Biomedical Information Retrieval

Jainisha Sankhavara
Information Retrieval and Language Processing Lab,
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat
jainishasankhavara@gmail.com

Prasenjit Majumder
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat
prasenjit.majumder@gmail.com

ABSTRACT

Retrieving relevant information from biomedical text data is a new challenging area of research. 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. Biomedical Text processing is different from others. It requires special kind of processing as it has complex medical terminologies. Medical entity identification and normalization itself is a research problem. Relationships among medical entities have the impact on any system. The Clinical Decision Support systems are aimed to provide assistance to the decision-making tasks in biomedical domain. The medical knowledge have the potential to impact considerably on the quality of care provided by clinicians. Medical field has various types of queries: short questions, medical case reports, medical case narratives, verbose medical queries, community questioning, semi-structured queries, etc. These diverse nature of medical data demands special kind of attention from IR and NLP.

CCS CONCEPTS

• **Information systems → Query reformulation**; **Document filtering**; *Clustering and classification*;

KEYWORDS

Biomedical Text Processing, Query Expansion, Clinical Decision support

1 MOTIVATION AND CHALLENGES

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. Major challenges in biomedical information retrieval are in handling complex, ambiguous, inconsistent medical terms and their ad-hoc abbreviations.

- There are many complex terms like 'nuclear factor kappa-light-chain-enhancer of activated B cells', 'NF-kB DNA binding with electromobility shift assay'. 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 'psoriasis arthritis' or 'poultry science association'.
- The rapid change in terminologies makes them inconsistent. For instance 'H1N1 influenza', 'H1N1 Virus', 'swine influenza', 'SI', 'Pig Flu' and 'Swine-Origin Influenza A H1N1 Virus', all refers to the same entity. Such different different representations of the same entity should be normalized to a single representation. This problem is known as entity normalization. This type of problems of acronym disambiguation leads to poor system performance.

Also, 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', 'Mouth ulcer' while medical practitioners/experts use the words 'Myocardial infarction', 'Cardiac arrhythmia', 'Mucosal ulcer' respectively. This leads to the problem of vocabulary mismatch where different people name the same thing or concept differently. As an effect of the characteristics of biomedical terminologies and user dependent query formulations, the problem of vocabulary mismatch between query and documents(relevant) arises in Biomedical Information Retrieval. 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. We need to construct such Biomedical search engines that can address the above issues.

2 DATA AND RESOURCES

Widely-used text collections in the biomedical domain are MEDLINE/PubMed, OHSUMED, and GENIA.

- The MEDLINE/PubMed database contains bibliographic references to journal articles in the life sciences with a concentration on biomedicine, and it is maintained by the U. S. National Library of Medicine (NLM). This MEDLINE/PubMed records can be downloaded for research.
- The OHSUMED [14] dataset contains all MEDLINE citations of 270 medical journals published over a five-year period (1987-1991).
- The TREC Genomics Track data [8] contains ten years of MEDLINE citations (1994-2003).
- The TREC Clinical Decision Support data [12, 13, 19] is the collection of 733,138 full-text articles from PubMed Central.
- The GENIA corpus [9] contains 1,999 MEDLINE abstracts retrieved using the MeSH terms. It is annotated for part-of-speech, syntax, coreference, biomedical concepts and events, cellular localization, disease-gene associations, and pathways.
- The Unified Medical Language System (UMLS) [3], a compendium of controlled vocabularies that is maintained by NLM, is the most comprehensive resource, unifying over 100

dictionaries, terminologies, and ontologies in its Metathesaurus. It also provides a semantic network that represents relations between Metathesaurus entries, a lexicon that contains lexicographic information about biomedical terms and common English words.

3 LITERATURE SURVEY

'Information Retrieval: A Health and Biomedical Perspective' [6] provides basic theory, implementation and evaluation of IR systems in health and biomedicine. The tasks of named entity recognition and relation and event extraction, summarization, question answering, and literature based discovery are outlined in Biomedical text mining: a survey of recent progress [18]. The original conception of literature-based discovery [20] was facilitated by the use of Medical Subject Headings (MeSH), which are controlled vocabulary terms added to bibliographic citations during the process of MEDLINE indexing.

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.

Automatic processing of biomedical text suffers from lexical ambiguity (homonymy and polysemy) and synonymy. Automatic query expansion (AQE) [11], [4] which has a long history in information retrieval can be useful to deal with such problems. For instance, medical queries were expanded with other related terms from RxNorm, a drug dictionary, to improve the representation of a query for relevance estimation [5].

The emergence of medical domain specific knowledge like UMLS can contribute to the retrieval system to gain more understanding of the biomedical documents and queries. Various approaches of information retrieval with the UMLS Metathesaurus have been reported: some with decline in results[7] and some with gain in results[2]. In [2], the pseudo-relevance feedback was used for query expansion where technique where the top retrieved documents are assumed to be relevant and used as feedback to the query and retrieval is performed using expanded query.

