# NLM_NIH at TREC 2016 Clinical Decision Support Track

**Asma Ben Abacha**

U.S. National Library of Medicine, Bethesda, MD.

## Abstract

In this paper, we present our approach for TREC 2016 Clinical Decision Support (CDS) track. We combined methods for question analysis, query expansion, document retrieval and result fusion to find relevant documents to a given clinical question. We submitted three automatic runs using the summaries and two automatic runs using the notes, provided for the first time at the CDS track. Our experiments showed that query expansion and rank-based result fusion led to the best performance. Our runs exploring the clinical notes used MeSH for topic analysis and achieved our best P10 score of 0.2533. Using the summaries, we obtained an infNDCG score of 0.1872 and a R-prec score of 0.1465 (score in the top 10 of 107 automatic runs submitted to the 2016 CDS track).

## 1 Introduction

The Clinical Decision Support (CDS) track[1] focuses on the retrieval of relevant biomedical articles for answering clinical questions about medical records. Like previous years, participants are tasked to retrieve full-text biomedical articles pertinent to answer questions related to three types of generic clinical questions: Treatment, Test and Diagnosis.

The topics are EHR admission notes curated by physicians from the MIMIC-III data. The notes are extracted from the history of present illness (HPI) section of the note. The HPI describes information related to the patient such as medical history, performed tests and the current diagnosis. 10 topics are provided for each question type.

For each topic, three versions of the patient records are provided (i) the EHR admission note (only the HPI section, which is the "case"), (ii) a description which removes much of the jargon and replaces clinical abbreviation and (iii) a summary which is a 1-2 sentence summary of the description. We present below an example from 2016 CDS topics:

- **Type:** Treatment
- **Note:** Mr. [**Known patient lastname 4075**] is a 63 yo man with h/o biphenotypic ALL, now Day + 32 from allogeneic SCT, who presents to clinc with one week of worsening SOB and two

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

days of a clear productive cough. The patient states his SOB occured when lying flat, but not with activity. Also admitted to chest pressure which would come and go in his left chest no related to the SOB. *(...)*

- **Description:** A 63 yo man with h/o biphenotypic ALL, now Day + 32 from allogeneic SCT, who presents with one week of worsening SOB and two days of a clear productive cough. The patient states his SOB occured when lying flat, but not with activity. Also admitted to chest pressure which would come and go in his left chest no related to the SOB. *(...)*

- **Summary:** A 63 year-old male with biphenotypic ALL, Day +32 after BMT, h/o CMV infection, aspergillus and Leggionare's disease, presents with acute onset of hypoxia accompanied by fever and two days of productive cough. His CXR showed an opacification of the left basilar lobe and also right upper lobe concerning for pneumonia.

## 2 Document Collection

The document collection for the CDS tracks (2014-2016) is the Open Access Subset[2] of PubMed Central (PMC), a free digital repository of full-text biomedical articles. The collection used for 2015 and 2014 tracks contained 733,138 full-text articles. For the 2016 CDS track, an updated collection is provided, a snapshot of the open access subset on March 28, 2016. The 2016 document collection contains a 1.25 million articles in NXML file format.

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

## 3 System Description

Figure 3 presents an overview of our retrieval system for the 2016 CDS track.

Semantic analysis of natural language questions is an important step towards the construction of relevant queries for information retrieval and question answering [2]. We explored the use of medical Subject Headings (MeSH) by (i) adding MeSH terms to the query and (ii) giving a higher weight to MeSH terms in the query.

We used the Terrier IR platform[3] for indexing and retrieving documents in the collection. Terrier implements various IR models such as the Okapi BM25 probabilistic model, In_expB2 (Inverse Expected Document Frequency model with Bernoulli after-effect and normalization), the classic tf-idf vector space model and Hiemstra's language model (Hiemstra_LM).

To combine results of IR models, two unsupervised approaches are usually used: rank fusion and score-based fusion [1,4]. Rank fusion aims at combining different ranked document lists into a single rank list in order to improve the rankings of individual systems. Several methods can be used such as CombMAX (max of individual similarities), CombMED (median of of individual similarities) or CombSUM (sum of individual similarities) [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) [3].

In our experiments on the CDS 2014 test set, CombSUM outperformed the other rank-based methods and also the RRF score-based method [1]. Table 1 summarizes these results.

