# *Report on the TREC 2005 Experiment: Genomics Track*

**Patrick Ruch**[1a], **Henning Müller**[a], **Samir Abdou**[b], **Gilles Cohen**[a], **Jacques Savoy**[b]

[a]*SIM, University ad University Hospital of Geneva, Geneva*
[b]*Université de Neuchâte*

*l*

## Summary

This year, for our participation to the TREC Genomics track, we participated in the two tasks: the ad hoc and the categorization task. In this notebook report, we do not detail our experiments, which will be described more precisely in the final proceedings. This papers focuses on the ad hoc task, while experiments conducted for task 2 are described in the Aronson and al. 2005.

**Task I**. For the ad hoc retrieval task, we used the easyIR tool, a standard vector-space engine developed at the University of Geneva.

Our approach uses thesaural resources together with a variant of the Porter stemmer for string normalization. Gene and Protein Entities (GPE) in queries are marked up by dictionary look up at retrieval time in order to be expanded using a gene and protein thesaurus. For indexing the Genomic collection, the following MEDLINE records were selected: article's titles, MeSH and RN terms, and abstract fields. Following observations made on MEDLINE documents regarding their length distribution, we decided to rely on a slightly modified dtu.dtn weighting schema. This constitutes our baseline run (Baseline=0.2312; Baseline+expansion=0.2373). Finally, we used a run provided by the University of Neuchâtel, which features thesaurus-based GPE expansion and automatic feed back (UniGeNe=0.2150) to produce a third run, which achieved our best results (UniGe2=0.2396).

## Introduction

Because, document length in MEDLINE is variable as shown in Figure 1, our investigations focused on exploring the document length parameter, using a slightly modified pivoted normalization factor (Singhal 1999, Fujita 2004): cf Table 2. We also decided to evaluate the effectiveness of using a set of terminological resources, such as a gene and protein thesaurus and other biomedical resources..

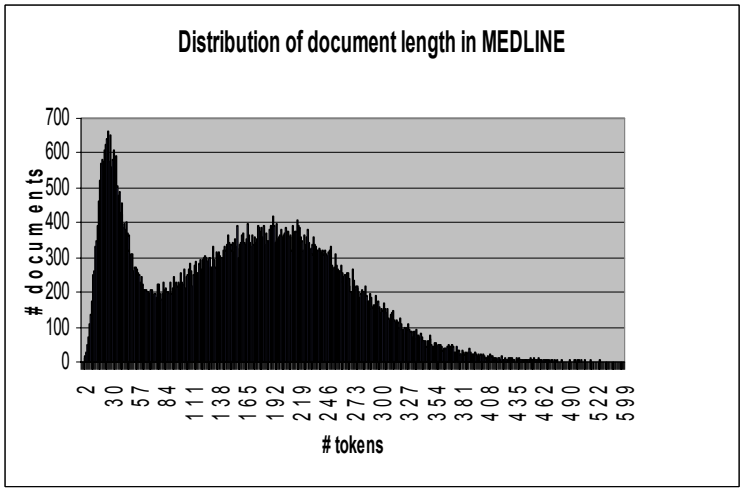


**Figure 1. Document length distribution in MEDLINE.**

Indeed, while document length, in general, document collections is normally distributed, MEDLINE documents length seems the results of a two-Gaussian mixture. While the largest set of MEDLINE records contains an average of 200 tokens, a few contain an average of about 30 tokens. The figure is computed using four fileds: Title, Abstract, MeSH and RN fields.

[1] Contact author. Email: patrick.ruch@sim.hcuge.ch

## Methods

We use the ten tuning topics provided by the organizers to tune our system. The tuning process aims at selecting the best parameters for the weighting schema and the best expansion strategy.

The best weighting was obtained using a slightly modified dtu.dtn formula, with slope = 13 and using a slightly modified Porter stemmer. Finally, we also use a specific tokenization module for the query in order to better handle hyphenation of biological and chemical words. The 'a' was removed from the stop words.

