TREC 2005 Genomics Track Experiments at DUTAI

Zhihao Yang, Hongfei Lin, Yanpeng Li, Baoyan Liu and Ye Lu
Department of Computer Science and Technology, Dalian University of Technology
No 2 LingGong Road Shahekou District, Dalian 116023, China.
{yangzh, hflin}@dlut.edu.cn, lyplyp2001@sina.com, lbyzmh1980@126.com, ldiye@sohu.com

Abstract

This paper describes the techniques we applied for the two tasks of the TREC Genomics track, i.e., ad hoc retrieval and categorization tasks. For the ad hoc retrieval task, we used query expansion, different scoring strategy on different parts of Medline record (Title, Abstract, RN, MH, etc.) and pseudo relevance feedback. Our submitted run DUTAdHoc2 obtained a MAP of 0.2349. For the categorization task, our system used a SVM classifier with TFIDF term weighting scheme. In addition concept replacing and filtering methods were adopted. Two of our submitted runs (eDUTCat1 and gDUTCat1) produced a Utility score of 0.8496 and 0.572 respectively ranking third and fourth out of 46 runs submitted for the categorization task.

1. Introduction

It is well understood that biomedical knowledge is growing at an astounding pace and these vast collections of publications offer an excellent opportunity for the discovery of hidden biomedical knowledge by applying information retrieval (IR) and related technologies. To foster the IR and related research in biomedical text, the Text Retrieval Conference (TREC) launched the genomics track in 2003 [1], which attracted the largest group of participants among all the tracks.

This is the first time that DUTAI (Artificial Intelligence laboratory of DaLian University of Technology) participated in TREC genomics track. We took part in both ad hoc retrieval task and categorization task. The following sections report our proposed methods and the results for the ad hoc retrieval and categorization tasks in turn.

2. Ad Hoc Retrieval Task

2.1 Overview

This is a conventional ad hoc retrieval task targeting the biomedical literature. Participants were provided with 50 topics, and for each topic, they were required to retrieve a set of relevant documents sorted according to the estimated relevance. In the 2005 ad hoc retrieval task, topics are more structured than the mostly free-form topics from the 2004 track. Five generic topic templates (GTTs) were developed, each of which have 10 instances, for a total of 50 topics. They are showed in Table 1. In order to get participating groups started with the topics, and in order for them not to "spoil" their automatic status of their official runs by working with the official topics, 10 sample topics were provided, with two coming from each GTT. The document collection for the 2005 ad hoc retrieval task is the same 10-year Medline subset using for the 2004 track.

Table 1: Five generic topic templates.

Generic Topic Type	Topic	Example Sample Topic
Find articles describing standard methods or protocols for doing some sort of experiment or procedure	100-109	Method or protocol: GST fusion protein expression in Sf9 insect cells
Find articles describing the role of a gene involved in a given disease	110-119	Gene: DRD4 Disease: Alcoholism
Find articles describing the role of a gene in a specific biological process	120-129	Gene: Insulin receptor gene Biological process: Signaling tumorigenesis
Find articles describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease	130-139	Genes: HMG and HMGB1 Disease: Hepatitis
Find articles describing one or more mutations of a given gene and its biological impact	140-149	Gene with mutation: Ret Biological impact: Thyroid function

2.2 Methods

Framework

Figure 1 depicts the overview of our retrieval system. To retrieve a set of relevant documents for each topic, the following step process is performed: NP extraction, query expansion, search and ranking, re-ranking and pseudo-relevance feedback.