4 BIOMEDICAL DOCUMENT RETRIEVAL

4.1 Preliminary Experiments

Query Expansion which uses the top retrieved relevant documents is known as Relevance Feedback since it uses the human judgement to identify the relevancy. Pseudo Relevance Feedback technique assumes the top retrieved documents relevant and uses as feedback documents. The Query expansion based approches for biomedical domain gives better results [16, 17].

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 ($p < 0.05$) as compared to no expansion.

4.2 Feedback Documents Discovery

Query expansion method largely rely on feedback documents and feedback terms. 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.

Algo1 : classification

For each query Q

(1) D_N - set of N top retrieved documents $\{d_1, d_2, ..., d_N\}$
(2) D_k - set of k top retrieved documents for which human judgements are available $\{d_1, d_2, ..., d_k\}$
(3) D_l - set of l=N-k top retrieved documents for which human judgements are not available $\{d_{k+1}, d_{k+2}, ..., d_N\}$
(4) D_F - set of feedback documents
(5) $D_F = \{d_i; relevance\ of\ d_i > 0, d_i \in D_k\}$
(6) Train a classifier C on D_k using relevance as a class label and generate model M_c
(7) For each document d_j in D_l, $k+1 \leq j \leq N$
(8) Predict the relevance r_j of d_j using trained model M_c
(9) If $r_j > 0$, then $D_F = D_F \cup \{d_j\}$

Second Algorithm. The second algorithm is an extension of first algorithm. The analysis of results of first algorithm shows that the feedback document set still contains some non-relevant docs

Table 1: Results of Standard Query Expansion

	CDS 2014		CDS 2015		CDS 2016	
	MAP	infNDCG	MAP	infNDCG	MAP	infNDCG
BM25	0.1012	0.1779	0.1039	0.2036	0.0371	0.1250
BM25+PRF	0.1448 (+43.1%)	0.2231 (+25.4%)	0.1650 (+58.8%)	0.2725 (+33.8%)	0.0401 (+8.1%)	0.1367 (+9.3%)
BM25+RF	0.2043 (+101%)	0.3127 (+75.7%)	0.1834 (+76.5%)	0.3034 (+49.0%)	0.0561(+51.2%)	0.1887 (+50.9%)
In_expC2	0.1167	0.1920	0.1118	0.2147	0.0445	0.1401
In_expC2+PRF	0.1483 (+27.1%)	0.2404 (+25.2%)	0.1634 (+46.1%)	0.2689 (+25.2%)	0.0606 (+36.1%)	0.1752 (+25.0%)
In_expC2+RF	0.2070 (+77.3%)	0.3431 (+78.6%)	0.1857 (+66.1%)	0.3145 (+46.4%)	0.0713 (+60.2%)	0.2118 (+51.1%)

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. K-means clustering is used with k=2. Since, the convergence of K-means clustering depends on the initial choice of cluster centroids, the initial cluster centroids are choosen as the average of relevant documents' vectors and the average of non-relevant documents' vectors from training data.

Algo2 : classification + clustering

For each query Q

(1) D_N - set of N top retrieved documents $\{d_1, d_2, ..., d_N\}$
(2) D_k - set of k top retrieved documents for which human judgements are available $\{d_1, d_2, ..., d_k\}$
(3) D_l - set of l=N-k top retrieved documents for which human judgements are not available $\{d_{k+1}, d_{k+2}, ..., d_N\}$
(4) D_F - set of feedback documents
(5) $D_F = \{d_i; relevance\ of\ d_i > 0, d_i \in D_k\}$
(6) Train a classifier C on D_k using relevance as a class label and generate model M_c
(7) $D_R = \phi, D_{NR} = \phi$
(8) For each document d_j in D_l, $k+1 \leq j \leq N$
(9) Predict the relevance r_j of d_j using trained model M_c
(10) If $r_j > 0$ then
$D_R = D_R \cup \{d_j\}$
(11) else
$D_{NR} = D_{NR} \cup \{d_j\}$
\\ D_R contains predicted relevant documents from D_l
(12) Perform K-means clustering on D_R with k=2 (relevant docs and non-relevant docs)
(13) $D_F = D_F \cup \{documents\ from\ relevant\ docs\ cluster\}$

The query expansion considers top N retrieved documents for feedback. Here, we have considered top 250 documents, from which the set of top 50 documents are used as training i.e. human judgements for top 50 documents are used in training and rest of 200 documents are taken for testing data. The relevance is predicted for those 200 documents and only relevant predicted documents are then used for feedback. The result of relevance feedback using top 50 documents is the baseline for other results. All the computed results are compared with the baseline.

