Shallow Semantic Annotation of Biomedical Corpora for Information Extraction *

Seth Kulick and Mark Liberman
Institute for Research in Cognitive Science
3401 Walnut Street, Suite 400A
Philadelphia, PA 19104

Martha Palmer and Andrew Schein
Dept. of Computer and Information Science
Levine Hall
3330 Walnut Street
Philadelphia, PA 19104

June 27, 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 (NLP) techniques that use full parsers ([11, 16]) and coreference to extract relations that span multiple sentences ([12, 6]) (For an overview, see [7]). 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 ([10]) led to the development of statistical language parsers (e.g., [3]).

To address this situation, we[1] are developing new linguistic resources in three categories: a large corpus of biomedical text annotated with syntactic structure (Treebank) and predicate-argument structure ("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 Philadelphia. These applications, worthwhile in their own right, provide excellent test beds for broader research efforts in natural language processing and data integration.

A guiding principle for this project is the annotation of a corpus with different levels of shallow semantics that will permit the development of NLP tools to extract the desired entities and relationships. These levels consist of entity tagging, reference and coreference, propbanking, and factbanking. Key to the approach is the integration of the different levels of semantic and syntactic annotation with an eye towards clear conceptual semantics, feasibility of implementation, and likelihood of practical benefit. This is a novel approach from the point-of-view of NLP since previous efforts at treebanking and propbanking have been independent of the special status of any entities, and previous efforts at entity annotation have been independent of corresponding layers of syntactic and semantic structure.

*This work was supported by NSF grant ITR-EIA-0205448 and the last author was also supported by NIH Training Grant in Computational Genomics T-32-HG00046.

[1]The authors are members of the project group at the University of Pennsylvania, which includes faculty and graduate students from the Computer Science and Linguistics departments, the Institute for Research in Cognitive Science, and associated staff.

2 Entity Tagging

We are developing guidelines for entity tagging in the areas of pediatric oncology and enzyme inhibition, in collaboration with domain experts. A key feature of entity tagging is establishment of standardized external reference using biomedical databases such as the Gene Ontology. We are currently concerned with the annotation of genes, malignancies, and chemicals. As has been noted by others, there are often ambiguities in the usage of the entity names. For example, it is sometimes unclear as to whether it is the gene or protein being referenced, or the same name might refer to the gene or the protein at different locations in the same document. Our approach to this problem is influenced by the named entity annotation in the Automatic Content Extraction (ACE) project ([4]), in which "geopolitical" entities can have different roles, such as "location" or "organization". Analogously, we consider a "gene" to be a composite entity that can have different roles throughout a document.

3 Reference and Coreference

Along the lines of some of the suggestions in [15], our view is that traditional notions of coreference in computational linguistics have conflated important distinctions, and we have separated out five distinct notions of reference or coreference for annotation:

Acronym Definition The usage of an acronym points back to the antecedent where it is defined. (See also [2]).

Acronym Linkage Acronyms are linked together, with the first occurrence in turn pointing to the definition of the acronym with an acronym definition link.

Anaphor This includes more traditional cases of coreference, with pronouns or definite NPs used as anaphors, as in *K-Ras is an oncogene. It is mutated in...* or *K-Ras is an oncogene. This gene....* It also includes more abstract cases, in which the antecedent is more abstract - e.g., *The translocation, t(1:12, p36:q28) was found...This variation...*

Is-a Relation This includes cases of predicate nominals and appositives, such as *C-kit, a tyrosine kinase which plays an important role....* By separating this out from "anaphor", we maintain the constraint that members of a coreference (anaphor or acronym) chain must be in an equivalence relation ([15]). We also depart from the ACE guidelines, which partition predicative cases into coreference and relation annotations.

Database Reference This is not a case of coreference to another entity in the text, but rather of reference to one of the standard databases or ontologies.

