Report on the TREC 2004 Experiment: Genomic Track

Patrick Ruch[1a], Christine Chichester[c], Gilles Cohen[a], Frédéric Ehrler[ad], Paul Fabry, Johan Marty[b], Henning Müller[a], Antoine Geissbühler[a]

[a]*SIM, University Hospital of Geneva, Geneva*
[c]*GeneBio SA, Geneva*
[d]*AI Laboratory, University of Geneva, Geneva*

Summary

Because of corruptions in the XML TREC Genomics collection, which were detected only some days before the submission deadline, we were not able to submit runs for the ad hoc retrieval task (task I), although relevance judgements made after polling were used to evaluate our approaches, and therefore this report mostly focuses on the text categorization task (task II: triage and annotation).

Task I. Our approach uses thesaural resources (from the UMLS) together with a variant of the Porter stemmer for string normalization. Gene and Protein Entities (GPE) of the collection were simply marked up by dictionary look up during the indexing in order to avoid erroneous conflation: strings not found in the UMLS Specialist lexicon (augmented with various English lexical resources) were considered as GPE and were moderately overweighed. Two different weighting schemas were tested: first, a standard tf-idf with cosine normalization, second a weighting based on the deviation from randomness model. For indexing the Genomic collection, the following MEDLINE records were selected: article's titles, MeSH and RN terms, and abstract fields. We investigated the use of high-precisions strategies and our system returned only highly reliable documents so that some queries were not answered by the system. Our best results achieved an average precision of 60%. The score was obtained using UMLS resources and GPE (Gene and Protein Entity) tagging together with a combination of a classical atc.ltn schema (following SMART notation) with a deviation from randomness [8] weighting.

Task II. We participated in both the triage and annotation tasks. For these tasks we attempted to adapt a Gene Ontology categorizer, which showed very effective results in the context of the BioCreative challenge, where training data were very sparse. The tool was completed by a naïve Bayes learner in order to take advantage of the TREC training data. The use of the last year GeneRIF extraction tool has also been evaluated.

Introduction

Systems for text mining are becoming increasingly important in biomedicine because of the exponential growth of knowledge. The mass of scientific literature needs to be filtered and categorized to provide for the most efficient use of the data. The problem of accessing this increasing volume of data demands the development of systems that first, can retrieve pertinent information from unstructured texts and second, can help professional curators to annotate high-quality DBs in the biomedical domain (as in SwissProt with Gene Ontology annotations [2, 6, 7, 11] or in MedLine with MeSH annotations [1]). The former task as been largely addressed in previous TREC studies, at least from a general point of view, however it is the second time that TREC investigates *ad hoc* retrieval in genomics. This year, the second task of the TREC 2003 Genomic track, has been discontinued and replaced by two different and complementary text categorization tasks. Task 1 (triage task) aims at deciding whether a given article is a good candidate for Gene Ontology annotation or not. Task 2 (annotation task) attempts to predict axe of the Gene Ontology is likely to be selected to support the annotation task of the curator. For these two tasks, full-text articles are also available; however following previous results [2, 11, 12], which tend to show that using full-text articles is no more effective than using abstracts, we decided to work on MEDLINE records only.

For this second participation in the TREC genomics track, we attempted a far as possible to reuse previously tested methods. The GeneRIF extractor and argumentative classifier were shown highly effective for TREC 2003 [9], while the Gene Ontology categorizer and passage retrieval module were successfully used in the context of the BioCreative campaign. Although results were disappointing in contrast to what was achieved in the previous competitions, it is still unclear to decide whether a completely different approach starting from scratch would have been more effective or if a better integration of the different tools would have brought better outcomes.

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

Methods

Because the annotation task is somehow related to the GeneRIF extraction, we attempted to use the GeneRIF extraction tools developed last year for the information extraction task of the Genomics track.

For the triage task, we rely on a set of Bayesian learners. Three one-class classifiers using three different features (stems, bigrams and trigrams) are linearly combined to get a final binary decision: relevant or not relevant for Gene Ontology annotation. Two runs were submitted: gt2 (official name: geneteam1) uses only stems and gt3 uses a combination of stems and bigrams (official name: geneteam3). A third official run, which was supposed to combine all features (geneteam1) was submitted but due to inappropriate data manipulation, it appears that the submitted file was in fact the same as gt2.

For the annotation task, we combine three serial steps: passage selection; Gene Ontology categorization; density estimation.

Passage selection

The first step consists in selecting a segment of text likely to support the 3-class annotation. To select the appropriate passage, we use the GeneRIF extractor [9] developed for TREC 2003. First, the system ranks sentences into four argumentative moves (PURPOSE and CONCLUSION sentences are preferred to RESULTS and METHODS), then a second ranking based on the targeted gene or protein is applied, finally, the selected sentence is "shortened" to remove non-content bearing stylistic phrases such as *in this paper we report that…* The original system has been refined to favour sentences where appears the targeted proteins. This last step combines exact match and fuzzy match based on a string-to-string edit distance calculus [11].

Gene Ontology categorization

The selected textual passage is then sent to the Gene Ontology categorizer. The tools combines three binary classifier, one for each axe of the ontology. Each basic classifier return a set of candidate concepts together with a class estimate based on the categorization status value (CSV). For each basic classifier, a an empirical threshold calculated on the BioCreative data is applied, so that the CSV can serve to return a binary decision.

