Structural Term Extraction for Expansion of Template-based Genomic Queries

Fabrice Camous[2,♦], Stephen Blott[2], Cathal Gurrin[1], Gareth J. F. Jones[1,2], Alan F. Smeaton[1,3]

[1]*Center for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, IRELAND*
[2]*School of Computing, Dublin City University, Glasnevin, Dublin 9, IRELAND*
[3]*Adaptive Information Cluster, Dublin City University, Glasnevin, Dublin 9, IRELAND*

This paper describes our experiments run to address the ad hoc task of the TREC 2005 Genomics track. The task topics were expressed with 5 different structures called Generic Topic Templates (GTTs). We hypothesized the presence of GTT-specific structural terms in the free-text fields of documents relevant to a topic instantiated from that same GTT. Our experiments aimed at extracting and selecting candidate structural terms for each GTT. Selected terms were used to expand initial queries and the quality of the term selection was measured by the impact of the expansion on initial search results. The evaluation used the task training topics and the associated relevance information. This paper describes the two term extraction methods used in the experiments and the resulting two runs sent to NIST for evaluation.

1. Introduction

This paper describes our experiments run to address the ad hoc task of the TREC 2005 Genomics track. The collection used for the task included a subset of the MEDLINE database, 50 test topics, and 10 additional sample topics with associated partial relevance information. The topics were expressed with 5 different structures called Generic Topic Templates (GTTs). We hypothesized that two kinds of terms were contained in title and abstract fields of documents relevant to an instance of a GTT: terms showing relevance to the GTT structure and terms showing relevance to the particular instance of the GTT, the topic. Terms relevant to a GTT are expected to express the generic information present in all instances of the GTT, such as interactions and relationships. Terms relevant to the instance of a GTT are expected to express the particular entities specific to the instance, i.e. the topic. We aimed at isolating terms specific to the structure of each GTT. GTT-specific terms were used for query expansion to seek to improve retrieval performance. Both relevance feedback and pseudo-relevance feedback were used to extract GTT-specific terms. Our methods were evaluated with the trec_eval program, using the partial relevance information available for the sample topics. The Físreál [1] search engine, developed at Dublin City University, was used to generate the rankings. The paper is organized in the following way: Section 2 introduces some background on the collection and relevance/pseudo-relevance feedback methods. Section 3 describes our experimental method and its analysis. Section 4 concludes on future work and experiments.

[♦] Contact author: fcamous@computing.dcu.ie

2. Background

2.1. Genomics track collection and the GTTs

The documents contained in the collection are the same as the ones used in the 2004 Genomics track. They consist of a 10-year subset of the MEDLINE biomedical abstract database (approximately 4.5 million documents). Most documents contain textual fields such as a title and an abstract.

This year 5 different types of query structures, or Generic Topic Templates (GTTs), were developed. They provided five different models of query expression. Table 1 gives a description of the five GTTs.

	GTT queries (entire GTT description)
1	Find articles describing standard methods or protocols for doing some sort of experiment or procedure.
2	Find articles describing the role of a gene involved in a given disease.
3	Find articles describing the role of a gene in a specific biological process.
4	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.
5	Find articles describing one or more mutations of a given gene and its biological impact in a given organism.

Table 1. GTT descriptions given by the task

The 50 test topics included 10 instances of each GTT and the sample topics 2 instances of each GTT. Table 2 gives the full narratives of the 10 sample topics. Partial relevance judgements associated with the sample topics were made available. They contained documents judged "probably relevant" and "definitely relevant" to the topics. They included a total of 596 judged documents with a range of 4-245 judged documents per topic.

