Combining Vector Space and Word-based Aspect Models for Passage Retrieval

Raymond Wan[1] **Vo Ngoc Anh**[2] **Ichigaku Takigawa**[1] **Hiroshi Mamitsuka**[1]

`rwan@kuicr.kyoto-u.ac.jp` `vo@csse.unimelb.edu.au` `takigawa@kuicr.kyoto-u.ac.jp` `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, Faculty of Engineering, University of Melbourne, Victoria, 3010, AUSTRALIA

Abstract

For the TREC 2006 Genomics Track, we devised a system of two parts for passage retrieval from a biomedical document collection. The first part was an existing document-level IR system. The second part was a newly-developed probabilistic word-based aspect model.

1. Introduction

Our system is composed of two distinct parts:

- An existing document-level IR system which outputs a ranked list of relevant documents [Anh and Moffat, 2005].
- A newly developed passage scoring system which is based on the aspect model [Hofmann et al., 1998]. The output of this second part is a ranked list of relevant passages.

Before the word-based aspect model is used to score a query against a passage, the scores are pre-calculated globally. Our entire system is illustrated in Figure 1.

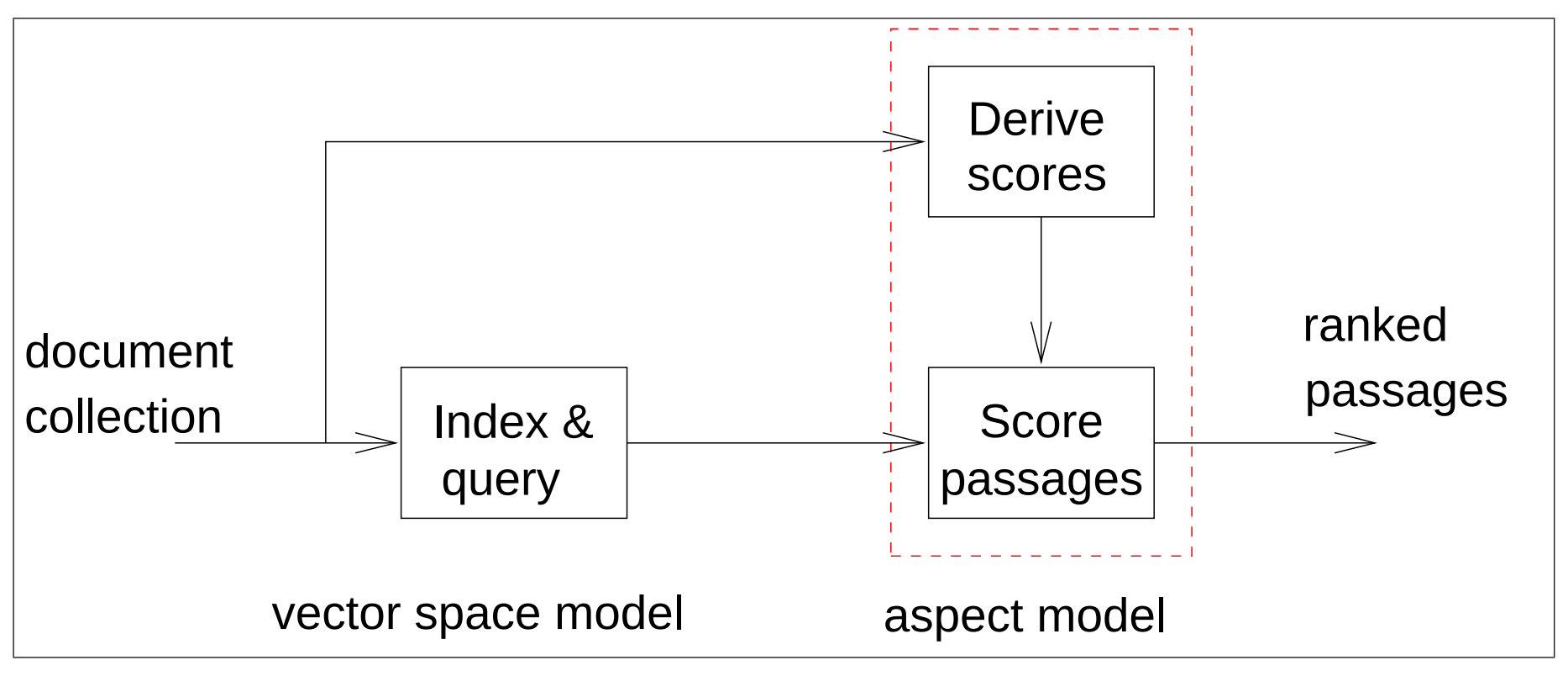

Figure 1: *Overview of our system of two models.*

2. System-wide Features

Our entire system incorporates the following:

- Words cannot be composed of a mix of alphabetic characters and digits.
- Punctuation marks are non-word characters.
- Case-folding is used.
- Stemming performed using Lovins algorithm.

