Vocabulary-driven Passage Retrieval for Question-Answering in Genomics

J. Gobeill [ab], F. Ehrler [ac], I. Tbahriti [ab], P. Ruch[ac]

[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

Abstract

This year, like in 2006, we use a collection of about 160000 full-text articles. The proposed task is a passage retrieval task. Several measures are applied on the returned data: passage mean average precision, document mean average precision, aspect mean average precision. This year, our efforts concentrated on combining knowledge-driven methods on top of a standard vector-space retrieval approach. We tested a passage selection methods based on vocabulary density estimation using several terminologies of the domain. We also attempted to improve standard retrieval approaches based on vector-space similarities by using a Boolean completion principle, which overweight documents containing all keywords. These combinations did not result in a significant improvement compared to the baseline system (document map ~ 0.20) and current results do not show much improvement compared to last year's reported results.

Introduction

The 2007 TREC Genomics track aims at evaluating automatic question-answering systems for biologists. Rather than finding an exact answer, as in factoid or definitional questions, the task was to extract passages containing answers to a list of user queries in a collection about 160 000 full-text articles published in 49 different genomics-related journals.

Data and Resources

In 2006, it has been shown that with such a "small" collection of articles, which hardly can cover the broad field of genomics, it is expected that several queries may not have answering passages in the collection. As in 2006, we used a pre-processed collection, see Demner and al. 2007 for a description of the pre-processing steps (HTML removal, UTF-8 conversions…).

Out of the 12,641,127 legal spans (passages) of the collection, passages shorter than a dozen words were simply filtered out of the collection. This step removed most section headings, subheadings, authors, author's affiliation, as well as most references. The resulting collection contained about 800,000 legal span only. We hypothesized that such short passages are likely to affect document length-related factors (Fujita 2004, Singhal 2001) in the computation of the retrieval status value. Moreover, we observe that the relevance of such passages is mostly a design artefact caused by using an artificially constructed collection. Such a partial collection is basically of poor interest since all MEDLINE and its 16 millions references are available for free. Indeed, in practice, both the abstract and its title should be indexed in the collection and so should be directly provided in the ranked document list.

Several terminologies were gathered to be used in our experiments (cf. Methods) such as: GPSDB for gene and protein names (Pillet and al. 2005), Swiss-Prot keywords for tissues, Gene Ontology categories for molecular functions. Others resources, mostly part of the Unified Medical Language System, such as the Medical Subject Headings or the International Classification of Diseases for pathological functions, were also used. An extended list (as proposed in Ruch et al. 2000) of semantic types based on the UMLS was used to classify entity types across the different vocabularies.

Methods

Our baseline run (GeneTeam1) uses the easyIR vector-space engine, with a modified dtu.dtn weighting schema (Ruch and al. 2006). Due to lack of time, we used the 2006 empirical settings (slope = 13). On top of this basic run, we investigate two complementary strategies: using available terminologies to shorten the selected passage; overweighting passages containing all keywords.

Basic run. All runs were generated using Porter-stemming with a specific handling of hyphens. The following string a-b-c would generate the following entries in the engine's index: *abc, ab, bc, a, b, c*. The original DF of the collection was combined with a DF list computed on the whole MEDLINE collection as successfully proposed for TREC 2006 (Ruch and al. 2006).

Boolean boosting. First, we try to avoid having highly-ranked passages, which do not contain all keywords. The phenomenon is well-known in modern IR and several strategies have been proposed to cop with such issues, see for example (). Our approach (run GeneTeaBB, Boolean Boosting) consists in overweighting documents, which contain all keywords for a given query. Given the lack of tuning, the boosting factor is very moderate (multiplicative factor = 1.1).

Shortening of passages. Second, we attempt to shorten the size of the passage by focusing only on high-density contents. Low-density contents are selected by targeting introductory expressions (*in this section…*; *in the following, we show…*) and argumentative expressions (*we conclude…, we aim at…*). The list of poorly-content bearing expressions is borrowed from TREC Genomics 2003 and related experiments (Ruch and al. 2005) to identify GeneRiFs (Gene References into Functions). The next reducing step is driven by the query. In each query the expected type of the target entity is provided by the organizers: e.g. molecular functions, diseases, drugs. For each entity-type a list of targets is extracted from the a set of pre-compiled vocabularies: MeSH for diseases and drugs; Gene Ontology for molecular functions ; Swiss-Prot tissues types for cell and tissues. In the ranked list of passages, we look at each passage. In each of these passages, we attempt to locate the first and last occurrence of a target entity using a generic text categorization tool (Ehrler and al, 2006, Ruch 2006). The text on the preceding the first occurrence of a target and the one following the last occurrence of a target is simply discarded.

Evaluation

Table 1, gives an overview of our official runs. Two non-official runs are also evaluated using different slopes and/or using the whole collection rather than the 700 000 passage collection.

Baseline	
Document	0.2
Passage	0.064
Aspect	0.175
Boolean Boosting (GeneTeam1)	
Document	0.196
Passage	0.06
Aspect	0.179
Passage reduction (GeneTeaBB)	
Document	0.196
Passage	0.04
Aspect	0.14
Baseline with the whole collection (GeneTeaPA)	
Document	0.21
Passage	0.067
Aspect	0.180
Baseline with the whole collection (slope = 30)	
Document	0.22
Passage	0.069
Aspect	0.184

Table 1. Official results (mean average precision).

Experiments made with non-submitted runs confirmed that using the whole collection (12 millions passages) rather than our reduced collection of 700,000 passages improve somehow modestly the retrieval effectiveness (document map = 0.21 vs. 0.2) of the system at the cost of a massive inflating the index. In contrast, we observe that changing the slope value (slope = 30) provides some modest improvement: from 0.21 to 0.22 regarding document map.

Regarding passage-specific measures, differences are even more moderated. In general, the baseline approach outperforms other more elaborated strategies, except for aspect, which is somehow improved using Boolean boosting. More experiments will be needed to confirm such a trend.

Conclusion

These results are similar to last year' results. Neither our template-based vocabulary-driven boosting approach using terminological knowledge repositories, nor our passage shortening strategies were effective in improving the task. Altogether, these preliminary results, as well as observations made by other groups (Demner and al. 2007) (Fautsch and Savoy 2007) suggest that usual run combination approaches (Fox and Shawn 1994) are very effective while currently the specificity of passage retrieval in life sciences is still at an early stage.

Acknowledgments

The study reported in this paper has been supported by the SNF (EAGL project 3252B0-105755).

References

C Fautsch and J Savoy (2006) IR-Specific Searches at TREC 2007: Genomics & Blog. TREC Proceedings, TREC 2007, Gaithersburg, MD, USA.

D Demner-Fushman, S M. Humphrey, J Lin, H Liu, P Ruch, M E. Ruiz, L H. Smith, L K. Tanabe, W J Wilbur, A R Aronson (2007). Combining resources to find answers to biomedical questions. TREC 2007, Gaithersburg, MD, USA.

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.

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

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

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

P Ruch, R Baud, AM Rassinoux, P Bouillon, and G Robert (2000) Medical Document Anonymisation with a Semantic Lexicon. *J Am Med Inform Assoc (Symposium Suppl).*

P Ruch and L Perret and J Savoy. Features Combination for Extracting Gene Functions from MEDLINE. 27th European Colloquium on Information Retrieval - ECIR, LNCS 3408, 2005.

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

V Pillet, M Zehnder, A K. Seewald, A-L Veuthey, J Petrak: GPSDB: a new database for synonyms expansion of gene and protein names. Bioinformatics 21(8): 1743-1744 (2005)