HLT-NAACL 2004 Workshop: Biolink 2004,
Linking Biological Literature, Ontologies and Databases, pp. 61-68.
Association for Computational Linguistics.

Integrated Annotation for Biomedical Information Extraction

Seth Kulick and **Ann Bies** and **Mark Liberman** and **Mark Mandel**
and **Ryan McDonald** and **Martha Palmer** and **Andrew Schein** and **Lyle Ungar**
University of Pennsylvania
Philadelphia, PA 19104
{skulick,bies,myl}@linc.cis.upenn.edu,
mamandel@unagi.cis.upenn.edu,
{ryantm,mpalmer,ais,ungar}@cis.upenn.edu

Scott Winters and **Pete White**
Division of Oncology,
Children's Hospital of Philadelphia
Philadelphia, Pa 19104
{winters,white}@genome.chop.edu

Abstract

We describe an approach to two areas of biomedical information extraction, drug development and cancer genomics. We have developed a framework which includes corpus annotation integrated at multiple levels: a Treebank containing syntactic structure, a Propbank containing predicate-argument structure, and annotation of entities and relations among the entities. Crucial to this approach is the proper characterization of entities as relation components, which allows the integration of the entity annotation with the syntactic structure while retaining the capacity to annotate and extract more complex events. We are training statistical taggers using this annotation for such extraction as well as using them for improving the annotation process.

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 (Park et al., 2001; Yakushiji et al., 2001) and coreference to extract relations that span multiple sentences (Pustejovsky et al., 2002; Hahn et al., 2002) (For an overview, see (Hirschman et al., 2002)). In this work we describe an approach to two areas of biomedical information extraction, drug development and cancer genomics, that is based on developing a corpus that integrates different levels of semantic and syntactic annotation. This corpus will be a resource for training machine learning algorithms useful for information extraction and retrieval and other data-mining applications. We are currently annotating only abstracts, although in the future we plan to expand this to full-text articles. We also plan to make publicly available the corpus and associated statistical taggers.

We are collaborating with researchers in the Division of Oncology at The Children's Hospital of Philadelphia, with the goal of automatically mining the corpus of cancer literature for those associations that link specified variations in individual genes with known malignancies. In particular we are interested in extracting three entities (Gene, Variation Event, and Malignancy) in the following relationship: Gene X with genomic Variation Event Y is correlated with Malignancy Z. For example, *WT1 is deleted in Wilms Tumor #5*. Such statements found in the literature represent individual gene-variation-malignancy observables. A collection of such observables serves two important functions. First, it summarizes known relationships between genes, variation events, and malignancies in the cancer literature. As such, it can be used to augment information available from curated public databases, as well as serve as an independent test for accuracy and completeness of such repositories. Second, it allows inferences to be made about gene, variation, and malignancy associations that may not be explicitly stated in the literature, both at the fact and entity instance levels. Such inferences provide testable hypotheses and thus future research targets.

The other major area of focus, in collaboration with researchers in the Knowledge Integration and Discovery Systems group at GlaxoSmithKline (GSK), is the extraction of information about enzymes, focusing initially on compounds that affect the activity of the cytochrome P450 (CYP) family of proteins. For example, the goal is to see a phrase like

> Amiodarone weakly inhibited CYP2C9, CYP2D6, and CYP3A4-mediated activities

with Ki values of 45.1–271.6 μM

and extract the facts

```
amiodarone inhibits CYP2C9 with
                    Ki=45.1-271.6
amiodarone inhibits CYP2D6 with
                    Ki=45.1-271.6
amiodarone inhibits CYP3A4 with
                    Ki=45.1-271.6
```

Previous work at GSK has used search algorithms that are based on pattern matching rules filling template slots. The rules rely on identifying the relevant passages by first identifying compound names and then associating them with a limited number of relational terms such as *inhibit* or *inactivate*. This is similar to other work in biomedical extraction projects (Hirschman et al., 2002).

