages even in the European Union, information constituting prior art might be described in a different language than the patent under investigation.

## 1.1 Dataset & Task

For the challenge, a collection of 1.9 million patent documents from the European Patent Office (EPO) was used. The documents in this collection correspond to approximately 1 million individual patents filed between 1985 and 2000 (thus one patent can have several files, with different versions/types of information). The patents are in the English, German, or French language. The language distribution is not equal: 69% of the patents are English, 23% are German, and 7% are French. The patents are given in an XML format and supply detailed information, e.g. title, description, abstract, claims, inventors, classification, abstract, etc. For more information about the dataset see the track web page[1].

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

Relevance assessments were compiled automatically using the citations to prior art documents found in the EPO files of the topic patent applications. The training data for the challenge consisted of 500 topics and relevant prior art. The evaluation was carried out on an unseen document set of size 500 (Small), 1.000 (Medium) and 10.000 (XLarge evaluation) topics, respectively.

The main goals of Patent Retrieval systems can be characterized as to achieve:

- high recall, as single result can invalidate a patent application, or
- high precision at top ranks to provide results that require less manual analysis by a patent expert.

For a more detailed description of the task, participating groups, the dataset and overall results, please see the challenge description paper: [15].

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

## 1.2 Related work

Patent retrieval has been studied extensively by the IR research community, within the scope of scientific workshops and patent retrieval shared tasks. In the early 2000s several workshops were organized at major Computational Linguistics conferences [11, 8] devoted to analyzing and discussing the applicability of IR techniques to patent document collections, and to assess the special characteristics of patent data compared to other genres.

Since 2003, Patent Information Retrieval has been studied within the scope of the NTCIR evaluation campaigns, mainly focusing to retrieval of patents in Japanese, and patent abstracts in English. At NTCIR-3, a patent retrieval task was given. The goal was to build technical surveys from two years of patent data [10]. Topics were given as newspaper articles and a memorandum from a person who is interested in a technical survey about a topic mentioned in the articles.

Invalidity Search was addressed at NTCIR-4 [4] using a document collection of Japanese patents published between 1993 and 2002. Using the claims in a topic patent, the task was to find a list of patents that invalidate the claims in the topic. Additionally, the passages that are in conflict with the topic had to be found. As a cross-lingual challenge, patents were also partially translated to English.

The same setup was given for the patent retrieval task at NTCIR-5, but with a larger number of topics. For some topics, passages that invalidate claims for a patent document had to be returned. More information about the task can be found in [5].

At NTCIR-6, the first English patent retrieval task was introduced [6]. The objective of the English task at NTCIR-6 was Invalidity Search using a collection of English patents from the US Patent & Trademark Office (USPTO). Topic documents were also patents from the USPTO. Relevance assessments were compiled using the citations in the topic documents, similarly to the CLEF-IP 2009 challenge.

The 2009 Intellectual Property challenge [15] at the Cross-Language Evaluation Forum campaign extended the scope of Invalidity Search to all the official languages of European Patent Office (i.e. English, German and French) using a collection of over 1 million patents in multiple languages. Another major difference between the NTCIR-6 and the CLEF-IP task was that here multiple manually assigned patent classification codes were available to each document, while at NTCIR-6 all patents were assigned a single label.

For the NTCIR-3 Workshop Iwayama et al. [9] indexed nouns, verbs, adjectives and out-of-dictionary words of patents and performed retrieval using newspaper articles as topics. They stated that patent retrieval was not significantly different from retrieval on newspaper items for the technical survey task. This made it promising to rely on the classical retrieval models also for invalidity search.

Fujii (2007) [3] employed word-based indexing for invalidity search and employed a citation based score to improve the ranking of retrieval results. The use of citation information was not allowed at the CLEF-IP challenge as the same information was used to compile relevance assessments for the challenge.

Subtopic analysis was performed by Takagi et al. (2004) [16] for invalidity search tasks. By analyzing and finding subtopics contained in target documents, multiple queries were built. For each query a search was carried out on the document database, resulting in a list of relevant patent documents. Unlike Takagi et al. we decided to use single queries for whole topic documents due to time constraints.

Several previous studies aimed at going beyond the traditional keyword-matching and apply a semantic retrieval approach for patents. For example, Patent Cafe[2] uses Latent Semantic Analysis to implement a semantic search engine for patent texts. The recent EU project called PATExpert [17] uses manually crafted ontologies for concept based retrieval of patents. On the other hand to our best knowledge Concept Based Query Expansion [14] has not yet been explored in Patent Retrieval.

