LIST at TREC 2015 Clinical Decision Support Track: Question Analysis and Unsupervised Result Fusion

Asma Ben Abacha and Saoussen Khelifi

Luxembourg Institute of Science and Technology (LIST), Luxembourg

Abstract

This paper describes our information retrieval approaches to the TREC 2015 Clinical Decision Support Track. We explore different question analysis methods in order to retrieve articles relevant to the given clinical questions. We particularly study the use of two knowledge sources: MeSH and DBpedia for question expansion and the simplification of questions by removing information about the patient and negation. We also compare single IR models with the fusion of results based on both ranks and scores. Our experiments conclude that (i) query expansion using Mesh and DBpedia improves the results and that (ii) the combination of IR results using the rank outperforms the fusion based on scores. For TREC 2015 CDS task A, our best results were obtained by using DBpedia for query expansion and by combining the 2 IR models Hiemstra LM and LGD using a rank-based method. Our best run achieved an infNDCG score of 0.2894 and was ranked second over 92 runs for task A.

1 Introduction

The 2015 Clinical Decision Support Track[1] focuses on the retrieval of biomedical articles relevant for answering clinical questions about medical records. The document collection is the Open Access Subset[2] of PubMed Central (PMC). The target collection contains 733,138 full-text articles[3], the same collection used for the 2014 track.

Participants are tasked with retrieving full-text biomedical articles useful for answering questions related to three types of generic clinical questions. The considered question types are: Diagnosis, Test and Treatment (10 topics are provided for each type).

The 2015 CDS Track includes two tasks. Task A is identical to the 2014 track. In Task B, a diagnosis field is provided for the treatment and test topics. Examples 1 and 2 show topics from each task.

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

[2] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[3] A snapshot of the open access subset on January 21, 2014.

Example 1 (Task A, CDS 2015):

- Question type = diagnosis
- Summary = A 65-year-old male presents with dyspnea, tachypnea, chest pain on inspiration, and swelling and pain in the right calf.

Example 2 (Task B, CDS 2015):

- Question type = treatment
- Summary = A 15 yo girl with fatigue, pale skin, low hemoglobin and ferritin.
- diagnosis = Iron-Deficiency Anemia

2 Methods

In our experiments, we explored several semantic analysis methods such as question simplification and query expansion. We also explored unsupervised methods for search result fusion. Figure 1 presents an overview of our retrieval system.

3 Semantic Question Analysis

Semantic analysis of natural language questions is an important step towards the construction of relevant queries for information retrieval [1]. We explored two different methods for question analysis: Question simplification and question annotation using external knowledge sources.

3.1 MeSH based Annotation vs. DBpedia based Annotation for Query Expansion

We explored the use of medical Subject Headings (MeSH) terms. We studied several configurations:

- Adding MeSH terms to the query.
- Giving a higher weight to MeSH terms in the query.
- Adding all the synonyms of MeSH terms to the query.
- Giving a higher weight to MeSH terms and their synonyms in the query.

We experimentally selected the following IR models as they provided the best results on the TREC 2014 corpus: *Hiemstra LM (Hiemstra's language model)*, *TF-IDF* and *LGD*.

We used the Terrier IR platform[4] for indexing and retrieving documents in the collection. The KODA system was used to annotate questions with DBpedia [4]. Table 1 presents a summary of our experiments on query expansion.

3.2 Question Simplification

We removed two types of information from the question: (i) Age and gender information about the patient and (ii) Negation.