Creating good pattern-action rules for an IE problem is far from simple. There are many complexities in the different ways that a relation can be expressed in language, such as syntactic alternations and the heavy use of coordination. While sufficiently complex patterns can deal with these issues, it requires a good amount of time and effort to build such hand-crafted rules, particularly since such rules are developed for each specific problem. A corpus that is annotated with sufficient syntactic and semantic structure offers the promise of training taggers for quicker and easier information extraction.

The corpus that we are developing for the two different application demands consists of three levels of annotation: the entities and relations among the entities for the oncology or CYP domain, syntactic structure (Treebank), and predicate-argument structure (Propbank). 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. The decomposition of larger entities into components of a relation, worthwhile by itself on conceptual grounds for entity definition, also allows the component entities to be mapped to the syntactic structure. These entities can be viewed as semantic types associated with syntactic constituents, and so our expectation is that automated analyses of these related levels will interact in a mutually reinforcing and beneficial way for development of statistical taggers. Development of such statistical taggers is proceeding in parallel with the annotation effort, and these taggers help in the annotation process, as well as being steps towards automatic extraction.

In this paper we focus on the aspects of this project that have been developed and are in production, while also trying to give enough of the overall vision to place the work that has been done in context. Section 2 discusses some of the main issues around the development of the guidelines for entity annotation, for both the oncology and inhibition domains. Section 3 first discusses the overall plan for the different levels of annotation, and then focuses on the integration of the two levels currently in production, entity annotation and syntactic structure. Section 4 describes the flow of the annotation process, including the development of the statistical taggers mentioned above. Section 5 is the conclusion.

2 Guidelines for Entity Annotation

Annotation has been proceeding for both the oncology and the inhibition domains. Here we give a summary of the main features of the annotation guidelines that have been developed. We have been influenced in this by previous work in annotation for biomedical information extraction (Ohta et al., 2002; Gaizauskas et al., 2003). However, we differ in the domains we are annotating and the design philosophy for the entity guidelines. For example, we have been concentrating on explicit concepts for entities like genes rather than developing a wide-range ontology for the various physical instantiations.

2.1 Oncology Domain

Gene Entity For the sake of this project the definition for "Gene Entity" has two significant characteristics. First, "Gene" refers to a composite entity as opposed to the strict biological definition. 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 (Consortium, 2002), 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. Standardization of "Gene" references between different texts and between gene synonyms is handled by externally referencing each instance to a standard ontology (Ashburner et al., 2000).

In the context of this project, "Gene" refers to a conceptual entity as opposed to the specific manifestation of a gene (i.e. an allele or nucleotide sequence). Therefore, we consider genes to be abstract concepts identifying genomic regions often associated with a function, such as MYC or TrkB; we do not consider actual instances of such genes within the gene-entity domain. Since we are interested in the association between Gene-entities and malignancies, for this project genes are of interest to us when they have an associated variation event. Therefore, the combination of Gene entities and Variation events

provides us with an evoked entity representing the specific instance of a gene.

Variation Events as Relations Variations comprise a relationship between the following entities: Type (e.g. *point mutation*, *translocation*, or *inversion*), Location (e.g. *codon 14*, *1p36.1*, or *base pair 278*), Original-State (e.g. *Alanine*), and Altered-State (e.g. *Thymine*). These four components represent the key elements necessary to describe any genomic variation event. Variations are often underspecified in the literature, frequently having only two or three of these specifications. Characterizing individual variations as a relation among such components provides us with a great deal of flexibility: 1) it allows us to capture the complete variation event even when specific components are broadly spaced in the text, often spanning multiple sentences or even paragraphs; 2) it provides us with a convenient means of tracking anaphora between detailed descriptions (e.g. *a point mutation at codon 14* and summary references (e.g. *this variation*); and 3) it provides a single structure capable of capturing the breadth of variation specifications (e.g. *A->T point mutation at base pair 47*, *A48->G* or *t(11;14)(q13;32)*).

Malignancy The guidelines for malignancy annotation are under development. We are planning to define it in a manner analogous to variation, whereby a Malignancy is composed of various attribute types (such as developmental stage, behavior, topographic site, and morphology).

