Query Refinement: Negation Detection and Proximity Learning Georgetown at TREC 2014 Clinical Decision Support Track

Christopher Wing and Hui Yang

Department of Computer Science,
Georgetown University, USA
cpw26@georgetown.edu, huiyang@cs.georgetown.edu

Abstract

In this paper we describe our efforts for TREC Clinical Decision Support Track 2014. Our system takes medical case narratives as input and returns relevant biomedical articles to answer clinical questions to determine the patient's diagnosis, the tests the patient should receive, and how the patient should be treated. We model each topic as highly representative keyword-based structured queries. Since both the topics and returned documents are written in highly technical language, we address the traditional vocabulary gap present within medical information retrieval, while also focusing on employing methodologies to refine our queries by detecting negation and applying query proximity learning. We hypothesized terms with high frequency among the topics, which are likely to create noise and impair the return of highly relevant documents. Our top two runs utilizing negation detection perform above the median for P@10, R-Prec, and infAP, and our other three runs that utilized proximity learning performed approximately consistent with the median. More research is required to explore the potential benefits of proximity learning over a more robust set of topics.

1. Introduction

Physicians can find relevant information in biomedical literature necessary to make clinical decisions. However, locating the most relevant and time-sensitive information for a particular clinical question can be a challenging and timely task. The TREC 2014 Clinical Decision Support track challenges participants to retrieve full-text biomedical articles to answer the three most common generic clinical questions faced by physicians on a daily basis: "what is the patient's diagnosis?", "what tests should the patient receive?", and "how should the patient be treated?" [1].

Thirty topics were created to serve as idealized representations of actual medical records, and include information such as medical history, current symptoms, diagnosis, and the physician's plan to treat the patient. Each topic is tagged with one of the specified questions mentioned above. For a given topic, the retrieved documents are then used to provide specific information to a physician relevant to the clinical question. Each case narrative was provided in two forms: a longer, more complete patient account referred to as "descriptions" and a simplified version referred to as "summaries." In our approach, we utilized the descriptions for all of our submitted runs.

2. Our Method

2.1 Indexing

The document collection used for the track is a snapshot of the Open Access Subset of PubMed Central (PMC), containing 733,138 articles in NXML format. We adopt the Lemur Search Engine[1] to build an index for the collection with stop word removal and Krovetz stemming [2]. We adopt the Language Modeling with Dirichlet smoothing [3] that Lemur implements as its default retrieval algorithm:

$$P(t|d) = \frac{tf_{t,d} + \mu P(t|M_C)}{\sum_{t' \in V} tf_{t',d} + \mu}$$

where $tf_{t,d}$ corresponds to the frequency of term t in document d, $M_C$corresponds to the corpus model, and V corresponds to the Vocabulary. Lemur's default value of $\mu = 2500$ was used.

2.2 Medical Concept Detection

Since the topics were written in natural language, we needed to generate representative keywords directly from the text. Each topic was queried on the National Library of Medicine's (NLM) MeSH on Demand,[2] which generates relevant MeSH terms using the NLM Medical Text Indexer (MTI).[3] MeSH, or Medical Subject Headings, are terminology used by the NLM to index articles, catalog books, and searching MeSH-indexed databases such as PubMed.

However, since many medical conditions may be expressed in varying terminology, a single representation of a medical concept in one case report may not match a given representation in another medical case document. Thus, we queried each detected MeSH term in MetaMap,[4] which maps text to the Unified Medical Language System (UMLS) Metathesaurus, in order to provide synonyms. A confidence score is given to each proposed synonym with a max value of 1000. Synonyms were only kept with candidate mapping scores of 1000. In addition, each synonym is also described by a classification. Any MeSH term not found in MetaMap were considered not relevant and were removed. Additionally, synonyms with categories belonging to the blacklist[5] were removed. For some runs, a stricter MetaMap filtering methodology was applied. Only synonyms with categories on the whitelist[6] were considered relevant, and the rest were removed. Subsequently, each list of MeSH terms and their synonyms was filtered by inverse document frequency (IDF). Terms with IDF < 1.2 were considered overly common and predicted to create too much noise, thereby causing irrelevant documents to be returned in the ranked list. Terms comprised of more than one word such as "myocardial infarction" were ignored as Lemur's indexing functionality breaks on spaces when creating inverted lists. These lists of terms and their synonyms formed the initial queries.

