Effects of Language and Topic Size in Patent IR: An Empirical Study

Florina Piroi, Mihai Lupu, and Allan Hanbury

Vienna University of Technology, Vienna, Austria
{piroi,lupu,hanbury}@ifs.tuwien.ac.at

Abstract. We revisit the effects that various characteristics of the topic documents have on the effectiveness of the systems for the task of finding prior art in the patent domain. In doing so, we provide the reader interested in approaching the domain a guide of the issues that need to be addressed in this context.

For the current study, we select two patent based test collections with a common document representation schema and look at topic characteristics specific to the objectives of the collections. We look at the effect of languages on retrieval and at the length of the topic documents. We present the correlations between these topic facets and their retrieval results, as well as their relevant documents.

1 Introduction

The large amounts of available digital information lead to research in large-scale IR engines. This, in turn, brings on questions such as how to evaluate IR engines in a context as realistic as possible. Creating large pools of documents is not a problem, but asking the right questions (topics) and, more importantly, providing the right answers (relevance judgements) is. Efforts to obtain humanly created relevance judgements are done either via a massively distributed online evaluation system (e.g., Amazon's Mechanical Turk webservice [1]), or by re-using specialized work done in some specific contexts [3]. For the data collections in this paper, the latter is the case, as we focus on patent search.

Independent of the tasks organized in an IR evaluation campaign with patent data, the main course in the campaigns of the last decade has been to make proper use of the extremely large amounts of work already done by professional patent searchers worldwide, rather than focusing on consistently reducing the number of topics. Most of the evaluation campaigns using patent data have relevance judgements based on search reports. Although the search report, just like an article's reference list, is never exhaustive, for comparison purposes, the evaluation is still valid and in line with current practice in standard evaluation campaigns. Some caveats in using the search report in this way exist and are specific to the patent domain and the way the intellectual property protection system is designed and functions. They are briefly explained in Section 1.2.

So far, the questions being asked in these evaluation campaigns have been more or less random and give an overview perspective of the performance of different systems. As patent-based test collections mature, we must grasp a better

T. Catarci et al. (Eds.): CLEF 2012, LNCS 7488, pp. 54–66, 2012.

understanding of the characteristics of the topics selected to evaluate systems, and their expected effects on the performance of such systems. Such an analysis can then be used either to act as a baseline (when one knows that, for instance, a particular kind of topic is easily answered by all systems), or to direct future evaluation campaigns into areas where more work is needed to achieve a satisfactory retrieval success.

Prompted by the use of patent search reports as a basis for generating relevance assessments, the CLEF–IP (for Cross-Lingual retrieval) and TREC–CHEM (for chemical retrieval) evaluation campaigns have taken patent application documents and used them as basis for topics in a "prior art" task. The objective: retrieve other patent documents related to the given application.

This study investigates how the results of these evaluation campaigns change when we vary the set of topics based on specific features of the documents used as topics in the evaluation. The scores we observe here are the Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG). The document features we take into consideration are: the document language, for the cross-lingual evaluation, and, for both evaluations, the length of the topic documents. While these have been studied to some extent in the literature, it is useful to see to what extent observations made before apply in the context of patent documents. The reasons to believe these observations may be different are as follows:

- most cross-lingual evaluations to date consider the query in one language and the result set in another. Patent retrieval considers the query in one language and the result set in several different languages.
- in the cases where the cross-lingual evaluation task does require a set of results in different language, there is little meta-information that the system can use to connect multilingual documents. In the patent domain, there are several explicit links between documents in different languages (e.g. family membership, inventor, assignee, etc.)
- "verbose" queries in general IR are a few tens or at most hundreds of words. Patent applications, i.e. the topics of prior art search, are thousands, up to hundreds of thousands of words in length and contain different language genres, not commonly found in studies of topical length.

1.1 Outline of the Paper

We continue this section with a compressed introduction to the patenting process, establishing, at the same time, the patent-specific terminology used throughout this paper. Related work on the influence of language and topic sizes, as well as the use of patent collections as IR test collections is described in Section 2. Section 3 describes the four collections used in this study and the methodology for the experiments and ensuing analysis. Section 4 represents the main body of this work, where we look at different aspects of the topic sets selected for this study. We summarize and provide directions for future work in Section 5.

1.2 Brief Survey of Patent Terminology

To facilitate understanding the characteristics of patent-based test collections, we need to establish the terminology used in the patent domain, terminology used throughout this work.

