Integrated Annotation of Biomedical Text: Creating the PennBioIE Corpus

Mark A. Mandel
(Institute for Research in Cognitive Science
Linguistic Data Consortium, University of Pennsylvania)

In 2001 the Institute for Research in Cognitive Science of the University of Pennsylvania received a grant from the National Science Foundation to develop "qualitatively better methods for automatically extracting information from the biomedical literature". In this presentation I will concentrate on the development of our corpora. Our texts consist of PubMed abstracts in two biomedical domains (the molecular genetics of cancer and the inhibition of cytochrome P-450 enzymes). All our annotation is standoff rather than in-line, using software developed at Penn. After paragraph, sentence, and word/token segmentation, we annotate each abstract for as many as about 20 different entity types before applying and manually correcting part of speech annotation, which in turn is necessary for manual syntactic annotation (treebanking). The highly technical "dialects" of these abstracts require correspondingly specialized tokenization and part of speech procedures, and the entity annotators are permitted and required to correct the results of the automatic tokenization as necessary. Similarly, all the development of our definitions and guidelines has involved continual intensive interaction between the biomedical specialists, the software developers, and the annotators, whose insights and feedback have proved essential to the process and thus to the success of the entire project. [This material is based upon work supported by the (US) National Science Foundation under Grant No. ITR-0205448.]