DUTIR at TREC 2006: Genomics and Enterprise Tracks

Zhihao Yang, Hongfei Lin, Yanpeng Li, Linhong Xu, Yu Pan and Baoyan Liu
Department of Computer Science and Technology, Dalian University of Technology
No 2 LingGong Road Shahekou District, Dalian 116023, China.
{yangzh, hflin}@dlut.edu.cn, lyp_8218@163.com, qingniao1203@163.com,
panyu_yu@yahoo.com.cn, lbyzmh1980@126.com

Abstract

This paper describes the techniques we applied for the two TREC 2006 tracks, i.e., Genomics and Enterprise track. For the Genomics Track, we used a Rocchio relevance feedback method to expand the terms and then performed passage retrieval by building dual index and using half overlapped windows passages. Several approaches to merge the results and rerank the passages are presented. For the Enterprise track, we stripped the non-letter character from documents and query, built the index by indri or lemur and established expert document pools.

1. Introduction

This is the second time that DUTIR (Information Retrieval laboratory of DaLian University of Technology) participated in TREC tracks. This time we took part in Genomics and Enterprise tracks.

This year's Genomics Track has a new single task that focuses on retrieval of passages (from part to sentence to paragraph in length) with linkage to the source document. For most information seekers, especially users of the biomedical literature, desire is a system that attempts to answer questions but put them in context while providing supporting information and linking to original sources. There are three levels of MAP used to measure the retrieval performance: passage level, aspect level, and document level.

Passage level - Precision will be computed as the fraction of characters overlapping with the gold standard passages divided by the total number of characters included in all nominated passages from this system for the topic up until that point.

Aspect level –This measure is used to normalize passages on the same answer, for most users prefer passages with different aspects. For each submitted run, the ranked passages will be transformed to two types of values, either the aspect of the gold standard or not-relevant.

Document level - Any PMID that has a passage associated with a topic ID in the set of gold standard passages will be considered a positive document for that topic.

Most of 2006's topics are derived from 2005's. For the topics that are the same we applied a Rocchio relevance feedback [1] method to expand query terms. Also in our run DUTgen3 a SVM classifier trained by 2005's gold standard was used to rerank the passages. According the 2006 track protocol all our three runs should be classified as interactive runs. We also experimented with 4 different kinds of passage ranking schemes.

As to Enterprise track we participated in both the Discussion task (DS) and Expert task (EX). The discussion search task is to search for messages pro and con in an argument or discussion regarding to a topic and the expert search is to look for a person or multiple people who were experts on a subject.

The following sections report our proposed methods and the results for Genomics and Enterprise tracks in turn.

2. Genomics Track

2.1 Preprocessing

The documents for this task are full-text biomedical corpus in HTML format which come from Highwire Press (http://www.highwire.org/). Our first step was to remove all the HTML tags and some other sections we thought that should not appear in the final retrieved passages including titles, authors and organizations, keywords, all texts within the HTML tags "<TABLE" and "</TABLE", acknowledgements and references etc.

Many gene names and other biomedical named entities contain Greek letters or other non-English tokens, while most of that letters in this corpus appear as pictures. For example, the Greek letter "α" in the HTML text is denoted as the following labels: <IMG SRC="/math/alpha.gif" ALT="{alpha}">. We replaced all the pictures in the ""/math/" directory by the tokens in the "ATL" fields. Again some HTML tokens in the format like "&#⋯;" are replaced by corresponding strings that should appear in the MEDLINE record (http://www.ncbi.nlm.nih.gov/entrez/query/static/entities.html). As is described in the papers of the previous year's participants[2][3], when using single terms as query, removing dash and some other tokens in the data set will enhance the retrieval performance. So we replaced all the non-digit and non-alphabetic tokens in the data collection by a white-space. At the same time we used simple rules to recognize the boundary of a sentence and tagged the offset of a paragraph and its every sentence.

2.2 Query Expansion

The topics for the 2006 track are expressed as questions. First, we extracted noun phrases from each question as initial query terms by using GENIA Tagger [4]. Then we have tried some query expanding scheme on 2005 track data using some biomedical databases such as Entrez Gene [5] and UMLS Metathesaurus [6], but we didn't find an effective automatic way to filter the "noizy" terms induced by synonyms expansion. But we found that using Rocchio relevance feedback based on 2005's gold standard lead to significant improvement of MAP of all topics. This year's topics are mostly identical to last year's, so for all topics from 160 to 187 except 177 and 180, a Rocchio feedback method was used to expand the terms. For every topic we selected top 20 most relevant terms by its score as expanded query in our submitted runs.