2.2 CYP Domain

In the CYP Inhibition annotation task we are tagging three types of entities:

1. CYP450 enzymes (*cyp*)
2. other substances (*subst*)
3. quantitative measurements (*quant*)

Each category has its own questions and uncertainties. Names like *CYP2C19* and *cytochrome P450 enzymes* proclaim their membership, but there are many aliases and synonyms that do not proclaim themselves, such as *17,20-lyase*. We are compiling a list of such names.

Other substances is a potentially huge and vaguely-delimited set, which in the current corpus includes *grapefruit juice* and *red wine* as well as more obviously biochemical entities like *polyunsaturated fatty acids* and *erythromycin*. The quantitative measurements we are directly interested in are those directly related to inhibition, such as *IC50* and *K(i)*. We tag the name of the measurement, the numerical value, and the unit. For example, in the phrase *...was inhibited by troleandomycin (ED50 = 1 microM)*, *ED50* is the name, *1* the value, and *microM* the unit. We are also tagging other measurements, since it is easy to do and may provide valuable information for future IE work.

3 Integrated Annotation

As has been noted in the literature on biomedical IE (e.g., (Pustejovsky et al., 2002; Yakushiji et al., 2001)), the same relation can take a number of syntactic forms. For example, the family of words based on *inhibit* occurs commonly in MEDLINE abstracts about CYP enzymes (as in the example in the introduction) 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. As discussed in the introduction, dealing with such complications in patterns can take much time and effort.

Our approach instead is to build an annotated corpus in which the predicate-argument information is annotated on top of the parsing annotations in the Treebank, the resulting corpus being called a "proposition bank" or Propbank. This newly annotated corpus is then used for training processors that will automatically extract such structures from new examples.

In a Propbank for biomedical text, the types of *inhibit* examples listed above would consistently have their compounds labeled as Arg0 and their enzymes labeled as Arg1, for nominalized forms such as *A is an inhibitor of B*, *A caused inhibition of B*, *inhibition of B by A*, as well the standard *A inhibits B*. We would also be able to label adjuncts consistently, such as the *with* prepositional phrase in *CYP3A4 activity was decreased by L, S and F with IC(50) values of about 200 mM*. In accordance with other Calibratable verbs such as *rise, fall, decline, etc.*, this phrase would be labeled as an Arg2-EXTENT, regardless of its syntactic role.

A Propbank has been built on top of the Penn Treebank, and has been used to train "semantic taggers", for extracting argument roles for the predicates of interest, regardless of the particular syntactic context.[1]

Such semantic taggers have been developed by using machine learning techniques trained on the Penn Propbank (Surdeanu et al., 2003; Gildea and Palmer, 2002; Kingsbury and Palmer, 2002). However, the Penn Treebank and Propbank involve the annotation of Wall Street Journal text. This text, being a financial domain, differs in significant ways from the biomedical text, and so it is

[1]The Penn Propbank is complemented by NYU's Nombank project (Meyers, October 2003), which includes tagging of nominal predicate structure. This is particular relevant for the biomedical domain, given the heavy use of nominals such *mutation* and *inhibition*.

necessary 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).

In this project, the syntactic and semantic annotation is being done on a corpus which is also being annotated for entities, as described in Section 2. Since semantic taggers of the sort described above result in semantic roles assigned to syntactic tree constituents, it is desirable to have the entities correspond to syntactic constituents so that the semantic roles are assigned to entities. The entity information can function as type information and be taken advantage of by learning algorithms to help characterize the properties of the terms filling specified roles in a given predicate.

This integration of these three different annotation levels, including the entities, is being done for the first time[2], and we discuss here three main challenges to this correspondence between entities and constituents: (1) entities that are large enough to cut across multiple constituents, (2) entities within prenominal modifiers, and (3) coordination.[3]

Relations and Large Entities One major area of concern is the possibility of entities that contain more than one syntactic constituent and do not match any node in the syntax tree. For example, as discussed in Section 2, a variation event includes material on a variation's type, location, and state, and can cut not only across constituents, but even sentences and paragraphs. A simple example is *point mutations at codon 12*, containing both the nominal (the type of mutation) and following NP (the location). Note that while in isolation this could also be considered one syntactic constituent, the NP and PP together, the actual context is *...point mutations at codon 12 in duodenal lavage fluid...*. Since all PPs are attached at the same level, *at codon 12* and *in duodenal lavage fluid* are sisters, and so there is no constituent consisting of just *point mutations at codon 12*.

