Full-texts representation with Medical Subject Headings, and co-citations network reranking strategies for TREC 2014 Clinical Decision Support Track

J. Gobeill [ab], A. Gaudinat[a], E. Pasche[c], P. Ruch[ab]

[a] *BiTeM group, University of Applied Sciences, Information Studies Department, Geneva*
[b] *SIBtex group, Swiss Institute of Bioinformatics, Geneva*
[c] *BiTeM group, University and Hospitals of Geneva, Geneva*

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

Abstract

In TREC 2014 Clinical Decision Support Track, the task was to retrieve full-texts relevant for answering generic clinical questions about medical records. For this purpose, we investigated a large range of strategies in the five runs we officially submitted. Concerning Information Retrieval (IR), we tested two different indexing levels: documents or sections. Section indexing was clearly below (-40% in R-Precision). In the domain of Information Extraction, we enriched documents with Medical Subject Headings concepts that were collected from MEDLINE or extracted in the text with exact match strategies. We also investigated a target-specific semantic enrichment: MeSH terms representing diagnosis, treatments or tests (relying on UMLS semantic types) were used both in collection and in queries to guide the retrieval. Unfortunately, the MeSH representation was not as complementary with the text as we expected, and the results were disappointing. Concerning post-processing strategies, we tested the boosting of specific articles types (e.g. review articles, case reports), but the IR process already tended to favour these article types. Finally, we applied a reranking strategy relying on the co-citations network, thanks to normalized references provided in the corpus. This last strategy led to a slight improvement (+5%).

Introduction

The Bibliomics and Text Mining group (BiTeM) in Geneva has a long history of participation in TREC campaigns, including TREC Genomics [1], TREC Medical Records [2] or TREC Chemical IR Tracks [3]. In parallel, the group has recently joined the Swiss Institute of Bioinformatics. Additionally, the group is currently involved in several translational medicine research project, including the MD-Paedigree project (EU FP7 Programme), where his task is to help clinicians to retrieve similar cases in a federated digital repository gathering data from 7 European clinical centres, for better personalised predictive medicine. The focus of the 2014 Clinical Decision Support Track was the retrieval of biomedical articles relevant for answering generic clinical questions about medical records [4]. This track provided a rare opportunity to investigate several approaches for linking medical cases to information relevant for patient care.

Indeed, a large range of strategies were implemented in the five runs we submitted. Concerning Information Retrieval (IR), we tested two different indexing levels: documents or sections. In the domain of Information Extraction, we enriched documents with Medical Subject Headings (MeSH) terms that were collected from MEDLINE or found in the text with exact match strategies. Depending on the runs, these metadata were added to the document representation, or exploited in a parallel index. We also investigated a target-specific semantic enrichment: MeSH terms representing diagnosis, treatments or tests (relying on UMLS semantic types) were used both in collection and in queries to guide the retrieval. Concerning post-processing strategies, we tested the boosting of specific articles types (e.g. review articles, case reports). Finally, we applied a reranking strategy relying on a co-citations network, thanks to normalized references provided in the corpus: articles that were cited by the top retrieved documents were added or boosted in the last run.

Data and strategies

We officially submitted five runs. As we had no training nor tuning data, all strategies were applied with a priori and intuitive settings. When results and qrel (gold file) were made available (what we call "in post-competition"), we also evaluated supplementary runs with different settings in order to have a better idea on the optimal performances of our strategies. Obviously, due to the nature of the gold file (pooling judgments), when we compute a post-competition run, we have to keep in mind that it can be under-evaluated, as it can contain a larger part of non-judged documents than official runs. Yet, improvements and comparisons remain valid.

All retrievals were generated with Terrier [5], an IR platform (in Java) which implements state-of-the-art indexing and retrieval functionalities, including a TREC format output. We used Terrier with a classical Okapi BM25 weighting scheme, with default settings, and an automatic relevance feedback query expansion: see [6] for more details about Terrier models. In the following, the *Retrieval Status Value* (RSV) is the relevance score attributed to a document by Terrier. Post-processing strategies were applied thanks to local scripts in Perl.

1) Full text indexing

Here are some statistics computed prior to the design of the full text search engine. The collection contained 733,328 documents from PubMed Central. Documents were in nxml format, and could be composed of an abstract, and/or different full-text sections. Most of them (79%) had an abstract and other sections, while 14% had sections but no abstract, and 7% had an abstract but no section. Documents had on average 11 sections, while a section contained on average 360 words, versus 160 for an abstract.

Document or section indexing. We decided to test two indexing units: document or section. For document indexing, all sections were merged into a unique representation. For section indexing, each section was indexed; then, in the output of the search engine, when multiple sections from the same document appeared in the ranking, the RSV of the document was the RSV of its first retrieved section.