A patent is a set of exclusive legal rights, for a limited period of time, for the use and exploitation of an invention in exchange for its public disclosure. The requirements for granting patents vary among patent offices, but a common first step is to file a patent application request with a patent office. For this, the applicant must supply a written specification of the invention (i.e. a *patent application document*) where the background of the invention, a description of the invention, and a set of claims which define the scope of protection, should the patent be granted, are given. Most of the time, applicants should name (patent) documents relevant to their invention in the text of the application.

To be granted, a patent application is examined by professionals who will analyze whether it meets certain patentability criteria. Of relevance to IR evaluation campaigns is the novelty criteria. A patent application satisfies the novelty requirement if no earlier patent or other kind of publication describing (parts of) the invention can be found in a reasonable amount of time. The search for novelty-relevant documents is called *prior art search*. Results of a prior art search, together with the patents named by the applicants themselves, are recorded in a *search report*. The documents listed in the search report of a patent are referred to as *patent citations* and, at least for European Patent Office (EPO) and US Patents and Trademarks Office (USPTO), are assigned degrees of relevance, which influence the course of the patent application within its life cycle.

Patent documents generated at the different stages of the patent's life-cycle are identified by a *country code* (denoting the patent office analyzing/granting the patent), a *numeric identifier*, and by a *kind code* together with a version number. Together, these three components form a unique *global* identifier - another very useful feature for IR evaluations.

The main types of patent documents are the ones mentioned above: application document, search report, granted patent. To these, depending on the legal procedures to which a patent is subjected, other patent documents may be added, e.g. additional search reports, documents marking a change in the owner of the invention, countries of applicability, etc. When regarding a patent, all its patent documents must be considered.

To protect an invention in several geographical areas, a patent application can be filed at more than one patent office. When the same invention is granted a patent by different patent offices, the two patents are said to belong to the same *patent family*. In certain conditions, a patent family may contain patents granted by the same patent office. This may happen for instance, when a patent office has a more granular patenting practice, and an invention which was granted one patent by a patent office is split into two or more inventions by another. The use of patent families in IR evaluation is generally useful, as it provides a more complete picture of the set of relevant documents to a topic, but the caveat is that, in situations where a patent application to one patent office is split into

several at another, or vice-versa, it is no longer clear which citations are relevant to which of the versions of the same invention.

For the cross-language evaluation campaigns, one feature of the EPO patenting process is of particular interest, namely, the mandatory translations of granted patents. It is a procedural step at the EPO to translate all claims of a granted patent into English, German and French.

2 Related Work

The first retrieval evaluation campaign on a collection of patent documents was organized in the frame of the NTCIR workshop series [5], based on seminal work done in the context of a workshop in 2000 [7].

The NTCIR patent collections contain a significant number of patents, over 3 million in the first year, mostly from the Japanese Patent Office. The tasks and their relevance judgements have changed over the years, including Prior Art, Classification and Machine Translation tasks. After initial experiments with manually evaluated topics, the NTCIR organizers moved to extracting relevance judgments from search reports. Later, in the work done by Fujii et al. [4], relevance judgements obtained from citation records are compared with judgements inferred from patent classification codes, showing considerably different ranking results, but without providing any insight as to which one may be better.

In the past, evaluation campaigns involving cross-language IR have shown various retrieval effectiveness results when looking at the different topic languages (see for example [2],[6] or [15]). Depending on the track settings and the type of data collections involved, retrieval results for English and German topics, for example, were similar [8],[9] or very different [2] with IR systems generally giving a better retrieval efficiency for English topics than other languages.

3 Methodology and Data

The analysis of the results of past evaluation campaigns provide useful insights into common and distinct features of retrieval systems. Existing data can however be overwhelming and difficult to analyze. A clearly defined methodology and target data are identified. The following subsections report on our instantiations.

3.1 Experimental Process

Here are the steps of our methodology, together with explanations and links to the afferent sections.

