Passage Retrieval from Genomic Texts: An Experience at TREC 2007

Raymond Wan[1] rwan@kuicr.kyoto-u.ac.jp

Vo Ngoc Anh[2] vo@csse.unimelb.edu.au

Hiroshi Mamitsuka[1] mami@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan

[2] Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia

Keywords: biomedical text collections, passage retrieval, vector space model, latent semantic analysis

1 Introduction

The Text Retrieval Conference* is an annual conference where researchers in information retrieval (IR) compare their systems on specified retrieval tasks through several tracks. This poster summarizes the work done by Kyoto University and the University of Melbourne for the 2007 Genomics Track. We begin by briefly describing the track and then continue with a description of our method. We conclude with our experimental results from our TREC participation.

2 TREC Genomics Track 2007

Similar to 2006, the sole task for the 2007 Genomics Track was passage retrieval from a biomedical text collection. The biomedical text collection was identical to the one used for 2006 and contains 162,259 articles from 49 genomics related journals provided by Highwire Press (see Hersh et al. [2006] for details). All articles were encoded in HTML format with HTML paragraph tags (<p>) to indicate paragraph boundaries. A passage was defined as a sequence of words that does not include any HTML paragraph tags.

Thirty six queries were used to evaluate each participating system. Two example queries are "Which [PATHWAYS] are mediated by CD44?" and "What [TUMOR TYPES] are found in zebrafish?" (Queries #221 and #231, respectively). Each system returns a ranked list of passages for each query. These passages were pooled together and evaluated by judges for relevance. Then, the effectiveness of each retrieval system was measured in terms of mean average precision (MAP) across all 36 queries.

Four forms of MAP were used in 2007: Document, Aspect, Passage, and Passage2. Document and aspect MAP measure the relevance of the documents and aspects (topics) of the passages from the ranked list. The remaining two metrics measure relevance at the character-level. While Passage was used in 2006, Passage2 was the primary evaluation measure for 2007 since it corrects some of the problems with Passage.

3 System Overview

Our system for passage retrieval is illustrated in Figure 1. It consists of two parts: an information retrieval (IR) system and a passage retrieval (PR) system. The IR system performs three tasks: indexing, querying, and generating frequency counts for the PR system. The PR system consists of two parts: score derivation and passage extraction. The two inputs to our system are a query and the document collection. The output is a ranked list of passages (1000 per query, as required by the TREC Genomics guidelines).

*http://trec.nist.gov/

ID	IR : PR		Document	Aspect	Passage	Passage2
kyoto1	100	0	0.1892	0.1208	0.0474	0.0209
kyoto2	0	100	0.1191	0.0302	0.0235	0.0054
kyoto3	50	50	0.1022	0.0312	0.0204	0.0065

Table 1: Some parameters and results for our three official runs submitted for evaluation.

The IR engine employs impact-based ranking, a variant of the vector space model (VSM). We index the document collection at the paragraph-level so that a list of ranked paragraphs are passed to the PR system for each query.

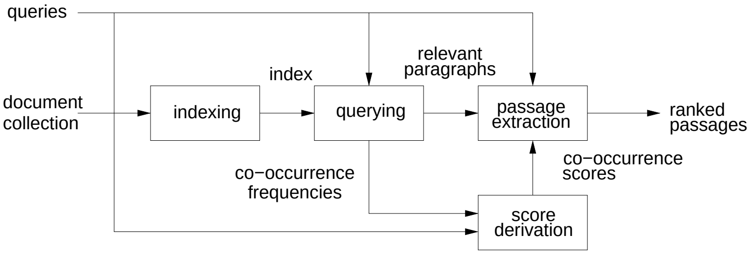

Figure 1: Our system for retrieving passages.

The PR system first uses the query and the pair-wise frequency counts to derive a matrix of co-occurrence scores. The initial counts take the form of a matrix of size $|Q|$ by $|D|$, where Q and D are the unique terms in the query and in the collection, respectively. This matrix is reduced to k clusters or latent states using latent semantic analysis (LSA). An iterative approach of LSA (probabilistic latent semantic analysis) applies the Expectation-Maximization (EM) algorithm to obtain a set of scores $s(i, j)$ where $i \in Q$ and $j \in D$ and $i \neq j$. This table is has the same dimensions as the original one, but with the noise removed to reach the "true" co-occurrences.

After these scores are calculated, each paragraph from the IR system is evaluated a word at a time. Each word is scored against the query using $s(i, j)$. An exact match is given a score of s_{max}, the maximum co-occurrence score in the entire table. An inexact match is given a score according to the table. These scores are further weighted by the inverse document frequency of the respective word in the paragraph. Wan et al. [2007] provides further details about this method, including the exact formulas used.

4 Results and Summary

The results from our submissions to TREC are shown in Table 1 as MAP for the four metrics. The median MAPs across the 66 submitted runs were 0.1897, 0.1311, 0.0565, and 0.0377 for Document, Aspect, Passage, and Passage2, respectively. Thus, while our best result is kyoto1, the MAP for all four measures still lie below the median.

Additional experiments (not shown) have demonstrated that the most important parameter affecting our performance is the parameter shown in the table. Each of the IR and PR systems assign a score to the paragraph or the passage, respectively. A final ranked list of passages is obtained by applying a weight to these two scores and then adding them. In the table, a weight of 100% indicates that that systems' score is used solely for the final ranking. As the table shows, the IR system appears to be the most important for passage scoring. While the three runs differ in other parameters, none of them appear to be as important.

This poster has summarized our joint work for the Genomics Track of TREC 2007. In the future, we need to further evaluate our method and, consequently, improve the performance and running time of our system.

References

W. Hersh, A. M. Cohen, P. Roberts, and H. K. Rekapalli. TREC 2006 Genomics track overview. In *Proc. 15th Text Retrieval Conference (TREC 2006)*, November 2006.

R. Wan, V. N. Anh, and H. Mamitsuka. Passage retrieval with vector space and query-level aspect models. In *Proc. 16th Text Retrieval Conference (TREC 2007)*, 2007. (To appear).