This question simplification approach improved the results when tested on the CDS 2014 test set. The TF-IDF and Hiemstra LM models obtained the following results using question simplification and query expansion with higher weights on DBpedia terms:

- Hiemstra LM: P@10 = 0.3167 & R-prec = 0.2378
- TF-IDF: P@10 = 0.3700 & R-prec = 0.2426

4 Unsupervised Approaches to Combine Search Results

[4] `terrier.org`

Figure 1: Overview of our retrieval system

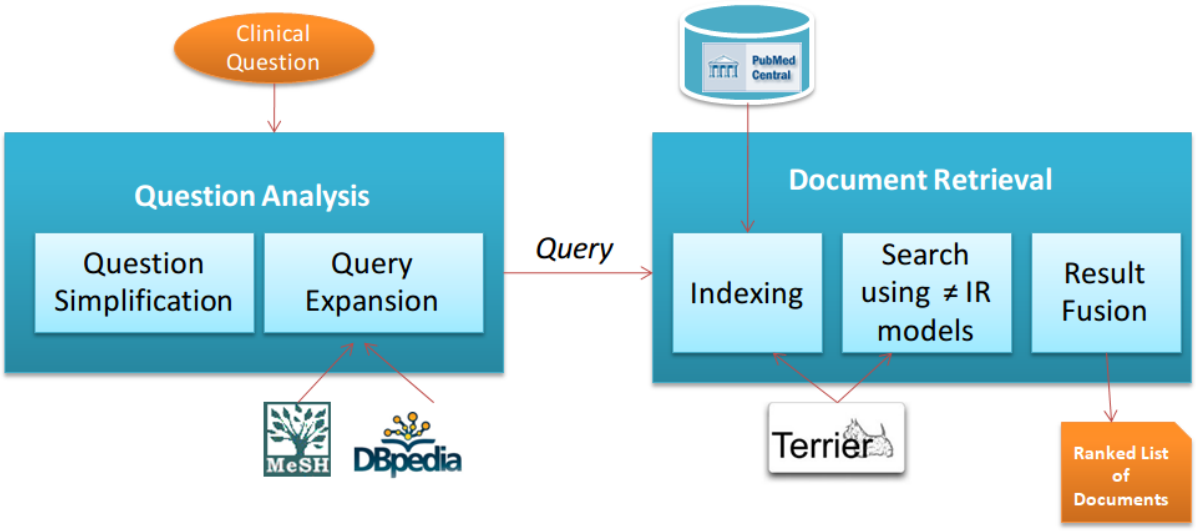

Table 1: Summary of our experiments on CDS 2014 test set: MeSH vs. DBpedia annotation for query expansion

IR Models	Query Expansion	P@10	R-prec	infAP	infNDCG
BM25	—	0.2800	0.2246	0.0180	0.1851
Hiemstra LM	—	0.3233	0.2193	0.0154	0.1671
Hiemstra LM	MeSH terms (+ higher weight)	0.3400	0.2374	0.0153	0.1641
Hiemstra LM	DBPedia terms (+ higher weight)	0.3433	0.2389	0.0152	0.1653
TF-IDF	—	0.3033	0.2234	0.0163	0.1737
TF-IDF	MeSH terms (+ higher weight)	0.3267	0.2400	0.0168	0.1796
TF-IDF	DBPedia terms (+ higher weight)	0.3367	0.2365	0.0165	0.1770
LGD	—	0.3067	0.2195	0.0164	0.1735
LGD	DBPedia terms (+ higher weight)	0.3633	0.2465	0.0166	0.1763

Combining different IR models can lead to a better overall performance [2, 3, 5]. We focus on exploring unsupervised methods for search result fusion and in particular the comparison of two unsupervised approaches: rank-based vs. score-based fusion. Rank fusion aims at combining different ranked document lists into a single list based on the rank. Several methods can be

used such as: CombMAX (max of similarity values), CombMED (median of similarity values) or CombSUM (sum of similarity values) [5].

Score-based fusion aims at combining different document lists into a single one based on the score. Different methods can be used such as Reciprocal Rank Fusion (RRF) [2]. Given a set D of documents to be ranked and a set of rankings R, each a permutation on 1..|D|,

$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k + r(D)} \quad (k = 60) \tag{1}$$

We tested the different methods for rank-based fusion. In our experiments on the CDS 2014 test set, CombSUM outperformed the other rank-based methods (such as CombMax) and the RRF score-based method. Table 2 summarizes these results.