1. **Initial data selection.** Before starting the process, filter out any parts of the data that cannot be used. This first step must also make sure that the data is large enough for the scope of the analysis. (Section 4.1)

2. **Identification of the investigation subject.** e.g. "How does the language of the topic patent application text influences the rankings and effectiveness measures?" The second step of our working plan identifies the document facets to investigate.
3. **Extract subsets of topics for the investigation subject.** For each of the facets identified above, we divide the entire topic set according to the particular facet under investigation.
4. **Experimentation.** Evaluate the selected retrieval experiments using the subsets. We use `trec_eval`[1] to compute the Mean Average Precision, MAP, and the Normalized Discounted Cumulated Gain, NDCG. In this same computation step we calculate the Kendall's Tau and the Spearman's Rho correlation coefficients between these results and the results using the undivided topic sets. We note here that there are at least two reasons to compute NDCG scores for these test collections: a) the relevance judgements are based on citation reports which have relevance degrees assigned to the relevant documents listed in them; and b) the average number of relevant documents per topic is low compared to other IR tracks.
5. **Analysis.** Observe results, compare to expectations, formulate new hypothesis. We analyze the MAP and NDCG scores and the correlation coefficients and conclude wether the formulated hypotheses are rejected or not. The effects observed on system rankings and retrieval effectiveness are presented in the subsections of Sections 4.2-4.3.

3.2 Test Collections

While consisting of different collections of data, both CLEF–IP and TREC–CHEM had a task named "Prior Art" in 2009 and 2010, where participants were invited to find patent documents in a certain relationship to other (topic) patent documents. Each such topic was phrased as "Please find all patent documents that would potentially invalidate patent X", where X was one of the topic patents. The selection of topics was done depending on the number of patent citations and the source of these citations (applicant, examiner, etc., sources being recorded in the patents' search report documents and available in the data collections).

The extraction of the relevance judgements for these topics was done using the citations, as described in [18], [11]. Characteristic to the patent data is that the number of citations[2] of any given patent is very small compared to the number of relevant documents for a topic in an ad-hoc evaluation campaign. To increase the number of relevant documents per topic, family members of the topic patent and of the citations were also included in the relevance judgements. By this procedure, for the set of CLEF–IP 09 topics in this study, the average number of relevant documents per topic was raised from 1.89 to 5.63, while for the TREC–CHEM 09 collection the average number of relevant document per

[1] `http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz`
[2] Recall that a patent *citation* refers to a document that is relevant to the patent.

topics increased from 30.8 to 48.9. Previous work [10] has shown that citation-based relevance judgments are indeed closer, in terms of ranking correlation, to manual judgments than fully automatic pseudo-judgments.

CLEF–IP 2009/2010 The CLEF–IP test collection is the first collection of patents with content in three main European languages addressed to the evaluation of cross-lingual IR. The patents are obtained from the EPO and contain text in German, English, and French [18]. The documents are assigned a 'document language', but parts of their content may additionally occur in one or both of the other languages. The documents in the collections refer to over 1, respectively 1.3 million patents published before the year 2000 (CLEF–IP 09) and 2001 (CLEF–IP 10). In both years the topics were extracted from a pool of documents different from the distributed corpus, a pool which contained over 0.7 million patent documents published after the documents in the corpus.

The Prior Art Search task organized in 2009 had a very large number of topics (10000). The topics were syntetic documents, created such that at least the claims were available in all 3 languages in the collection [18]. In 2010 the number of topics was reduced to 2000, the topic documents were patent application documents with the claims usually present in one language only [16].

TREC–CHEM 2009/2010 The TREC–CHEM test collection has been used for the Chemical IR track of TREC [19]. It contains both patent documents and scientific articles, all chemistry-related. The task of interest for us in this article, the Prior Art Task, used only a subset with 1.1 million patent documents that contained claims and either an abstract or a description of the invention, or both. The documents are obtained from the USPTO and the EPO.

The 2010 version of the collection has added more content to the scientific articles sub-collection, and added a set of images and chemical structure files to the collection. For the prior art task, the patent sub-collection was increased to approximately 1.3 million documents, not only from the USPTO and EPO sources, but also from the World Intellectual Property Organization (WIPO).

Each TREC–CHEM 2009 and 2010 contained a set of 1,000 topics for the Prior Art Search task.

4 Analysis

In this section we instantiate the methodology described in Section 3.1. We want to examine whether the observations made in other retrieval contexts with respect to the influence of language and size of the topic apply in the case of patents. We look at these two aspects in sections 4.2 and 4.3, respectively. For each of them, we outline the lessons learned for both participants and evaluation campaign organizers.