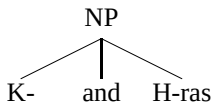
Casting the variation event as a relation between different component entities allows the component entities to correspond to tree constituents, while retaining the capacity to annotate and search for more complex events. In this case, one component entity *point mutations* corresponds to a (base) NP node, and *at codon 12* is corresponds to the PP node that is the NP's sister. At the same time, the relation annotation contains the information relating these two constituents.

Similarly, while the malignancy entity definition is currently under development, as mentioned in Section 2.1, a guiding principle is that it will also be treated as a relation and broken down into component entities. While this also has conceptual benefits for the annotation guidelines, it has the fortunate effect of making such otherwise syntax-unfriendly malignancies as *colorectal adenomas containing early cancer* and *acute myelomonocytic leukemia in remission* amenable for mapping the component parts to syntactic nodes.

Entities within Prenominal Modifiers While we are for the most part following the Penn Treebank guidelines (Bies et al., 1995), we are modifying them in two important aspects. One concerns the prenominal modifiers, which in the Penn Treebank were left flat, with no structure, but in this biomedical domain contain much of the information - e.g., *cancer-associated autoimmune antigen*. Not only would this have had no annotation for structure, but even more bizarrely, *cancer-associated* would have been a single token in the Penn Treebank, thus making it impossible to capture the information as to what is associated with what. We have developed new guidelines to assign structure to prenominal entities such as *breast cancer*, as well as changed the tokenization guidelines to break up tokens such as *cancer-associated*.

Coordination We have also modified the treebank annotation to account for the well-known problem of entities that are discontinuous within a coordination structure - e.g., *K- and H-ras*, where the entities are *K-ras* and *H-ras*. Our annotation tool allows for discontinuous entities, so that both *K-ras* and *H-ras* are annotated as genes.

Under standard Penn Treebank guidelines for tokenization and syntactic structure, this would receive the flat structure



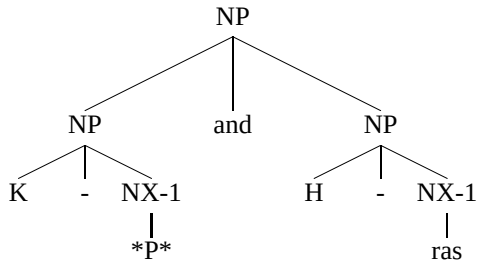
in which there is no way to directly associate the entity *K-ras* with a constituent node.

We have modified the treebank guidelines so that *K-ras* and *H-ras* are both constituents, with the *ras* part of *K-ras* represented with an empty category co-indexed with *ras* in *H-ras*:[4].

[2] An influential precursor to this integration is the system described in (Miller et al., 1996). Our work is in much the same spirit, although the representation of the predicate-argument structure via Propbank and the linkage to the entities is quite different, as well as of course the domain of annotation.

[3] There are cases where the entities are so minimal that they are contained within a NP, not including the determiner, such as *CpG site* in the NP *a CpG site*. entities. We are not as concerned about these cases since we expect that such entity information properly contained within a base NP can be associated with the full base NP.

[4] This is related to the approach to coordination in the GENIA project.



4 Annotation Process

We are currently annotating MEDLINE abstracts for both the oncology and CYP domains. The flowchart for the annotation process is shown in Figure 1. Tokenization, POS-tagging, entity annotation (both domains), and treebanking are in full production. Propbank annotation and the merging of the entities and treebanking remain to be integrated into the current workflow. The table in Figure 2 shows the number of abstracts completed for each annotation area.

