TREC 2005 Genomics Track at I2R

Nie Yu, Yang Lingpeng, Ji Donghong, Zhang Jie,
Su Jian, Yang Xiaofeng, Soon-Heng Tan, Xiao Juan, Zhou Guodong
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
{ynie,lpyang,dhji,zhangjie,sujian,xiaofengy,soonheng,stuxj,zhougd}@i2r.a-star.edu.sg

Abstract

This paper describes the methods we used for the Ad Hoc task of TREC Genomics Track. Synonym dictionary for genes and pseudo relevance feedback are used to expand queries. BM25 model is implemented to retrieve relevant documents. We also tried to exploit name entities and their co-references in the retrieval. Results of submitted runs are listed and discussed.

1 Introduction

The enormous amount of biological literature makes the strong expectation of efficient retrieval ways for biological information. This motivated various research on information retrieval from large scale of information or corpus. The Text Retrieval Conference (TREC) provides a platform for testing and experiments of retrieval methods. In this year, the genomics track of TREC prepares two tasks. One consists of ad hoc retrieval, while the second involves text categorization. We attend the ad hoc retrieval task.

2 Ad Hoc Retrieval Task

In Ad Hoc task of TREC 2005 Genomics Track, topics are developed from generic templates to provide systems with better defined queries for finding genomics information. Five generic topic templates (GTTs) are developed derived from an analysis of the topics from the 2004 track and other known biologist information needs, each of which has 10 instances, for a total of 50 topics. Following is an example for a GTT and an instance of it:
GTT: *Find articles describing the role of a gene involved in a given disease.*
Instance: *Find articles describing the role of `Interferon-beta` involved in `Multiple Sclerosis`.*

The document collection for this task was a 10-year (1994-2003) subset of the MEDLINE bibliographic database of the biomedical literature with a total of 4,591,008 records. Each record consists of many fields. We used fields of abstract (AB), title (TI), and MeSH Terms (MH) only for this task.
The target of this task is to submit a ranked list of documents for each topic. Relevance judgments are done by TREC organizer using the similar pooling method with TREC 2004 Genomics Track. Topic ranking documents from attending groups' runs will be

pooled and given to biologists to make judgments. According to how to generate queries from topics, runs are grouped into "automatic", "manual" and "interactive". We submitted two automatic runs for the task of this year.

3 Methods

We use a gene synonym dictionary and pseudo relevance feedback to expand queries. Okapi BM25 [1][2] is implemented to retrieve relevant documents. Single words and previously extracted terms are used with BM25 method. We also tried to exploit name entities and their co-references in the retrieval process.
Figure 1 describes the framework of our system.

