CRP Henri Tudor at TREC 2014: Combining Search Results for Clinical Decision Support

Duy Dinh, Asma Ben Abacha
{*duy.dinh, asma.benabacha*}*@tudor.lu*

CR SANTEC, Public Research Centre Henri Tudor,
6 avenue des Hauts-Fourneaux, Esch-sur-Alzette, Luxembourg

1 Introduction

This paper presents the first participation of the Luxembourgish Public Research Center Henri Tudor in the TREC 2014 Clinical Decision Support (CDS) Track. At the Resource Centre for Healthcare Technologies (SANTEC) department, we focus our research activities on healthcare technologies. Our mission consists primarily in improving healthcare by developing methods, tools, services and solutions that can be applied by healthcare professionals, patients and citizens on a daily basis.

In this research work, we present an approach to combining search results using data fusion techniques. The focus of the 2014 Clinical Decision Support Track was the retrieval of relevant biomedical articles for answering generic clinical questions about medical records. Each question consists of a case report and one of three generic clinical question types, such as "What is the patient's diagnosis?". Retrieved articles are judged relevant if they provide information of the specified type that is relevant to the given case.

The remainder of this report is organized as follows. Section 2 gives a brief description about the CDS Task. Section 3 describes our methodology for combining search results. Our submitted runs and the official results of TREC CDS Track are described in Section 4. Finally, Section 5 draws our conclusions and outlining directions for future work.

2 Overview of the CDS Task

The TREC Clinical Decision Support (CDS) Track investigates the performance of systems that search a static set of documents obtained from short case reports, such as those published in biomedical articles, as idealized representations of actual medical records. The goal of the task is to rank the documents in the collection in decreasing probability of relevance.

2.1 Document collection

The target document collection for the track is the Open Access Subset of PubMed Central (PMC). PMC is an online digital database of freely available full-text biomedical literature. Because documents are constantly being added to PMC, to ensure the consistency of the collection, we obtained a snapshot of the open access subset on January 21, 2014, which contained a total of 733,138 articles. The full text of each article in the open access subset is represented as an NXML file (XML encoded using the NLM Journal Archiving and Interchange Tag Library), and images and other supplemental materials are also available.

Each article in the collection is identified by a unique number (PMCID), which is specified by the <article-id> element within each article's NXML file.

2.2 Topics

The topics for the track are medical case narratives created by expert topic developers that serve as idealized representations of actual medical records. The case narratives describe information such as a patient's medical history, the patient's current symptoms, tests performed by a physician to diagnose the patient's condition, the patient's eventual diagnosis, and finally, the steps taken by a physician to treat the patient.

There are many clinically relevant questions that can be asked for a given case narrative. In order to simulate the actual information needs of physicians, the topics are annotated according to the three most common generic clinical question types : diagnosis, test and treatment.

2.3 Evaluation protocol

The track received a total of 102 runs from 26 different groups. All the runs contributed to the judgment sets, which were constructed to be compatible with the computing of the inferred measures. In particular, the judgment sets were created using two strata: all documents retrieved in ranks 1-20 by any run in union with a 20% sample of documents not retrieved in the first set that were retrieved in ranks 21-100 by some runs.

Documents in the judgment set were judged on a three-point scale of 0: *not relevant*, 1: *possibly relevant*, 2: *definitely relevant*. The evaluation measures were computed by conflating the possibly relevant and definitely relevant sets into a single relevant set. The exception to this is the inferred NDCG measure which makes use of the different relevance grades. The measures reported by sample_eval are inferred average precision (infAP), inferred NDCG (infNDCG), inferred precision at 10, 50, and 100 documents retrieved (iP10, iP50, iP100), inferred number of relevant retrieved (inum_rel_ret), inferred number of relevant (inum_rel) and number retrieved (num_ret).

3 Methodology

In our previous work [6, 7, 9], we have examined several data fusion techniques, term weighting models and query expansion models. In this work, we propose another strategy for merging search results into a list of results. We first present the document indexing and retrieval architecture (Section 3.1) and then we present our approach for combining results (Section 3.2).

3.1 Document indexing and retrieval

Our indexing and retrieval framework is based on an open source search engine, which has been widely used for research in IR. More specifically, we used the Terrier IR platform for indexing and retrieving documents in the collection [10].

The indexing aims to organize, structure and store statistical and/or linguistic information about terms and documents in the collection allowing a rapid and efficient search. During the indexing stage, stop-words are removed from documents before stemming using the Porter algorithm [11].

