Report on the TREC 2009 Experiments: Chemical IR Track

J. Gobeill [a], D. Teodoro[b], E. Patsche[b], P. Ruch[ab]

[a] *BiTeM group, University of Applied Sciences, Information Studies Department, Geneva*
[b] *BiTeM group, University and Hospitals of Geneva, Geneva*

contact: {julien.gobeill;patrick.ruch}@hesge.ch

Abstract

The goal of the first TREC Chemical track was to retrieve documents relevant to a given patent query, within a large collection of patents in chemistry. Regarding this objective, for the Prior Art subtask, our runs performed significantly better that runs submitted by other participating teams. Baseline retrieval methods achieved relatively poor performances (Mean Average Precision = 0.067). Query expansion, driven my chemical named entity recognition resulted in some modest improvement (+2 to 3%). Filtering based on IPC codes did not result in any significant improvement. A re-ranking strategy, based on claims only improved MAP by about 3%. The most effective gain was obtained by using patent citation patterns. Somehow similar to feed-back but restricted to citations, we used patents cited in the retrieved patents in order to boost the retrieval status value of the baseline run. This strategy led to a remarkable improvement (MAP 0.18, +168 %). Nevertheless, as official topics were sampled from the collection disregarding their creation date, our strategy happened to exploit citations of patents which were patented after the topic itself. From a user perspective, such a setting is questionable. We think that future TREC-CHEM competitions should address this issue by using patents filed as recently as possible.

Introduction

The first TREC Chemical competition provided a large testbed to evaluate, the state of the art in information retrieval for chemistry in a patent repository [1]. The collection consisted of about 1.2 million patents files from the chemical domain, covering patents until 2007. In addition to the patent corpus, the collection contained 59 000 scientific articles. The competition included two sub-tasks. In the Prior Art subtask, the queries were patents sampled from the collection, and participants had to retrieve only relevant patents. No human assessors were used and a retrieved patent was considered as relevant when it was cited in the original query. Patents cited in the original query were to be ignored by the participants. They also needed to be removed from the collection. As patents contain a citations field, the goal was then to rebuild this state of the art.

In the Technology Survey sub-task, the queries were natural language expression of an information need, often dealing with a chemical compound. Participants had to retrieve relevant patents and articles. Human assessors were used to generate relevance judgements by pooling methods as traditionally done in Cranfield-like evaluations.

In 2009, the BiTeM group [2] participated in a similar competition in the Cross-Language Evaluation Forum (CLEF). The Intellectual Property track [3] was similar to the Prior Art subtask used a multi-lingual collection and, which was covering all patent domains, without restriction to chemistry. In the TREC-CHEM 2009 Track, we sometimes exploited results obtained in CLEF in order to select or skip strategies and data; see [4] for more information about our work at CLEF-IP 2009.

Strategies and Methods

As there were more than 1 million patent documents, and as these patent documents were large files, often exceeding several megabytes, the task was firstly to be considered as a very large scale Information Retrieval task. Size reduction was performed not only to make the collection manageable with our tools but mainly to obtain decent tuning in a relatively short time. Three full-

time equivalents worked for about three weeks on the task, including a biologist of our team. The pre-processing was eased thanks to the patent structure, which is well normalized and stable across patents. Of particular interest were the IPC code section and the citation section. IPC codes are keywords assigned to patents, as for instance Medical Subject headings or CAS Registry Number are assigned to Literature articles in the MEDLINE digital library; see TREC Genomics track reports [5].

1) Document Representation

The first step was to decide upon a strategy to merge several patent documents belonging to the same patent into a unique file. We decided to keep all information contained in the different files and to concatenate it in a unique patent file.

The second step was to determine which fields to keep in the indexed patent files. Each patent document was a XML file containing structured data; different fields were delimited by specific tags. Fields that retained our attention were :

- *Title*
- *Description*
- *Abstract*
- *Claims*
- *Applicants*
- *Inventors*
- *IPC codes*
- *Patent references*

In our works during the CLEF-IP 2009 campaign, we evaluated several document representations. Our retained document representation used *title*, *abstract*, *claims* and *IPC codes* fields. We also used these fields for document representation in TREC-CHEM, adding *inventors* and *applicants* fields. Despite our efforts, we never were able to take benefit from the *description* field in CLEF-IP, which is often a very huge field, and we finally discarded it. Thus, we arbitrarily decided to discard it for TREC-CHEM as well. Moreover, we used *IPC codes* in two different formats: 4-digits codes (e.g. D21H) and complete codes (e.g. D21H 27/00). *Citations* were not used for building the patent representation, but were investigated for post processing purposes.

Concerning patent representation for patent queries, we decided to keep the *Description* field. Concerning articles representation, we simply used the *abstract* and *title* fields, considering that information density in abstracts is significantly higher than in full-text.

2) Indexing and Retrieval Model

Runs were generated with Terrier [6]. During the CLEF-IP campaign, we evaluated a wide range of weighting schemas and query expansion strategy with a similar patent collection. The same weighting model was applied using generally good parameters, which were not returned for the track [7].