Query representation. In the official test set (30 queries), each query contained a full description of the information need, and a summary. For the official runs, we only used the summary part, but in post-competition we also evaluated retrieval with full queries. For all experiments, we removed numbers from the queries.

Boosting based on article types. Still for full text indexing, we also investigated a boosting strategy depending on article types. Our initial hypothesis was that review articles and case reports were more likely to be relevant for clinical decision support, thus we decided to apply a +20% boost on RSV for these article types. This +20% boost was applied to all official runs, but in post-competition we evaluated different boosting values.

2) MeSH recognition and indexing

In parallel with full text indexing, we also investigated MeSH recognition and indexing. MeSH indexing can offer a complementary representation in some retrieval tasks, such as with clinical captions in previous CLEF evaluation campaigns [7]. MeSH concepts were recognised in documents using a classical strict mapping (Rabin-Karp algorithm). See [8] for further details and evaluation of MeSH recognition by Rabin-Karp algorithm. MeSH concepts manually assigned by human indexers also could be collected when the PubMed Central document had a corresponding citation in MEDLINE. Thus, each document (or section) could be represented and indexed with its MeSH concepts. Then, the same extraction was performed with queries.

MeSH concepts for document representation. We extracted an average of 422 MeSH concepts per document. Dealing with MeSH concepts collected from MEDLINE, 92% of the documents in the collection had an associated PMID, and only 53% had MeSH terms assignments in MEDLINE (usually around 10 MeSH concepts). Thus, for official submissions, we had the choice between building one index for text and one for MeSH, and then combining the rankings, or building a unique entity for each document, merging text and MeSH terms.

MeSH concepts for query representation. The same extraction was performed for queries. The Figure 1 shows the MeSH concepts that were extracted from summaries of queries 1,11 and 21.

```
<topic number="1" type="diagnosis">
 <summary>58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.</summary>
 <MeSH_in_summary>Women ; Hypertension ; Obesity ; Exercise ; Thorax ; Pain ; Back ; Chest Pain ; Pain</MeSH_in_summary>
 </topic>
<topic number="11" type="test">
```

<summary>40-year-old woman with severe right arm pain and hypotension. She has no history of trauma and right arm exam reveals no significant findings.</summary>
<MeSH_in_summary>Women ; Arm ; Pain ; Hypotension ; History ; Wounds and Injuries ; Arm</MeSH_in_summary>
</topic>
<topic number="21" type="treatment">
<summary>21-year-old female with progressive arthralgias, fatigue, and butterfly-shaped facial rash. Labs are significant for positive ANA and anti-double-stranded DNA, as well as proteinuria and RBC casts.</summary>
<MeSH_in_summary>Female ; Arthralgia ; Fatigue ; Butterflies ; Exanthema ; DNA ; Proteinuria ; DNA</MeSH_in_summary>
</topic>

Figure 1. Test set in the official format, with MeSH concepts that were extracted in the summary.

MeSHtargets. Queries dealt with one of these three categories: diagnosis, tests or treatments. A particularly promising investigated strategy was to identify in documents the extracted MeSH concepts that belonged to the corresponding category, and to over-weight them. For instance, for queries dealing with tests, documents that have a lot of MeSH concepts related to tests should be favoured. Thanks to the UMLS Semantic Types [9], we designed sets of Semantics Types for each category. For instance, for tests, we selected T060 Diagnostic Procedure and T059 Laboratory Procedure. Thus, for each document, for each MeSH concept belonging to the test category, we added the word *MeSHtargetTest* in the document representation. There was an average of 14 *MeSHtargetTest* in documents, versus 37 for *MeSHtargetDiagnosis* and 22 *MeSHtargetTreatment*. The same MeSHtargets were used in queries. For instance, for queries dealing with tests, we added *MeSHtargetTest* three times in the query representation.

3) Co-citations network

At last, we explored post-processing strategies dealing with co-citations. The idea was to start from a ranking, and then to promote the citations of the top retrieved documents. This strategy achieved leading results with patents (see TREC Chem campaigns [3], with up to +150% for MAP), but it was the first time we applied this to the medical literature. Formula 1 gives the final score of a document *d* after re-ranking. *E* is the set of retrieved documents (1000 by default), $is\_cited_{d,e}$ is 1 if document *d* is cited in document *e*, 0 otherwise. α is a setting variable.

$$Score_d = RSV_d + \sum_E is\_cited_{d,e} \times \alpha \times RSV_e$$

Formula 1. Co-citations network boosting.