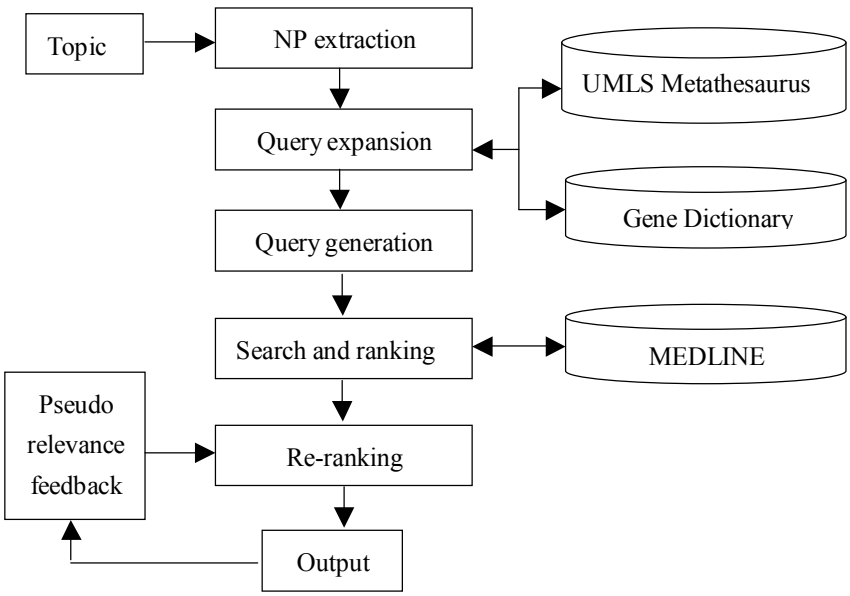

Figure 1: Framework of our IR system.

The following paragraphs describe the components of our IR system in more detail.

NP extraction

This process is accomplished by manual. Given a topic, we first extracted noun phrases (NPs) from it as potential query terms. For the verbs in the topic, if possible, they were transformed into corresponding noun phrases.

Query expansion

The extracted NPs were then expanded using two sources of information to form the final query. Firstly, a gene name dictionary was consulted to find synonyms. The dictionary is compiled from the Entrez Gene database on Pubmed（http://www.ncbi.nih.gov /entrez /query.fcgi）. Every record includes a gene/protein name and its aliases. Gene and protein names follow few, if any, true naming conventions and are subject to great variation in different occurrences of the same name. For example, the protein name "Interferon-beta" has many spelling variants such as "IFN-beta," "IFN-B," and "beta interferon." Experiments showed that query expansion by via of gene name dictionary could improve recall rate greatly.

The second source of information used in query expansion is UMLS Metathesaurus [2]. The purpose of NLM's Unified Medical Language System (UMLS) is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. There are three UMLS Knowledge Sources: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. The Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health-related concepts, their various names, and the relationships among them. In test topics, there are some disease names, each of which may have some synonyms. For example, "Alzheimer's disease" has synonyms such as "Alzheimer disease," and "AD." By via of UMLS Metathesaurus, the diseases' synonyms were found and used for query expansion.

In addition, other dictionaries were built to perform query expansion. The words expressing method or protocol (such as method, protocol, approach, and technique) were collected in a dictionary, which was used for query expansion in topics 100-109. The words expressing interactions between two or more genes (such as inhibit, suppress, promote, regulate) and their nominal forms were collected in another dictionary, which was used for query expansion in topics 130-139.

Search and ranking

In this process, terms obtained through query expansion were concatenated by Boolean OR operators, forming the final query. The query was then fed to Zettair [3] to retrieve a list of candidate abstracts from the corpus, which were used as baseline. We limited the number of retrieved documents to the first 8000 in our test runs. The ranking score produced by Zettair was denoted as score1.

Re-ranking

Zettair doesn't take account of the query terms' position (such as in TI, AB, MH or RN fields) in a Medline record. In fact these different positions can influence a Medline record's relevance. For example, a Medline record in which a query term appears in TI field tends to be more related than one in which a query term appears in AB field.

Therefore, we designed a set of scoring rules to re-rank the baseline records. Firstly, query terms were divided into two kinds: Necessary term and Optional term. Necessary terms are those that are necessary in related Medline records, i.e., they are directly related and must appear in related records; while Optional query terms are those that are optional in related Medline records, i.e., they are related but not necessary to appear in related records. The examples of Necessary terms and Optional terms are showed in Table 2. Secondly, these two kinds of terms were given different ranking scores according to their positions in Medline records. If there are no Necessary terms found in any of TI, AB, MH or RN fields in a record, the record will be given a score of -200. These scoring rules are showed in Table 3.