We also employed synonym expansion automatically. In general, topics are of the form:

- “What is the role of APC (adenomatous polyposis coli) in colon cancer?” (query #163)

where the biological object (gene) and biological process (disease) are indicated.
Synonym expansion follows the flowchart of Figure 2 for each query term:

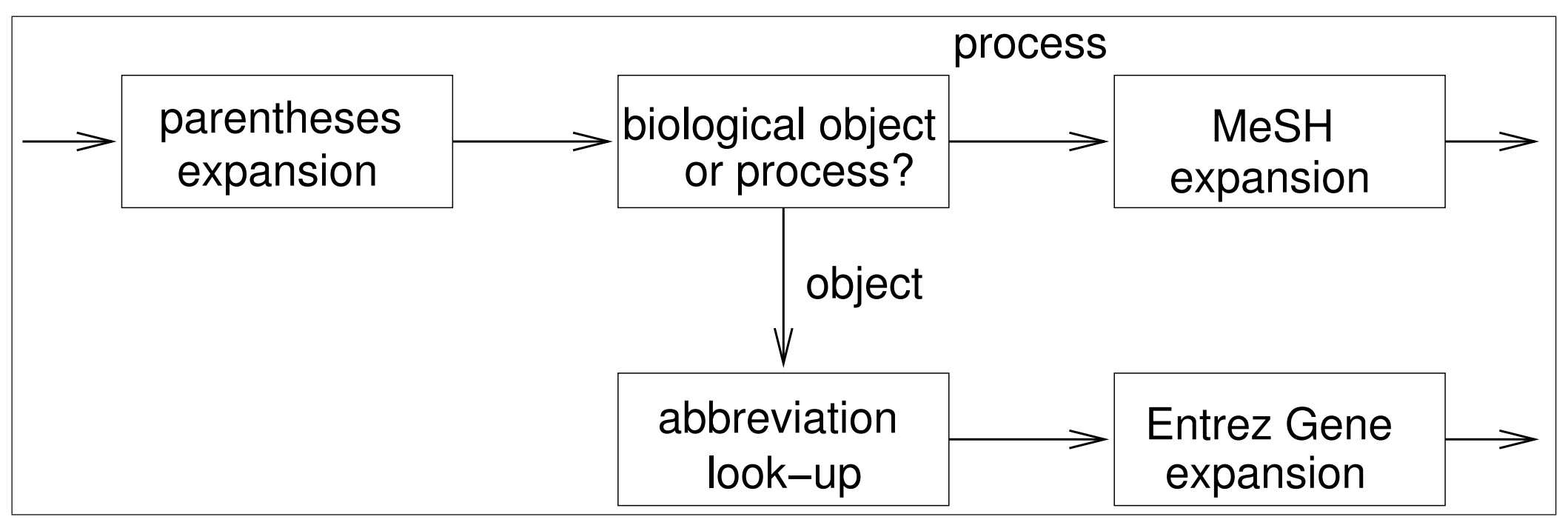

Figure 2: *Flowchart for synonym expansion.*

- Parentheses expansion would equate “APC” to “adenomatous polyposis coli”.
- Abbreviation expansion searches a Biomedical abbreviation server[1].
- Both Entrez Gene and MeSH (Medical Subject Headings) expansion are done using NCBI[2].

3. Vector Space Model

Our vector space model (VSM):

- Represents a document, paragraph, or a query as a vector.
- Uses impact-based ranking where coordinates are integers instead of floating point.
- Calculates similarity using the inner product.

This system has been applied to document-level retrieval [Anh and Moffat, 2005].

