BiTeM group report for TREC Chemical IR Track 2011

J. Gobeill [a], A. Gaudinat [a], E. Pasche[b], D. Teodoro[b], D. Vishnyakova[b], P. Ruch[a]

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

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

Abstract

For the third year, the BiTeM group participated in the TREC Chemical IR Track. For this campaign, we applied strategies that already showed their effectiveness, as the Citations Feedback, which takes benefit from the citations of the retrieved documents in order to re-arrange the ranking. But we also investigated a new inter-lingua model built with chemical annotations with concepts that we automatically mapped into documents. We used the MeSH controlled vocabulary for this purpose. For the Technology Survey task, the fusion of the MeSH and Text models led to a remarkable improvement (+71% for MAP) compared to the Text model alone. The most interesting aspect is that both models are highly complementary as the MeSH model brings 70% of new relevant documents that were not retrieved by the Text model. For the Prior Art task, we showed that there exist patterns of chemical patents that are interconnected (i.e. linked together with direct citations) and that are more likely to be present together in a prior art. Such patterns are efficiently retrieved with our Citations Feedback strategy. On the other hand, we pointed out that the less the prior art of a given topic is interconnected, the less efficient is the Information Retrieval. We hypothesize that such patents have a larger technical focus, maybe represented by a larger set of IPC codes, and then have a lower textual similarity with their prior art documents. These topics should gain to be recognized in order to be treated with complementary techniques.

Introduction

For the third year, the BiTeM group [1] participated in the TREC Chemical IR Track. This competitions aims at providing benchmarks to evaluate, in a realistic scenario, the state of the art in chemical information retrieval and extraction tools [2]. In 2011, organizers maintained the Prior Art task – that is, from a patent given as topic, rebuilding the state of the art – and the Technology Survey task – an ad hoc search task in a collection of patents and journal articles. This year there was an Image-to-structure task, but we didn't participated.

For this campaign, we chose to apply strategies that achieved very good results in previous years [3], especially our Citations Feedback strategy. As a novelty, we chose to complement our Text model with a second inter-lingua model, built with semantic descriptors that were automatically mapped in documents. Such semantic annotation had showed to be efficient in order to enhance the performance of Information Retrieval [4,5]. We chose to use the Medical Subject Headings (MeSH) because it contains more 200'000 chemical compounds and is well designed for automatic mapping.

Compared to the two last campaigns, organizers kept the same corpus, but made some adjustments for topics. For the PA task, they chose to limit topics to application documents and no more to granted patents: this is – they claimed – a lesson learned from last campaign [2]. Such an adjustment could have impact in our Citations Feedback strategy. For the TS task, organizers chose to focus on biochemistry.

Data

The corpus was the same as last year. For the Prior Art task, here was 1.3M of patent files from the chemical domain in the collection. 180'000 scientific articles were added for the Technology Survey task.

The only differences were for topics. As mentioned in the introduction, the PA topics were no more granted patents. For the TS task, topics were biochemistry-oriented.

Strategies

Except for the MeSH model, all the strategies were already deployed in the previous campaigns, and thus will not be fully described below. Please report to [3] for longer descriptions.

1) Text model

The Text model remained the core of our system. Only titles, abstracts and claims were indexed. Citations that pointed to a patent in the collection were extracted in order to perform the Citations Feedback strategy. Other metadata were discarded. The collection was indexed using the Terrier platform [6]. We applied no stemming.

2) MeSH model

For this campaign, we deployed a new strategy based on a medical controlled vocabulary: Medical Subject Headings (MeSH). MeSH contains 26'000 general concepts, and 200'000 supplementary concepts that are principally chemical compounds. Moreover, the MeSH contains more than 1.5M of synonyms.

MeSH concepts were automatically mapped into documents. We chose to apply naïve word matching [7] as this technology needs no learning data, is not time-consuming, shows a good precision, and do not need a threshold [8]. Thanks to their semantic type, we distinguished MeSH concepts dealing with chemistry from those dealing with general domain and applied a different weight to them.

On average, 35 general and 61 chemical MeSH concepts were mapped into a patent. We apply the same process for articles for the TS task. We then built a second index – in such an inter-lingua – and performed the same process on topics.