In this way, the baseline records were re-ranked and the ranking score produced by our scoring rules was denoted as score2. The final score of a related record is the sum of score1 and score2.

Table 2: Necessary term and optional term.

Topic	Necessary	Optional
131	L1, L2, HPV11	Virus, viral, capsid
141	Huntington, mutation	Role

Table 3: Scoring rules.

	Title	Abstract	MeSH, RN	No Found
Necessary	20	5	10	-200
Optional	10	4	6	0

Pseudo-relevance feedback

For the submitted run DUTAdHoc2, we applied a simple pseudo-relevance feedback (PRF) method: The pseudo-relevance feedback module assumed the top *n* ranked documents to be relevant and used MeSH terms (in MH fields) in these documents to refine the query. As a measure of significance, we used TFIDF values. The *m* MeSH terms with highest TFIDF were sent back to the re-ranking module to add the records including them with a score of 10. We experimentally set $n=m=10$.

2.3 Results

We submitted two runs (DUTAdHoc1 and DUTAdHoc2) for this task; only the difference between them is that DUTAdHoc2 used pseudo-relevance feedback, while DUTAdHoc2 did not. DUTAdHoc1 obtained a MAP of 0.2344, while DUTAdHoc2 obtained a little better one of 0.2349.

Table 4: Ad hoc task results.

Run	MAP	R-Prec	B-Pref	P@10	P@30
Baseline	0.1457	0.1571	0.6242	0.2469	0.2184
DUTAdHoc1	0.2344	0.2718	0.6625	0.402	0.3163
DUTAdHoc2	0.2349	0.2678	0.6616	0.3939	0.315

Through the experiments, we observed that:

- Query expansion helped a lot to improve the retrieval recall rate.
- The position that query terms appear influenced a Medline record's relevance.
- Our pseudo-relevance feedback method contributed little to the retrieval performance. We will introduce more complex pseudo-relevance feedback methods in the future.

3. Categorization Task

3.1 Overview

This year, categorization task was divided into four subtasks: allele, expression, GO and tumor. We participated in all the four subtasks and submitted two runs for each subtask. Our system used a SVM classifier with TFIDF term weighting scheme. A concept replacing approach was used in our first runs of the allele, expression and tumor subtasks, which proved to enhance the performance slightly by the official result from TREC. In GO subtask, documents were classified three times using different feature selection schemes each time, and the results were processed using a decision algorithm. Finally, the positive instances obtained were filtered according by the numbers of biomedical named entities appearing in it, which improved the normalized utility by 5% in GO subtask.

3.2 Methods

Text Processing

The documents provided by TREC were full-text articles in SGML format. We converted them into plain texts, by removing all the SGML tags and replacing non-English characters (e.g., &agr;) by corresponding strings (e.g., alpha) that should appear in the Medline record (http://www.ncbi.nlm.nih.gov/entrez/query/static/entities.html). We also downloaded the documents' corresponding Medline records from Pubmed.

Framework

Different approaches were applied in the four subtasks, while the general step process can be described as: concept replacing, feature selection, training, classifying, decision and filtering. The final two steps were performed only in GO subtask. Figure 2 depicts the overview of our categorization system.