Density estimation

While the Gene Ontology (GO) categorizer estimate the relevance of each returned GO candidate term, the density estimator provides a synthetic measure for each of the three axes. The density estimator depends on two parameters: 1) a term-based factor (TBF) based on the CSV assigned to each of the top-N considered terms; 2) a voting factor (VF): the number of terms (N) returned by the categorizer. The values for N as well as its linear combination with VF were established based on the training set for each Gene Ontology axe.

Results

Results for the triage task are reported in Table 1. Results for the annotation task are reported in Table 2.

Run #	Precision	Recall	F-score	Utility
1 (gt2)	0.1333	0.1333	0.1333	0.0900
2 (gt3)	0.1829	0.1833	0.1831	0.1424
Best U.	-	-	-	0.6512

Table 1. Results for the triage task (ranking based on the utility measure).

Run #	Precision	Recall	F-score	Utility
1 (gta5)	0.2274	0.7859	**0.3527**	0.6523
2 (gta4)	0.2090	0.9354	0.3417	0.7584
3 (gta2)	0.2025	0.9535	0.3340	0.7658
4 (gta1)	0.2090	0.9778	0.3248	0.7757
5 (gta3)	0.1938	0.9798	0.3235	**0.7760**
Best U.	-	-	-	0.7842

Table 2. Results for the annotation task (ranking based on the utility measure).

As for the metrics, it is interesting to note that F-score and utility measures are inversely ranked: the best run regarding the F-score (0.3527) is the worst regarding the utility measure (0.6523) and the best run regarding the utility measure (0.7760) obtain the lowest F-score (0.3235). Regarding the utility defined for the task, our best results are statistically similar to those obtained by the best TREC runs but the resulting system tends to classify all instances as positive !

Conclusion

From a general perspective, current classification power of the triage and annotation tool are obviously insufficient. In particular, when the system tries to estimate the similarity between the input text and the cellular component axe of the Gene Ontology, the argumentative classification, which tends to select CONCLUSION and PURPOSE passages should be refined to take advantage of METHODS segments, since cellular components and tissues are often given in METHODS and MATERIALS sections of articles [13]. Following experiments made for searching similar documents in MedLine [14], we plan to evaluate the impact of global argumentation [15] and local rhetorical moves in biomedical abstracts [3] for ad hoc IR tasks.

Acknowledgments

The study reported in this paper has been supported by the SNF (MEDTAG, Grant 3200-065228.01 and an EU/OFES grant (SemanticMining, IST Grant 507505/03.0399).

References

[1] [1] A Aronson, O Bodenreider, H Chang, S Humphrey, J Mork, S Nelson, T Rindflesch, W Wilbur, The NLM Indexing Initiative. Proc AMIA Symp. 2000. p. 17-21.

[2] [2] F Ehrler and P Ruch. Report on the BioCreative Experiment: Task Presentation, System Description and Preliminary Results. In *Notebook of BioCreative 2004*, Granada, March 2004.

[3] S Teufel and M Moens, Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: I. Mani, M. Maybury (eds.), Advances in automatic text summarization, MIT Press, 1999.

[4] J Savoy, Y Rasolofo and L Perret. Report on the TREC 2003 Experiment: Genomics and Web Searches. In *Notebook of the TREC-2003*, Gaithersburg, p. 686-697.

[5] W Hersh and RT Bhupatiraju. TREC genomics track overview. In *Notebook of the TREC-2003*, 148-157.

[6] Blaschke, C., Andrade, M.A., Ouzounis, C.A., and Valencia, A. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, 1999, p. 60-67.

[7] [3] PB Dobrokhotov, C Goutte, AL Veuthey, E Gaussier. A Probabilistic Information Retrieval Approach to Medical Annotation in SWISS-PROT. *Stud Health Technol Inform*.2003;95:421-6.

[8] [4] G Amati, and C van Reijsbergen, Probabilistic models of information retrieval based on measuring the divergence from randomness**,** *TOIS* 20 (4), 2002, p. 357-389.

[9] [5] P Ruch, C Chichester, G Cohen, G Coray, F Ehrler, H Ghorbel, H Müller, V Pallotta. Report on the TREC 2003 Experiment: Genomic Track, *TREC* 2003, 2004.

[10] [6] W Hersh, R Bhupatiraju, TREC genomics track overview, *TREC 2003* , 14-23, 2004.

[11] [7] F Ehrler, A Yepes, P Ruch , Data-poor Categorization and Passage Retrieval for Gene Ontology Annotation in Swiss-Prot, *BMC Bioinformatics*, Special Issue on BioCreative (to appear).

[12] [8] M.J. Schuemie, M. Weeber, B.J.A Schijvenaars, E.M. van Mulligen, C.C. van der Eijk, R. Jeliert B. Mons, and J. A. Kors. Distribution of information in biomedical abstracts and full text publications. *Bioinformatics*. 2004. (to appear).

[13] [9] E Camon. BioCreative report. *BMC Bioinformatics*, Special Issue on BioCreative. (personnal communication).

[14] [10] I Tbarhiti, C Chichester, F Lisacek and P Ruch, Using Argumentation to Retrieve Articles with Similar Citations from MEDLINE, *Joint Workshop on Natural Language Processing in Biomedical Applications (JNLPBA)*, COLING 2004.

[15] [11] Y Mizuta and N Collier, Zone Identification in Biology Articles as a Basis for Information Extraction, *Joint Workshop on Natural Language Processing in Biomedical Applications (JNLPBA)*, COLING 2004.