5 Runs and Results

In CDS 2015, participants may submit a maximum of three automatic or manual runs. Each run consists of a ranked list of up to one thousand PMCID. Retrieved articles are judged relevant if they provide information of the specified type that is pertinent to the given case. The evaluation of submissions follow standard TREC evaluation procedures.

For task A, we submitted 3 automatic runs using the topics summaries:

- Run1DBpSimp: IR model TF-IDF. Query expansion using 30 expanded terms within top 20 documents. Semantic annotation of queries using DBpedia. Query simplification by deleting information about patients.
- Run4HLM: IR model Hiemstra LM. Query expansion using 30 expanded terms within top 20 documents.
- Run2DBpComb: Combination of 2 IR models Hiemstra LM and LGD. Query expansion using 30 expanded terms within top 20 documents. Semantic annotation of queries using DBpedia.

Table 3 presents our official results for task A. The run Run2DBpComb combining the IR models Hiemstra LM and LGD and using DBpedia for query expansion obtained the best results. This run was ranked second over 92 submitted runs for task A.

For task B, we tested abstract-only indexation. The run Run5DBpAbs consisted in indexing only the titles and the abstracts with the TF-IDF model. Query expansion was also applied using 30 expanded terms within top 20 documents and semantic annotation of queries using DBpedia and query simplification by deleting information about patients. We obtained the following results for run Run5DBpAbs:

- R-prec: 0.2114
- P10: 0.4000
- infAP: 0.0801
- infNDCG: 0.2855

6 Conclusion

Retrieving biomedical articles relevant for answering clinical questions is becoming more and more critical with the exponential growth of biomedical literature. In this paper we described our experiments in the scope of the clinical decision support track at TREC 2015.

Table 2: Summary of our experiments on CDS 2014 test set: Combining IR models using a score-based (RRF) vs. a rank-based method (CombSUM).

IR Model Fusion + Query Expansion (QE)	P@10	R-prec	infAP	infNDCG
RRF (In_expB2, Hiemstra_LM) + MeSH for QE	0.3200	0.2100	0.0155	0.1677
RRF (TF_IDF, Hiemstra_LM) + MeSH for QE	0.3233	0.2103	0.0154	0.1671
CombSUM (In_expB2, Hiemstra_LM) + MeSH for QE	0.3267	0.2480	0.0168	0.1789
CombSUM (TF_IDF, Hiemstra_LM) + MeSH for QE	0.3300	0.2268	0.0164	0.1743
CombSUM (TF_IDF, Hiemstra_LM) + DBpedia for QE	0.3567	0.2534	0.0162	0.1755
CombSUM (Hiemstra_LM, LGD) + DBpedia for QE	0.3733	0.2654	0.0170	0.1807

Table 3: CDS 2015: Our Official Results for Task A vs. Median (over 92 automatic task A runs)

–	MEDIAN	Run2DBpComb	Run4HLM	Run1DBpSimp
R-prec	0.1615	**0.2299**	0.2152	0.2033
P10	0.3433	**0.4533**	0.4500	0.4100
infAP	0.0414	**0.0787**	0.0746	0.0715
infNDCG	0.2109	**0.2894**	0.2768	0.2571

Our results show that (i) query expansion using open-domain knowledge bases such as DBpedia and (ii) search result fusion using a rank-based method such as CombSUM led to a noticeable improvement. In future work we plan to enhance further our IR approach by annotating semantically the abstracts of all documents in the test collection.

References

[1] 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* (2012).

[2] Cormack, G. V., Clarke, C. L. A., and Büttcher, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA* (2009), pp. 758–759.

[3] Dinh, D., and Ben Abacha, A. CRP henri tudor at TREC 2014: Combining search results for clinical decision support. In *Proceedings of The Twenty-Third Text RE-*

trieval Conference, TREC 2014, Gaithersburg, Maryland, USA (2014).

[4] Mrabet, Y., Gardent, C., Foulonneau, M., Simperl, E., and Ras, E. Towards knowledge-driven annotation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA* (2015), pp. 2425–2431.

[5] Shaw, J. A., and Fox, E. A. Combination of multiple searches. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA* (1994), pp. 105–108.