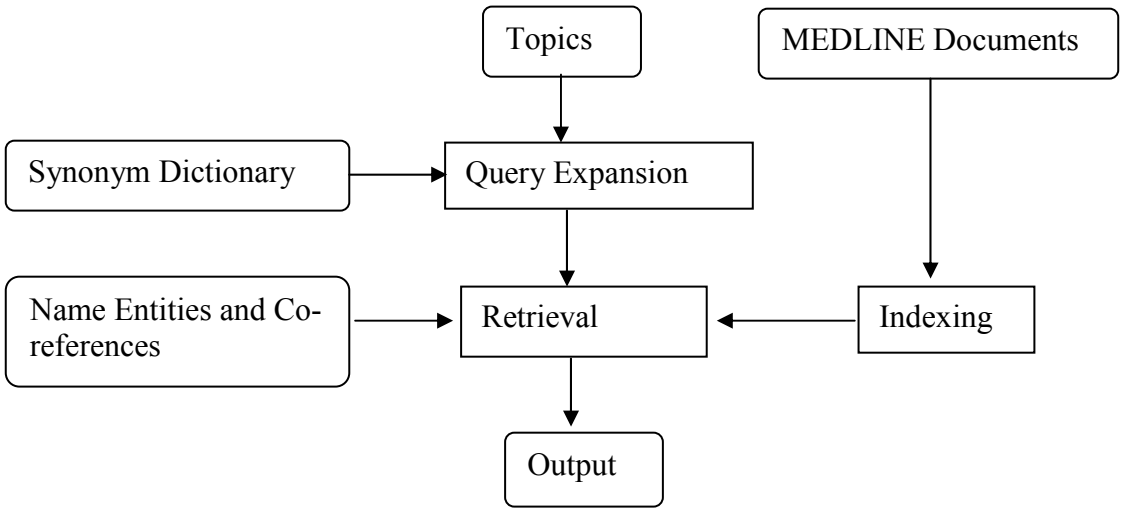

Figure 1: Framework of our system

3.1 Indexing

All documents are indexed before experiments. We only take the fields of title, abstract and MeSH Terms into account in indexing and later in document retrieval for each topic. Giving different weight to different field might bring about better performance, though it takes time to get weighting parameters by training. Since this is the first time we attend the track and time is quite tight, we didn't focus on that and simply combined information from these three fields of each document together into a plain text.
Indexing is made for words, terms, name entities and their co-references. About 1.4 million terms are extracted from the whole document set [4].

We use a seeding-and-expansion mechanism to extract key terms from the texts. The procedure of term extraction consists of two phases, seed positioning and term determination. Intuitively, a seed for a candidate term is an individual word, seed positioning is to locate the rough position of a term in the text, while term determination is to figure out which string covering the seed in the position forms a term.

To determine a seed needs to weigh a word to reflect their significance in the text. We make use of a very large corpus r (LDC's English corpus) as a reference corpus. Suppose d is a text, w is a word in the text, let Pr(w) and Pd(w) be the probability of w occurring in r and d respectively, we adopt relative probability or salience of w in d with respect to r, as the criteria for evaluation of seeds.

$$Pd(w) / Pr(w) \qquad (1)$$

We call w a seed if $Pd(w) / Pr(w) \geq \delta$ (δ=10 in this work). That is, its probability occurring in document must be 10 times higher than its average probability in the reference corpus. Although it is difficult to give out the definition of terms, we have the following assumption bout a key term in a document.

i) A term contains at least one seed.
ii) A term occurs at least L (L=3 in this work) times in the document.
iii) A maximal word string meeting i) and ii) is a term.
iv) For a term, a real maximal substring meeting i) and ii) without considering their occurrence in all those terms containing the sub-string is also a term.

Here a maximal word string meeting i) and ii) refers to a word string meeting i) and ii) while no other longer word strings containing it meet i) and ii). A real maximal substring meeting i) and ii) refer to a real substring meeting i) and ii) while no other longer real substrings containing it meet i) and ii). The above assumptions tell us a term is an independent maximal string which must contain a seed and occur at least 3 times in a document.

Name entities [7][8][9] and their co-references [5][6] are also extracted and indexed for later use.

3.2 Query Expansion

We expand queries in two ways. For the first method we built a small synonym dictionary for genes from Entrez Gene (http://www.ncbi.nih.gov/entrez/). Synonyms of genes occur in topics are taken from the dictionary to expand queries. The second way is pseudo-relevance feedback. Rocchio feedback [3] for BM25 is adopted.

3.3 Retrieval

Okapi BM25 [8][9] is implemented to retrieve the top 1000 documents for each query, where a score of each document is calculated as following formula and ranked.

$$\sum_{T \in Q} w^{(1)} \frac{(k_1 + 1)tf}{K + tf} \frac{(k_3 + 1)qtf}{k_3 + qtf}$$

Here $w^{(1)}$ is the Robertson/Spark Jones weight of T in Q:

$$\log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}$$

By respectively using single words or the terms we extracted previously as term units in BM25 method, there will be a word-based retrieval result and a term-based retrieval result. To fuse results got from these 2 ways, a parameter α is imported:

$$R = \alpha * R_{word} + (1-\alpha) * R_{term}$$

Here R is the final rank a document, which is got upon word-based rank result R_{word} and term-based rank result R_{term} of that document.

Name entities extracted could be considered as a single word or term. Co-references of a name entity could be considered as occurrences of that name entity. These replacements change frequency distribution of terms in BM25 method. Intuitively the using of name entity produces a better context cutting, and the using of co-reference strengthens the impact of name entities in retrieval process, though a high quality of name entity and co-reference extraction is required to promote the performance while exploiting them in the retrieval process. In the track of this year we did some experiments on that. A similar fusion strategy is used to fuse the result got with or without name entities and co-references.