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.

The classification results are not significant to the baseline results. We investigated the matter and came to know that the relevant classified documents in relevance class are not all actually relevant. The feedback document set also contains some irrelevant documents (misclassification). 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. The results of manually removing false classified documents from feedback set and automatic clustering approach are also shown in the table 2.

The same experiments are performed on CDS 2015 and 2016 datasets. The results of both the algorithms using six different classifiers are shown in tabel 3. For CDS 2015 dataset second algorithm performs better than baseline but the difference is not significant. For CDS 2016 dataset, both the algorithms perform similar to the baseline.

REFERENCES

[1] Gianni Amati, Cornelis Joost, and Van Rijsbergen. 2003. Probabilistic models for information retrieval based on divergence from randomness. (2003).
[2] Alan R Aronson and Thomas C Rindflesch. 1997. Query expansion using the UMLS Metathesaurus.. In *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 485.
[3] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl 1 (2004), D267–D270.
[4] Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)* 44, 1 (2012), 1.
[5] Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell F Loane, Bastien Rance, François-Michel Lang, Nicholas C Ide, Emilia Apostolova, and Alan R Aronson. 2011. A Knowledge-Based Approach to Medical Records Retrieval.. In *TREC*.
[6] William Hersh. 2008. *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media.

Table 2: Results of different classifiers on CDS 2014 dataset

	CDS 2014		
MAP	classification	classification + manually removing false Relevant docs	classification + clustering
Baseline (RF_50)	0.2768	0.2768	0.2768
Nearest-Neighbors	**0.2761**	**0.2815** (p = 0.048)	**0.2794** (p = 0.305)
Linear-SVM	0.2736	0.2760	0.2750
RBF-SVM	0.2736	0.2760	0.2750
Gaussian-Process	0.2736	0.2762	0.2753
Decision-Tree	0.2496	0.2788	0.2725
Random-Forest	0.2733	0.2760	0.2747
Neural-Net	**0.2790**	0.2808	0.2790
AdaBoost	0.2618	0.2806	0.2741
Naive-Bayes	0.2614	0.2792	0.2661

Table 3: Results of different classifiers on CDS 2015 and CDS 2016 dataset

	CDS 2015		CDS 2016	
MAP	classification	classification + clustering	classification	classification + clustering
Baseline (RF_50)	0.2283	0.2283	0.1456	0.1456
Nearest-Neighbors	0.2234	**0.2324** (p = 0.115)	0.1456	**0.1459** (p = 0.895)
Decision-Tree	0.2065	0.2218	0.1138	0.1370
Random-Forest	0.2130	0.2281	0.1450	0.1458
Neural-Net	0.2295	0.2299	0.1460	0.1466
AdaBoost	0.2092	0.2213	0.1255	0.1345
Naive-Bayes	0.2172	0.2269	0.1436	0.1468

[7] William Hersh, Susan Price, and Larry Donohoe. 2000. Assessing thesaurus-based query expansion using the UMLS Metathesaurus.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 344.

[8] William R Hersh and Ravi Teja Bhupatiraju. 2003. TREC genomics track overview.. In *TREC*, Vol. 2003. 14–23.

[9] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*. Association for Computational Linguistics, 1–9.

[10] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*. Vol. 1. Cambridge university press Cambridge.

[11] Melvin Earl Maron and John L Kuhns. 1960. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)* 7, 3 (1960), 216–244.

[12] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. 2016. Overview of the TREC 2016 Clinical Decision Support Track. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*.

[13] Kirk Roberts, Matthew S Simpson, Ellen M Voorhees, and William R Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track.. In *TREC*.

[14] Stephen E Robertson and David A Hull. 2000. The TREC-9 Filtering Track Final Report.. In *TREC*. 25–40.

[15] Stephen E Robertson and K Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the Association for Information Science and Technology* 27, 3 (1976), 129–146.

[16] Jainisha Sankhavara and Prasenjit Majumder. 2016. Team DA IICT at Clinical Decision Support Track in TREC 2016: Topic Modeling for Query Expansion.. In *TREC*.

[17] Jainisha Sankhavara, Fenny Thakrar, Prasenjit Majumder, and Shamayeeta Sarkar. 2014. Fusing manual and machine feedback in biomedical domain. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*.

[18] Matthew S Simpson and Dina Demner-Fushman. 2012. Biomedical text mining: a survey of recent progress. In *Mining text data*. Springer, 465–517.

[19] Matthew S Simpson, Ellen M Voorhees, and William Hersh. 2014. *Overview of the trec 2014 clinical decision support track*. Technical Report. LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD.

[20] Don R Swanson. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine* 30, 1 (1986), 7–18.