Exploiting Domain Thesaurus for Medical Record Retrieval

Miguel A Callejas P, Yue Wang and Hui Fang
University of Delaware
{wangyue,hfang}@udel.edu

Abstract

InfoLab at the University of Delaware participated in the TREC 2012 Medical Records Track. This paper explains our method and describes experiment results. One limitation of existing keyword matching based retrieval functions is the problem of vocabulary mismatch. To overcome this limitation, we propose to first map topics and visits to bags of concepts using domain thesaurus, and then model the relevance based on the similarities between those concepts.

1. Introduction

The task of this year's medical records track is same as the one from last year, i.e., content-based access to the free-text fields of electronic medical records. Specifically, a topic describes an information need from a comparative effectiveness study, and the task is to find the relevant population based on the records of patient visits. Note that the topics are related to diseases/conditions or treatments/interventions.

Although traditional keyword-matching-based retrieval functions can be directly applied to solve the problem, they are unlikely to perform well due to the inability of bridging the vocabulary gap for all the terms describing the same concept. To solve this problem, we propose to leverage the domain knowledge to represent topics and visits as bags of concepts rather than bags of terms. The proposed method is able to standardize the language used in topics and visits based on UMLS[1] and translate them into a language based on semantic codes provided by the thesaurus. The relevance is then computed based on the similarity between two bags of concepts.

2. Our Method

Natural languages are ambiguous, and this claim still holds in the medical domain. A medical concept, such as a disease, may be described in multiple ways. And traditional keyword-based retrieval functions cannot bridge the vocabulary gap between the terms from the same concept. To solve the problem, we propose to use concept-based representation for both topics and visits.

First, we map the visits and topics from bags of terms to bags of concepts. The mapping is done by the MetaMap[2] system, which is developed by the U.S. National Library of Medicine. The following is an example output of MetaMap system using the input string "hearing loss":

```
|: Established connection to Tagger Server on localhost.
Processing 00000000.tx.1: hearing loss
Phrase: "hearing loss"
Meta Candidates (6):
1000  C0011053:hearing loss (Deafness) [Disease or Syndrome]
1000  C0018772:hearing loss (Hearing Loss, Partial) [Finding]
1000  C1384666:Hearing Loss (hearing impairment) [Finding]
```

[1] http://www.nlm.nih.gov/research/umls/
[2] http://metamap.nlm.nih.gov/

861 C0018767:Hearing [Physiologic Function]

861 C1455844:hearing (Hearing examination finding) [Finding]

861 C1517945:Loss [Quantitative Concept]

Meta Mapping (1000):

1000 C0011053:hearing loss (Deafness) [Disease or Syndrome]

It is clear that a text string can be mapped to more than one concept in the UMLS and those concepts are represented as codes. Thus, instead using the term-based representation, we use code-based representation for both topics and visits. For example, the topic "patient with hearing loss" is converted to the following concepts in our system: "C0030705 C0030705 C0011053 C0018772 C1384666 C0018767 C1455844 C1517945 C0011053 C0018772 C1384666".

It is also worth to note that different phrases or words can be mapped to the same concepts. In the previous example, both "Hearing Loss" and "Deafness" will be mapped to the UMLS concept C0011053. Thus, by mapping them from terms to concept codes, different terms with the same meaning are mapped to the same concept, so the potential gap between the topic and the visit is also fixed.

With the mapping strategy described earlier, we can then represent both topics and visits as bags of codes. We then use a state of art retrieval function, i.e., language model retrieval function with two-stage smoothing [1], to retrieve relevance visits based on the new representation. The two stage smoothing method captures the influence of the query and document on the smoothing parameter setting. In the first stage, the document language model is smoothed using Dirichlet prior, while in the second stage, the query background model is used to further smooth the language model. Based on the previous experiments, the two-stage language model is chosen because it outperform than other methods.

To further improve the performance, we study how to use gender and age information to retrieve records of the patients that better match the profile given in the topic. We first start with a set of seed terms related to gender and ages, such as young, teenagers, female, he, etc, and retrieve a set of related concepts from UMLS. We then add the new terms from the related concepts to the seeds, and then submit them to UMLS again. We keep doing this until no more concepts can be found. By doing that, we got two lists of concepts about age and gender, respectively. These two lists are used to interpret the visit records. When mapping the visit records into codes, we add two new fields into the original file, which are <GENDER> and <AGE>. If the returned code is known in either list, we will put the code in the specific field instead of put them together with other codes. These two fields are used for query language in Indri[3]. If the concepts from the topic match the ones in these two fields, it will have a higher weight than it match the concepts in the remaining place. For example, given the concept C0011053, the query language we build in Indri is:

#wsum(5.0 C0011053.(GENDER) 5.0 C0011053.(AGE) 1.0 C0011053)

By giving these fields higher weight, the topics which the gender or age information is specified could be distinguished with others, and then improve the performance.