2.3 Indexing and Retrieval

Passages in our experiment are defined as follow:

Passage1 = (Part1, Part2)   Passage2 = (Part2, Part3)

Where Passage1 is a passage which consists of two parts: Part1 and Part2, and its following passage is denoted as Passage2 which is half overlapped with Passage1. A Part is composed of complete sentences:

Part = (sentence1, sentence2, ⋯, sentenceN)

WordsCount (sentence1, sentence2, ⋯, sentenceN-1) < PartLength

WordsCount (sentence1, sentence2, ⋯, sentenceN) >= PartLength

Where WordsCount () is the number of words of all the sentences in one Part. For PartLength we set 30 in our submitted runs. After the results and evaluation tools were distributed we found this length was larger than the average length of passages in the gold standard.

Previous research [7] has shown that when documents are very long, methods based on passage-level retrieval can give much higher document-level MAP than document-level retrieval. Our retrieval method is also inspired by the work of York University in 2004 HARD Track [8], which built two levels of index and combined the two results into one. In our experiment, we built two types of index with Indri [9]: paragraph-level and passage-level. Porter stemmer and Indri's stop word list were used. Each of the two types of index was ranked respectively with BM25 algorithms which parameters are adopted from Lemur's [9] default setting. Then we merged the results into one using four different methods:

Method1 – Paragraph-first scheme: paragraphs were ranked by their BM25 scores, and for every paragraph we chose the passage with the highest score。

Method2 – Passage-first scheme: rank passages according to passage scores. If two passages were overlapped, the one with higher score was selected as final result.

Method3 - Combining scheme: we combined the passage and paragraph score by giving weights to them as the following function:

S = W paragraph * S paragraph + W passage * S passage

Where W paragraph, W passage are the weights of passage and paragraph score which were set 3 and 1 separ in our submitted run DUTgen2. Then we ranked passages by the final score S, and chose passages with higher score for overlapped passages.

Method 4 - SVM reranking: in this experiment, we treated the task as a binary text classification problem. Documents of each topic were classified into two classes: relevant or irrelevant. Training data was last year's gold standard and classifier was SVMlight [10] with TFIDF term weighting scheme. First we selected top 2000 paragraphs by Method1, and then the paragraphs were reranked by the classifier. The method of passage extraction is the same as Method1.

2.4 Results

Table1: Performance of official and unofficial runs

Method	Passage MAP	Aspect MAP	Document MAP	Run ID
Method1	0.07066621	0.18569347	0.36335342	DUTgen1
Method2	0.05491280	0.15559669	0.30699291	DUTgen4 (unsubmitted)
Method3	0.07302423	0.16477437	0.36005838	DUTgen2
Method4	0.04467985	0.13790016	0.29021274	DUTgen3

From Table1, we can see that Method1 and Method3 have better overall performances, which indicate that paragraph-first ranking (Method1) is more effective than passage-first ranking (Method2) in each of the three measures, while by combining the two results (Method3), we can get an improvement in Passage MAP but a decrease in both Aspect MAP and Document MAP. In our run DUTgen3 the weights for paragraph and passage are 3 and 1. From the results, it can be seen that the paragraph weight we have chosen is somewhat large, so there is not much difference between DUTgen1 and DUTgen2. Performance of Method4 is much lower than Method1 and Method3. That is what we have not expected. There is a large gap between its performances in the training data (2005 track data) and implement data. It indicates that for retrieval tasks it is difficult to get a higher precision applying a categorization method than the state-of-art searching algorithms such as BM25.

3. Enterprise track

3.1 Preprocessing

As in TREC 2005, TREC 2006 still used the W3C collection. The collection is a crawl of the public W3C (World Wide Web Consortium) sites in June 2004. It comprises 331,037 documents, retrieved via multithreaded breadth first crawling [11]. The collection contains six different types of web pages which were lists (emails), dev, www, esw, other, and people. The discussion search utilizes the emails, while the expert search utilizes the whole collection.

The documents provided by TREC are full-text articles in HTML format. To improve performance, we chose the cleaned W3C collection provided by Daqing He, who repackaged Wu's parsed w3c-lists collection, and also cleaned w3c-www, w3c-esw and w3c-people collections by removing HTML tags [12]. Based on this collection, we cleaned the collection further including removing the special character, such as “–”, “/,” etc. This could be superior to matching the “if-else” and the “if else”. Moreover, considering the encoding of the text, we parsed documents in ISO-8859-1, which could promise some non-English, such as “é”, to be identified correctly.