4 Results and Discussions

We made a series of experiments with retrieval based on single words or extracted terms, with pseudo relevance feedback or without it, with entities and co-references or not. Because at most only two runs can be submitted, after testing on topics of TREC 2004, we chose two runs with the best performance to submit: one run (i2r1) implements single words based retrieval without pseudo relevance feedback, the other run (i2r2) implements single words based retrieval with pseudo relevance feedback. Table 1 lists the MAP of 2 runs evaluated by TREC organizer.

Runs	MAP
I2r1	0.2391
I2r2	0.2375

Table 1: MAP of our submitted runs

With the evaluation program the organizer provided after all submitted runs are evaluated, we evaluated our other runs. We list the results of term used run (rwt) and of name entities and co-references used run (rwne) in table 2 and table 3. Parameter α is set to 0.8 for both runs.

Runs	MAP
Rwt(α=1)	0.2375
Rwt(α=0.9)	0.2342
Rwt(α=0.8)	0.2264
Rwt(α=0.7)	0.2158

Table 2: MAP of runs with terms used

Runs	MAP
Rwt(α=1)	0.2375
Rwt(α=0.9)	0.2351
Rwt(α=0.8)	0.2296
Rwt(α=0.7)	0.2215

Table 3: MAP of runs with Name Entities and Co-references used

We can see that from the two tables, while the fusion parameter α decreases, which means that the impact from terms or name entities in the retrieval process increases, the MAP value decreases.
From the results table we can find that using extracted terms doesn't promote the performance. This might be because that the way we extracted terms is not suitable for MEDLINE records which are usually quite short. Meanwhile we extracted noun terms only, though verb terms are also quite important in some cases. We didn't do any special dealing for MeSH Terms of each record either.
According to experiments results, the name entities and co-references don't help to promote the performance either. Further investigations are needed to show the effectiveness to the task.

5 Acknowledgement

We'd like to acknowledge the Institute of High Performance Computing of Singapore providing computing platform for running Name Entity recognition and Co-reference resolution module on TREC data.

6 References

[1] Robertson, S.E. and Walker S. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In Proceedings of the 1994 ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 232-241.

[2] Robertson, S.E., Walker S., Jones S., Hancock-Beaulieu, M.M. and Gatford, M. 1995. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference(TREC-3), NIST Special Publication 500-225, Washington D.C., 109-126.

[3] Rocchio, J.J. 1971. Relevance feedback in information retrieval, In The SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton ed. Prentice-Hall, Englewood Cliffs, NJ, 313-323.

[4] Lingpeng Yang, Donghong Ji. Improving Retrieval Effectiveness by Using Key Terms in Top Retrieved Documents. 27th European Conference on Information Retrieval, ECIR 2005. LNCS 3408 pp.169-184.

[5] Yang XiaoFeng, Zhou GuoDong, Su Jian and Tan ChewLim. Coreference Resolution Using Competition Learning Approach. Proceedings of ACL 2003, Sapporo, Japan, 7-12 July 2003, pp176-183.

[6] Xiaofeng Yang, Jian Su, Guodong Zhou and Chew Lim Tan. A NP-Cluster Based Approach to Coreference Resolution. P226-232, Proceedings of 20th International Conference on Computational Linguistics (COLING'2004). Aug 23-27, 2004, Geneva, Switzerland.

[7] Zhou Guo Dong, Zhang Jie, Su Jian, Shen Dan and Tan Chew Lim. Recognizing Names in Biomedical Texts: a Machine Learning Approach. Bioinformatics, 20(7):1178-1190, DOI: 10.1093/bioinformatics/bth060, 2004, ISSN: 1460-2059, pp.1178-1190.

[8] Zhou GuoDong and Su Jain. Named Entity Recognition Using a HMM-based Chunk Tagger. Proceedings of ACL 2002, Philadelphia, US, July 2002.

[9] Zhang Jie, Shen Dan, Zhou GuoDong, Su Jian and Tan Chew Lim. Enhancing HMM-based Biomedical Named Entity Recognition by Studying Special Phenomena. Journal of Biomedical Informatics, 37(6), 2004, ISSN: 1532-0464, pp.411-422.