4.1 Selecting the Data

For the purpose of this study, we have used the same topic sets as the evaluation campaigns in the case of the TREC–CHEM collections. In the CLEF–IP case we

chose a random subset of 1,000 topics out of the largest set of topics in 2009 (10,000), and 1937 topics out of 2000 topics in 2010 (it was later found that 63 topics had faulty relevance judgements).

To test our hypotheses we have made evaluations on experiments submitted to the two campaigns. The evaluation results described in this section are obtained by evaluating the data in 15 runs of the TREC–CHEM09 track, 9 runs of TREC–CHEM10, 24 runs of CLEF–IP09, and 18 of the CLEF–IP10 track for the above mentioned sets of topics, respectively. Although the number of submissions to both tracks is larger, we selected only the runs that actually provide results for all of the topics in each set.

4.2 The 'Document Language' Feature

When examining a patent application for the novelty criteria a patent professional has to look for prior art in all collections available to her, regardless of language. It is often the case that for a patent application written in, for example, German there are relevant documents written in English or French. The CLEF–IP collection includes these cases, and since the collection was meant for cross-language retrieval, it is expected to look at how the topic's document language reflect upon the retrieval results.

We split each of the CLEF–IP topic sets in this study in three, based on the document language. For both years, approximately 60% of the topics have the document language English, 30% German, and 10% French. (This distribution faithfully reflects the distribution of document language in the whole CLEF–IP collection.)

Effects On System Rankings and Effectiveness Measures. We have computed the MAP and NDCG scores (not displayed here) and the correlation coefficients between the system rankings (Table 1).

As it can be seen, the rankings are generally highly correlated. The lower correlation scores for the German language, the most different of the three, due to its compounding, reflect the fact that experiments which did not take this into account suffered a significant drop in performance. Regarding the MAP (highest values 0.35 in 2009, 0.38 in 2010) and NDCG (highest values 0.57 in 2010, 0.54 in 2009) scores we have found that about half of the runs were better in finding documents for English language topics, the other half is better for the German ones, while French topics get the lowest scores for almost all runs. This is due to the various and particular methods that the participating systems involved to treat the multilingual aspect of the CLEF–IP data collection (see [17] Appendix). For CLEF–IP as for other cross-language retrieval tasks, the language issue is not only to differently treat the documents with different languages, but also to care for the language difference between topic documents and relevant documents. To analyse how well this was done, we split the two sets of topics into difficulty bins based on the number of experiments that retrieved their relevant documents. To each relevant document of a topic T is assigned a score equal to the number of

Table 1. Correlations of language based topic subsets with the full topic set in CLEF–IP

	CLEF–IP09				CLEF–IP10			
	MAP		NDCG		MAP		NDCG	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ
EN	0.79	0.88	0.79	0.91	0.93	0.98	0.93	0.97
DE	0.6	0.8	0.64	0.82	0.90	0.96	0.78	0.81
FR	0.69	0.84	0.69	0.87	0.96	0.99	0.78	0.78

experiments that retrieved it. The topic T is in the difficulty bin b_j when the median of its relevant documents scores is j.

Fig. 1 shows the number of topics in each bin, and the percentages of relevant documents having the same document language as the topic document. It is easy to see from the two figures that the fewer relevant documents with a different document language than the topic's, the more systems are able to find them.

Surprisingly, there are no topics in bins b_0 to b_{11} for CLEF–IP 2009, which means that at least 12 runs have found most of the relevant documents for all the 1000 selected topics for this study. One reason behind this is that different type of judgements were provided: in 2009 the relevance judgements were at the patent level, while in 2010 they were given at patent *document* level. The effect of this is that in 2009 more patents were returned as relevant compared to the 2010 results when looking at patents and not patent documents[3]. The second reason behind empty bins is to be found in differences in the origins of the topic document for the two campaign years [18], [16]. The 2009 topic files were artificially created to have a large amount of replicated content in three languages, which made the cross-lingual retrieval problem easier by using monolingual searches.

We note, though, that the margin bins, b_{12} and b_{24} in 2009, and b_1 and b_{18} in 2010 could be joined into their next neighboring bins, as they contain too few topics to draw any believable conclusion.