For the 2016 CDS track, we selected three IR models (*Okapi BM25*, *TF-IDF* and *In_expB2*)

[3] `terrier.org`

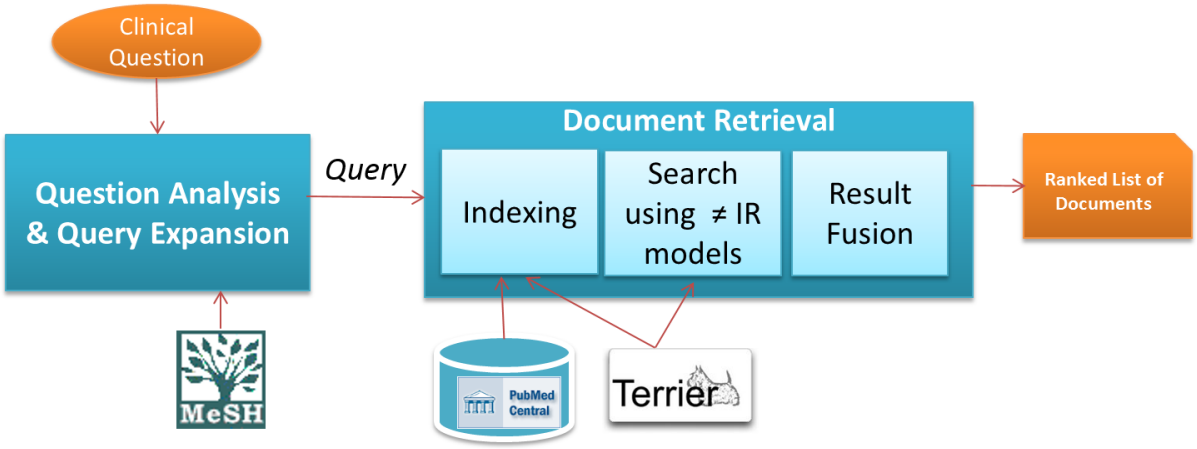


Figure 1: Overview of our retrieval system

Table 1: Summary of our experiments on 2014 test data: Combining IR models using a score-based (RRF) vs. a rank-based method (CombSUM) [1]

| IR Model Fusion | P@10 | R-prec | infAP | infNDCG |
|---|---|---|---|---|
| RRF (In_expB2, Hiemstra_LM) | 0.3200 | 0.2100 | 0.0155 | 0.1677 |
| RRF (TF-IDF, Hiemstra_LM) | 0.3233 | 0.2103 | 0.0154 | 0.1671 |
| CombSUM (In_expB2, Hiemstra_LM) | 0.3267 | **0.2480** | **0.0168** | **0.1789** |
| CombSUM (TF-IDF, Hiemstra_LM) | **0.3300** | 0.2268 | 0.0164 | 0.1743 |

and the *CombSUM* method to combine the individual ranks.

## 4 Runs

This year, participants are required to use only EHR notes, only descriptions, or only summaries for any given run submission and allowed to submit a maximum of five automatic or manual runs. At least two runs must use the EHR note. Each run consists of a ranked list of up to one thousand PMCIDs.

We submitted five automatic runs to 2016 CDS track. Two runs use only the notes and the other three runs use only the summaries:

- Run1: uses the summaries, query expansion performed using 30 expanded terms within top 20 documents, and combines the results of TF-IDF and In_expB2 models using the CombSum method.

Table 2: TREC CDS 2016 results for our submitted runs

| Measure | NLMrun1 (Summary) | NLMrun2 (Summary) | NLMrun3 (Summary) | NLMrun4 (Note) | NLMrun5 (Note) |
|---|---|---|---|---|---|
| R-prec | 0.14 | **0.1465** | 0.1405 | 0.0849 | 0.0866 |
| P10 | 0.25 | 0.2467 | 0.24 | 0.24 | **0.2533** |
| infAP | **0.0239** | 0.0230 | 0.0228 | 0.0146 | 0.0202 |
| infNDCG | **0.1872** | 0.1853 | 0.1871 | 0.1477 | 0.1687 |

- Run2: uses the summaries, MeSH for query expansion and the Okapi BM25 model.
- Run3: uses MeSH terms extracted from the summaries and the Okapi BM25 model.
- Run4: uses MeSH terms extracted from the notes and the Okapi BM25 model.
- Run5: uses MeSH terms extracted from the notes and combines TF-IDF and In_expB2 ranks using the CombSum method.

# 5 Results

The evaluation of submissions followed standard TREC evaluation procedures. Runs are scored according to precision at 10 (P@10), R-precision (R-prec) and two inferred retrieval measures infNDCG and infAP.

Table 2 presents the official results at CDS 2016 for our submitted runs. Bold values show best results. Our best infNDCG and infAP are obtained using the summaries (NLMrun1 using query expansion and combining TF-IDF and In_expB2 ranks). The best R-prec is obtained using the summaries too (NLMrun2) and is ranked in the top 10 over 107 automatic runs. Interestingly, the best P@10 is obtained using the notes (NLMrun5). The infNDCG scores for each topic are presented in Figure 2 (NLMrun1 vs. median, using summaries) and in Figure 3 (NLMrun5 vs. median, using notes).

# 6 Conclusion

In this paper, we described our participation in the TREC 2016 CDS Track. We submitted five automatic runs using summaries or notes. Our experiments showed that query expansion and result fusion led to the best performance in our runs. The runs exploring the clinical notes achieved a P10 score of 0.2533 using MeSH for topic analysis and the combSUM method to combine TF-IDF and In_expB2 ranks. Using the summaries, we achieved an infNDCG score of 0.1872 and a R-prec score of 0.1465 ranked in the top 10 over 107 automatic runs submitted to the CDS track.