GTT#	**Topic #**	**Full narratives of sample topics**
1	**90**	Describe the procedure or methods for quality control in microarray experiments.
	91	Describe the procedure or methods for GST fusion protein expression in Sf9 insect cells.
2	**92**	Provide information about the role of the gene Ribosomal Protein L11 in the disease Cancer.
	93	Provide information about the role of the gene DRD4 in the disease Alcoholism.
3	**94**	Provide information on the role of the gene HMG in the process of chromatin restructuring and transcriptional regulation.
	95	Provide information on the role of the gene Insulin receptor gene in the process of signaling tumorigenesis.
4	**96**	Provide information about the genes HMG and HMGB1 in hepatitis.
	97	Provide information about the genes MyD88, TRAM and TRIF in autoimmunity.
5	**98**	Provide information about Mutation of Ret in thyroid function.
	99	Provide information about Mutations of thiopurine S-methyltransferase in metabolism of drugs.

Table 2. Full narratives of sample topics

2.2. Query expansion, relevance and pseudo-relevance feedback

Expanding a query means adding new terms to it. The terms can be extracted from documents judged relevant, in the case of relevance feedback. They can also be extracted from documents that are assumed to be relevant, as in pseudo-relevance feedback.

When relevance feedback is used, the number of documents used for term extraction is the number of judged documents available for the topic. However, when pseudo-relevance feedback is used, an arbitrary number of documents located at the top of an initial ranking can be assumed relevant. In our experiment, we assumed the top 5 documents of an initial ranking to be relevant.

Terms need to be scored and ranked according to their association with the relevance of the documents. We used Robertson's Offer Weight method [2] to score the terms contained in title or abstract fields of documents assumed or judged relevant. The term score is given by the following formula:

$$term\_score = r.\log\left(\frac{(r+0.5).(N-n-R+r+0.5)}{(n-r+0.5).(R-r+0.5)}\right) \qquad (1)$$

where r is the number of relevant documents containing the term, n the number of documents containing the term, and N= 4,591,008, the total number of documents in the collection. After the terms are scored and ranked, the top n are selected to expand the query. In our experiments, n took the values 5, 10, 15, and 20.

3. Experimental method and analysis

Partial relevance information was only available for the sample topics. As a consequence, our experiment aimed at improving baseline rankings for the sample topics.

If two distinct topics are instances of a same GTT, relevant documents retrieved in the two distinct rankings are expected to contain title/abstract terms specific to each topic. Our hypothesis is that they will also contain common terms that are specific to the GTT. For example, topic 92 and 93 from table 2 are both instances of GTT 2. This query structure is a basis for formulating an information need relative to the role of a gene in a disease. Topic 92 and 93 concern distinct genes and diseases and relevant documents retrieved for each topic should contain terms that relate to each topic separately. Following our hypothesis, documents relevant to both topics are expected to contain common generic terms that express relationships between genes and diseases. Those terms can be used to expand the initial query and push up the ranks of relevant documents that mention a gene and a disease as well as the GTT relationship. Two methods were used to extract the GTT-specific structural terms. The first method used the relevance information available for the sample topic. It is described in section 3.2. The second method used pseudo-relevance feedback on rankings obtained with Físreál [1] search engine for the 50 test topics. This method is described in section 3.3.

3.1. Baseline sample rankings

A first objective was to find a baseline ranking for the 10 sample topics by generating queries from the full narratives of table 2. Two options were investigated. First, the full narratives were sent unmodified as queries to Físreál search engine to generate a ranking for each sample topic. Those queries were called "full narratives". Secondly, another set of queries was produced from the full narratives by keeping only terms unique to the individual topic (not present across a GTT). Those queries were called "basic narratives". As an example, the "basic" version of the full narrative "Provide information about the role of the gene DRD4 in the disease Alcoholism" is "DRD4 Alcoholism". No GTT structural information is present in the "basic narratives". Table 3 shows a list of all the sample "basic narratives". They were sent to Físreál to generate a ranking for each topic.

Table 4 shows the results for both full and sample basic narratives in terms of Mean Average Precision (MAP) and recall. As the sample basic narratives produced better rankings, we chose them as our baseline rankings.

3.2. Using the sample topics relevance information.

Formula 1 was used to score and rank the terms contained in relevant documents for each sample topic. Table 5 shows the amount of documents judged relevant per topic. 10 term rankings, 2 per GTT, were generated. GTT-specific structural terms were assumed to be contained in the two rankings obtained for both instances of the GTT. For each GTT, after score normalization (all scores in each topic ranking are divided by the highest score of the topic ranking), terms in common in the two topic rankings were kept and their scores were added. The common-term list was re-ranked to give the GTT-specific structural term ranking.