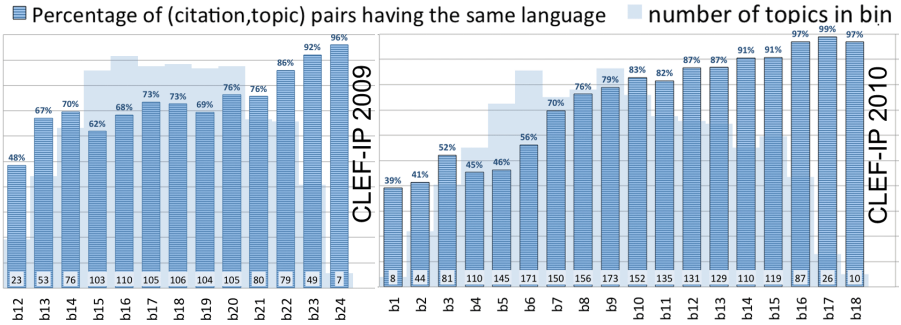

Fig. 1. Language as a sources of difficulty in topics, CLEF–IP 09 and CLEF–IP 10

[3] See section 1.2 for the difference between patents and patent documents.

Lessons Learned. The results we have seen while looking at the effect of the document language on IR scores confirm the research published in the CLIR tracks at TREC[4] or CLEF[5] related papers.

In the case of the CLEF–IP track, the better scores for the English topics are most likely due to the English documents being over-represented in the target collection. This makes it such that even systems that effectively discarded non-English document were able to obtain good scores.

We do remark, though, that the top ranking runs performed better for non-English than for English topics. This is to be attributed to the use of further patent specific data and patent expert know-how in the respective retrieval experiments. This pleads in favor of at least augmenting 'off-the-shelf' retrieval solutions with implementations of patent specific know–how in order to obtain IR systems that better perform in a setting like CLEF–IP.

Participants must be aware that without specific language processing, they will not reach the best scores.

4.3 The 'Document Size' Feature

In general IR, it is common knowledge that a longer narrative is easier to answer and evaluate, therefore systems tend to perform better. We ask whether this is still the case in the patent domain, where longer documents are usually associated with having a verbose, legal document—not necessarily useful for retrieval.

In order to observe potential differences, we have divided the TREC–CHEM and CLEF–IP topics into 10 bins of equal sizes, based on their number of words. In the ascending order of their word count, bin 1 contains 10% of the topics with the fewest words, bin 2 the next 10% more verbose topics, and so on until bin 10 which contains the top 10% most verbose topics. Each bin contains 100 topic documents, with all but one CLEF–IP2010 bins containing 193 topics, the last one containing 200 topics.

Effects on System Rankings. Figures 2 to 5 show the different (average) scores per bins for MAP and NDCG, while Table 2 and shows the correlation figures. With some exceptions in TREC–CHEM 2009, systems maintain their ranks. There is more variation in the correlation results for CLEF–IP 09, and, at a first inspection, we conclude that this is due to the different content type of the two collections. TREC–CHEM contains documents that are more homogeneous regarding their technological content (chemistry) compared to the CLEF–IP collection which contains patents from all technological areas. Participants to TREC–CHEM use retrieval systems tuned for finding chemical documents, while participants to CLEF–IP have to deal with a more general collection. Still, considering the significance values the overall rankings remain unchanged.

Effects on Effectiveness Measures. For TREC–CHEM, we observe in Figures 2 and 3 that MAP and NDCG do tend to be higher for bins of topics with higher

[4] `http://trec.nist.gov`
[5] `http://www.clef.org`

Table 2. Correlations for subsets of topics based on the size of the topics

	TREC–CHEM09				TREC–CHEM10				CLEF–IP09				CLEF–IP10			
	MAP		NDCG		MAP		NDCG		MAP		NDCG		MAP		NDCG	
bin	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
1	0.85	0.96	0.90	0.97	0.83	0.93	1.0	1.0	0.68	0.83	0.69	0.84	0.92	0.98	099	1.0
2	0.83	0.95	0.89	0.98	0.89	0.97	1.0	1.0	0.6	0.75	0.67	0.85	0.99	1.0	0.96	0.99
3	0.94	0.99	0.98	1.0	0.89	0.95	1.0	1.0	0.69	0.84	0.84	0.93	0.99	1.0	0.97	1.0
4	0.94	0.99	1.0	1.0	0.89	0.97	1.0	1.0	0.9	0.92	0.9	0.98	0.89	0.97	1.0	1.0
5	0.98	1.0	1.0	1.0	0.89	0.95	0.94	0.98	0.75	0.88	0.82	0.94	0.91	0.97	0.95	0.98
6	0.94	0.98	0.94	0.99	0.94	0.98	1.0	1.0	0.65	0.8	0.78	0.93	0.93	0.98	0.93	0.99
7	0.94	0.98	0.92	0.98	0.89	0.97	1.0	1.0	0.63	0.79	0.67	0.84	0.84	0.91	0.84	0.91
8	0.89	0.95	0.90	0.98	1.0	1.0	0.89	0.97	0.67	0.84	0.75	0.9	0.71	0.83	0.82	0.91
9	0.81	0.92	0.89	0.97	0.94	0.98	0.83	0.93	0.71	0.84	0.73	0.88	0.69	0.83	0.7	0.79
10	0.71	0.85	0.83	0.93	0.89	0.97	0.83	0.93	0.72	0.88	0.8	0.9	0.83	0.89	0.75	0.79