The annotation sequence begins with tokenization and part-of-speech annotating. While both aspects are similar to those used for the Penn Treebank, there are some differences, partly alluded to in Section 3. Tokens are somewhat more fine-grained than in the Penn Treebank, so that *H-ras*, e.g., would consist of three tokens: *H*, -, and *ras*.

Tokenized and part-of-speech annotated files are then sent to the entity annotators, either for oncology or CYP, depending on which domain the abstract has been chosen for. The entities described in Section 2 are annotated at this step. We are using WordFreak, a Java-based linguistic annotation tool[5], for annotation of tokenization, POS, and entities. Figure 3 is a screen shot of the oncology domain annotation, here showing a variation relation being created out of component entities for type and location.

In parallel with the entity annotation, a file is treebanked - i.e., annotated for its syntactic structure. Note that this is done independently of the entity annotation. This is because the treebanking guidelines are relatively stable (once they were adjusted for the biomedical domain as described in Section 3), while the entity definitions can require a significant period of study before stabilizing, and with the parallel treatment the treebanking can proceed without waiting for the entity annotation.

However, this does mean that to produce the desired integrated annotation, the entity and treebanking annotations need to be merged into one representation. The consideration of the issues described in Section 3 has been carried out for the purpose of allowing this integration of the treebanking and entity annotation. This has been completed for some pilot documents, but the full merging remains to be integrated into the workflow system.

[5] `http://www.sf.net/projects/wordfreak`

As mentioned in the introduction, statistical taggers are being developed in parallel with the annotation effort. While such taggers are part of the final goal of the project, providing the building blocks for extracting entities and relations, they are also useful in the annotation process itself, so that the annotators only need to perform correction of automatically tagged data, instead of starting from scratch.

Until recently (Feb. 10), the part-of-speech annotation was done by hand-correcting the results of tagging the data with a part-of-speech tagger trained on a modified form of the Penn Treebank.[6] The tagger is a maximum-entropy model utilizing the opennlp package available at `http://www.sf.net/projects/opennlp`. It has now been retrained using 315 files (122 from the oncology domain, 193 from the cyp domain). Figure 4 shows the improvement of the new vs. the old POS tagger on the same 294 files that have been hand-corrected. These results are based on testing files that have already been tokenized, and thus are an evaluation only of the POS tagger and not the tokenizer. While not directly comparable to results such as (Tateisi and Tsujii, 2004), due to the different tag sets and tokenization, they are in the same general range.[7]

The oncology and cyp entity annotation, as well as the treebanking are still being done fully manually, although that will change in the near future. Initial results for a tagger to identify the various components of a variation relation are promising, although not yet integrated into annotation process. The tagger is based on the implementation of Conditional Random Fields (Lafferty et al., 2001) in the Mallet toolkit (McCallum, 2002). Briefly, Conditional Random Fields are log-linear models that rely on weighted features to make predictions on the input. Features used by our system include standard pattern matching and word features as well as some expert-created regular expression features[8]. Using 10-fold cross-validation on 264 labelled abstracts containing 551 types, 1064 lo-

[6] Roughly, Penn Treebank tokens were split at hyphens, with the individual components then sent through a Penn Treebank-trained POS tagger, to create training data for another POS tagger. For example `(JJ York-based)` is treated as `(NNP York) (HYPH -) (JJ based)`. While this works reasonably well for tokenization, the POS tagger suffered severely from being trained on a corpus with such different properties.

[7] The tokenizer has also been retrained and the new tokenizer is being used for annotation, although although we do not have the evaluation results here.