We then were able to normalize the chemical compounds. For example, the topic PA-9 dealt with "1,4-butanediol" which is the MeSH concept C039681. 4'498 patents contained "1,4-butanediol", but 294 other patents contained "1,4-butylene glycol" which is given as a synonym in MeSH. All these forms of the same chemical compounds were normalized into "C039681".

For the PA task, we submitted one run computed with the Text model, and one run computed with the MeSH model. Then, we applied a linear combination in order to make an official Combo run. Based on the 2010 test set, we chose to add 10% of the MeSH run scores to the Text run scores. For the TS task, we chose to merge both representations in a unique model and then make a Fusion run.

3) Citations Feedback strategy

In previous years, the most powerful strategy we applied in the Prior Art task was Citations Feedback [2]. This strategy consists on re-ranking a run by exploiting the citations of the retrieved patents. A fourth run for the PA task was then computed from the Combo run and was submitted.

Results & Discussion

1) Official results

a) Prior Art task

Run	MRR	MAP
Text	0.415	0.059
MeSH	0.247	0.030
Combo	0.409	0.059
Combo+CitFB	**0.427**	**0.082**

Table 1 : official results for the PA task.

The performance of the Text model (MAP 0.059) remains quite low, and is between performances obtained in 2009 and 2010 (0.043 in 2010, 0.067 in 2009). As the Text model remained the same, these differences are strictly due to the test set.

The new MeSH model that we introduced this year did not perform as well as the Text model. Their combination raised the MAP to an equivalent level. Yet, we will detail in part 4 how the Combination model has a different coverage from the Text model.

Finally, our Citations Feedback strategy achieved the best results. However, while it led to a +507% improvement for MAP in 2010, it only led to a +72% improvement this year. We will try to explain this underachievement in part 3.

The impact of the publication and application dates filtering strategy is the highest: for the Combo+CitFB run, MAP reached from 0.038 without the filtering to 0.082 (+118%).

b) Technology Survey task

Run	NDCG	MAP
Text	0.253	0.063
MeSH	0.275	0.089
Fusion	**0.327**	**0.108**

Table 2: official results for the TS task.

Due to the new biochemical orientation of this task, it is difficult to compare the results with the 2010 results which were very low (MAP 0.011). The most interesting point lies in the MeSH model's

performance (MAP 0.089) which is higher than Text (MAP 0.063), and their Fusion that was still better (MAP 0.108, +71%). Contrary to the linear combination used for the PA task, the models' fusion achieved to take benefit from both models. Once again we will see in Part 4 how complementary they are.

2) Weighting of MeSH descriptors

The results presented in Table 3 were computed after the competition with the official gold file of the Prior Art task. This setting aims at finding the good weight for general and chemical MeSH descriptors in the MeSH model.

General descriptors weight	Chemical descriptors weight	MAP
x	1	0.021
1	1	0.023
1	**3**	**0.030**
1	5	0.025

Table 3: best weighting for general and chemical MeSH descriptors for the MeSH model in the PA task.

These results are consistent with what the setting we did on last year's PA test set. The power of the MeSH model is obviously more in chemical descriptors, but the general descriptors – even for the PA task which was not biochemistry oriented – are useful.

Other complementary results are : the MeSH model is less performant when concepts are mapped in full text rather than abstract and claims (-3% for MAP). And the MeSH model is less performant when the concepts are expressed in text rather than in identifiers (-15% for MAP).

3) Patents interconnection

Last year, we introduced the notion of *interconnection* for Prior Art. The interconnection is, in the set of patents belonging to a topic's prior art, the rate of patents that are connected to at least one other patent (i.e. there exists a direct citation between both patents).

Interconnection in chemical patents is the cause of the high impact of our Citations Feedback strategy. However, in 2011, this impact was lower than last year. This may be the effect of the new orientation of the PA task (i.e. not choosing granted patents as topics). A deeper analysis of the gold file reveals salient differences between 2010 and 2011 topics. If we focus in relevant patents that are highly connected (i.e. that are connected to more than 50% of the other relevant patents for this topic): there were 410 highly connected patents in the 2011 test set instead of 345 in 2010. Yet, the Text model only retrieves 30% of them in 2011, instead of 74% in 2010. It means that in 2011 there still were highly connected patents, but they were less textually similar to the topics than in 2010. This fact is hard to interpret except by considering it is due to the new orientation of the test set.