4. Aspect Model – Derive Scores

Passages are scored by doing a pair-wise comparison of the words in the query with the words in a prospective passage. If q_j is a query word and r_i is a passage word, then the score is:

$$\text{score}(r_i, q_j) = \begin{cases} c & \text{if } r_i = q_j \\ p(r_i, q_j) & \text{if } r_i \neq q_j \,. \end{cases} \tag{1}$$

In this case, c is a fixed constant which indicates how much value do we place on an exact match. For inexact matches, we use an aspect model to derive the scores $p(r_i, q_j)$. The aim is to regulate the number of words around the exact matches should be included.

Earlier uses of the aspect model maps words and documents to k latent states (clusters) [Hofmann, 2001, Hofmann et al., 1998]. Here, we map the n distinct words to the k clusters (where $k \ll n$) by using a co-occurrence window of a paragraph. Then, the score between two words w_x and w_y is the sum of their (pair-wise) scores across the Z clusters:

$$p(w_x, w_y) = \sum_{z \in Z} p(w_x|z)p(w_y|z)p(z), \tag{2}$$

Our implementation has these characteristics:

- It is based on the probabilistic latent semantic analysis (PLSA) of Hofmann [2001].
- Probability parameters are estimated using the maximum likelihood (ML) and the Expectation-Maximization (EM) algorithm.
- The log-likelihood was maximized until two consecutive iterations did not differ by $> 1\%$.

We considered two possible values for c (where $p_{\max}$ is the maximum score):

$$c = \begin{cases} 1 \\ 2 \times p_{\max} \,. \end{cases} \tag{3}$$

5. Aspect Model – Score Passages

The query q is scored against a prospective passage r (bounded by punctuation marks) by:

$$\text{score}(r, q) = \frac{\sum_i^s \alpha_{w_i} \times (\sum_i^s \sum_j^t (\text{score}(r_i, q_j)))}{s \times t} \,. \tag{4}$$

The parameter α_{w_i} (not shown) uses the following to affect the scoring mechanism:

- passage length (s)
- query length (t)
- word frequency in the current paragraph
- overall document frequency

6. Results

In total, seven runs were performed. The parameters used and the mean average precisions are shown in Table 1[3] and Table 2, respectively

Run	Level	c	(VSM:AM)
PARAGRAPH-1	paragraph	1	$\frac{1}{2} : \frac{1}{2}$
DOCUMENT-1	document	$2 \times p_{\max}$	$0 : 1$
DOCUMENT-2	document	$2 \times p_{\max}$	$0 : 1$
PARAGRAPH-2	paragraph	1	$1 : 0$
PARAGRAPH-3	paragraph	1	$0 : 1$
PARAGRAPH-4	paragraph	$2 \times p_{\max}$	$\frac{1}{2} : \frac{1}{2}$
PARAGRAPH-5	paragraph	1	$\frac{1}{2} : \frac{1}{2}$

Table 1: *Some parameters for the seven runs.*

Run	Document	Passage	Aspect
PARAGRAPH-1	0.2248	0.0248	0.1217
DOCUMENT-1	0.1231	0.0075	0.0610
DOCUMENT-2	0.1297	0.0071	0.0692
PARAGRAPH-2	0.1744	0.0131	0.0348
PARAGRAPH-3	0.2067	0.0261	0.1081
PARAGRAPH-4	0.1558	0.0091	0.0955
PARAGRAPH-5	0.2369	0.0258	0.1235

Table 2: *Mean average precision for our runs.*

In all cases, the number of clusters was $k = 128$. Our best result (PARAGRAPH-5) indexed the document collection at the paragraph-level and set $c = 1$. The ranking scheme employed gave equal weight to both systems.

7. Future Work

We plan to investigate the following:

- Reduce the execution time of the aspect model.
- Re-evaluate equations such as Equation (4).
- Consider the effect of adjusting the various parameters.

References

V. N. Anh and A. Moffat. Simplified similarity scoring using term ranks. In *Proc. 28th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 226–233, 2005.

T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1–2):177–196, January–February 2001.

T. Hofmann, J. Puzicha, and M. I. Jordan. Learning from dyadic data. pages 466–472, 1998.

[1] `http://abbreviation.stanford.edu/`
[2] `http://www.ncbi.nlm.nih.gov/`
[3] Differences between certain runs involve parameters not shown in this table. See our notebook paper for further details.

Text Retrieval Conference (TREC 2006), November 14-17, 2006, Gaithersburg, Maryland, USA