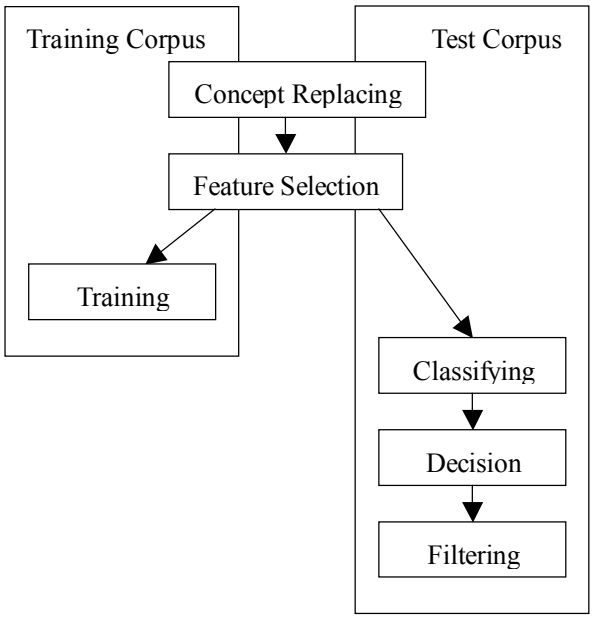

Figure 2: Framework of our categorization system.

Concept Replacing

Articles in the training and test corpus were papers of biomedicine, so there were a large number of named entities in them such as protein and DNA names. However they were not good features for classifier, because many of them appeared very few times in all the articles and some had aliases. So we used ABNER (A Biomedical Named Entity Recognizer [4]) to find proteins, DNAs, RNAs, cell lines and cell types in all the documents and then replaced them by concept names. We defined five concept names: CONCEPT_PROTEIN, CONCEPT_DNA, CONCEPT_RNA, CONCEPT_CELLLINE, CONCEPT_CELLTYPE. For example, the named entity "type-II transmembrane protein" was replaced by "CONCEPT_PROTEIN". In our submitted runs, this method was used in aDUTCat1, eDUTCat1 and tDUTCat1, but not in aDUTCat2, eDUTCat2 and tDUTCat2. From Table 1, we can see that the former group performed a little better than the latter. In addition, the numbers of feature terms were reduced by 20%.

Feature Selection, Classifying and Decision

We applied Joachims' SVM^{light} [5] classifier and set weights using TFIDF [6] scheme. Methods of feature selection in GO subtask were different from other subtasks. First, full-text articles that had been processed before were divided into the following parts: titles (denoted by f1), abstracts (f2), bodies (f3), subtitles (f4), references (f5) and MeSH (f6). Then we did a great deal of experiments, in which single part or multiple parts were selected as feature part(s).

We found that MeSH (f6) performed much better than any other kind of features, as was mentioned in last year's papers [7]. Furthermore, we processed MeSH terms in several different ways, and found that using the main headings only produced the best result. For example, for the

MeSH term “Ataxia Telangiectasia/*genetics/metabolism”, we used only “Ataxia Telangiectasia” and discarded all its subheadings. To make full use of full-text articles, we classified the documents three times respectively using f6, f1+f2+f3 (join f1, f2 and f3), and f1+f2 as features and got three results, denoted by r1, r2 and r3. We defined final score as the following Equation (1):

$$final\_score = (score1 - thres1) * 1.0 + (score2 - thres2) * 0.7 + (score3 - thres3) * 0.5 \quad (1)$$

where score1,score2 and score3 were the result scores of r1, r2 and r3 given by the SVM classifier, and thres1,thres2 and thres3 were the threshold of the above results, which were assigned -1. Their weights were 1.0, 0.7 and 0.5. One document was judged as a positive instance only if its final_score bigger or equal to 0, otherwise, it was judged as a negative instance. This procedure was denoted by d (f6, f1+f2+f3, f1+f2).

Our approaches in allele, expression, and tumor subtasks were almost the same. Titles, abstracts and bodies were selected as feature parts to represent the documents (denoted by f1+f2+f3). However our official run tDUTCat2 use titles and abstracts only, but achieved a normalized utility of 0.8807, which was almost equal to tDUTCat1. It indicated that in this subtask titles and abstracts contained enough information for classification. Decision algorithm was not used in these three subtasks, for MeSH terms didn’t perform as well as did in the GO task.

Filtering

In GO subtask, after extracted the five types of named entities, we computed the numbers of protein and DNA names that appeared in each article of the training set. We found that this number in the positive instances was much higher than that in the negatives. In the preceding stage, a large number of positive instances were obtained through the SVM classifier and decision algorithm. If we filter out the instances including less numbers of protein and DNA names from positive instances, the recall rate may be lost while the precision rate could be significantly improved.