Gene and protein names can be highly variables; therefore we thought using a thesaurus could help retrieving more relevant articles. We use a set of online resources (UniProt/SwissProtKB, GPSDB...) to build our thesaurus. Detection of entities to be expanded in queries was done using two different strategies: using an automatic text categorizer (Ehrler and al. 2005, Ruch 2005), 2) using exact match. Similar expansion strategies have been tested for disease categories, based on the UMLS and internal resources (Lovis and al. 1997). From these comprehensive set of experiments it is worth observing that we have never been able to improve the retrieval effectiveness of our engine even when expansion was based on exact matches of entities.

| | |
|---|---|
| Baseline | 0.1751 |
| **Expanding chemicals, diseases, species and body parts and removing documents not containing the species [EXP1]** | **0.1775** |
| Expanding genes and proteins names [EXP2]:<br>6 x original query<br>3 x best representative<br>1 x other synonyms | 0.1632 |
| Expanding genes and proteins names [EXP3]:<br>10 x original query<br>1 x best representative<br>1 x other synonyms | 0.1717 |

**Table 1: results of easyIR on test data.**

We also evaluate the impact of expanding other types of entities: chemicals (calcium), diseases (cancer), species (rats), and body parts (spleen). For species, results reported for TREC Genomics 2003 showed that filtering out documents based on medical subject headings was also effective. Results are reported in the following table:

Expansion based on the gene and protein names has been tested using different weighting combination in order to overweight the importance of the original query. On the opposite, synomyms were underweighted, as well as the preferred acronym in the thesaurus. Thus, for the topic 93 ('Provide information about the role of the gene DRD4 in the disease Alcoholism.'), the following synomyms and the preferred acronym are added with different weights in the query:

**Prefered symbol**: DRD4

**Synonyms**: D4DR, d4 dopamine receptor, DRD IV, d dopamine iv receptor, 4 d dopamine receptor, drd4, d4dr, drd-4, drd 4.

## Results

Separately, the run produced with pivoted normalization (EXP1) performed slightly better than other state-of-the-art systems on the official data (cf. Aronson and al. 2005), achieving a map = 0.2373. When combined (using a CombSUM linear combination, cf. Fox and al. 1994) with a massively expanded run provided by Jacques Savoy from the University of Neuchatel (UniGeNe, map = 0.2150), which was produced with the gene and protein names expansion (Abdou and al. 2005), the combined system (UniGe2) obtained a **map = 0.2396**, which is our best official submission. Post-competitions experiments and results are reported in Abdou and al. 2005.

## Conclusion

Pivoted length normalization seems effective for retrieval in MEDLINE, but further experiments are needed to establish the effectiveness of our modified version. On the opposite, thesaurus-based gene expansion seems rather ineffective. Thesaurus-based expansion for other entities resulted in some modest improvement at least on the ten tuning queries. Run fusion based on linear combinations seems also effective (Ruiz 2005).

## Acknowledgments

The study reported in this paper has been partially supported by the SNF (EAGL project 3252B0-105755). The first author was supported by a visiting faculty grant (ORAU) at the Lister Hill Center of the National Library of Medicine in 2005. We would like to thank Alan 'anti-B' Aronson, Dina Demner-Fushman, Susanne M. Humphrey, Jimmy Lin, Hongfang Liu, Miguel E. Ruiz, Lawrence H. Smith, Lorraine K. Tanabe, W. John Wilbur for the fruitful discussions we had during our joint TREC meeting at the NLM.

| |
|---|
| dtu $w_{ij}$ = $\frac{\left(\ln\left(\ln((tf_{ij}) * K_{Length(Feature)}) + 1\right) + 1\right) \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$ |
| dtn $w_{ij}$ = $[\ln(\ln(tf_{ij}) + 1) + 1] \,.\, idf_j$ |

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

## References

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

[2] Alan R. Aronson, Dina Demner-Fushman, Susanne M. Humphrey, Jimmy Lin, Hongfang Liu, Patrick Ruch, Miguel E. Ruiz, Lawrence H. Smith, Lorraine K. Tanabe, W. John Wilbur (2005) Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. TREC Proceedings. TREC 2005, Gaithersburg, MD, USA.

[3] Christian Lovis, Robert H. Baud, Anne-Marie Rassinoux, P. A. Michel, Jean-Raoul Scherrer (2007) Building Medical Dictionaries for Patient Encoding Systems: A Methodology. AIME 1997: 373-380

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

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

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

[7] Ruiz, M.E. (2005) Experiments on Genomics ad hoc Retrieval. TREC Procedings, TREC 2005. Gaithersburg, MD.

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

[9] P Ruch (2005) *Automatic Assignment of Biomedical Categories: Toward a Generic Approach*, Bioinformatics (Advance Access published online). 2005.