GTT#	Topic #	Basic sample narratives
1	90	quality control in microarray experiments.
	91	GST fusion protein expression in Sf9 insect cells.
2	92	Ribosomal Protein L11 Cancer
	93	DRD4 Alcoholism
3	94	HMG chromatin restructuring and transcriptional regulation.
	95	Insulin receptor gene signaling tumorigenesis.
4	96	HMG and HMGB1 in hepatitis.
	97	MyD88, TRAM and TRIF in autoimmunity.
5	98	Mutation of Ret thyroid function.
	99	Mutations of thiopurine S-methyltransferase metabolism of drugs.

Table 3. Basic sample narratives

	Full sample narratives	Basic sample narratives
Mean Average Precision	0.1144	0.1268
Recall	0.5772	0.5856

Table 4. MAP and recall results for basic and full sample narratives

GTT#	Topic #	Number of relevant docs R
1	90	49
	91	58
2	92	4
	93	37
3	94	72
	95	43
4	96	4
	97	9
5	98	245
	99	75

Table 5. Number of relevant documents, R, per topic.

The top n terms from each GTT were used to expand the basic narratives of the sample topics related to the GTT. In the expanded query, terms were combined in the following way: Topic-specific terms (from basic narrative) were given a weight of value 3 and GTT-specific terms (from relevance feedback) were given a weight of value 1. This weighting policy gave us good results in preliminary experiments. The results are given in table 6 for different values of n. For all n values, there is an improvement in Mean Average Precision. However, recall decreases as n grows. As table 5 shows, the number of relevant documents per topic can vary from 4 to 245. Therefore, high variation of recall for a topic with a high number of relevant documents, such as topic 98, can strongly influence the average result across the 10 topics. Figure 1 shows that a 10-term expansion improves the recall for 5 topics, leaves it unchanged for 2 topics, and decreases it for 3 topics. The 3 topics concerned by the decrease in recall are 95, 98 and 99. They contain 43, 245 and 75 relevant documents, respectively. Therefore the drop of recall for topics 95, 98 and 99 may have strongly contributed to the drop of the average recall for the 10 topics. The best MAP/recall combination is obtained with an expansion of 10 terms.

		Topic Term weight=3, GTT term weight=1			
	Basic sample narratives	top 5 terms kept	top 10 terms kept	top 15 terms kept	top 20 terms kept
Mean Average Precision	0.1268	0.1280	0.1303	0.1309	0.1303
Recall	0.5856	0.5772	0.5721	0.5621	0.5604

