Report on the TREC 2006 Experiment: Genomics Track

P. Ruch[ac], A. Jimeno Yepes[c], F. Ehrler [ac], J. Gobeill [ab], I. Tbahriti [ab]

[a] *Medical Informatics Service, University and University Hospital of Geneva, Geneva*
[b] *Swiss-Prot, Swiss Institute of Bioinformatics, Geneva*
[c] *Artificial Intelligence Lab., University of Geneva Geneva*

contact: patrick.ruch@sim.hcuge.ch

Introduction

In previous TREC Genomics competition, *ad hoc* experiments were based on MEDLINE corpora (about 4.5 millions in 2005). This year, the collection has been replaced by a collection of about 160000 full-text articles. The proposed task is a passage retrieval task. Because document length in MEDLINE follow a binomial distribution (Figure 1), our previous investigations were focused on exploring the document length parameter, using a slightly modified pivoted normalization factor (Singhal 1999, Fujita 2004). This year, our efforts concentrated on combining knowledge-driven methods to a standard vector-space retrieval system.

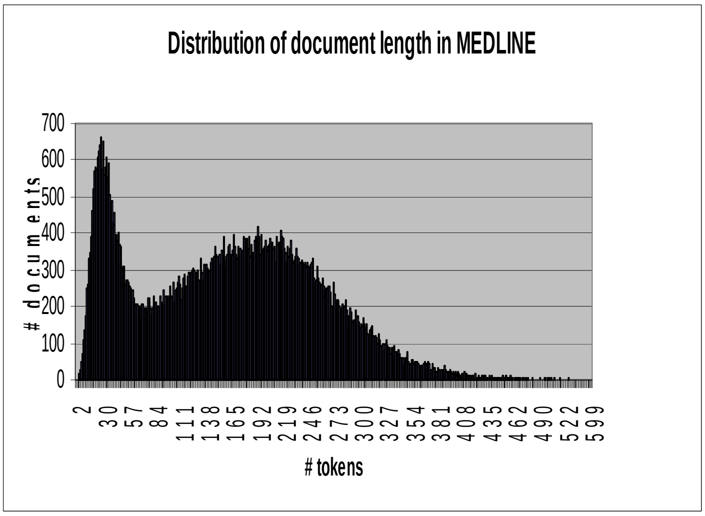

Figure 1. Document length distribution in MEDLINE.

Methods

From the official topics of the 2005 Genomics track, a subset was used to obtain the official 2006 topics. Topics from 2005, which were not official this year, were used to tune the parameters of our engine. It means that tuning was based on document retrieval (abstract) rather than on passages, which can explain why tuning was particularly difficult this year and should improve next year… if the task is reconducted !

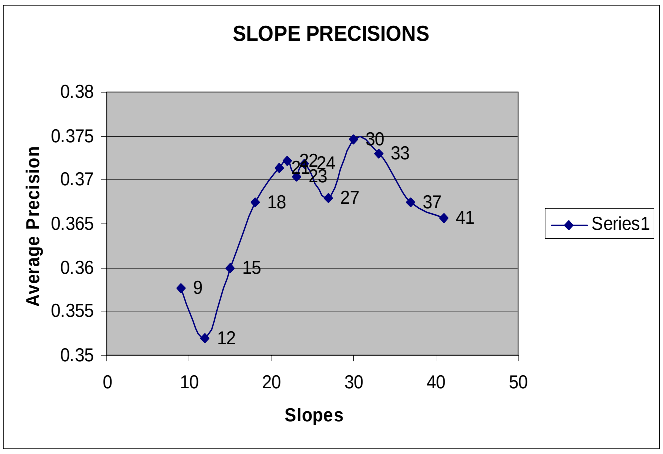

Figure 2. Tuning of the slope parameter.

Indexes were generated based on a pre-processed document collection. The document collection was transformed to obtain a collection of documents based on sections (cf. Demner-Fushman and al. 2006). Short documents of the collection (i.e. less than 50 bytes) are simply removed from the index. Thus, sections containing titles, bibliographical references, or list of keywords cannot be retrieved.

The best weighting was obtained using a slightly modified dtu.dtn (cf. Table 2 for a formal description) schema, with slope=30 and using a modified Porter stemmer (cf. Figure 2): the idf parameter is smoothed using the length distribution of the stem in the collection. The tuning of the mixture factor for two different slope values (22 and 30) is given in Figure 3.

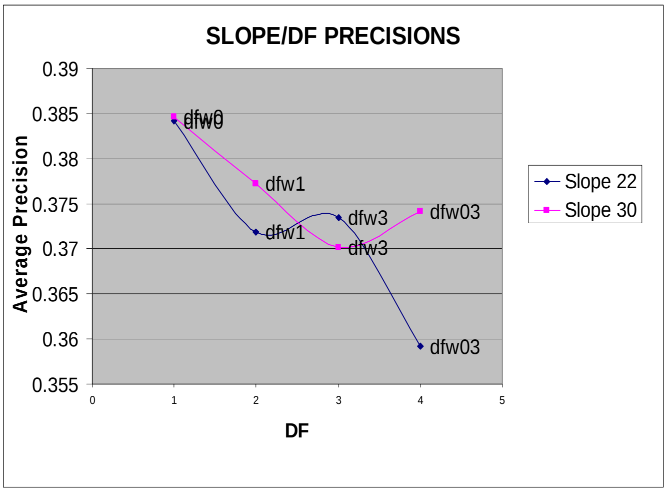

Figure 3. Tuning of the mixture parameter.