The main findings of recent evaluation campaigns were that traditional IR methods work reasonably for patent collections, but the special language used in patent texts and the use of

[2] http://www.patentcafe.com/

different terminology might pose problems to keyword-based retrieval. Many studies point out the importance of exploiting the manually assigned topic labels (i.e. the patent classification codes assigned to applications by patent experts) for more efficient retrieval. The task overview papers of the above mentioned evaluation campaigns, the state of the art survey [2] of PATExpert and a recent survey on Patent informatics [1] provide a good overview of work related to this study.

# 2 Our Approach

In this section, we discuss our system submitted to the challenge.

## 2.1 Data Characterization

For most patents, several files were available, corresponding to different versions of the patent (an application text is subject to change during the evaluation process).

We decided not to use all the different versions available for the patent, but used the most up-to-date version. We considered the latest version to contain the most authoritative information. If a specific field used by our system was missing from that version, we extracted the respective information from the latest source that included the particular field. In our system, we used the information provided under the *claims*, *abstract*, *description*, *title* and *IPC codes* fields only.

Exploiting other, potentially useful sections of patent applications such as authors or date was omitted so far.

## 2.2 Preprocessing

To create the indices, we employed Lucene and performed the following preprocessing steps:

- *Sentence splitting* based on the Java BreakIterator implementation.
- *Tokenization* based on the Java BreakIterator (for the French documents we also used apostrophes as token boundaries: e.g. *d'un* was split to *d* and *un*).
- *Stopword removal* using manually crafted stopword lists. We started with general purpose stopword lists containing determiners, pronouns, etc. for each language, and appended them with highly frequent terms manually. We considered each frequent word (appearing in several hundreds of thousand of documents) a potential stopword and included it in the list, if we judged it as a generic term or a domain specific stopword, that is not representative of the patent content. For example, a large number of patent documents contain words like *figure* (used in figure captions and also to refer to the pictures in the text), or *invention* (it usually occurred in the 1st sentence of the documents). Since we lacked the necessary domain expertise to evaluate each term properly, stopword lists compiled by experts could easily improve our system to some extent.
- for the German language, we applied dictionary-based *compound splitting* [12][3].
- *Stemming* using the Porter algorithm[4].

The preprocessing pipeline was set up using the *Unstructured Information Management Architecture (UIMA)*, a framework for the development of component based *Natural Language Processing (NLP)* applications. We used the DKPro Information Retrieval framework [13], which provides efficient and configurable UIMA components for common NLP and Information Retrieval tasks.

[3] `http://www.drni.de/niels/s9y/pages/bananasplit.html`
[4] `http://snowball.tartarus.org`

## 2.3 Retrieval

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

### 2.3.1 Indices

In order to employ Lucene for patent retrieval, we created a separate index for each language using only fields for the corresponding language. That is, for example, for the German index, only fields with a language attribute set to *DE* were used.

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

Topic documents were indexed similarly in a separate topic index, in order to have the topic texts preprocessed in the same manner as the document collection. We created topic indices using the *title* and *claim* fields in each language. All the text in these fields was used to formulate the queries, without any particular further filtering. This way our system ranked documents according to their lexical overlap with the topic patent.

To exploit the IPC codes assigned to the patents, a separate index was created containing only the IPC categories of the documents. This information could provide a language independent ranking measure of the domain overlap between the query and documents.

### 2.3.2 Queries

In this section, we describe how the query is formulated for each topic.

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

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

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

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

### 2.3.3 Result Fusion

As a result, our system retrieved three ranked lists of patent documents, one result list for each of the three language indices. Since the majority of the true positive documents for the training topics shared at least one full IPC code[5] with the topic patent, we decided to filter our results to contain only such documents that shared an IPC code with the topic. Additionally, we acquired one result list from the IPC code index. We normalized each single list to have a maximum relevance value of 1.0 for the top ranked document in order to make the scores comparable to each other.

To prepare our system output for the language specific subtasks, we added the relevance scores returned by the IPC and the textual query and ranked the results according to the resulting

[5]For example *A61K-6/027*, corresponding to *Preparations for dentistry – Use of non-metallic elements or compounds thereof, e.g. carbon.*

relevance score. For the combination of results, we normalized the lists and then used the following formula: $Score(d) = \frac{Score_{IPC}(d)+Score_{text}(d)}{2}$

For the *Main* task submission, the three language-specific lists had to be combined in order to end up with a single ranked list of results. To do this, we took the highest language specific result from the three individual lists for each document. That is, each document was ranked according to its highest relevance score in the *Main* task submission: $Score_{main}(d) = MAX(Score_{EN}(d), Score_{DE}(d), Score_{FR}(d))$.