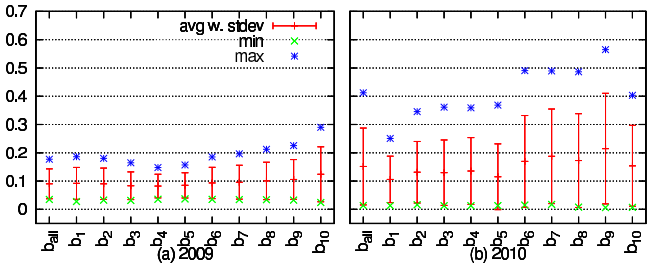

Fig. 2. TREC–CHEM MAP results for topic subsets based on the size of the topics

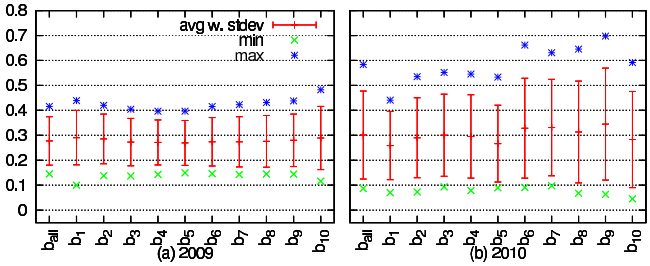

Fig. 3. TREC–CHEM NDCG results for topic subsets based on the size of the topics

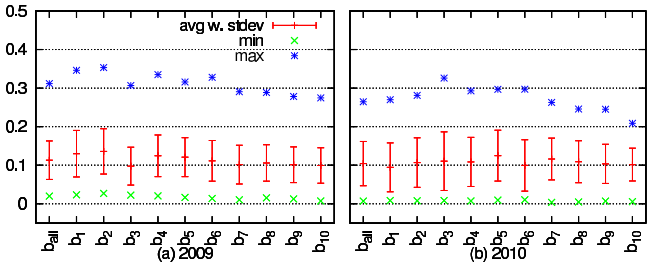

Fig. 4. CLEF–IP MAP results for topic subsets based on the size of the topics

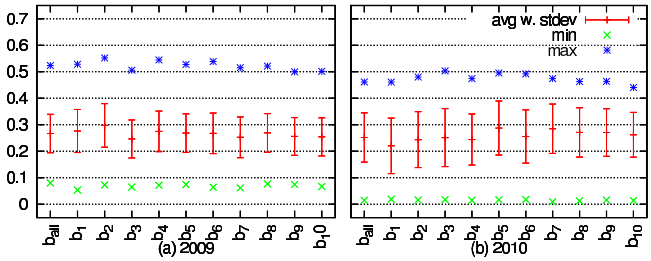

Fig. 5. CLEF–IP NDCG results for topic subsets based on the size of the topics

word counts. This is particularly so for the systems which perform better also on average, as they are able to properly process the input documents and extract the necessary query terms. In 2010, the trend of increasing performance with the increase of the topic size stops at the second last bin (b_9), which is not the case in 2009. This is because the last bin contains documents which are extremely large (the largest document is over 400000 words long—about 400 pages of text) and therefore extremely problematic not only for the IR engines (see for example [12]), but also for the evaluators who created the search reports.

However, overall, the differences observed are too small to draw any link between topic size and effectiveness scores for prior-art search. It appears that in this context the particular content of the request for information (i.e. the patent application text) overweighs the length differences, especially since, compared to standard IR campaigns, the topic file lengths are extreme.