3. Experiments

We submitted 3 runs this year with the methods we described above:

UDInfoMed1: Traditional keyword-based retrieval. We used language model method with

[3] http://www.lemurproject.org/indri/

two-stage smoothing. The parameter is the default one in Indri. The Porter stemmer is also used. This run serves as the baseline method.

UDInfoMed12: Concept-based retrieval. We use the same retrieval function as in the baseline method. No stemmer is used in this run.

UDInfoMed123: Concept-based retrieval. We use the same retrieval function, and treat the concepts related to gender and ages as separate fields.

Figure 1 shows the comparison of the submitted runs with the median score of this year's runs on infAP measure. It is clear that all of our runs are above median for most topics. The same trend can be observed using the other measures.

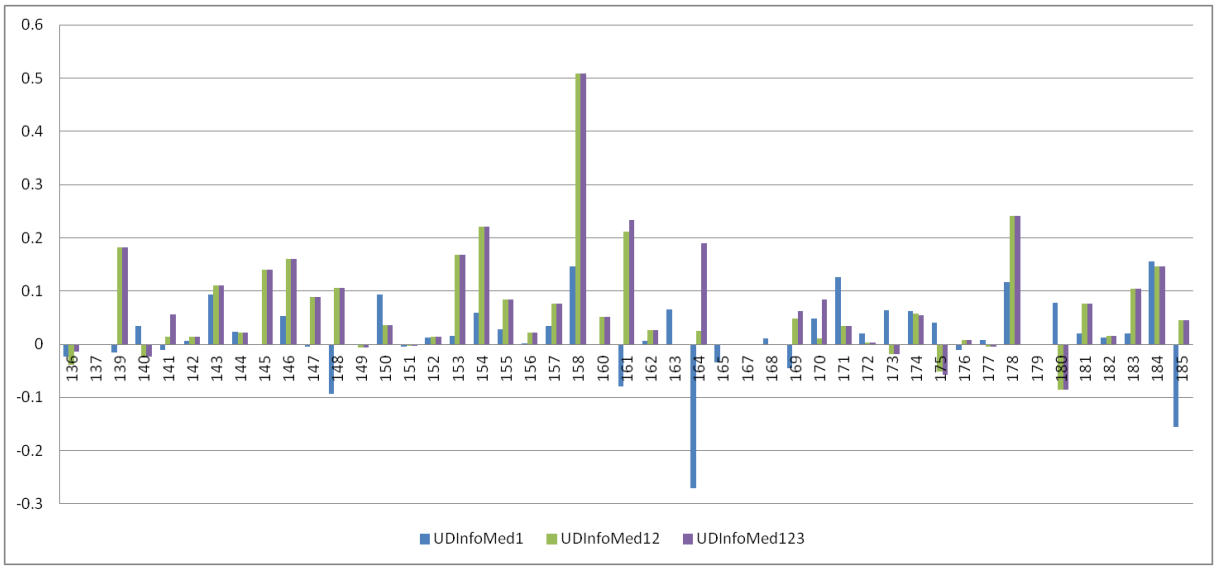

Figure 1 Improvement of each run compared with median (InfAP)

Table 1 summarizes the results of the three runs. Both the UDInfoMed12 and UDInfoMed123, which are utilizing the concept retrieval method, are better than the baseline method. Moreover, the improvement of UDInfoMed123 over UDInfoMed12 is not very significant, which may indicate that the effectiveness of representing the visits and topics using multiple fields is limited.

Table 1 Performance of each run

	infAP	**infNDCG**	**R-prec**	**P@10**
UDInfoMed1	0.1905	0.4492	0.3240	0.4872
UDInfoMed12	0.2293	0.5082	0.3618	0.5213
UDInfoMed123	**0.2364**	**0.5167**	**0.3675**	**0.5277**

4. Conclusion

To sum up, by using the concept retrieval method, the performances of submitted runs could improve the performance at least 10% on every measure. Using the tags to specify the gender and age information from other part also increases the performance. However, due to the topics of this year, the improvement is not significant.

5. Reference

1. Zhai, C. and Lafferty, J., "Two-stage language models for information retrieval," Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '02), 49-56, 2002.