[8] e.g., `chr|chromosome [1-9]|1[0-9]|2[0-2]|X|Y p|q`

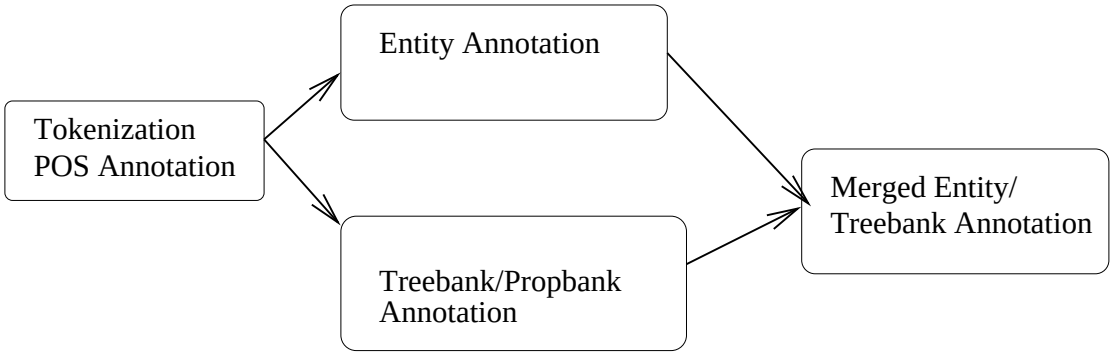


Figure 1: Annotation Flow

Annotation Task	Start Date	Annotated Documents
Part-of-Speech Tagging	8/22/03	422
Entity Tagging	9/12/03	414
Treebanking	1/8/04	127

Figure 2: Current Annotation Production Results

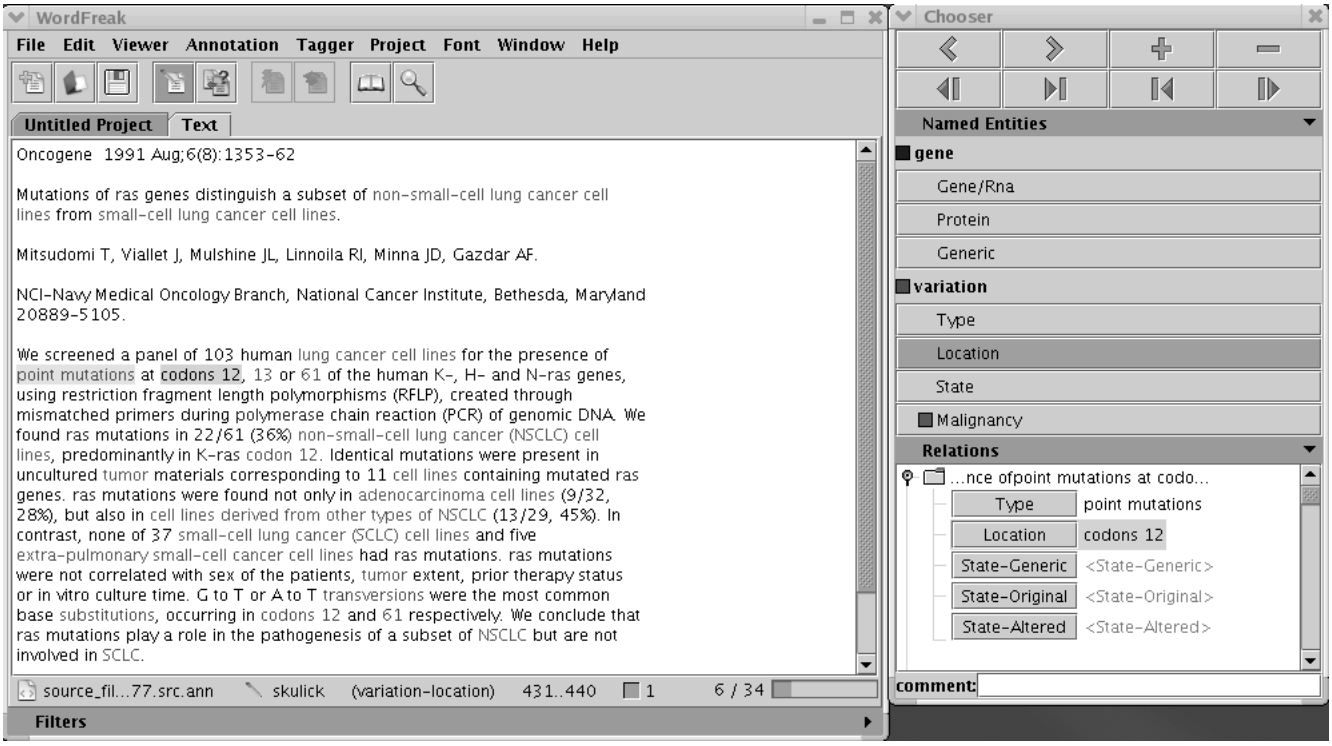


Figure 3: Relation Annotation in WordFreak

Tagger	Training Material	Token Instances
Old	Sections 00-15 Penn Treebank	773832
New	315 abstracts	103159