4 Treebanking, Propbanking, and Factbanking for Relation Extraction

As has been noted (e.g., [12, 16]), the same relation can be take a number of syntactic forms. For example, the family of words based on *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.

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. Recent work in information extraction ([14]) has instead used a "semantic tagger" to extract argument roles for the predicates of interest, with the same argument roles resulting regardless of the particular syntactic context. For

example, with the “inhibit” relation, the “inhibitee” would always be annotated as “ARG1”, whether it appears as the object in an active construction or the subject in a passive. As a result, the IE patterns become much simpler to create, and also potentially more accurate, whether for human rule writers or for machine learning algorithms.

Such semantic taggers have been developed by using machine learning techniques trained on the Penn Propbank ([5, 8]). The basic idea is that the Propbank corpus contains a level of shallow semantics that normalizes the predicate-argument structure of all the occurrences of some verb. Crucially, the Propbank is designed as an additional layer of annotation on top of the syntactic structures in the Penn Treebank. This allows the semantic taggers to make use of the results from a statistical syntactic parser that has been trained on the Treebank, in addition to training on the same corpora annotated for the shallow semantic structure in the Propbank.

However, the Penn Treebank and Propbank involve the annotation of Wall Street Journal text, and so parsers and semantic taggers trained on those corpora will not be very successful when applied to the biomedical domain. It is therefore essential for this approach to have a corpus of biomedical texts such as MEDLINE articles annotated for both syntactic structure (Treebanking) and shallow semantic structure (Propbanking).

Recent work in Propbanking has been extended to the annotation of nominals that have argument structure. Interestingly, this includes nominalizations such as “mutation” and relational nouns such as “inhibitor” that cover some of the relations that are of concern in this domain. As [12] notes, it is possible for an entity to incorporate relational information - e.g., *Tissue inhibitors of metalloproteinase*, a property that will be captured within this annotation framework.

At a higher level of abstraction than the predicate-argument structure annotated in the Propbank, the Factbank will include information on the “variation” relation, covering the mutation (translocation, deletion, etc.) for any chromosomal abnormality. That is, any instances of such an abnormality will be recorded with the information regarding the type of the variation, the location, and the original and altered states of the amino acid(s) and/or nucleic acid(s).

There is a close link between the concerns of the predicate-argument structure of the Propbank and this more abstract level of information, since the annotation of nominals such as *mutation*, *translocation*, etc. will necessarily involve annotation of their arguments which in turn coincide with the desired variation information (type, location, etc.). We are therefore very hopeful that this approach to annotation for different levels of shallow semantics will provide the basis for the training of improved tools for relation extraction.

5 Software Tools

Coinciding with the publication of our entity and coreference annotations will be the release of state-of-the art software for entity tagging and coreference resolution. Our taggers will be based on various statistical methods for tagging sequences of data (See [9, 13, 1] for prior work). Our goal is to provide open source entity taggers accessible for field use in projects requiring entity tagging. These tools will provide a baseline for performance using the annotated corpus. Though the release of entity tagging software will be of immediate benefit to those working in field, we view as our primary contribution the publication of semantic and syntactic annotation that will pave the way for the next generation of entity tagging tools, developed by the community at large.

References

[1] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Nyu: Description of the mene named entity system as used in muc-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*,

1998.

[2] J. Castano, J. Zhang, and J. Pustejovsky. Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution for Natural Language Procesing*, 2002.

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

[4] Linguistic Data Consortium. Entity detection and tracking - phase 1 - EDT and metonymy annotation guidelines version 2.5 20021205, 2002. http://www.ldc.upenn.edu/Projects/ACE/PHASE2/Annotation/.

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

[6] 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.

[7] 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.

[8] Paul Kingsbury and Martha Palmer. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain, 2002.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

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

[11] 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.

[12] 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.

[13] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey, 1996.

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

[15] K. van Deemter and R. Kibble. On coreferring: Coreference annotation in MUC and related schemes. *Computational Linguistics*, 26(4):615–623, 2000.

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