Whenever our system retrieved less then 1000 individual documents using the above described procedure, we appended the result list with documents retrieved by the same steps, but applying a less restrictive IPC code filter. This means that at the end of the list, we included such documents that shared only a higher level IPC category[6], but not an exact code with the topic.

# 3 Experiments and Results

In this section we present the performance statistics of the system submitted to the CLEF-IP challenge and report on some additional experiments performed after the submission deadline. We provide Mean Average Precision (MAP) as the main evaluation metric, in accordance with the official CLEF-IP evaluation. Since precision at top rank positions is extremely important for systems that are supposed to assist manual work, like prior art search, we always indicate Precision at 1 and 10 retrieved documents (P@1 and P@10) for comparison[7].

## 3.1 Challenge submission

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

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

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

The bold line in Table 1 represents our submitted system. The same configuration gave the best scores on the training topic set for each individual language. Table 2 shows the scores of this system configuration for each language and the Main task on the 500 training and on the 10000 evaluation topics.

[6]Here we took into account only the 3 top levels of the IPC hierarchy. For example *A61K*, corresponding to *Preparations for dentistry*.

[7]During system development we always treated all citations as equally relevant documents, so we only present such evaluation here. For more details and analysis of performance on highly relevant items (e.g. those provided by the opposition) please see the task description paper [15].

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

Table 1: Performance on Main task, 500 train topics.

| | Train 500 | | | Evaluation 10k | | |
|---|---|---|---|---|---|---|
| **Task** | **MAP** | **P@1** | **P@10** | **MAP** | **P@1** | **P@10** |
| English | 0.1157 | 0.2160 | 0.0876 | 0.1163 | 0.2025 | 0.0876 |
| German | 0.1067 | 0.2140 | 0.0818 | 0.1086 | 0.1991 | 0.0813 |
| French | 0.1034 | 0.1940 | 0.0798 | 0.1005 | 0.1770 | 0.0774 |
| Main | 0.1212 | 0.2160 | 0.0896 | 0.1186 | 0.2025 | 0.0897 |

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

### 3.2 Post submission experiments

After the submission deadline, we ran several additional experiments to gain a better insight to the performance limitations of our system. We only experimented with the English subtask, for the sake of simplicity and for time constraints. First, we experimented with different weightings for accumulating evidence from the text- and IPC-based indices. The results are summarized in Table 3. The results indicate that slightly higher weight to text-based results would have been beneficial to performance in general. Using 0.6/0.4 weights, as suggested by Table 3, would have given 0.1202 MAP score for English, on the 10k evaluation set – which is a 0.4% point improvement.

| Weight | MAP |
|---|---|
| 0.0 / 1.0 | 0.0684 |
| 0.1 / 0.9 | 0.0771 |
| 0.2 / 0.8 | 0.0821 |
| 0.3 / 0.7 | 0.0939 |
| 0.4 / 0.6 | 0.1046 |
| 0.5 / 0.5 | 0.1157 |
| **0.6 / 0.4** | **0.1198** |
| 0.7 / 0.3 | 0.1180 |
| 0.8 / 0.2 | 0.1102 |
| 0.9 / 0.1 | 0.1046 |
| 1.0 / 0.0 | 0.0997 |

Table 3: MAP scores on Train 500 with different weightings of the combined text / IPC results.

We also examined retrieval performance using different document length thresholds. That is, we extracted the first 400, 800, 1600 or 3200 words of the concatenated claim, abstract and description fields to see whether more text improves the performance. Table 4 shows the MAP scores for these experiments. The results show that only a slight improvement could be reached by using more text for indexing documents. Using 1600 words as the document size threshold, as suggested by Table 4, would have given 0.1170 MAP score for English, on the 10k evaluation set – which is only a marginal improvement over the submitted configuration.

The best performing configuration we obtained during the submission period included a filtering step to discard any resulting document that did not have any IPC code shared with the topic. This

| Method | MAP |
|---|---|
| 400 words max | 0.1116 |
| 800 words max | 0.1161 |
| **1600 words max** | **0.1184** |
| 3200 words max | 0.1177 |
| fulltext | 0.1147 |

Table 4: MAP scores for English on Train 500 with different document length thresholds.

way, retrieval was actually constrained to the cluster of documents that had some overlapping IPC labeling. A natural idea was to evaluate whether creating a separate index for these clusters (and thus having in-cluster term weighting schemes and ranking) is beneficial to performance. Results of this cluster-based retrieval approach are reported in Table 5.