We observe that the mixture bring a very modest improvement (+3%).

dtu w$_{ij}$ =

$$\frac{\left(\ln\left(\ln((\mathrm{tf}_{ij}) * \mathrm{K}_{\mathrm{Length(Feature)}}) + 1\right) + 1\right) \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$$

dtn w$_{ij}$ =

$[\ln(\ln(\mathrm{tf}_{ij}) + 1) + 1]\ .\ \mathrm{idf}_j$

$\mathrm{idf}_{\mathrm{mix}} = \alpha.\mathrm{idf}_{\mathrm{collection}} + \beta.\mathrm{idf}_{\mathrm{Zippf}}$

$\mathrm{idf}_{\mathrm{Zippf}} = \mathrm{Length}(\mathrm{w}_{ij})$

Table 2: Formula for dtu.dtn, modified to take into account the Length of the feature.

Further, we also use a specific tokenization module for the query in order to better handle hyphenation of biological and chemical words. Following observations from 2005, gene and proteins were neither normalized not expanded. In contrast, disease, chemical substances and body parts were normalized and expanded, using the UMLS resources and the categorizer described in Ehrler and al. (2005) and Ruch (2006)

Furthermore, using the categorizer we also attempt to boost documents having a category-based similarity to the queries. The idea is to re-rank the run generated by the basic retrieval system (Baseline in Table 1) by computer a category-based similarity. In our approach the selection of categories is template-dependent because topics have been pre-categorized into five different sets.

As the categorizer cannot be applied to all paragraphs in the document collection, only the top 2000 articles retrieved by the engine are considered for re-ranking. The approach is somehow similar to the method used by Voorhees (1994). The same strategy is applied to rerank documents containing MeSH and GO categories, which also appear in queries. This strategy was effective on our tuning queries (+ 8% regarding mean average precision, from 0.385 to 0.416) for document retrieval.

Results

We submitted three runs: a baseline run, a run with MeSH (Medical Subject Headings) boosting, and a run with GO (Gene Ontology) boosting (Table 1).

Baseline	
Document	0.27551924
Passage	**0.03429375**
Aspect	**0.17019397**
GO boosting	
Document	0.27061541
Passage	0.03093560
Aspect	0.13856591
MeSH boosting	
Document	**0.28142632**
Passage	0.03284667
Aspect	0.15771459

Table 1. Official results (mean average precision).

We observe that MeSH boosting is moderately effective for document retrieval but for passage retrieval (and aspect) the baseline system outperforms systems which overweights MeSH and GO categories. While it is established that MeSH categories can help information search in MEDLINE, see for example Abdou and al. 2006, it is still unclear why it did not help passage retrieval. A possible explanation could be found in the metrics used this year for evaluating passage retrieval. Metrics were heavily biased toward systems outputting short passages so that additional experiments are needed to clarify these issues.

Conclusion

Out template-based category-specific boosting approach seems ineffective for passage retrieval, as well as for aspect retrieval, but interestingly, it seems to have some effectiveness for document retrieval (ad hoc). This last result is consistent with the state-of-the-art, which tends to confirm that using MeSH categories can significantly improve retrieval effectiveness in MEDLINE (Srinivasan, 1996). Finally, in 2005, pivoted length normalization seemed effective for retrieval in MEDLINE, but let us note that this year; the collection exhibits a normal distribution, so that length normalization could be less effective than with MEDLINE abstracts. Our baseline run was also used to generate a fusion run (Demner-Fuchman 2006) and promising results were reported.

Acknowledgments

The study reported in this paper has been supported by the SNF (EAGL project 3252B0-105755) and the EU-IST/CH-OFES SemanticMining Network of Excellence (507505/03.0399).

References

[1] S Abdou, P Ruch, J Savoy (2006) Searching in MEDLINE: Stemming, Query Expansion and Manual Indexing Evaluation. TREC Proceedings, TREC 2005, Gaithersburg, MD, USA.

[2] Dina Demner-Fushman, Susanne M. Humphrey, Jimmy Lin, Hongfang Liu, Patrick Ruch, Miguel E. Ruiz, Lawrence H. Smith, Lorraine K. Tanabe, W. John Wilbur, Alan R. Aronson (2006). Finding relevant passages in scientific articles: fusion of automatic approaches vs. an interactive team effort. TREC Proceedings. TREC 2006, Gaithersburg, MD, USA.

[3] F Ehrler, A Geissbühler, A Yepes, P Ruch , Data-poor Categorization and Passage Retrieval for Gene Ontology Annotation in Swiss-Prot, *BMC Bioinformatics*, Special Issue on BioCreative.

[4] Fox E.A. and Shaw J.A. (1994). Combination of multiple searches. In Proceedings TREC-2, (pp. 243-249). Gaithersburg: NIST Publication.

[5] S Fujita (2004) Revisiting Again Document Length Hypotheses: TREC-2004 Genomics Track Experiments at Patolis. The Thirteenth Text Retrieval Conference, TREC-2004, Gaithersburg, MD.

[6] A. Singhal (2001) Modern Information Retrieval: A Brief Overview. IEEE Data Eng. Bull. 24. p 35-43

[7] P Ruch (2006) Automatic Assignment of Biomedical Categories: Toward a Generic Approach, Bioinformatics 2006.

[8] P Srinivasan, Optimal document-indexing vocabulary for MEDLINE, Inf. Process. Manage, 32 (5), 1996, 503-514.

[9] E Voorhees, Query expansion using lexical-semantic relations, SIGIR 1994, 61-69.