3.2 Email Discussion Search

This task is to search some emails which contain a discussion about the topic, and emails could have pro or con point about given topic. The emails, about 198,275, are utilized in this task.

Overview

Firstly, we preprocessed the cleaned W3C collection, based on which an index was built by indri or lemur [13]. Then we processed the query topic the same way as cleaning the documents, i.e. Stripping the special character and stopping word. Ultimately, relevant documents were retrieved by indri or lemur. Figure 1 depicts the framework of our retrieval system.

Discussion runs

For the discussion task, we did some experiments on TREC 2005 discussion topic, changing different ranking method and stem. Through experiments, we found the following facts: Firstly, stripping the stop-word from the field #title of the query directly was superior to composing the query by bigram method obviously. Secondly, it made results more precise to appending the field #narrative for query. Thirdly, removing non-letter character from query grew 1 percent. Fourthly, by stripping non-letter character from documents, it had a 3 percent increase in results. Finally, BM25 was superior to other ranking methods.

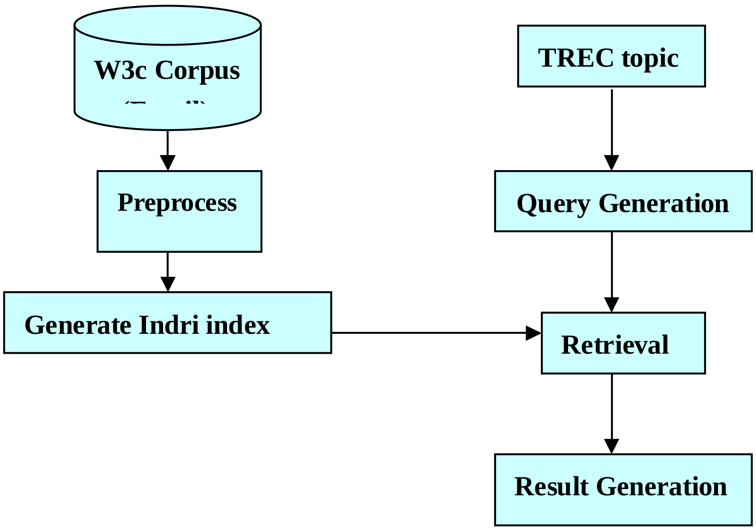

Figure 1: Framework of DS track IR system.

Results

The detail information of our four submitted runs is displayed in Table 2 and the results of these runs are displayed in Table 3.

Table 2: Detail information of four runs

Run ID	Index type	query	Ranking method	remark
DUTDS1	Indri	Title	BM25	Auto
DUTDS2	Indri	Title, narrative, description	BM25	Manual
DUTDS3	Indri	Title	Indri	Manual
DUTDS4	Indri	Title	Indri	Auto

Table 3: Results for Discussion Search

Run ID	MAP	R-prec	Bpref	Reciprocal rank	p@10
DUTDS1	0.2252	0.2603	0.2390	0.4963	0.3087
DUTDS2	0.2166	0.2501	0.2334	0.4837	0.2913
DUTDS3	0.2808	0.3110	0.2958	0.6483	0.4022
DUTDS4	0.2714	0.3066	0.2856	0.5433	0.3826

In Table 3, the first column is the run identifier, the second column is the mean average precision (MAP), other columns display other important factor. In terms of the MAP measure, DUTDS2 (whose query text was taken from title field, narrative field and description field.) is the lowest. DUTDS3 which ranked by Indri increased the MAP by about 7% over DUTDS1 (ranked by BM25), which is contrary to the results we obtained on the training topics (TREC 2005 discussion topics). We will explore the reasons in the next step.

3.3 Expert Search

In this task participants should retrieve a list of candidate experts on a subject. This year the topics and relevance judgments are created by the participant and all 331,037 documents can be used.

Overview

Based on the cleaned W3C collections, we created a correlative document pool for each candidate. We gained the expert list and the support document with the pool. The framework of our approach is depicted in figure 2. We collected the identity for all the

1092 candidates, including name, email, nick, phone, homepage and so on.

Correlative document pool generation

Firstly, we collected the identities of each candidate, including his name, email, phone, nick, personal main page and so on [14]. There were two stages in this process: automatic and manual. In the automatic stage we made several rules for identity extraction combining the technique of named identity recognition, then adjust and recruit the result in the manual stage.