Tagger	Overall Accuracy	Number Token Instances Unseen in Training Data	Accuracy on Unseen	Accuracy on Seen
Old	88.53%	14542	58.80%	95.53%
New	97.33%	4096	85.05%	98.02%

(Testing Material: 294 abstracts from the oncology domain, with 76324 token instances.)

Figure 4: Evaluation of Part-of-Speech Taggers

cations and 557 states, we obtained the following results:

Entity	**Precision**	**Recall**	**F-measure**
Type	0.80	0.72	0.76
Location	0.85	0.73	0.79
State	0.90	0.80	0.85
Overall	0.86	0.75	0.80

An entity is considered correctly identified if and only if it matches the human labeling by both category (type, location or state) and span (from position a to position b). At this stage we have not distinguished between initial and final states.

While it is difficult to compare taggers that tag different types of entities (e.g., (Friedman et al., 2001; Gaizauskas et al., 2003)), CRFs have been utilized for state-of-the-art results in NP-chunking and gene and protein tagging (Sha and Pereira, 2003; McDonald and Pereira, 2004) Currently, we are beginning to investigate methods to identify relations over the variation components that are extracted using the entity tagger.

5 Conclusion

We have described here an integrated annotation approach for two areas of biomedical information extraction. We discussed several issues that have arisen for this integration of annotation layers. Much effort has been spent on the entity definitions and how they relate to the higher-level concepts which are desired for extraction. There are promising initial results for training taggers to extract these entities.

Next steps in the project include: (1) continued annotation of the layers we are currently doing, (2) integration of the level of predicate-argument annotation, and (3) further development of the statistical taggers, including taggers for identifying relations over their component entities.

Acknowledgements

The project described in this paper is based at the Institute for Research in Cognitive Science at the University of Pennsylvania and is supported by grant EIA-0205448 from the National Science Foundation's Information Technology Research (ITR) program.

We would like to thank Aravind Joshi, Jeremy Lacivita, Paula Matuszek, Tom Morton, and Fernando Pereira for their comments.

References

M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing guidelines for Treebank II Style, Penn Treebank Project. Tech report MS-CIS-95-06, University of Pennsylvania, Philadelphia, PA.

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

Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. 2001. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *ISMB (Supplement of Bioinformatics)*, pages 74–82.

R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett. 2003. Bioinformatics applications of information extraction from journal articles. *Journal of Bioinformatics*, 19(1):135–143.

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

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

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

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

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. 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.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Ryan McDonald and Fernando Pereira. 2004. Identifying gene and protein mentions in text using conditional random fields. In *A Critical Assessment of Text Mining Methods in Molecular Biology workshop*. To be presented.

Adam Meyers. October, 2003. Nombank. Talk at Automatic Content Extraction (ACE) PI Meeting, Alexandria, VA.

Scott Miller, David Stallard, Robert Bobrow, and Richard Schwartz. 1996. A fully statistical approach to natural language interfaces. In Aravind Joshi and Martha Palmer, editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 55–61, San Francisco. Morgan Kaufmann Publishers.

Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, and Jun'ici Tsuji. 2002. The GENIA corpus: An annotated corpus in molecular biology domain. In *Proceedings of the 10th International Conference on Intelligent Systems for Molecular Biology*.

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

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

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceeds of Human Language Technology-NAACL 2003*.

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

Yuka Tateisi and Jun-ichi Tsujii. 2004. Part-of-speech annotation of biology research abstracts. In *Proceedings of LREC04*. To be presented.

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