Biomedical Information Retrieval

Introduction

Thousands of articles are being added into biomedical literature each year and this large collection of publications offer an excellent opportunity for discovering hidden biomedical knowledge by applying information retrieval (IR) and Natural Language Processing (NLP) technologies.

The recent statistics shows that 70% of total web search queries are of medical and healthcare category. Biomedical Information Retrieval(BIR) is a special type of information retrieval.

Literature survey

Major challenges in biomedical information retrieval are in handling complex, ambiguous medical terms and their ad-hoc abbreviations. The average length of biomedical entities is much higher than general entities. Identifying such medical entities is a preliminary subtask. Physicians use ad-hoc abbreviations very frequently and they are ambiguous like ’PSA’ can be ’prostate specific antigen’ or ’poultry science association’. Such different different representations of the same entity should be normalized to a single representation. This problem is known as entity normalization

There can be two types of users of the healthcare related search systems : experts(clinicians) and laymen(other than clinicians). The query formulations of both the users are different for the same information need. For example, general people use the words ’heart attack’, ’Irregular heartbeat’, while experts use the words ’Myocardial infarction’,’Cardiac arrhythmia’, respectively.

This leads to the problem of vocabulary mismatch where different people name the same thing or concept differently. Missing synonyms causes low retrieval recall i.e. out of all relevant documents in the collection, very few relevant documents get retrieved. Also, ambiguous terms cause low precision i.e. out of all retrieved documents, very few are relevant.

PubMed is a biomedical search engine which accesses primarily the MEDLINE database of abstracts and references on biomedical topics and life sciences and is maintained by the United States National Library of Medicine (NLM)1 at the National Institutes of Health (NIH). PubMed does binary matching[15] and is useful for short queries only.

On the contrary medical and healthcare related queries are longer than general queries since people used to describe the symptoms, tests and ongoing treatments. For verbose and longer queries, biomedical IR systems should deal properly with ambiguous, complex and inconsistent biomedical terminologies which is difficult to handle. synonymy.

Automatic query expansion (AQE) [11], [4] which has a long history in information retrieval can be useful to deal with such problems.

Query Expansion which uses the top retrieved relevant documents is known as Relevance Feedback since it uses the human judgement to identify the relevancy.

PRF

Pseudo Relevance Feedback technique assumes the top retrieved documents relevant and uses as feedback documents. Table 1 shows the results of standard retrieval, Pseudo-Relevance Feedback (PRF) based Query Expansion and Relevance Feedback (RF) based Query Expansion with BM25 [1] and In\_expC2 [1] retrieval models. Terrier tool has been used for all these experiments. MAP and infNDCG are used as evaluation metrics [10]. Higher the value of evaluation measure, better the retrieval result of system. The result improves when Query expansion is used. PRF based query expansion and RF based query expansion give statistically significant results as compared to no expansion.

Automatic query expansion methods based on pseudo relevance feedback uses top retrieved documents as feedback documents.[10] [4] Those feedback documents might not be all relevant. The feedback document set might contain non-relevant docs along with truly relevant documents. The retrieval system gets harm with these non-relevant documents in feedback set. They are like noise in the feedback system. One attempt is to learn the truly relevant documents for feedback by using minimum human intervention. The approach uses human judgements for a small set of feedback documents and then it tries to learn identifying true relevant documents from rest of the documents. Then the documents identified relevant are used for feedback and query expansion is performed. Two approaches for this learning based on classification and clustering are presented here. First Algorithm: The first proposed algorithm is based on classification. If we have human judgements available for some of the feedback documents, then it will serve as a training data for classification. The documents are represented as a collection of bag-of-words, the TF-IDF scores of the words represent features and human relevance scores provides the classes. By using this as a training data, we want to predict the relevance of other top retrieved feedback documents.

and it is responsible for insignificant improvement. This approach further removes non-relevant documents from relevant document class identified by classification approach. The idea is to perform clustering on the relevant identified documents with number of clusters two: one from actually relevant documents and second from non-relevant documents.

The experiments are performed using nine different classifiers for classification in first algorithm. The table 2 shows the results in terms of MAP score for CDS 2014 dataset. Neural-Net gives best result among all nine classifiers. Also, the result of classification with Nearest-Neighbors is comparable to the baseline.

For all the 30 queries of CDS 2014, classification Nearest-Neighbours classified 625 documents as relevant out of all 200\*30 documents. Out of 625 documents used for feedback, 244 documents were actually relevant while other 381 documents were wrongly classified as relevant. So, these 381 irrelevant documents are noise to the system. The second approach takes this matter into consideration and further refine the feedback document set by performing 2-cluster clustering on 625 documents. Manually removing 381 irrelevant documents from feedback document set shows significant improvement over baseline.