---

[1] http://www.lemurproject.org/
[2] http://www.nlm.nih.gov/mesh/MeSHonDemand.html
[3] http://ii.nlm.nih.gov/MTI/index.shtml
[4] http://metamap.nlm.nih.gov/
[5] Blacklist (remove these categories): geographic area, population group, activity, human, <string> activity, human rights, occupation or discipline, functional concept, bird, temporal concept, intellectual product, profession or occupation group, quantitative concept, qualitative concept, manufactured object, food, family group
[6] Whitelist (keep these categories): Amino Acid, Peptide, or Protein, Biologically Active Substance; Anatomical abnormality; Body Location or Region; Body Part, Organ, or Organ Component; Body System; Cell; Disease or Syndrome; Health Care Activity; Injury or poisoning; Finding; Laboratory procedure; Medical Device; Mental or Behavioral Dysfunction; Mental Process; Neoplastic Process; Organic chemical; Pathologic function; Pharmacologic Substance; Sign or Symptom; Therapeutic or Preventative Procedure

2.3 Topic Sub-Classification and Query Expansion

Topics were already classified by diagnosis, test, or treatment. Yet, our system was heavily reliant on the MeSH on Demand and MetaMap APIs to detect relevant medical concepts. To reduce this dependence and increase the robustness of our keyword-based queries, we developed a sub-classification methodology to further expand queries as necessary.

The mean number of MeSH terms per query was then determined, ignoring synonyms. If a given query's number of terms was less than or equal to one-half the standard deviation from the mean, its corresponding topic was classified as a hard topic. Using the Stanford NLP parser,[7] nouns from the original topics were added to expand those queries. Duplicate terms already present in the queries were ignored. These nouns were also filtered by IDF and queried in MetaMap for synonym expansion and category filtering. The IDF threshold used varied according to the original size of the queries. For queries with an average number of terms less than or equal to one standard deviation from the mean, an IDF threshold of 0.7 was used. Otherwise, a threshold of 1.1 was used. After filtering by IDF, a second layer of filtering was applied to the nouns using the classification blacklist or whitelist, as described above. All non-hard topics were not expanded to prevent introducing additional noise.

2.4 Query Learning Approach

A novel approach we implement in our methodology is that of allowing queries to learn from one another. We argue that if words are very common among queries, then their weights should be reduced. Such words are not likely specific enough to return highly relevant documents and are more likely to create additional noise, thereby introducing less relevant documents. After applying the topic classification and query expansion methodologies above, we built an index of the queries. The IDF of each query term was computed using the index of the queries. Terms with IDF < *threshold* had their weight reduced in the final queries.

For example, certain symptoms such as fever and cough were very common among the topics. These symptoms are relevant to the diagnosis or pertinent clinical question, but they are also secondary symptoms to a large number of diseases. We believe it is more important to give greater weight to less common symptoms such as the presence of bilateral lung infiltrates, which gives greater confidence in diagnosing pneumonia.

While our runs that utilized this approach actually performed slightly worse than our other runs, we believe this may be due to the limitation of the topics rather than a limitation of the approach. We hypothesize that this approach may have greater effectiveness over a larger sample of topics or in a real-time application when used by health care professionals. As this is only the first year of the track and there were only 30 static topics present, the ability of queries to learn from one another may be limited.

2.5 Negation Detection

The topics discuss both the presence and absence of symptoms. However, we observe that the documents to be retrieved more often discuss the symptoms present rather than symptoms that are not present. Consequently, we propose that the positive symptoms present in the topics have slightly greater importance. This can be illustrated by a simple example: The patient presented with abdominal pain but no fever. The MeSH terms detected to form the initial keyword query may be "abdominal pain" and "fever". However, to aid in diagnosing and treating the patient it is more relevant to read documents which primarily discuss patients experiencing abdominal pain rather than those patients experiencing fever. Yet, MeSH on Demand detects both positive and negative symptoms in the topics without discrimination. Thus, to decrease the probability of false positives, we reduced the weight of the MeSH terms and their synonyms that came from

[7] http://nlp.stanford.edu/software/tagger.shtml

negated phrases. We utilized NegEx [4] to detect negated phrases in each topic. Any MeSH term mapping from the negated phrase had its weighting in the query reduced.

3. Experimental Results

Table 1 presents our results compared to the median and max values obtained across all systems. Overall, our runs are fairly consistent with the median, with GuHSINeg, GuHSINegL, and GuHSNegProxL presenting a small improvement over the median for three of the four metrics.