The best paramater settings of our system (i.e. one using 1600 words as document length treshold, 0.6/0.4 weights for text/IPC indices, respectively and separate indices for the cluster of documents with a matching IPC code for each topic – the bold line in Table 5.) showed a MAP score of 0.1243, P@1 of 0.2223 and p@10 of 0.0937 on the 10.000 documents size evaluation set, for English. This is a 0.8% point improvement in MAP compared to our submitted system.

| Method | MAP |
|---|---|
| 800 words 0.5/0.5 weights (text/IPC) | 0.1203 |
| 800 words 0.6/0.4 weights (text/IPC) | 0.1223 |
| 1600 words 0.5/0.5 weights (text/IPC) | 0.1202 |
| **1600 words 0.6/0.4 weights (text/IPC)** | **0.1252** |

Table 5: MAP scores for English on Train 500 with different document length thresholds.

# 4 Discussion

In the previous section, we introduced the results we obtained during the challenge timeline, together with some follow-up experiments. We think our relatively simple approach gave fair results, our submission ended 6th out of 14 participating systems on the Small evaluation set of 500 topics[8] and 4th out of 9 systems on the larger evaluation set of 10000 topics. Taking into account that only one participating system achieved remarkably higher MAP scores, and the simplicity of our system, we find these results promising.

We attribute these promising results to the efficient use of IPC information to enhance keyword-search. Our experiments demonstrated that the several ways we employed IPC codes to restrict/focus text search (i.e. filtering according to IPC, retrieval based on IPC codes, local search in the cluster of patents with matching IPC) all improved retrieval performance. We also tried to incorporate the whole IPC taxonomy to extend the traditional vector space model based retrieval similarly to Qui and Frey's [14] Concept Based Query Expansion technique. Unfortunately, this approach did not improve the performance of our system, most probably due to the very short descriptive information given in the IPC taxonomy for each category. We think that this approach would be particularly promising if a version of the taxonomy with a reasonable amount of descriptive text for its categories were available.

We also discussed in detail that during the challenge development period, we made several arbitrary choices regarding system parameter settings and that (even though we chose reasonably

[8]Since the larger evaluation set included the small one, we consistently reported results on the largest set possible. For more details about performance statistics on the smaller sets, please see [15]

well performing parameter values), tuning these parameters could have improved the accuracy of the system to some extent. The limitations of our approach are obvious though:

- First, as our approach mainly measures lexical overlap between the topic patent and prior art candidates, such prior art items that use significantly different vocabulary to describe their innovations are most probably missed by the system.
- Second, without any sophisticated keyword / terminology extraction from the topic claims, our queries are long and probably contain irrelevant terms that puts a burden on the system's accuracy.
- Third, the patent documents provided by the organizers were quite comprehensive, containing detailed information on inventors, assignees, priority dates etc. Out of these information types we only used the IPC codes and some of the textual description of patents.
- Last, since we made the compromise to search among documents with a matching IPC code (and only extend to documents with a matching main category when we had insufficient number of retrieved documents in the first step), we obviously lost the chance of retrieving such prior art items that have different IPC classification from the patent being investigated. We think these patents are possibly the most challenging and important items to find – since they are more difficult to discover for humans as well.

## 4.1 Error analysis

We examined a few topics manually to assess the typical sources of errors produced by our system. Our findings nicely agree with the claims above. Restricting the search for the cluster with a matching IPC code reduces the search space to a few thousand of documents on average. On the other hand, our system configuration results in losing 11% of the prior art items entirely (i.e. those that do not have even a main IPC category shared with the topic patent) and in very low chances of retrieving another 10% of the prior art (i.e. those that share only a main IPC category but not an exact IPC code). Besides these issues, the system is able to retrieve the majority of the remaining relevant documents within the top ranked 1000 results.

Poor ranking (that is, having relevant items ranked low in the list) comes from the lack of selection of meaningful search terms from the topic patents. Since many patents discuss very similar topics, usually there are a number of documents with substantial overlap, but relevance is defined more by the presence or absence of a very few very specific terms or phrases (we only consider unigrams for retrieval). This is the most obvious place for improvement regarding our system.

