Combining Light-Weight Retrieval Strategies for Robust Text Categorization

Patrick Ruch
University and University Hospitals of Geneva

Motivation

We report on the development of a general purpose text categorization system designed to automatically assign biomedical categories to any input text. Unlike usual automatic text categorization systems, which rely on data-intensive models extracted from large sets of training data, our categorizer is largely data-independent and so it can be used when training data are not available provided that a small set of instances is available for tuning the system. Like it is usual with information retrieval engines, the tool provides a ranked list of categories, which can then be interactively filtered by the user.

Methods

In order to evaluate the robustness of our approach we test the system on two different biomedical terminologies: the Medical Subject Headings (MeSH) and the Gene Ontology (GO). MEDLINE abstracts are used as input documents in each case. Our lightweight classifier combines a pattern matcher with a vector space retrieval engine, and uses both stems and linguistically-motivated indexing units. We also evaluate the impact of using synonyms on the categorization task.

Results

Results show the effectiveness of phrase indexing (+5%, measured by mean average precision) for both GO and MeSH categorization. In general, the categorization power of the tool depends on the controlled vocabulary: precision at high ranks ranges from above 90% for MeSH to less than 20% for GO. Synonyms are useful for GO categorization, while for MeSH categorization it is observed that the use of such resources does not help.

Conclusion

From a general perspective, it is concluded that approaches based on retrieval methods can provide an effective categorization strategy when training data are not available but that effectiveness is directly related to the controlled vocabulary. Finally, standard metrics, which are used to measure categorization effectiveness, are questioned.

Full article forthcoming in *Bioinformatics*.

Address for Correspondence:
Patrick Ruch
University Hospital of Geneva
Medical Informatics Division
CH-1211 Geneva 14
phone: +41 22 372 61 64
fax: +41 22 382 86 80
patrick.ruch@sim.hcuge.ch

Acknowledgements:
The study has been supported by the Swiss National Foundation (Grant 3200-065228) and by the EU (SemanticMining Grant 507505 - Swiss OFES Grant 03.0399).