Run	**P@10**	**R-Prec**	**infAP**	**infNDCG**
Median	0.10	0.1001	0.1053	0.0169
Max	0.70	0.3164	0.1522	0.4863
GuHNegProxL	0.10	0.0956 (−4%)	0.1005 (−4%)	0.0120 (−29%)
GuHSINeg	0.15 (+50%)	0.1023 (+2%)	0.1183 (+12%)	0.0148 (−12%)
GuHSINegL	0.15 (+50%)	0.1050 (+5%)	0.1085 (+3%)	0.0148 (−12%)
GuHSNegProxH	0.10	0.0956 (−4%)	0.1126 (+7%)	0.0184 (+9%)
GuHSNegProxL	0.15 (+50%)	0.0923 (−8%)	0.1169 (+11%)	0.0184 (+9%)

Table 1: Comparison of our approaches to the median result obtained across all TREC systems. The percent delta compared to the median is also shown.

The structured queries were written and formatted in Lemur Query Language. Unless otherwise specified, the weight of a given term = 1. Our five submitted automatic runs include various aggregations of the above methodologies and are summarized here:

1. GuHNegProxL
   - Medical Concept Detection
   - Topic Classification and Query Expansion
   - NegEx (weight of negated terms = 0.5)
   - Query Learning (IDF threshold = 0.7; weight of common terms = 0.7)
2. GuHSINeg
   - Medical Concept Detection
   - Topic Classification and Query Expansion
   - Second level stricter filtering of MetaMap classifications using the whitelist
   - NegEx (weight of negated terms = 0.5)
3. GuHSINegL
   - Same as GuHSINeg, but weight of negated terms = 0.3
4. GuHSNegProxH
   - Medical Concept Detection
   - Topic Classification and Query Expansion
   - Second level stricter filtering of MetaMap classifications using the whitelist
   - NegEx (weight of negated terms = 0.5)
   - Query Learning (IDF threshold = 0.78; weight of common terms = 0.7)
5. GuHSNegProxL
   - Same as GuHSNegProxH, but query learning IDF threshold = 0.7

There were several topics for which the median and all five of our runs performed consistently poor: topics 3, 9, 18, 23, and 25 with P@10 ≈ 0 and R-Prec ≈ 0. Some generalizations regarding topics 3, 9, and 18 reveal a few limitations of our approach. Topics 3 and 9 were generally shorter and contained less technical, specific medical nouns. Thus the utility of MeSH on Demand, MetaMap and query expansion approaches to construct the base query was very limited. Topic 18 contained a lot of relevant information in numeric test results rather than in natural language. Our approach did not account for numeric text.

For topics 2, 5, 6, 7, 17, 24, and 29 all five of our runs generally outperformed the median for P@10 and R-Prec. These results suggest that the common features among our runs were successful for these topics. We observe that the queries corresponding to these topics were generally longer in length than our queries, due to strong performance of MeSH on Demand, MetaMap, and noun query expansion. The refinement of these queries by detecting negation and query proximity played less importance since there were enough medically specific keywords in the query to retrieve highly relevant documents.

4. Conclusion

In our approach for the TREC 2014 Clinical Decision Support Track, we attempt to generate refined, medical keyword queries which accurately represent the provided topics: medical case reports written in natural language. Inferring from the topics for which our runs perform well, we recognize the importance of query length and robustness, which seemed to have more significant impact compared to our refinement strategies of negation detection and query learning. In future work, we would like to develop additional techniques to generate keywords from the topics before refining the queries with negation and query learning. Also, we would like to explore if query proximity learning can introduce greater improvements by utilizing a greater number of topics within a single year or queries spanning over multiple years of the track.

5. Acknowledgements

The research is supported by NSF grant CNS-1223825 and DARPA Memex program. This material is based on research sponsored by DARPA under agreement number FA8750-14-2-0226. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

6. References

1. Ely John W, Osheroff Jerome A, GormanPaul N, Ebell Mark H, Chambliss M Lee,Pifer Eric A et al. A taxonomy of generic clinical questions: classification study BMJ2000; 321:429.
2. R. Krovetz. Viewing morphology as an inference process. SIGIR 1993.
3. W. B. Croft, D. Metzler, and T. Strohman. Search engines: Information retrieval in practice. Addison-Wesley Reading, 2010.
4. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 34(5), 301-10.