Weijian Xuan, Stanley J. Watson, and Fan Meng
Mental Health Research Institute
University of Michigan
mengf@umich.edu

Disambiguating Sentence Boundaries in Biomedical Corpus

Sentence Boundary Disambiguation (SBD) is a prerequisite for virtually any syntactic analysis of text corpus, such as part-of-speech tagging, sentence alignment and text segmentation. SBD may appear to be an easy task. In reality, however, achieving high accuracy is not a trivial problem due to the ambiguity of punctuation marks. For example, a period mark can signal a decimal point, a sentence boundary, an abbreviation, or even an abbreviation at a sentence boundary. In 6.5 million Medline abstracts we investigated, about 33% of the periods are ambiguous.

Biomedical literature is growing at unprecedented rate. Medline [1] database alone already contains over 13 million citations from 4000 biomedical journals. This vast corpus is of tremendous value for researchers and students from a range of fields for research and learning purposes. While extensive efforts have been devoted to mining knowledge from biomedical text, few attempts are targeted at SBD in such corpus, which has some unique features that can degrade the accuracy of algorithms [2,3] designed for general English genre. For example, significant more abbreviations and proper names, lack of naming convention, frequent inline citations, etc., will considerably hamper SBD.

We developed an efficient method using a combination of heuristic and statistical strategies. We incorporated linguistic resources and knowledge bases, e.g. WordNet and Unified Medical Language System [4], to assist in disambiguating abbreviations. We also took advantage of a number of linguistic features, such as n-grams, sub-section segmentations, sequential groups and patterns of citations which are common in academic text.

Our approach does not require POS taggers or training procedures. Experiments with biomedical test corpora show our system significantly outperforms existing sentence boundary determination algorithms, particularly for full text biomedical literature. Our system is very fast and it can be used as a component for large-scale biomedical text mining systems.

References:

1. PubMed, Interface for Medline and other databases, http://www.ncbi.nlm.nih.gov/entrez.
2. Palmer, D.D. and Hearst, M.A., Adaptive Sentence Boundary Disambiguation. In Proceedings of the 4th Conference on Applied Natural Language Processing. 1994. Stuttgart, Germany.
3. Reynar, J.C. and Ratnaparkhi, A., A Maximum Entropy Approach to Identifying Sentence Boundaries. In Proceedings of the 5th Conference on Applied Natural Language Processing. 1997. Washington, D.C.
4. Humphreys, B.L., et al., The Unified Medical Language System: An Informatics Research Collaboration. Journal of the American Medical Informatics Association, 1998. 5(1):1-11.