Table 6. Results with expansion using relevance feedback

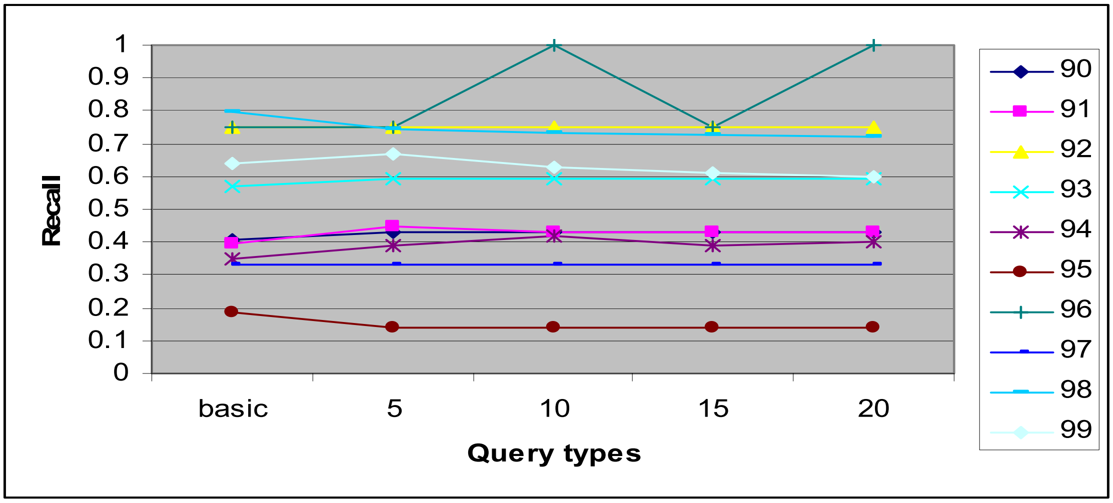

Figure 1. Recall by sample topic and query type (basic, basic + expansion terms)

3.3. Pseudo-relevance feedback with test topic rankings.

Basic narratives were produced for the 50 test topics. The basic narratives were sent to the Físreál search engine to generate 50 document rankings. For each document ranking, the top 5 documents were assumed to be relevant. Formula 1 was used to score and rank title/abstract terms contained in the documents assumed relevant.

To extract each GTT-specific structural term ranking, 10 topic term rankings had to be combined. To obtain a reasonable amount of terms per GTT (more than 20), a term was assumed GTT-specific if it appeared in at least 4 topic term rankings. As before, term scores were normalized in each topic ranking, normalized score were added and GTT-specific terms were re-ranked.

The GTT-specific terms generated with this method gave poor retrieval results when used to expand basic narratives from the sample topics. However, the terms were used to expand basic narratives from the test topics. Our assumption was that the terms could still work well with more topics and more relevance judgments.

3.4. Runs sent to NIST

Two runs were sent to NIST, dcu1 and dcu2. dcu1 used the GTT-specific terms generated with the method described in section 3.2 to expand basic narratives of the test topics. Only the top 10 terms of the GTT-specific term rankings were used. dcu2 used the GTT-specific terms generated with the method described in section 3.3 to expand basic narratives of the test topics. As before, only the top 10 terms were used for expansion.

Relevance information for the test topics was made available at the end of September 2005. Table 7 shows the impact of dcu1 and dcu2 expansion methods on a baseline search with basic narratives of the test topics. dcu1 gave lower MAP and recall values. This can be explained by the few topics per GTT (only 2) and the limited relevance information available. In the future, the relevance information released for the test topics will be used to generate GTT-specific structural terms. dcu2 gave a lower MAP value but maintained the recall level. The drop in MAP can be explained by the noise introduced when relevance is assumed for the top 5 documents from the initial rankings.

	Basic narratives	**dcu1**	**dcu2**
Mean Average Precision	0.1947	0.1851	0.1844
Recall	0.6374	0.6169	0.6405

Table 7. Impact of dcu1 and dcu2 methods on baseline

4. Conclusion

In this paper we have presented two methods to extract textual structural terms from MEDLINE documents relevant to the structure of template topics. The extraction methods were evaluated by measuring the impact of expanding initial queries with the extracted terms. Although query expansion results were encouraging with the sample topics, the two methods did not impact positively on the test topics results. The negative impact can be explained by the partial relevance information used in the first method, and the noisy pseudo-relevance information used in the second method.

In future work the relevance information available for the test topics will be used to generated template-specific structural terms. The position of GTT-specific structural terms relative to the topic-specific terms will be integrated in the extraction process.

Structural information can also be extracted from other fields of MEDLINE documents. The Medical Subject Headings (MeSH) are used to represent the conceptual content of MEDLINE records. This standardized conceptual information will be used to generate structural information relating to relationships present in documents.

Acknowledgments

This work is funded by Enterprise Ireland under the Basic Research Grants Scheme, project number SC-2003-0047-Y, and by Science Foundation Ireland under grant number 03/IN.3/361.

References

[1] Ferguson, P. *et al*. (2005). “Físreál : A Low Cost Terabyte Search Engine”. In Proceedings of 27th European Conference in Information Retrieval, March 2005.

[2] Robertson, S. E. & Spark Jones, K. (1996). “Simple, proven approaches to text retrieval”. Technical report 356, Computer Laboratory, University of Cambridge.