Other interesting facts are revealed focusing on interconnection. Table 4 shows different measures when the test set is split into four equal parts of 250 patents according to their interconnection rate.

	1st quarter	2nd quarter	3rd quarter	4th quarter
Average interconnection	**90%**	**69%**	**50%**	**23%**
Average relevant patents in prior art	32	29	27	23
different words in prior art	1510	1450	1440	1510
shared words in prior art	**48%**	48%	46%	**41%**
Average topic length	14'500	13'500	14'100	15'800
Average IPC codes in topic	11.6	11.5	12.5	**16.3**
MAP for the Combo model	**0.083**	**0.075**	**0.047**	**0.031**
MAP after citFB	0.148 **(+78%)**	0.097 **(+69%)**	0.047 **(+50%)**	0.031 **(-2.3%)**

Table 4: different measures for a partition of the test set according to interconnection. Interesting results are in bold.

Table 4 shows that the performance of the textual Information Retrieval (Text MAP) is highly correlated with the interconnection. We make this intuitive hypothesis: a prior art that is highly interconnected contains patents that are more likely to focus on the same technical domains, and then are more likely to have a higher textual similarity between us (from 41% of shared words to 48%), and with the topic. On the other hand, it is difficult to find features for predicting that a topic's focus is large, and then its prior art is lowly interconnected and harder to retrieve with textual similarity: this could lead to recognize such a topic and to apply a different treatment. A high number of IPC codes in topics (16.3 for the fourth quarter) could be an indicator, but more experiments are needed.

Obviously, the performance of the Citations Feedback is higher with highly interconnected prior art (+78% for

the first quarter) than with lowly interconnected; yet, it does not decrease the performance too much (-2.3% for the fourth quarter).

4) Complementarity of both models

At last, we analysed, for both tasks, the coverage of both Text and MeSH model. That is: how many relevant documents were retrieved by the Text model only, how many by the MeSH model only, and how many by both. Figure 1 gives a graphical view of this coverage.