# Acknowledgements

This research was supported by the Intramural Research Program at the U.S. National Library of Medicine, National Institutes of Health.

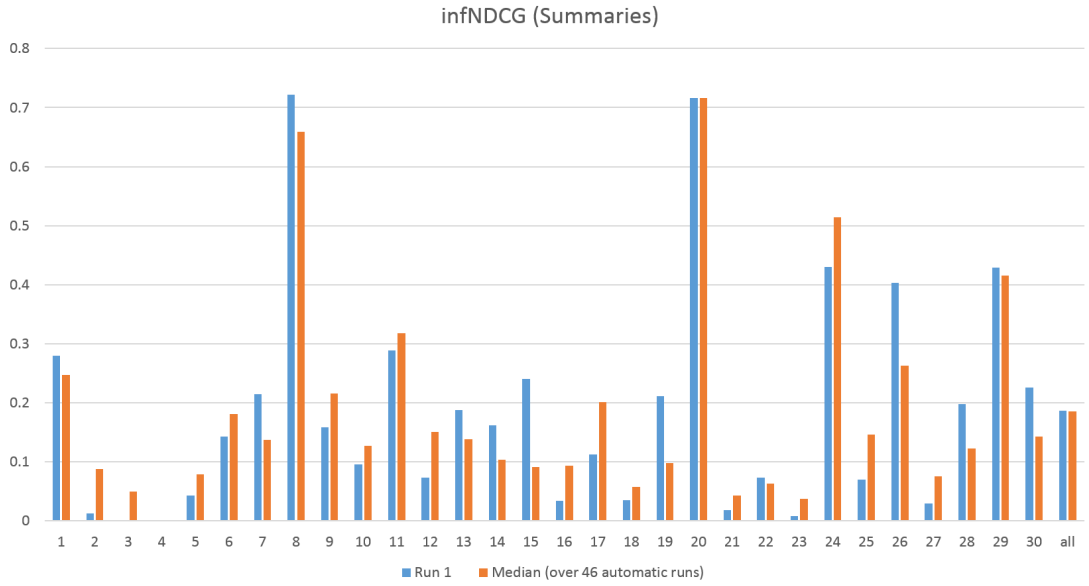


Figure 2: infNDCG scores for each topic – using the summaries only

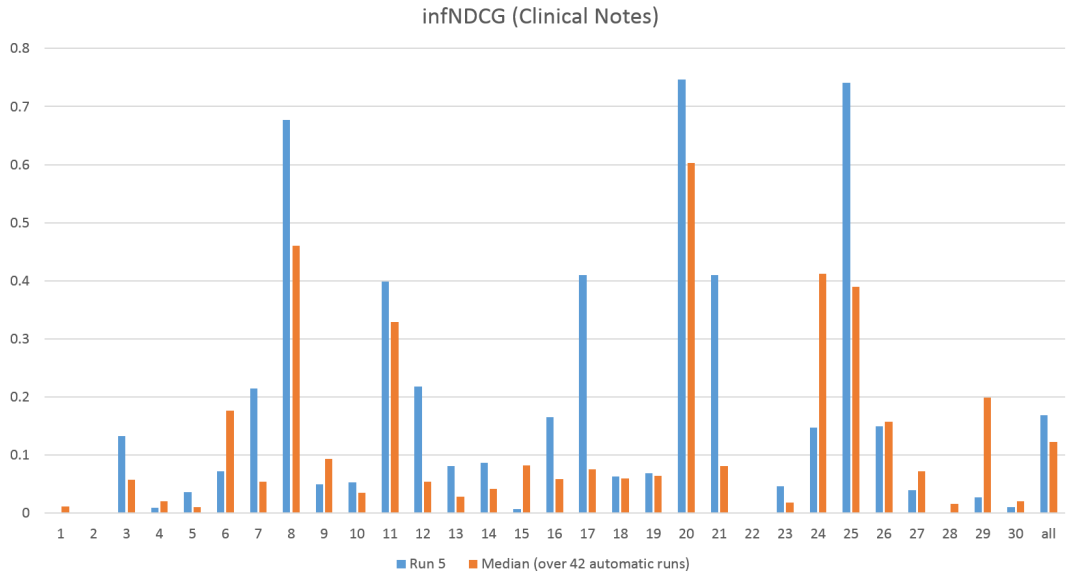


Figure 3: infNDCG scores for each topic – using the notes only

## References


[1] Ben Abacha, A., and Khelifi, S. LIST at TREC 2015 clinical decision support track: Question analysis and unsupervised result fusion. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015* (2015).

[2] 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).

[3] 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.

[4] 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 REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA* (2014).

[5] Fox, E. A., and Shaw, J. A. Combination of multiple searches. In *Proceedings of The Second Text REtrieval Conference, TREC 1993, Gaithersburg, Maryland, USA, August 31 - September 2, 1993* (1993), pp. 243–252.