3) Exploiting Citations Network

We explored post-processing strategies dealing with patent citations. Few studies addressed the co-citations issue in the patent domain. Li and al. [8] used citations information in order to design a citation graph kernel; evaluating their work with a retrieval task, they obtained better results exploiting citation network rather than only direct citations. Directly related to the prior art task, Tbahriti et al. [9] used citations network to automatically acquire relevance judgements in a subset of MEDLINE dealing with peptides.

We extracted the patent references of all patents contained in the TREC-CHEM collection and computed the citations network. The combination with the baseline was fairly simple: for each retrieved patent, we boosted the score of its referenced patents, when found in the collection. For a given patent i contained in the collection, its final score $Final\_Score_i$ is computing by adding its initial Retrieval Status Value (IR_Score_i) and a fraction of the RSV of all patents j citing the patent i. The IR_Score_i equals 0 if the patent is not in the top J first retrieved patents.

$$Final\_Score_i = IR\_Score_i + \sum_J is\_cited_{i,j} \times \alpha \times IR\_Score_j$$

Formula 1. Co-citations factor.

The empirical value for the constant α was 0.1 in CLEF-IP, which was re-used for TREC.

4) Query Expansion using chemical annotation

Patents belonging to the TREC-CHEM collection belong to the chemical domain. Given the highly specialized language in this filed, we explored a Query Expansion strategy (Figure 1), based on a named-entity recognizer for chemical compounds. First, we used the Oscar3 tool [10], an open-source chemistry analysis tool for chemical annotation, in order to detect entity boundaries (e.g. *C2H5*). Second, we normalized the identified entities using the MeSH categorizer [11], powered with supplementary concepts from the UMLS, PubChem, chEBI and DrugBank. The objective was to attribute an unambiguous identifier to each chemical entity. Finally, we queried PubChem [12] with the

MeSH identifier or with the PubChem term when the normalization was unsuccessful. The PubChem database returned a set of compounds, corresponding to the given entity or its children. If the set of compounds was too large (e.g. *hydrocarbon* covers 63 compounds), the query expansion was given up.

5) Filtering based on IPC codes

In an expert patent searching context, Stemitzke [13] assumed in his abstract that "*patent searches in the same 4-digits IPC class as the original invention reveal the majority of all relevant prior art in patent*". Another study assumed that it is between 65% and 72% – whether citations were added by the applicant of the examiner – of European patent citations that are in the same technology class [14]. Moreover, dealing with various IPC granularities – whether 4-digits or complete codes – used in patent searches, the EPO best practices guidelines indicate that "*for national searches […] the core level is usually sufficient*" [15].

Therefore, we decided to explore IPC filtering strategies that consisted in downweighting retrieved patents that did not share any IPC code with the query. We evaluated this strategy for both 4-digits and complete codes.

6) Re-ranking based on claims

A more advanced strategy that we wanted to evaluate in TREC-CHEM 2009 was inspired by the works done during the patent tracks of the NTCIR campaigns [16]. In particular, Mase and al. [17] proposed to re-arrange a classic Information Retrieval run by considering only claims. Thus, for each query, we re-indexed only the claims of the 1000 retrieved patents, and then re-computing a run with only the claims of the query. A linear combination, using a β constant was performed in order to merge both runs into a single one.

Results and Discussion

We submitted 6 official runs, but some of the following results were obtained after the competition, using the gold file provided by the organizers.

1) Document Representation

Table 1 shows how much each field contributed to the final performance of the baseline run, which was officially submitted for the PA sub-task. Document representation for the official baseline run (*BiTeM09PAbl*) was computed using the *Title*, *Abstract*, *Claims*, *Inventors* and *IPC codes* fields.

Discarded field	MAP	Improvement
Official baseline run	0.067	
Inventors	0.062	+ 7%
Applicants	0.067	+ 0%
Claims	0.057	+ 15 %

Table 1. Results for Documents Representation (Mean Average Precision).

Using *Claims* led to a + 86% in CLEF-IP [4], while it only led to a +15 % in TREC-CHEM. As CLEF-IP patents were dealing with general domains, it could be a domain-specific feature, applying only to chemistry. Interestingly, *Inventors* are more content-bearing than *Applicants* for this particular field.

2) Indexing and Retrieval Model

The collection was simply indexed with Okapi BM25. No specific tuning was performed.

3) Exploiting Citations Network

An official run (*BiTeM09PAcit*) was submitted using the Citation network strategy with α set to 0.1. Computed from the baseline run (MAP 0.067), it led to an impressive improvement with MAP 0.1798 (+ 168 %). Similar strategy in the CLEF-IP campaign resulted in a + 3% improvement.

This reranking strategy was applied also on patents that were posterior to the patent query, which is by no mean a realistic task model for prior art search. Discarding patents filed after the topic, resulted in a MAP of 0.148. When fine tuning α, a maximal MAP of 0.158 was obtained.

