Assignment of ontology-based broad semantic classes to biomedical text

K. Bretonnel Cohen
Computational Bioscience Program, U. Colorado School of Medicine
Department of Linguistics, U. Colorado at Boulder
kevin.cohen@gmail.com

Abstract

Natural language processing of biomedical text benefits from the ability to recognize broad semantic classes, but the number of semantic types is far bigger than is usually treated in newswire text. A method for broad semantic class assignment using lightweight linguistic analysis is described and evaluated using traditional and novel methods.

1 Introduction

Experience with coreference resolution, information extraction, and document classification indicates that these natural language processing tasks, and probably others, benefit from the ability to assign semantic classes to entities in text. However, the set of semantic classes in biomedical text is much larger than the traditional MUC categories of PERSON, ORGANIZATION, and LOCATION that are normally dealt with in newswire text. This talk addresses the hypothesis that it is possible to label as many as twenty semantic types in biomedical text, with those semantic classes being grounded in the topic areas of ontologies from the Open Biomedical Ontologies collection at the National Center for Biomedical Ontology. Belonging to the topic of an Open Biomedical Ontology is then considered to constitute membership in that semantic class. We determined membership in semantic classes through a lightweight linguistic analysis consisting of four techniques of decreasing levels of stringency.

2 Method

Twenty ontologies thought to be relevant to mouse genomics were selected from the Open Biomedical Ontologies collection. Four methods were used to match these to pre-extracted phrases in text. In the first technique, a simple exact match was attempted. In the second technique, a normalized form of the term, with whitespace and punctuation removed, was used to attempt a match. (Identical normalization was applied to the input text.)

In the third technique, a lightweight linguistic technique was applied. Simple heuristics were used to extract the head noun from each term in the ontology, and from the input text phrase. A match of the headwords was then attempted. If this failed, both headwords were stemmed and a match was attempted again.

The four techniques were applied sequentially. This allowed for modular evaluation of each technique, and at runtime, it allows for the user to select a minimum level of stringency.

The method was evaluated by three techniques:

1. Measuring precision, recall, F-measure, and accuracy with a unique corpus of full-text journal articles marked up with a number of biomedical ontologies.
2. Using the ontologies themselves as input. This is a novel evaluation technique, and it was shown to be robust in detecting errors in the method.
3. Running the method against a structured test suite (Cohen et al. 2010) used for ontology concept recognition.

Proceedings of the Workshop on Biomedical Natural Language Processing, pages 1–2,
Hissar, Bulgaria, 15 September 2011.

A micro-averaged F-measure of 72.32 was achieved on the corpus. The macro-averaged F-measure was 75.31. Accuracies of 77.12 to 95.73% were achieved, depending on the ontology. The lightweight linguistic technique of head noun extraction was found to make a significant contribution to efficacy, sometimes a very large contribution.

As would be expected, the evaluation against the ontologies themselves typically achieved very high performance numbers, but in two cases it uncovered significant lapses in our processing of the ontologies themselves, indicating that this novel evaluation method is robust.

Finally, the structured test set yielded considerable insight into the strengths and weaknesses of the method.

A full description of the materials, method, and evaluation can be found in Cohen et al. (2011).

References

Cohen, K. Bretonnel; Christophe Roeder; William A. Baumgartner Jr.; Lawrence E. Hunter; and Karin Verspoor (2010) Test suite design for biomedical ontology concept recognition systems. *Languages Resources and Evaluation Conference*, pp. 441-446.

Cohen, K. Bretonnel; Tom Christiansen; William A. Baumgartner Jr.; Karin Verspoor; and Lawrence E. Hunter (2011) Fast and simple semantic class assignment for biomedical text. *Biomedical Natural Language Processing 2011*, pp. 38-45.