NovaSearch at TREC 2014 Clinical Decision Support Track

André Mourão, Flávio Martins and João Magalhães

Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Caparica, Portugal
a.mourao@campus.fct.unl.pt, flaviomartins@acm.org,
jm.magalhaes@fct.unl.pt

Abstract. This paper describes the participation of the NovaSearch group at TREC Clinical Decision Support 2014. As this is the first edition of the track, we decided to assess the performance of multiple information retrieval techniques: retrieval functions, re-ranking, query expansion and classification of medical articles into question categories. The best performing run was based on an ensemble of state-of-the-art retrieval algorithms combined with unsupervised fusion.
Our best run was based on the late fusion of runs using MeSH query expansion, pseudo-relevance feedback with terms from top retrieved results and multiple retrieval functions (BM25L, BM25+, TF-IDF and Language Models) combined with RRF fusion algorithm.
We also tested an algorithm to measure article relevance to the target medical questions (diagnosis, test and treatment articles), based on the frequency of words to some categories. An additional experiment was based on pseudo relevance feedback based on each article's journal reputation. Although some techniques did not increase our baseline performance, we are satisfied with our global performance.

1 Introduction

TREC Clinical Decision Support Track is a new track focused on the "retrieval of biomedical articles relevant for answering generic clinical questions about medical records."[1] These medical records consist on cases summarizing patients medical records and conditions, presented on well-formed natural English text. For each record, there are two formulations: descriptions, containing a detailed account of the patients' visits and summaries, simplified versions of the descriptions with less irrelevant and negative information.

Our participation on this track follows our work on the textual component of our ImageCLEF Med 2013 system [3, 4], adapted to answer the specific types of questions on this track. Section 2 details our usage of general and domain specific IR techniques. Section 3 describes how we measured article relevance to the Generic Clinical Question types. Section 4 describes our pseudo-relevance

[1] http://www.trec-cds.org/

feedback algorithm based on journal reputation. Section 5 contains the results and discussion.

2 Medical retrieval system

2.1 Text indexing and retrieval

This section summarizes the indexing and retrieval techniques applied in our medical retrieval system. For a more detailed explanation, see the text and fusion sections of [4]. The retrieval function BM25L [1] is our systems baseline. We experimentally found it to be the best model in the medical domain [2], in particular because it handles long documents (i.e. article full text) better than other retrieval functions. We indexed and searched the full document text (all chapters including image captions), abstract and title. We ran pseudo-relevance feedback using the top 3 results retrieved using the initial query. We added a maximum of 25 new query terms to the initial query.

At query time, we expanded the initial query with preferred and alternative terms sourced from a SKOS formatted version of MeSH using Lucene-SKOS.

2.2 Rank fusion

One of the hypothesis we wanted to test was that some retrieval functions can give better results for certain queries, and that an unsupervised combination of multiple functions can improve the results from individual functions. Rank fusion aims at combining ranked document lists (ranks) from multiple sources into a single (combined) ranked list. Previous works explored the effectiveness of different types of unsupervised rank fusion. and we have chosen RRF, as it generally achieves good, all-round performance. Due to a limited number (5) of submissions, we only submitted two combinations:

- a combination using only one fusion algorithm (RRF) and a retrieval function set (TF-IDF, BM25L, BM25+ and Language Models with Dirichlet priors), with the techniques described in the previous section; no weighted PRF nor query type weighting;
- a combination of the baseline run and the query type weighting run.

3 Answering medical questions with relevant articles

Other hypothesis we wanted to test was: can we measure article relevance to questions (question type weighting, QTW) using word frequency patterns (e.g. articles relevant to the diagnosis questions would have more diseases or symptoms names, treatment articles would have medications name, ...).

For each category, we empirically selected the terms from top level MeSH hierarchy:

- **Diagnosis:** B03, B04, C

– **Test:** E01
– **Treatment:** D02, D04, D06, D26, D27, E02, E04

At indexing time and for each category, we assigned each document a weight derived from the percentage of total words from that category. For example, an abstract with 100 words containing 10 words from the "diagnosis" category would receive an unnormalized score of 0.1. These scores were then Min-Max normalized in relation to other scores for the category. Each document was weighted for each category (meaning each document has 3 weights). At retrieval time, document score (as returned by the retrieval function) was multiplied by the weight of that document for that question's category.

4 Reputation based re-ranking

We tested an alternative of weighing articles for PRF, weighted PRF, that computed an article keyword weight based on its journal's "Impact Factor" from Thomson Reuters' Journal Citation Reports 2013 [2]. Instead of using terms from the top 3 documents, we extracted the terms from the top 10 documents, giving higher weights to terms from documents with higher impact factor.

5 Results and discussion

Table 1 contains run summaries detailing the techniques used in each run.

Table 1. NovaSearch run summary

Run id	Retrieval functions	MeSH	PRF	W-PRF	QTW	Notes
1	BM25L	×	×			Baseline
2	BM25L	×	×		×	Question Type Weighting experiment
3	BM25L	×	×		*	RRF of NS1 and NS2
4	BM25L, BM25+, TF-IDF, LM	×	×			RRF of runs with multiple ret. functions
5	BM25L	×		×		Journal reputation weighting experiment

Our best performing run was based on the combination of multiple retrieval functions with query expansion and pseudo-relevance feedback, Table 2. QTW and W-PRF runs performed worse than the baseline. In our experiments, we found that relevant documents were being penalized excessively, due to normalization and weight tunning. Regarding W-PRF, we also performed additional experiments with the documents in the relevance judgments and found that

[2] `http://thomsonreuters.com/journal-citation-reports/`

Table 2. TREC CDS 2014 results for NovaSearch runs. Bold values represent best result for summary-based manual runs

Run id	infAP	infNDCG	R-prec	P@10
4	0.0757	**0.2631**	0.2165	**0.3900**
1	0.0727	0.2504	0.1971	0.3667
3	0.0686	0.2418	0.1906	0.3333
2	0.0669	0.2360	0.1843	0.3333
5	0.0579	0.2101	0.1691	0.3033

there are many articles from journals that did not appear on the JCR list, and thus, receiving zero weight.

References

1. Lv, Y., Zhai, C.: When documents are very long, BM25 fails! In: SIGIR '11. pp. 1103–1104 (Jul 2011), http://dl.acm.org/citation.cfm?id=2009916.2010070
2. Martins, F., Haslhofer, B., Magalhães, J.: Query expansion using open web-based skos vocabularies. In: ACM SIGIR Workshop on Health Search and Discovery: Helping Users and Advancing Medicine. pp. 35–38 (2013)
3. Mourão, A., Martins, F., Magalhães, J.: NovaSearch on medical ImageCLEF 2013. In: CLEF 2013 Online Working Notes/Labs/Workshop. pp. 1–10 (2013)
4. Mourão, A., Martins, F., Magalhães, J.: Multimodal medical information retrieval with unsupervised rank fusion. Computerized Medical Imaging and Graphics 39, 35 – 45 (2015), http://www.sciencedirect.com/science/article/pii/S0895611114000664, medical visual information analysis and retrieval