Linguistic Corpora for Biomedical Information Extraction

May 16, 2003

1 Introduction

Work over the last few years in literature data mining for biology has progressed from linguistically unsophisticated models to the adaptation of natural language processing techniques that use full parsers ([6, 9]) and coreference to extract relations that span multiple sentences ([7, 3]) (For an overview, see [4]). However, there has been a lack of annotated corpora that can fuel further work in this direction in the same way that the development of syntactically annotated corpora such as the Penn Treebank ([5]) led to the development of statistical language parsers ([1]).

To address this situation, we are developing new linguistic resources in three categories: a large corpus of biomedical text annotated with syntactic structures (Treebank) and shallow semantic structures ("proposition bank" or Propbank); a large set of biomedical abstracts and full-text articles annotated with entities and relations of interest to researchers, such as enzyme inhibition, or mutation/cancer connections (Factbanks); and broad-coverage lexicons and tools for the analysis of biomedical texts. We are also developing and adapting software tools that allow human experts to annotate biomedical texts for entity tagging, as well as for treebanking and propbanking. We are focusing initially on two applications: drug development, in collaboration with researchers in the Knowledge Integration and Discovery Systems group at GlaxoSmithKline, and pediatric oncology, in collaboration with researchers in the eGenome group at Children's Hospital of Pennsylvania. These applications, worthwhile in their own right, provide excellent test beds for broader research efforts in natural language processing and data integration.

2 Entity Tagging

Say something about entity tagging, how we're approaching the issue of metonomy, coreference.

What is the status of annotation? Perhaps some numbers on what's been done, etc.?

something about tools.

Are we annotating for entities and relations that haven't been done before? I know people have been concerned with "inhibit", but how about variations? I'm not sure how the connects with the protein-protein interactions that have been of concern in other papers.

In the full paper we discuss in further detail our schemes for handling metonomy and coreference.

3 Treebanking and Propbanking for Relation Extraction

As has been noted (e.g., [7, 9]), the same relation can be take a number of syntactic forms. For example, the family of words based on the morpheme *inhibit* occurs commonly in MEDLINE abstracts about CYP enzymes in patterns like *A inhibited B, A inhibited the catalytic activity of B, inhibition of B by A,* etc. As [7] notes, it is even possible for an entity to incorporate relational information - e.g., *Tissue inhibitors of metalloproteinase*.

Such alternations have led to the use of pattern-matching rules (often hand-written) to match all the relevant configurations and fill in template slots based on the resulting pattern matches. However, by using "semantic taggers" ([2]), such IE patterns become much simpler to crate, and also potentially more accurate, whether for human rule writers or for machine learning algorithms. The basic idea is that the Propbank corpora contain a "shallow semantic" analysis that normalizes the predicate-argument structure of all the occurrences of some verb. For example, the "inhibitee" would always be annotated as "ARG1", no matter what particular expression the "inhibit" relation takes in each case.

Such semantic taggers have recently been used for information extraction from financial domains ([8]). These taggers use the results from a statistical syntactic parser trained on the Penn Treebank, together with training on the same corpora annotated for shallow semantic structure (The Penn Propbank). However, since the Penn Treebank and Propbank contain Wall Street Journal text, parsers and semantic taggers trained on those corpora will not be very successful when applied to the biomedical domain, and it is therefore essential for this approach to have a corpus of MEDLINE abstracts and articles annotated for both syntactic structure (Treebanking) and the shallow semantic structure (the Propbank). New issues for such annotation arise when applied to this domain, and in the full paper we discuss these aspects in further detail.

References

[1] Mike Collins. Three Generative, Lexicalised Models for Statistical Parsing. In *Proc. of ACL-1997*, 1997.

[2] Daniel Gildea and Martha Palmer. The Necessity of Syntactic Parsing for Predicate Argument Recognition. In *Proc. of ACL-2002*, 2002.

[3] U. Hahn, M. Romacker, and S. Schulz. Creating knowledge repositories from biomedical reports: The medsyndikate text mining system. In *Proceedings of the Pacific Rim Symposium on Biocomputing*, pages 338–349, 2002.

[4] Lynette Hirschman, Jong C. Park, Junichi Tsuji, Limsoon Wong, and Cathy H. Wu. Accomplishments and challenges in litearture data mining for biology. *Bioinformatics Review*, 18(12):1553–1561, 2002.

[5] M. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 1993.

[6] J. Park, H. Kim, and J. Kim. Bidirectional incremental parsing for automatic pathway identication with combinatory categorial grammar. In *Proceedings of the Pacific Rim Symposium on Biocomputing*, pages 396–407, 2001.

[7] J. Pustejovsky, J. Castano, and J. Zhang. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Proceedings of the Pacific Rim Symposium on Biocomputing*, pages 362–373, 2002.

[8] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*, Sapporo, Japan, 2003.

[9] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomecial papers using a full parser. In *Proceedings of the Pacific Rim Symposium on Biocomputing*, pages 408–419, 2001.