Co-reference Resolution in Biomedical Texts: a Machine Learning Approach

Xiaofeng Yang [1], Jian Su [1], Huaqing Hong [2], Yuka Tateisi [3], and Jun'ichi Tsujii [3]

[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613.
[2]National Institute of Education, 1 Nanyang Walk, Singapore, 637616.
[3]University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, JAPAN, 113-0033.

ABSTRACT

Motivation: Coreference resolution, the process of identifying different mentions of an entity, is a very important component in a text-mining system. Compared with the work in news articles, the existing study of coreference resolution in biomedical texts is quite preliminary by only focusing on specific types of anaphors like pronouns or definite noun phrases, using heuristic methods, and running on small data sets. Therefore, there is a need for an in-depth exploration of this task in the biomedical domain.

Results: In this article, we presented a learning-based approach to coreference resolution in the biomedical domain. We made three contributions in our study. Firstly, we annotated a large scale coreference corpus, MedCo, which consists of 1,999 medline abstracts in the GENIA data set. Secondly, we proposed a detailed framework for the coreference resolution task, in which we augmented the traditional learning model by incorporating non-anaphors into training.

Lastly, we explored various sources of knowledge for coreference resolution, particularly, those that can deal with the complexity of biomedical texts. The evaluation on the MedCo corpus showed promising results. Our coreference resolution system achieved a high precision of 85.2% with a reasonable recall of 65.3%, obtaining an F-measure of 73.9%. The results also suggested that our augmented learning model significantly boosted precision (up to 24.0%) without much loss in recall (less than 5%), and brought a gain of over 8% in F-measure.

Dagstuhl Seminar Proceedings 08131
Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives
http://drops.dagstuhl.de/opus/volltexte/2008/1522