In our run aDUTCat1, we sorted these instances twice (in an descending ordering), firstly, by the rate of protein names (number of proteins / length of the article) and secondly, by the final scores given by the decision algorithm. Instances that had ranked in the first 1/3 both times were removed from positive group as negative instances. As can be seen from Table 5, using this method, the normalized utility increased from 0.5428 to 0.5720 in GO subtask. Among the 214 instances that were filtered out, only four were true positive instances and the rest were negative ones.

3.3 Results

We participated in all the four subtasks and submitted two runs for each subtask. From Table 5 and Table 6, we can see most our submitted runs were above the median, and the first group of results (aDUTCat1, eDUTCat1, gDUTCat1 and tDUTCat1) was better than the second one (aDUTCat2, eDUTCat2, gDUTCat2 and tDUTCat2). It suggests that using Named Entity Recognition (NER) technique is an effective way to improve the performance of biomedical

Table 5: Results of our official runs.

subtask	runID	feature	Concept Replacing	Filtering	Precision	Recall	Unorm
allele	aDUTCat1	f1+f2+f3	Yes	No	0.2858	0.9307	0.7939
allele	aDUTCat2	f1+f2+f3	No	No	0.2620	0.9217	0.7690
expression	eDUTCat1	f1+f2+f3	Yes	No	0.2383	0.9429	0.8496
expression	eDUTCat2	f1+f2+f3	No	No	0.1104	0.9429	0.8241
GO	gDUTCat1	d(f6,f1+f2+f3,f1+f2)	No	Yes	0.1914	0.9286	0.5720
GO	gDUTCat2	d(f6,f1+f2+f3,f1+f2)	No	No	0.1779	0.9363	0.5428
tumor	tDUTCat1	f1+f2+f3	Yes	No	0.0745	0.9500	0.8989
tumor	tDUTCat2	f1+f2	No	No	0.0350	1.0000	0.8807

Table 6: Results of all runs.

subtask	Unorm (Best)	Unorm (Median)	Unorm (Worst)	Unorm (Our best)
allele	0.8710	0.7785	0.2009	0.7939
expression	0.8711	0.6548	-0.0074	0.8496
GO	0.5870	0.4575	-0.0342	0.5720
tumor	0.9433	0.7610	0.0413	0.8989

document classification. Different from newswire domain, biomedical articles contain more named entities, which play an important role in document classification. So how to recognize and handle these entities is a key problem.

4. Conclusion

This is the first time that DUTAI participated in TREC genomics track. We took part in both ad hoc retrieval task and categorization task. For the ad hoc retrieval task, we used query expansion, different scoring strategy on different parts of Medline record and pseudo relevance feedback. Our submitted run DUTAdHoc2 obtained a MAP of 0.2349. We found that query expansion helped a lot to improve the retrieval recall rate and the position information of query terms could influence a Medline record's relevance. For the categorization task, our system used a SVM classifier with TFIDF term weighting scheme combined with concept replacing and filtering methods. Most our submitted runs were above the median, and the first group of results (aDUTCat1, eDUTCat1, gDUTCat1 and tDUTCat1) was better than the second one (aDUTCat2, eDUTCat2, gDUTCat2 and tDUTCat2). It suggests that using Named Entity Recognition (NER) technique is an effective way to improve the performance of biomedical document classification.

Acknowledgement

This work is supported by grant from the Natural Science Foundation of China (60373095).

References

[1] William Hersh. Report on TREC 2003 genomics track first-year results and future plans. *SIGIR Forum*, 38(1): 69–72, 2004.

[2] Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med, 32(4): 281 – 291, 1993.

[3] The Zettair Search Engine. http://www.seg.rmit.edu.au/zettair/

[4] Burr Settles. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics, 21(14): 3191-3192, 2005

[5] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[6] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5): 513–523, 1988.

[7] Kazuhiro Seki. TREC 2004 Genomics Track Experiments at IUB. NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004).