The document retrieval aims to match the user query and document representations in order to retrieve a list of results that may satisfy the user information need. A document D containing terms used for formulating a query Q is weighted by summing the score of each term appearing in document D:

$$RSV(D,Q) = \sum_{t \in Q} score(t \in D) \tag{1}$$

where $score(t \in D)$ is the query term weight calculated using a particular term weighting model. For evaluating the performance of current state-of-the-art weighting models, we chose three different term weighting models used in our experiments, namely BM25 [12], In_expB2 [2] and LGD [4]. We then applied several state-of-the-art pseudo-relevance feedback techniques using statistical measures such as the Bose-Einstein (Bo) statistics [1] in order to select most related terms for enriching the original query.

3.2 Document fusion

Let $L = \{L_1, L_2, ..., L_k\}$ be the set of different lists of documents returned by k IR methods, where L_i is the result list obtained by the IR method i. Formally:

$$L_i = \{D_{i1}, D_{i2}, ..., D_{iz}\} \tag{2}$$

where z is the size of the list L_i and D_{ij} is document j belonging to the list L_i.

Each IR method can be configured by a set of parameters (e.g., the term weighting model, term's properties (stopword, low IDF, term length), etc.).

When the result lists for each query are combined together, we build a set of runs of documents. Each run contains documents with the same ID (docno) but with different scores returned by different IR methods.

Inspired by our previous work cited in [5,8,9], we aim to define novel combination techniques based on several means.

$$\begin{cases} CombH = \frac{n}{a_1^{-1} + a_2^{-1} + ... + a_n^{-1}} & (\texttt{harmonic mean}) \\ CombG = \sqrt[n]{a_1 * a_2 * ... a_n} & (\texttt{geometric mean}) \\ CombA = \frac{a_1 + a_2 + ... + a_n}{n} & (\texttt{arithmetic mean}) \\ CombQ = \sqrt{\frac{a_1^2 + a_2^2 + ... + a_n^2}{n}} & (\texttt{quadramatic mean}) \end{cases} \tag{3}$$

where n is the number of documents in each run and a_i is the score of document D_i.

4 TREC CDS Submissions

4.1 Run description

We submitted five official runs to the TREC CDS track. Our submitted runs are divided into two groups: the first one includes four automatic runs and the second one includes one manual run. For each group of runs, we aim to evaluate the performance of our data fusion techniques in comparison to state-of-the-art retrieval models. The description of the five submitted runs are as follows:

- *tudorComb1:* query terms are only in the *description* field (long query). We ignore low IDF terms and run the retrieval using three IR models namely BM25, LGD and In_expB2 with query expansion (top 30 terms from top 20 returned documents) using the Bo1 model. Finally, the results are combined using the CombA fusion technique (see Formula 3.)

- *tudorComb2:* this run is different from the previous run in the sense that query terms are extracted only from the *summary* field.
- *tudorComb3:* this run is similar to the first run except that we consider also low IDF terms for retrieval.
- *tudorComb4:* this run is similar to the third run with the exception that we consider query terms in the *summary* field.
- *tudorCombm:* the last run is similar to the first run but is classified as manual because we expand the original query using long forms of abbreviations and give a more important weight for terms denoting medical concepts.

We use the default term processing pipeline in Terrier: stop-words are removed from documents and queries before stemming using the Porter algorithm. For manual runs, we further removed query terms that are not present in the stop-word list but that we believe are not quite informative. For example, in the query "Patients with hearing loss", the term "with" is recognized as a stop-word and is therefore automatically removed from documents/queries. We compare the combined results with the baseline results obtained by each of the three IR models namely BM25, LGD and In_expB2 using query expansion (top 30 terms from top 20 returned documents) using the Bo1 model.

In what follows, we present the results of our official runs submitted to TREC CDS 2014. Afterwards, we present the results obtained unofficially on the small set of documents that have been judged either relevant or irrelevant.

4.2 Official results

Table 1 shows the official results of our runs submitted to TREC CDS 2014. According to the results, we observe that there is no significant difference when using long (description) and short (summary) queries. This is probably due to the poor performance of the IR models on the underlying collection. Indeed, the baseline IR performance is very low : $bpref \approx 0.10, MAP \approx 0.05, infAP \approx 0.09, infNDCG \approx 0.11$ and $P10 \approx 0.25, iP10 \approx 0.18$.

We also notice that the combined results outperform the results obtained by the baseline, i.e. without combination. For example, the $infNDCG$ of *tudorComb*2 run is 0.1640 which is quite better than the best baseline LGD+Bo1 ($infNDCG$ = 0.1175) with an improvement rate of +39.57%. There is also an improvement rate of +38.19% in terms of iP10 when comparing run *tudorComb*1 ($iP10$ = 0.2533) and the best baseline BM25+Bo1 ($iP10$ = 0.1833). Therefore, the results show the evidence of combining results for improving the IR performance.

