Team DA_IICT at Clinical Decision Support Track in TREC 2016: Topic Modeling for Query Expansion

Jainisha Sankhavara and Prasenjit Majumder

IRLP Lab
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, India

Abstract. Clinical Decision Support (CDS) task aims to find the biomedical literature articles related to medical case reports. These articles should help to get answers to the questions of generic clinical types. This paper reports the results on query expansion using topic modeling on CDS-2016 data.

Keywords: Clinical Decision Support, Information Retrieval, Topic Modeling

1 Introduction

CDS 2016: Clinical Decision Support[1] task focuses on retrieving biomedical documents, from biomedical literature, related to generic clinical questions about medical records. The task this year is quite similar to previous years 2014[1] and 2015[2] but actual electronic health records (EHR) are used in the task of 2016 instead of synthetic version of medical case reports. The challenge is to retrieve full-text biomedical articles that address the questions for a given EHR note. Each topic will consist of a note and one of three generic clinical question types: diagnosis, treatment and test. This paper describes 5 submitted runs which all are automatic runs which are based on query expansion technique using topic modeling[3] [4]. The standard query expansion techniques have also been tried applied on CDS data and it shows improvement[5]. The description of data is provided in section 2. The experiments and results are described in section 3 and section 4 respectively and we conclude in section 5.

2 Data Statistics

Documents: The document collection for this year, like previous years, is the Open Access Subset of PubMed Central (PMC) which is an online digital database of freely available full-text biomedical literature. For 2016 task, the document

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

collection was updated by taking a new snapshot of the open access subset on March 28, 2016 and it contains 1.25 million articles represented using NXML file (XML encoded using the NLM Journal Archiving and Interchange Tag Library) format.

Topics: The admission notes from MIMIC-III are used as topics. It describes a patient's medical history, current complaint, tests performed by a physician to diagnose the patient's condition, possibly the patient's current diagnosis, and finally, any steps taken by medical professionals to treat the patient. Specifically, MIMIC-III focuses on ICU (Intensive Care Unit) patients and these notes are extracted from the history of present illness (HPI) section of the note, which most resembles the narrative cases used in previous tracks. These admission notes are the actual data generated by clinicians (mostly physicians, including residents, and nurses) and contain a significant number of abbreviations as well as other linguistic jargon and style.

There were 30 topics provided this year and they are annotated according to the three most common generic clinical question types (Ely et al., 2000) shown in the table below.

Type	**Generic Clinical Question**	**Number of Topics**
Diagnosis	What is the patient's diagnosis?	10 to 15
Test	What tests should the patient receive?	10 to 15
Treatment	How should the patient be treated?	10 to 15

Each topic consist of three versions of medical case report. First, the EHR admission note (only the HPI section, which is the "case"). Second, a more layman-friendly "description" similar to previous tracks, which removes much of the jargon and replaces clinical abbreviation for better readability. Third, a "summary" similar to previous tracks, which is a 1-2 sentence summary of the description.

The participants has to use only EHR notes, only descriptions, or only summaries for any given run submission. Since the note section is the actual real representation of the medical report, It is must to utilize the note in a subset of their submissions. At most five run submissions were allowed from which at most three runs may use description or summary versions of the topics.

3 Experiments

The experiments are done using terrier[6] and mallet[7] tool-kits which are openly available. The experiments focuses on query expansion using topic modeling. Submission consists of five runs which are describes here.

1. **Run DAsummTM:**
   This is an automatic run using summary as a query and pseudo-relevance feedback based query expansion where the expansion terms are chosen using

topic modeling on pseudo-relevant documents (top retrieved documents). On the top retrieved documents, topic modeling is performed using mallet and retrieval is performed using terrier with In_expC2 retrieval model.

2. **Run DAdescTM:**
   This run is similar to run DAsummTM but uses description part of the topics as query instead of summary part.
3. **Run DAnoteTM:**
   This run is again similar to runs DAsummTM and DAdescTM but uses note of the topics as query.
4. **Run DAnoteRoc:**
   This run uses standard Rocchio model for pseudo-relevance based query expansion with In_expC2 retrieval model and note as query.
5. **Run DAnote:**
   The run DAnote is a standard retrieval run using In_expC2 without using query expansion on note.

4 Results

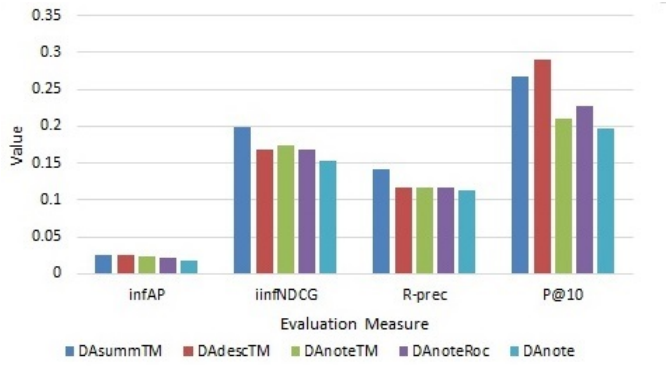
The querywise results infAP, infNDCG, R-Prec and p@10 are provided for each submitted run by the officials of TREC-CDS task. Using the relevance judgement provided by them, the overall scores of infAP, infNDCG, R-Prec and p@10 for all the runs are given in table 1. The highest among five runs are marked as bold in the table.

Run	**infAP**	**infNDCG**	**R-prec**	**P@10**
DAsummTM	0.0253	**0.1988**	**0.1416**	0.2667
DAdescTM	**0.0255**	0.1692	0.1175	**0.2900**
DAnoteTM	0.0227	0.1734	0.1160	0.2100
DAnoteRoc	0.0214	0.1676	0.1173	0.2267
DAnote	0.0186	0.1536	0.1127	0.1967

Table 1. Evaluation results for all runs

5 conclusion

The paper describes results of topic modeling on summary, description as well as note. Topic modeling gives better results on summary and description as compared to note. When comparing topic modeling with Rocchio based query expansion and without expansion, it outperforms the other two. The detailed study of topic modeling for query expansion in biomedical can be done in future.



References

1. Simpson, Matthew S., Ellen M. Voorhees, and William Hersh. Overview of the trec 2014 clinical decision support track. LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD, 2014.
2. Roberts, Kirk, et al. "Overview of the TREC 2015 Clinical Decision Support Track."
3. Brett, Megan R. "Topic modeling: a basic introduction." Journal of Digital Humanities 2.1 (2012): 12-16.
4. Steyvers, Mark, and Tom Griffiths. "Probabilistic topic models." Handbook of latent semantic analysis 427.7 (2007): 424-440.
5. Sankhavara, Jainisha, et al. Fusing manual and machine feedback in biomedical domain. DHIRUBHAI AMBANI INST OF INFORMATION AND COMMUNICATION TECHNOLOGY GANDHINAGAR (INDIA), 2014.
6. Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald and Douglas Johnson: Terrier Information Retrieval Platform. In Proceedings of the 27th European Conference on Information Retrieval, ECIR 2005.
7. McCallum, Andrew Kachites: "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.