Lessons Learned. The size of the topic document is not an as important factor for this task as it is for information retrieval in general. It is not the case that longer or shorter topics perform better, but rather that extreme topic sizes perform worse. Therefore, such cases should either be handled separately, or methods aware of extreme cases should be developed. Campaign organisers must make sure that all topic sizes are represented in the test collection, not necessarily following any distribution in the corpora.

5 Conclusion

Although it is known that the disclosures made in published patents constitute a large corpus of technological know-how and development, patent data is hardly used in a researcher's work. A main reasons for this is that scientists in the academic communities find it difficult to retrieve data out of patent repositories. Evaluation campaigns that use patents are an important step in bringing this kind of data closer to the them. They incite research about how IR methods perform on this data. The present work contributes to this research focus.

In general terms, this study illustrates how topic feature analysis can be done in the context of a test collection. To this end we have designed a methodology which we apply to revisit some of the observations made in a general IR context, for two patent-based test collections.

In the case of the multilingual collections, we found that while English queries tend to perform, on average, marginally better when considering all runs, the system rankings are sensitive to the use of a different language in the query. Still, patent know-how is a deciding factor in the performance of a system, able to overweight the cross-lingual deficiencies of the system. Another hypothesis, that documents with more text tend to perform better on average, could not be verified. The topic length also did not change the ranking of the systems.

A further analysis of other document characteristics in the patent domain would be useful, and the decision as to which of the many potential facets of the documents should be investigated lies with the organizers of such evaluation campaigns, as a function of the target audience (both in terms of participants and end-users).

References

1. Alonso, O., Mizzaro, S.: Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In: Proc. of SIGIR IR Evaluation Workshop (2009)
2. Ferro, N., Peters, C.: CLEF 2009 Ad Hoc Track Overview: TEL and Persian Tasks. In: Peters et al. [14]
3. Fujii, A.: Enhancing patent retrieval by citation analysis. In: Proc. of SIGIR (2007)
4. Fujii, A., Iwayama, M., Kando, N.: Overview of the Patent Retrieval Task at the NTCIR-6 Workshop. In: Proc. of EVIA (2007)
5. Iwayama, M., Fujii, A., Kando, N., Takano, A.: Report on the patent retrieval task at NTCIR workshop 3. SIGIR Forum 38(1), 22–24 (2004)
6. Kishida, K., Chen, K.-H., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., Myaeng, S.H., Eguchi, K.: Overview of CLIR Task at the Fourth NTCIR Workshop. In: Proc. of the NTCIR Workshop (2004)
7. Kando, N., Leong, M.-K.: Workshop on Patent Retrieval (Workshop Report). SIGIR Forum 34(1) (2000)
8. Kürsten, J., Wilhelm, T., Eibl, M.: The Xtrieval Framework at CLEF 2008: Domain-Specific Track. In: Peters, et al. [13]
9. Larson, R.: Back to Basics - Again - for Domain-Specific Retrieval. In: Peters et al. [13]
10. Lupu, M., Piroi, F., Hanbury, A.: Aspects and analysis of patent test collections. In: Proc. of PaIR (2010)
11. Lupu, M., Piroi, F., Huang, J., Zhu, J., Tait, J.: Overview of the TREC Chemical IR Track. In: Proc. of the 18th Text Retrieval Conference (2010)
12. Lv, Y., Zhai, C.: When documents are very long, BM25 fails! In: Proc. of SIGIR (2011)
13. Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.): CLEF 2008. LNCS, vol. 5706. Springer, Heidelberg (2009)
14. Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.): CLEF 2009. LNCS, vol. 6241. Springer, Heidelberg (2010)
15. Petras, V., Baerisch, S.: The Domain-Specific Track at CLEF 2008. In: Peters et al. [13]

16. Piroi, F.: CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In: CLEF 2010 LABs and Workshops, Notebook Papers (2010)
17. Piroi, F., Zenz, V.: Evaluating Information Retrieval in the Intellectual Property Domain: The CLEF-IP Campaign. In: Current Challenges in Patent Information Retrieval. The Information Retrieval Series, vol. 29 (2011)
18. Roda, G., Tait, J., Piroi, F., Zenz, V.: CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In: Peters et al. [14]
19. Voorhees, E., Buckland, L. (eds.): Proc. of TREC, volume Special Publication 500–278. NIST (2009)