In simple words, this reranking consists in scanning the retrieval ranking, and for each document *e* and its RSV_e, adding $\alpha \times RSV_e$ to its citations. This means that documents that were not retrieved by IR can appear if they are cited by most top retrieved documents.

Results and Discussion

In the following, we describe results in light of R-Prec. Figure 2 contains different official R-Prec of TREC CDS 2014. BiTeM runs from 1 to 5 investigated different strategies that are discussed in the following.

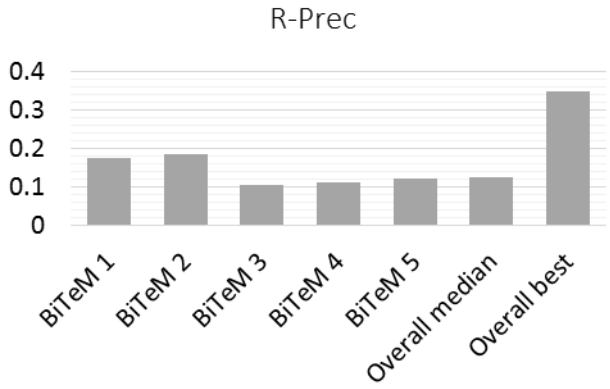

Figure 2. Different official R-Prec values, for the BiTeM runs and the other teams' runs (median and best).

1) Full text indexing

All the submitted runs were computed with both text and MeSH indexes, but in post-competition we investigated text-only indexes.

Document or section indexing. Official runs 1 and 2 were computed with document indexing, while official runs 3 et 4 were with section indexing. The official run 5 was supposed to be our optimal run, and was computed from the run 4: this illustrates how we thought that section indexing would have better results. Unfortunately, runs 1 and 2 were much better than runs 3 and 4. In particular, runs 2 and 4 only differed on the indexing, and run 2 had a R-Prec of 0.187, versus 0.114 for run 4 (-39%). In post-competition, no further experiments were done with section indexing.

Query representation. All submitted runs were computed using only the summary part of the queries. In post-competition, we compared the value of description and summary, evaluated with the document indexing, without automatic query expansion. In terms of R-Prec, descriptions obtained 0.169, summaries obtained 0.170, while a query representation with both fields obtained

0.185 (+9%). With automatic query expansion in Terrier, R-Prec reached 0.211 (+14%).

Boosting based on article types. In post-competition, we analysed the qrel in order to find which article types were overrepresented in the qrel compared with the collection, i.e. which article types are more likely to be relevant for this task.

Article type	Distribution		
	in qrel	in collection	in our runs
research-article	52.2%	74.3 %	37.9 %
case-report	20.4 %	4.0 %	41.5 %
review-article	17.9 %	6.9 %	10.9 %
Other	3.2 %	2.6 %	3.6 %
brief-report	1.5 %	1.1 %	0.9 %

Table 1. Distribution of article types in qrel (only relevant documents), in collection, and in one of our runs.

Our intuition was good, as review-articles and case-reports are much more represented in qrel compared to the collection. For all our official runs, we applied a +20% boosting for RSV for these article types. Unfortunately, post-competition experiments did not confirm the effectiveness of this strategy. Starting from the previous post-competition run (R-Prec 0.211), the +20% boosting degraded the run (R-Prec 0.195, -8%). Actually, no tested value for boosting led to better results. Table 1 shows the distribution in our run, and it seems that the IR engine already returns a larger number of case-reports.

2) MeSH indexing

All the submitted runs were built with both text and MeSH indexes, but in post-competition we investigated MeSH-only indexes.

MeSH concepts for document representation. For the query representation in competition, we only used MeSH terms extracted from the summary. In post-competition, like for text, we compared the value of description and summary. With the MeSH indexing, querying with MeSH terms extracted from description led to R-Prec of 0.123, versus 0.125 with MeSH terms extracted from summary. Like for text, both values are equivalent, and like for text the result is slightly better when using both sources: R-Prec of 0.143 (+14%). Unlike text, automatic query expansion is not useful for MeSH representation.

Yet, the optimal performances of the MeSH representation are lower than text: R-Prec 0.143 versus 0.211. We then wanted to know how complementary both representations were. We thus analysed the qrel and our both runs (text and MeSH) and looked at the proportion of relevant documents that were found by each. The following Venn diagram (Figure 3) illustrates the distribution.

