Meilan Zhang
School of Education
University of Michigan

Weijian Xuan
Mental Health Research Institute
University of Michigan
wxuan@umich.edu

Towards Discovering Linguistic Features from Scientific Abstracts

The sheer volume of scientific literature is growing at unprecedented rate. For example, in biomedical domain, Medline alone contains over 12 million abstracts. The increasing availability of electronic abstracts not only facilitates literature research, but also provides a good opportunity to compile academic text corpus and discover linguistic knowledge for teaching, learning purposes.

We are developing a Natural Language Processing (NLP) system to automatically discover linguistic features from scientific text. This paper describes our initial results on extracting feature terms and sentence patterns from scientific abstracts. We built a text corpus using 6,400 full-length biomedical papers, which consists of 24 million words. Frequency profiling was used to extract feature terms that differentiate the abstract section from other sections. In order to filter out domain-specific terms, we utilized Brown corpus and a corpus we curated from education field consisting of 30 million words. Log likelihood statistic was calculated to identify domain keywords. As a result, 325 terms and their inflections were selected, e.g., investigate, develop, purpose, describe, etc. We scan each abstract and keep only sentences containing one or more feature words. We use Brill's tagger to assign POS tags and apply heuristic rules to extract noun phrases (NP). The sentences were then sent to Link Grammar Parser (LGP) to parse its syntactic structure. If LGP fails to parse a complex sentence, we will replace NPs and precompiled collocations with capitalized proper nouns to simplify the sentence and rerun LGP. After parsing all selected sentences, we group sentence patterns by feature terms and syntactic structures. For example, "We developed [NP[DET (a|an)] [JJ(analytica|new|...)?] [NN(model|algorithm|...)]] for [NP]". Patterns of relatively high frequency can be more interesting to learners and developers. Such feature terms and patterns can guide students in writing scientific abstracts and enhance the effectiveness of automated systems that score student writings. Furthermore, the linguistic characteristics discovered by our system can be used to boost text mining systems in text segmentation and fact extraction.

References:

1. PubMed (2004). Interface for Medline and other databases http://www.ncbi.nlm.nih.gov/entrez.
2. Brill E (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. Computational Linguistics, 21(4):543-565
3. Sleator D, Temperley D (1993). Parsing English with a Link Grammar, Third International Workshop on Parsing Technologies.