4.3 Performance analysis

We study the performance of different data fusion techniques for combining search results. Table 2 depicts the IR results obtained on the pool of documents that have been judged either relevant or not relevant. Here, we only use documents that have been judged, i.e. 29,969 documents, assuming that all documents in the collection must be judged. The results confirm that the **arithmetic** mean yields the best performance with $P10 = 0.2967$ and

Performance Run	**bpref**	**MAP**	**P10**	**infAP**	**infNDCG**	**iP10**
Submitted runs						
tudorComb1	0.1045	0.0503	**0.2533**	0.0324	0.1508	**0.2533**
tudorComb2	**0.1130**	0.0513	0.2500	0.0335	**0.1640**	0.2500
tudorComb3	0.0320	0.0120	0.1167	0.0100	0.0655	0.1167
tudorComb4	0.0394	0.0168	0.1433	0.0151	0.0819	0.1433
tudorCombm	0.1070	**0.0531**	0.2467	**0.0349**	0.1618	0.2467
Baseline						
BM25+Bo1	0.1049	0.0502	0.2500	0.0092	0.1170	0.1833
LGD+Bo1	0.1033	0.0483	0.2433	0.0089	0.1175	0.1833
In_expB2+Bo1	0.1163	0.0486	0.2467	0.0077	0.1017	0.1700

Table 1: IR effectiveness obtained by each run on the TREC CDS 2014 collection. Bold numbers correspond to the best performance of submitted run while underlined numbers correspond to the best performance of the baselines.

$iP10 = 0.2000$ w.r.t. the other means (harmonic, geometric and quadramatic). However, even if the document space was dramatically reduced, i.e. unjudged documents were removed from the evaluation, the overall IR effectiveness in terms of $MAP, infAP$ and $infNDCG$ is quite small ($MAP_{best} = 0.1803, infAP_{best} = 0.0168, infNDCG_{best} = 0.1756$). Also, the overall effectiveness of our system is close to the performance of the IR model B25 in a smaller collection (only documents that have been judged).

Figure 1 depicts the percentage of relevant *vs.* irrelevant documents that have been judged either relevant or irrelevant. We observe that for each query, there are a high number of irrelevant documents in comparison to relevant documents.

A first analysis of false positives showed that some irrelevant documents were well ranked by our system due to irrelevant keywords. For example, the word "trip" in the query 2[1] led to several irrelevant documents. In future work, It would be interesting to investigate whether the semantic analysis of topics could improve the performance of our system. We plan to use NLP techniques to analyze queries semantically (e.g. [3]) in order to give different weights to keywords according to their importance w.r.t. the focus of the query.

5 Conclusion

We presented our participation to the Clinical Decision Support track in TREC 2014, which is a biomedical retrieval *ad hoc* task. The underlying IR platform of our experiments is the Terrier search system. Our participation focused on the evaluation of several IR models for term weighting as well as state-of-the-art query expansion models. We have also evaluated several combination techniques.

[1] Diagnosis. 8-year-old boy with 2 days of loose stools fever and cough after returning from a trip to Colorado Chest x-ray shows bilateral lung infiltrates.

Performance Comb.	**bpref**	**MAP**	**P10**	**infAP**	**infNDCG**	**iP10**
CombAVGA *(tudorComb1)*	0.1902	**0.1803**	**0.2967**	**0.0168**	**0.1756**	**0.2000**
CombAVGG	0.0848	0.0456	0.0400	0.0043	0.0821	0.0833
CombAVGH	**0.1910**	0.1443	0.2400	0.0104	0.1211	0.1267
CombAVGQ	0.0848	0.0456	0.0400	0.0043	0.0821	0.0833
Baseline						
BM25+Bo1	0.1965	0.1846	0.3233	0.0179	0.1869	0.2033
LGD+Bo1	0.1852	0.1728	0.3133	0.0174	0.1832	0.2067
In_expB2+Bo1	0.1955	0.1828	0.3000	0.0161	0.1737	0.1933

Table 2: IR effectiveness obtained by each of the combination techniques on the judged documents (either relevant or irrelevant) extracted from TREC CDS 2014 collection.

