Automatic Term List Generation for Entity Tagging

Ted Sandler   Andrew I. Schein   Lyle H. Ungar
University of Pennsylvania, Department of Computer and Information Science

Abstract

Entity tagging is perhaps the most basic tasks in information extraction. Consequently, if any sophisticated information extraction is to be done, this task must be done well. Currently, state of the art taggers are trained with hand-labeled training data that resembles the text to which they will be applied. However, such training data is rarely abundant enough to contain more than a modest subset of entities from the target entity class. To remedy this, designers of information extraction systems frequently augment their systems with large lists of known members of the entity class so as to cover most of the prototypical, and perhaps not so prototypical, positive instances that the training data may not contain. Such entity lists, or term lists as we refer to them, contribute significantly to entity tagging performance.

Unfortunately, term lists are not available for every domain, and where they do exist, they are usually incomplete. Therefore, finding cheap and efficient means of generating term lists is of importance to the information extraction community, particularly in so far as such methods can generate lists that improve the performance of information extraction systems.

Unsupervised methods for bootstrapping domain-specific lexicons from large corpora have existed for quite some time, as have methods for partitioning words into pseudo-semantic classes based on their distributional properties. While not perfect, these methods work surprisingly well and are straightforward to implement. Therefore, it is natural to ask whether such methods can be used to generate lists of terms, and whether the lists they generate can improve the functionality of a larger information extraction system. This raises the question: how well do such lists fair in comparison to lists compiled through manual means, and how well do they fair in comparison to lists learned through supervised methods?

In this work we investigate these questions by demonstrating how term lists can be built using a shallow parser to extract syntactic contexts from a large body of domain-relevant text. We show that a number of different feature weighting schemes and clustering methods can be used to different effect. Experiments on generating lists of terms from MEDLINE abstracts and using them as features in a state-of-the-art CRF-based gene tagger show that the automatically generated lists boost system performance, and do so to a degree competitive with a hand curated list. However, our results also indicate that when abundant hand labeled data exists, supervised learning can generate lists even more beneficial than either unsupervised learning or hand curation.

Full article forthcoming in *Bioinformatics*.

Address for Correspondence:
Ted Sandler
University of Pennsylvania
Department of Computer and Information Science
3330 Walnut Street
Philadelphia, PA 19104
tsandler@seas.upenn.edu

Acknowledgements:
This work was supported in part by NIH Training grant T32HG 0000461. The authors also wish to thank Ryan T. McDonald for his helpful advice and assistance.

References

1. Dekang Lin. Automatic retrieval and clustering of similar words. In COLING-ACL, pages 768-774, 1998.

2. Ryan McDonald and Fernando Pereira. Identifying gene and protein mentions in text using conditional random fields. In A critical assessment of text mining methods in molecular biology (BioCreAtIvE), 2004.

3. Lorraine Tanabe and W. John Wilbur. Tagging gene and protein names in full text articles. In Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, pages 9-13, Philadelphia, July 2002. Association for Computational Linguistics.

4. Lorraine Tanabe and W. John Wilbur. Generation of a large geneprotein lexicon by morphological pattern analysis. Journal of Bioinformatics and Computational Biology, 1(4):611-626, 2004.