After candidate identity extraction was finished, an index was built based on the cleaned W3C collections and utilized the candidate identities to query. We singled out a number of words around the candidate identity to form the correlative document pool. Using this method, a pool of 1092 correlative documents was built. We experimented with different number of word: 10, 50, 100, 200, and 300 and found that the performance was best when the number was 200.

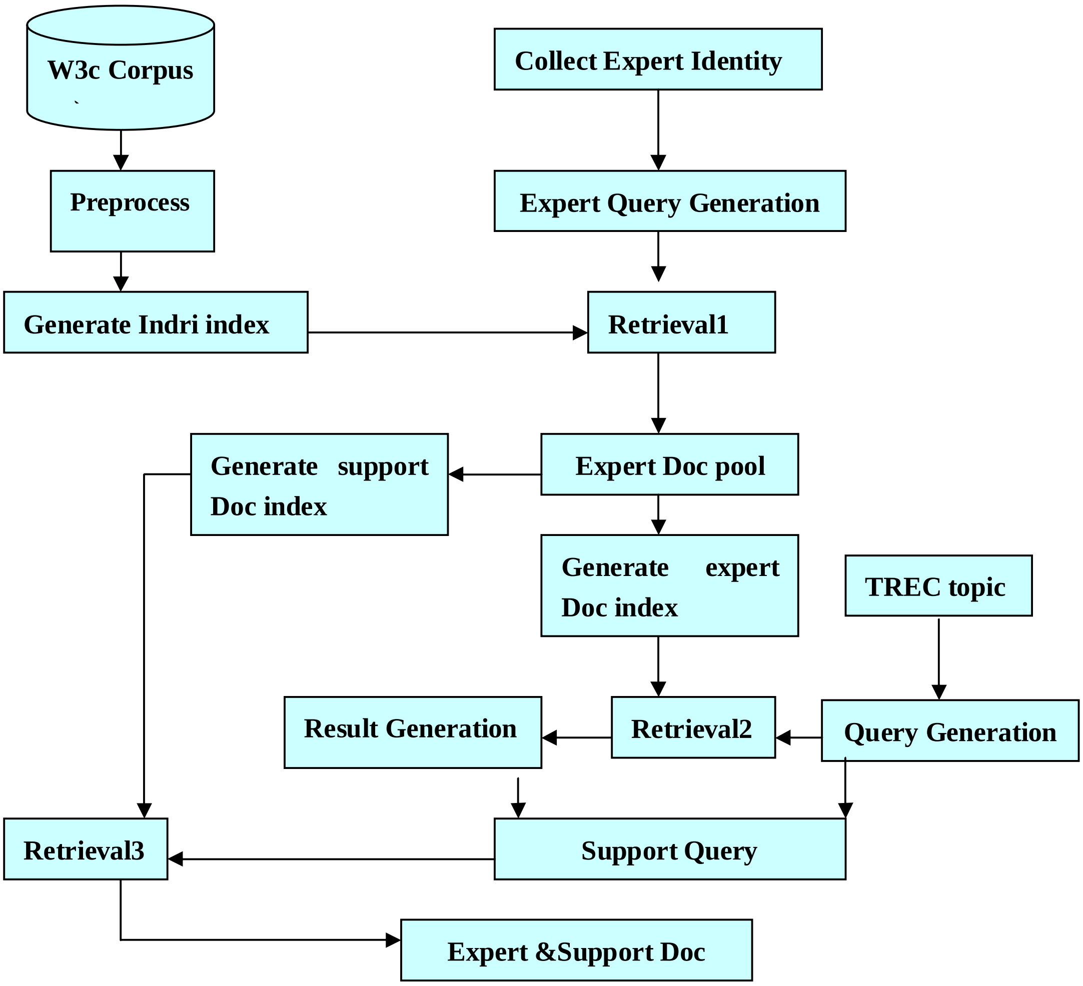

Figure 2: Framework of ES track IR system.

Expert list and supporting document generation

In this process, an index was built based on the correlative pool firstly. We attempt to compose the query in several ways for each topic and introduced the query to the indri. The expert list was gained through the retrieved indri score.

Different from last year, every retrieved expert should be provided with corresponding supporting documents which can explain why the candidate is an expert in this subject. Accordingly, we dealt with the correlative document pool. We took the "document ID-candidate ID" as the supporting document ID, in this way the correlative document pool of a candidate was divided into some supporting documents [15]. Then we added the candidate identities to the original query and utilized indri to gain the supporting documents of the expert.