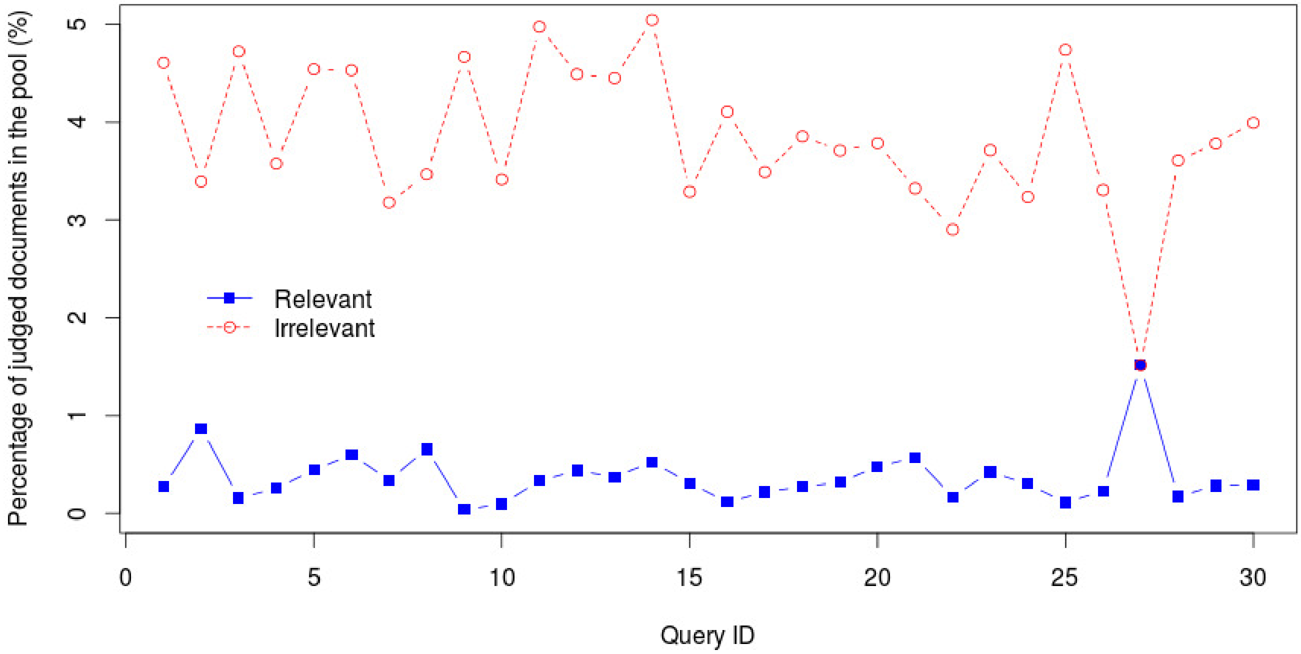

Figure 1: Some statistics about the documents submitted by all participants. Those documents are then merged into a pool and judged by TREC as relevant or irrelevant.

The combination of search results showed an improvement of the IR performance for large document collections.

In future work, we aim to investigate the adequate linguistic features extracted from relevant documents in order to better promote relevant documents. For example, we can study the semantic similarity between relevant documents and derive an IR model to rank documents based on their pairwise semantic similarity.

Acknowledgments

This work has been done in the context of the GECAMed[2] R&D project ("Gestion de Cabinets Médicaux"), which was developed at the SANTEC department of the Public Research Center Henri Tudor, with the help of general practicioners in Luxembourg. The project is sponsored by the Luxembourgish E-Santé Agency.

References

[1] Amati, G. *Probability models for information retrieval based on divergence from randomness.* PhD thesis, University of Glasgow, 2003.

[2] Amati, G., and Van Rijsbergen, C. J. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Trans. Inf. Syst. 20*, 4 (Oct. 2002), 357–389.

[3] Ben Abacha, A., and Zweigenbaum, P. Medical question answering: Translating medical questions into sparql queries. In *ACM SIGHIT International Health Informatics Symposium (IHI 2012)* (Miami, FL, USA, January 2012).

[4] Clinchant, S., and Gaussier, E. Information-based Models for Ad Hoc IR. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2010), SIGIR '10, ACM, pp. 234–241.

[5] Dinh, D. *Accès à l'information biomédicale: vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques.* PhD thesis, Paul Sabatier University - Toulouse, 2012.

[6] Dinh, D., and Tamine, L. IRIT at TREC 2011: Evaluation of Query Expansion Techniques for Medical Record Retrieval. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, November 15-18, 2011* (2011).

[7] Dinh, D., and Tamine, L. Voting Techniques for a Multi-terminology Based Biomedical Information Retrieval. In *Artificial Intelligence in Medicine - 13th Conference on Artificial Intelligence in Medicine, AIME 2011, Bled, Slovenia, July 2-6, 2011. Proceedings* (2011), pp. 184–193.

[8] Dinh, D., and Tamine, L. Towards a context sensitive approach to searching information based on domain specific knowledge sources. *J. Web Sem. 12* (2012), 41–52.

[9] Dinh, D., Tamine, L., and Boubekeur, F. Factors affecting the effectiveness of

[2] `http://www.gecamed.lu`

biomedical document indexing and retrieval based on terminologies. *Artificial Intelligence in Medicine 57*, 2 (2013), 155–167.

[10] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Lioma, C. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)* (2006).

[11] Porter, M. F. An Algorithm for Suffix Stripping. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 313–316.

[12] Robertson, S., Walker, S., Beaulieu, M., and Willett, P. Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive track. *In 21* (1999), 253–264.