A typical example is the topic patent with id *EP-1474501*[9]. This patent had 16 true positive (TP) documents in the collection. Out of these, we retrieved 12 and missed 4 – 3 having not even a main IPC category shared with the topic[10] and 1 sharing only main category[11]. We also got a TP document top ranked (*EP-0767824*), which is indeed very similar to the topic: IPC codes and whole phrases and sentence parts match between them. The rest of the TPs on the other hand came ranked low (under the 100th). We saw the main reason for this in many documents having a large vocabulary overlap with the topic, but having different technical terms like materials, names of chemicals, etc. - aspects that really make the difference in a prior art search setting. We think that an intelligent selection of query terms would have resulted in ranking the relevant documents higher here.

[9] *Lubricating compositions / IPC:C10M* as main topic

[10] Having *detergent compositions / IPC:C11D*; *macromolecular compounds obtained by reactions only involving carbon-to-carbon unsaturated bonds / IPC:C08F* and *shaping or joining of plastics / B29C and containers for storage or transport of articles or materials / IPC:B65D* as their main topics

[11] Both on *lubricating compositions*, but the topic categorized as a mixture, while the prior art document categorized according to its main component (different 4th and 5th level classification)

# 5 Conclusion

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

On the other hand, our experiments conducted within the scope of the CLEF 2009 Intellectual Property challenge might not provide a good insight to the precision of such systems: to our knowledge only such topics were selected for evaluation that were actually opposed by third parties (in other words only such patents were used for evaluation purposes that actually were questionable regarding their novelty). This also emphasizes that our system probably would be usable only for assisting manual search.

## 5.1 Future work

As follow up research, we plan to extend our system in several different ways. We showed that local and global term weightings behave differently in retrieving prior art documents. A straightforward extension would be therefore to incorporate both to improve our results further. Similarly, experimenting with other weighting schemes than the one implemented in Lucene is another straightforward way to extend our system.

More important, we plan to further investigate the possibilities of incorporating semantic similarity measures to the retrieval process, complementary to lexical overlap. For this – since we don't have access to an IPC taxonomy with sufficient textual descriptions – we plan to experiment with the concept based query expansion model when Wikipedia is used as a source of background knowledge [7] for constructing the concept-based text representation.

# 6 Acknowledgements

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

# References

[1] D. Bonino, A. Ciaramella, and F. Corno. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics (in press). *World Patent Information*, 2009.

[2] S. Brügmann. Patexpert - state of the art in patent processing. Technical report, ISJB, 2006.

[3] A. Fujii. Enhancing patent retrieval by citation analysis. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 793–794. ACM New York, NY, USA, 2007.

[4] A. Fujii, M. Iwayama, and N. Kando. The patent retrieval task in the fourth NTCIR workshop. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 560–561. ACM New York, NY, USA, 2004.

[5] A. Fujii, M. Iwayama, and N. Kando. Overview of patent retrieval task at NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 269–277, 2005.

[6] A. Fujii, M. Iwayama, and N. Kando. Overview of the patent retrieval task at the ntcir-6 workshop. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 359–365, 2007.

[7] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007.

[8] M. Iwayama and A. Fujii, editors. *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[9] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. An empirical study on retrieval models for different document genres: patents and newspaper articles. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 251–258, New York, NY, USA, 2003. ACM.

[10] M. Iwayama, A. Fujii, N. Kando, and A. Takano. Overview of patent retrieval task at NTCIR-3. In *Proceedings of the NTCIR-3 Workshop Meeting*, 2003.

[11] N. Kando and MK. Leong. Workshop on Patent Retrieval SIGIR 2000 - Workshop Report. volume 34, pages 28–30, New York, NY, USA, 2000. ACM.

[12] S. Langer. Zur Morphologie und Semantik von Nominalkomposita. In *Tagungsband der 4. Konferenz zur Verarbeitung naturlicher Sprache*, pages 83–97, 1998.

[13] C. Müller, T. Zesch, MC. Müller, D. Bernhard, K. Ignatova, I. Gurevych, and M. Mühlhäuser. Flexible UIMA Components for Information Retrieval Research. In *Proceedings of the LREC 2008 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*, pages 24–27, Marrakech, Morocco, May 2008.

[14] Y. Qiu and HP. Frei. Concept based query expansion. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169, New York, NY, USA, 1993. ACM.

[15] G. Roda, J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In *Working Notes of the 10th Workshop of the Cross Language Evaluation Forum (CLEF)*, Corfu, Greece, 2009.

[16] T. Takaki, A. Fujii, and T. Ishikawa. Associative Document Retrieval by Query Subtopic Analysis and its Application to Invalidity Patent Search. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 399–405. ACM New York, NY, USA, 2004.

[17] L. Wanner, R. Baeza-Yates, S. Brügmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, et al. Towards Content-oriented Patent Document Processing. *World Patent Information*, 30(1):21–33, 2008.