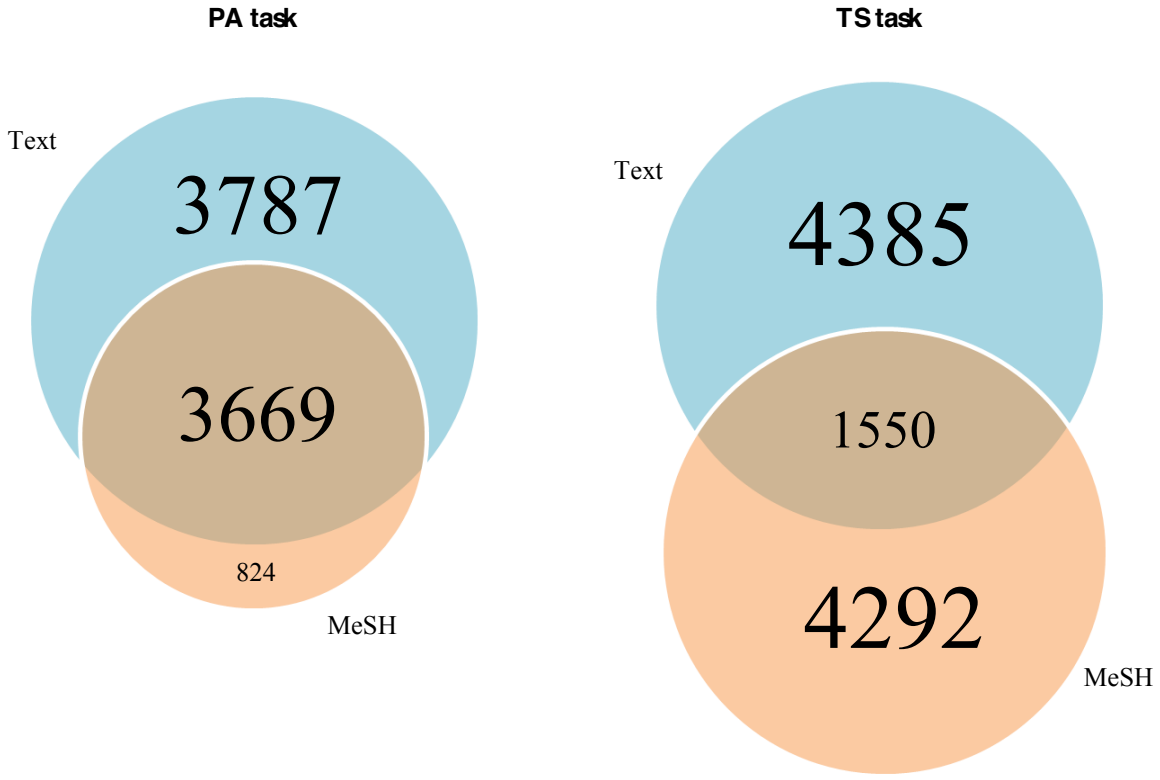

Figure 1 : complementary of the Text and MeSH models for the PA and the TS tasks. Blue discs stand for the set of relevant documents retrieved by text, orange circles stand for the set of relevant documents retrieved by MeSH, the numbers stand for the cardinals of each section (Text only, MeSH only, and intersection).

For the PA task, both models are not highly complementary, as the set of the relevant documents retrieved by the MeSH model is 60% smaller and nearly (80%) included in the set of relevant documents retrieved by the Text model. The complementarity is more interesting for the TS task that was clearly biochemical-oriented and then much designed for the MeSH vocabulary: both sets are equivalent and the intersection is small. The MeSH model actually retrieved 4292 new relevant documents (+70%). This is obviously of great help for users focusing on recall.

Conclusion

From this 2011 TREC Chemical IR Track, we will retain two main conclusions.

First of all, the interest of representing the documents with a chemical terminology (the MeSH) was proved. For the Technology Survey task that was especially biochemical-oriented, the fusion of the Text and the MeSH models led to a 71% better performance. Moreover, the MeSH model showed complementarity with the Text model, as the MeSH model retrieved 70% of new relevant documents compared to the Text model.

Second, we pointed out a high disparity in terms of interconnection in patents' prior arts. Partition of the Prior Art test set showed that, for a given patent topic, the less interconnected is the prior art, the less efficient is the Information Retrieval, and then the less efficient is our Citations Feedback strategy. We make the hypothesis that patents whose prior art is lowly interconnected deal with a larger technical focus, and

thus the textual similarity is lower in their prior art, making the Information Retrieval less efficient. Such patents need a different or complementary treatment based on metadata. Unfortunately, there are few hypotheses about how to recognize them.

Acknowledgments

The study reported in this paper has been partially supported by the European Commission Seventh Framework Program (DebugIT project grant no. FP7-ICT 217139 and Khreshmoi project grant no. FP7-ICT-2009-5).

References

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

[2] TREC Chemical IR Track 2011 Guidelines

[3] J Gobeill, A Gaudinat, E Pasche, D Teodoro, D Vyshnyakova and P Ruch, "BiTeM site report for TREC Chemistry 2010: Impact of Citations Feedback for Patent Prior Art Search and Chemical Compounds Expansion for Ad Hoc Retrieval" in TREC 2010.

[4] J Gobeill, D Teodoro, E Pasche and P Ruch. "Taking Benefit of Query and Document Expansion using MeSH descriptors in medical imageclef 2009". In Proceedings of CLEF 2009.

[5] D Trieschnigg, P Pezik, V Lee et al. "MeSH Up: effective MeSH text classification for improved document retrieval." in Bioinformatics 2009;25:1412–18.

[6] I Ounis, C Lioma, C Macdonald and V Plachouras, "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] Y Yang and JO Pedersen, "A comparative study on feature selection in text categorization", in Proc. 14th International Conference on Machine Learning, 412--420, 1997.

[8] J Gobeill, "Modèles de Question / Réponse pour la Biomédecine", PHD Thesis, University of Geneva.