Results

The detail information of our four submitted runs is displayed in Table 4 and the results of these runs are displayed in Table 5.

Table 4: Detail information of four runs

Run ID	Index type	query	Ranking method	Words number
DUTEX1	Indri	Title	Indri	200
DUTEX2	Indri	Manual	Indri	200
DUTEX3	Indri	Title	Indri	50
DUTEX4	Indri	Title, Narrative	Indri	200

Table 5: Results for Expert Search

Run ID	MAP	R-prec	Bpref	Reciprocal rank	p@10
DUTEX1	0.3033	0.3343	0.3205	0.6007	0.4184
DUTEX2	0.3779	0.4175	0.4077	0.8094	0.5184
DUTEX3	0.3267	0.3662	0.3637	0.6931	0.4857
DUTEX4	0.2834	0.3392	0.3953	0.4430	0.3796

In Table 5, the first column is the run identifier, the second column is the mean average precision (MAP), other columns display other important factor. In terms of the MAP measure, DUTEX3 is better than DUTEX1. The only difference between them is the number of words in correlative document pool. We can see that it is better when the number is 50. The performance is not consistent with the results obtained on the training topics (TREC 2005 Expert topics). DUTEX2 gains the best result in all the runs since we modified its queries by manual. So we can conclude that it is effective to apply manual interfere in the process.

4. Conclusion

In TREC 2006 we took part in Genomics and Enterprise tracks. For Genomics track, due to insufficient of training data, our methods and parameters of the experiment are mostly chose empirically. In the future, we should focus on the method that retrieves passages with more variable length. In addition, syntactic parsing, domain specific knowledge and machine learning approaches will be used to enhance the retrieval performance.

For Enterprise track, we found that structured information, such as thread structure, was not useful in the discussion search, and data preprocess such as removing special characters from W3C collection increased the MAP by about 3%. Expert search task is different from the traditional search problem. To resolve this problem, a new method which we called it correlative document pool, was applied and the result indicates the effectiveness of this method. In addition, there are many pronouns in the document and they are important to identify the expert. Therefore, it may help to improve the performance by introducing anaphora resolution technology.

Acknowledgement

This work is supported by grant from the Natural Science Foundation of China (60373095). In addition, we are grateful to Maoshu Ni, Baodong Wu, Baojin Cui for their contributions to our work.

References

[1] Allan J., Incremental Relevance Feedback for Information Filtering. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich, 270-280. 1996.

[2] Tzong-Han Tsai, Chia-Wei Wu. Enhance Genomic IR with Term Variation and Expansion: Experience of the IASL Group at Genomic Track 2005. Proceedings of the 14th Text Retrieval Conference, 2005.

[3] Rie Kubota Ando, Mark Dredze. TREC 2005 Genomics Track Experiments at IBM Watson. Proceedings of the 14th Text Retrieval Conference, 2005.

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[5] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

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

[7] Liu X, Croft W B. Passage Retrieval Based on Language Models. In Proceedings of the 11th International Conference on Information and Knowledge Management. ACM Press, 2002:375-382.

[8] X. Huang, Y.R. Huang, M.Wen and M.Zhong. York University at TREC 2004: HARD and Genomics Tracks. Proceedings of the 13th Text Retrieval Conference, 2004.

[9] J. Allan, J. Callan, K. Collins-Thompson, B. Croft, F. Feng, D. Fisher, J. Lafferty, L. Larkey, T. N. Truong, P. Ogilvie, L. Si, T. Strohman, H. Turtle, and C. Zhai. The lemur toolkit for language modeling and information retrieval. http://www.cs.cmu.edu/~lemur/, 2003.

[10] 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.

[11] Nick Craswell, Arjen P. de Vries, Ian Soboroff. Overview of the TREC-2005 Enterprise Track. Proceedings of the 14th Text Retrieval Conference, 2005.

[12] http://www.ins.cwi.nl/projects/trec-ent/wiki/index.php/Main_Page

[13] http://www.lemurproject.org

[14] Yupeng Fu, Wei Yu, Yize Li, Yiqun Liu, Min Zhang, Shaoping Ma. THUIR at TREC 2005: Enterprise Track. Proceedings of the 14th Text Retrieval Conference, 2005.

[15] Conglei Yao, Bo Peng, Jing He, Zhifeng Yang, CNDS Expert Finding System for TREC2005. Proceedings of the 14th Text Retrieval Conference, 2005.