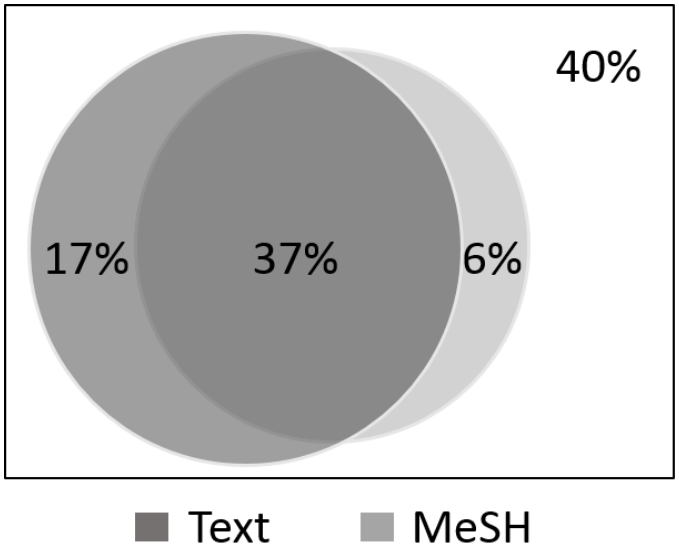

Figure 3. Complementarity of text and MeSH representations. 17% of relevant documents were only retrieved by the text index (at rank 1000), 6% only by the MeSH index, 37% by both. 40% were not retrieved.

Hence, starting from the text index, it seems hard to combine the MeSH index and to take benefit from the 6% of relevant documents only retrieved by MeSH. We tested different linear combinations but only achieved a little gain (R-Prec from 0.211 to 0.213) using 10% of the MeSH RSVs.

Finally, the impact of the MeSH concepts collected from MEDLINE was weak: when indexing only these MeSH concepts, the computed run had R-Prec 0.028.

MeSHtargets. In the official submissions, runs 3 and 4 only differed in the application of the MeSHtargets strategy. We observe a slight improvement in R-Prec (+6%). Unfortunately, these official runs were computed with section indexing. In post-competition, we explored a wide range of settings for applying this strategy to runs computed with document indexing, but we did not observed any significant gain.

3) Co-citations network

Table 2 gives the result of the co-citation network strategy applied to the best run described (text representation + 10% MeSH representation, R-Prec 0.213).

α	0	0.01	0.05	0.1	0.2	0.3
R-Prec	0.213	0.214	0.224	0.224	0.218	0.209

Table 2. R-Prec after the co-citations network strategy with different values of alpha. 0 is the baseline.

With $\alpha = 0.05$ we observe a slight improvement (+5%). In the official runs, we applied this strategy to the run 4 and arbitrarily set α to 0.10. Unfortunately, run 4 was computed with section indexing and was far from being the best one. Yet, a gain was also observed from run 4 to run 5 (R-Prec from 0.114 to 0.124, +8%).

Conclusion

For this TREC CDS 2014 campaign, we explored a wide range of strategies, such as :

- document or section indexing,
- MeSH representation,
- article-type boosting,
- co-citations network.

Section indexing was clearly a weak approach. The article-type boosting was counter-productive, but it appeared that the IR process already tends to favour reviews and case reports. MeSH representations, extracted from the full-text or collected from MEDLINE, led to very slight improvement and did not show great complementarity with the text. Finally, the co-citations network strategy led to significant improvements (+5%).

Acknowledgments

This work was partially funded by the European Commission under grant agreement 600932 (FP7-ICT-2011-9), MD-Paedigree project.

References

[1] Gobeill, J., Tbahriti, I., Ehrler, F., & Ruch, P. (2007). Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics. In TREC.

[2] Gobeill, J., Gaudinat, A., Ruch, P., Pasche, E., Teodoro, D., & Vishnyakova, D. (2011). BiTeM Group Report for TREC Medical Records Track 2011. In TREC.

[3] Gobeill, J., Theodoro, D., & Ruch, P. (2009). Exploring a wide range of simple pre and post processing strategies for patent searching in CLEF IP 2009. CLEF working notes, 2009.

[4] TREC 2014 CDS Track website: http://www.trec-cds.org/2014.html

[5] Ounis I, Amati G, Plachouras V, et al. (2006) Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval. Terrier: A High Performance and Scalable Information Retrieval Platform.

[6] Amati G. (2009) Probabilistic Models for Information Retrieval based on Divergence from Randomness. Ph.D. thesis. Science University of Glasgow, Department of Computing.TREC-CHEM Track Guidelines.

[7] Gobeill, J., Theodoro, D., Patsche, E., & Ruch, P. (2009). Taking benefit of query and document expansion using mesh descriptors in medical imageclef 2009. Working Notes of CLEF.

[8] Gobeill, J. (2012) Modèles de Question/Réponse pour la Biomédecine. Doctoral dissertation, PHD Thesis, University of Geneva.

[9] Burgun, A., & Bodenreider, O. (2001). Mapping the UMLS Semantic Network into general ontologies. In Proceedings of the AMIA Symposium (p. 81). American Medical Informatics Association.