4) Query Expansion using chemical annotation

An official run (*BiTeM09PAqe*) was submitted using the Query Expansion strategy. The MAP increased from 0.179 to 0.182 (+ 2%). Although modest, such a strategy is regarded as promising considering that tuning data were very sparse. It is thus expected fine-tuned terminology-driven expansion could help improve recall.

5) Filtering based on IPC codes

No official runs were submitted using filtering based on IPC codes. In experiments made after the competition, we applied both 4-digits codes and complete codes strategy to the previous run: from a MAP of 0.18, both strategies led respectively to a degradation of 3% and 7%, while similar strategies led respectively to an improvement of 5% and 11% in the CLEP-IP campaign [4]. The explanation of such a phenomenon is unclerar, but as the TREC-CHEM collection is focused on chemistry, the retrieved patents already deal with the same domain and share a smaller range of IPC codes.

6) Re-ranking based on claims

Finally, we submitted two official runs with the *re-ranking based on claims* strategy, with β set to 0.1 and 0.3 (*BiTeM09PAcba* and *BiTeM09PAcbb* runs). Official runs and further experiments showed an improvement of about + 3% when using this strategy.

Discussion and Conclusion

Beyond the competitive results obtained by our team, our performance needs to be discussed. The pure retrieval step led to a relatively weak baseline. *Query Expansion using chemical annotation and normalization*as well as *re-ranking based on claims* brought small – yet probably statistically significant – improvements. More important, the effectiveness of the *Citation network* strategy is clearly a (positive) surprise. However a part of the positive effect may be an artifact caused by the set of topics chosen for the competition.

Finally, some retrieved patents can share up to 224 relevant citations with the query, because they had been applied by the same inventors, with a nearly similar state of the art. Patents visibility depends on several dates contained in the file, and is not easily understandable for a non Intellectual Property expert. Last but not least, user requirements can vary depending on whether the user is an applicant submitted a new patent, an applicant that has already applied similar patents, a patent officer, or even an opponent. Finally, working with EPO and USPTO also leads to ambiguous results, as some patent queries may have their equivalent in the other database, i.e. when a patent has been filed in both patent libraries.

The TREC-CHEM 2009 campaign was a good starting point, providing a large collection of structured data. Yet, we think that future TREC-CHEM campaigns need to carefully define realistic task models. Thus, we would like to suggest the following: 1. define clearer date filtering rules (maybe exluding queries from the collection); provide topics as recent as possible with respect to the collection.

Acknowledgments

The study reported in this paper has been supported by the European Commission Seventh Framework Programme (DebugIT project grant no. FP7-ICT 217139).

References

[1] TREC-CHEM 2009 Track Guidelines.

[2] http://eagl.unige.ch/bitem/

[3] http://www.ir-facility.org/research/evaluation/clef-ip-09/overview

[4] J Gobeill, D Theodoro and P Ruch, “Exploring a wide Range of simple Pre and Post Processing Strategies for Patent Searching in CLEF IP 2009”, CLEF 2009 Working Notes.

[5] P Ruch, A Jimeno Yepes, F Ehrler, J Gobeill and I Tbahriti, “Report on the TREC 2006 Experiment: Genomics Track”, TREC 2006 Working Notes.

[6] Ounis I., Lioma C., Macdonald C. and Plachouras V.. “Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web”, Novatica/UPGRADE Special Issue on Next Generation Web Search, vol 8, pp 49-56, 2007.

[7] http://ir.dcs.gla.ac.uk/terrier/doc/configure_retrieval.html

[8] Li X., Chen H, Zhang Z. and Li J., “Automatic patent classification using citation network information: an experimental study in nanotechnology”, Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp 419-427, 2007.

[9] I Tbahriti, C Chichester, F Lisacek and P Ruch, “Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the MEDLINE digital library”, I. J. Medical Informatics vel. 75, pp 488-495, 2006

[10]P Corbett and P Murray-Rust, “High-Throughput Identification of Chemistry in Life Science Texts”, CompLife 2006, LNBI 4216, pp. 107 – 118, 2006.

[11]http://eagl.unige.ch/EAGL/

[12] Y Wang, J Xiao, TO Suzek, J Zhang, J Wang and SH Bryant, “PubChem: a public information system for analyzing bioactivities of small molecules", Nucleic Acids Res, 2009.

[13] Sternitzke C, “Reducing uncertainty in the patent application procedure – insights from malicious prior art in European patent applications”, World patent Information, vol.31, pp 48-53, 2009.

[14] Criscuolo P and Verspagen B, “Does it matter where patent citations come from? Inventor versus examiner citations in European patents”, Research Policy, vol.37, pp 1892-1908, 2008.

[15] http://www.epo.org/patents/patent-information/ipc-reform/faq/levels.html

[16] http://research.nii.ac.jp/ntcir

[17] H Mase and al., “Two-Stage Patent Retrieval Method Considering Claims Structure”, Proceedings of NTCIR-4, Tokyo, 2004.