Medstract: Creating Large-scale Information Servers for biomedical libraries

J. Pustejovsky, J. Castaño, R. Saurí, A. Rumshinsky, J. Zhang, W. Luo
Department of Computer Science, Brandeis University
415 South St., Waltham, MA 02454, U.S.A.
{jamesp,jcastano,roser,arum,jyzhang,weiluo}@cs.brandeis.edu

Abstract

The automatic extraction of information from Medline articles and abstracts (commonly referred to now as the *biobibliome*) promises to play an increasingly critical role in aiding research while speeding up the discovery process. We have been developing robust natural language tools for the automated extraction of structured information from biomedical texts as part of a project we call MEDSTRACT.[1] Here we will describe an architecture for developing databases for domain specific information servers for research and support in the biomedical community. These are currently comprised of the following: a *Bio-Relation Server*, and the Bio-Acronym server, *Acromed*, which will include also aliases. Each information server is derived automatically from an integration of diverse components which employ robust natural language processing of Medline text and IE techniques. The front-end consists of conventional search and navigation capabilities, as well as visualization tools that help to navigate the databases and explore the results of a search. It is hoped that this set of applications will allow for quick, structured access to relevant information on individual genes by biologists over the web.

[1] http://www.medstract.org.

1 Introduction

A vast amount of new biological information is made available in electronic form on a regular basis. Medline contains over 10 million abstracts, and approximately 40,000 new abstracts are added each month. Although there are growing numbers of sequence databases and other hand-constructed databases, most new information is unstructured text in Medline and full-text journals. This information, which is coming to be referred to as the "biobibliome", is a repository of biomedical knowledge that is larger and faster growing than the human genome sequence itself (Stapley and Benoit, 2000). In this age of genomics and proteomics, the ability to process this natural language based information computationally is becoming increasingly important. It is now not uncommon for biologists to study protein complexes and pathways composed of dozens of dynamically interacting proteins. With the recent advent of high sensitivity methods to rapidly identify components of multiprotein complexes (Link et al., 1999), the extent of this complexity is likely to grow exponentially in the next few years. For this reason, the automatic extraction of information from Medline articles and abstracts will play an increasingly critical role in aiding in research and speeding up the discovery process.

The use of computational linguistic techniques for automatically extracting information from biological texts (in particular from Medline) has received increasing attention lately e.g., (Sekimizu et al., 1998; Hishiki et al., 1998; Blasche et al., 1999; Craven and Kum-

Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, July 2002, pp. 85-92. Association for Computational Linguistics.

lien, 1999; Rindflesch et al., 2000; Pustejovsky et al., 2002a). Much of the work reported on thus far has focused on specific protein-protein interactions, and in particular, on predicates implicated in binding activities (Sekimizu et al., 1998; Blasche et al., 1999; Rindflesch et al., 2000; Ono et al., 2001). Craven et al. (1999) use a relational learning algorithm to induce pattern-matching rules on shallow parsed trees for protein-location relations.

Most of this work has been focused on particular instances of information "targets", and does not scale up to large-scale information needs, often because it is necessary to redo an important part of the machinery to define new patterns. Our approach looks at the problem from the general perspective of creating usable information servers for a specific domain. To this end, we have architected some fairly generic language processing modules that focus on the creation of databases of bio-entities and the relations they enter into. This requires several modifications and extensions to conventional NLP systems architecture. First the normalization of data, so that the output of NLP processing is consistent, coherent and easily accesible across an entire data collection derived from the document collection. This normalization has to be carried out throughout different domains and makes apparent the need to solve the naming problem (viz. resolving aliases and acronyms/abreviations). Secondly also a document model is necessary to provide an account of the intersentential relations as expressed through anaphoric relations. Consequently, we present here a global picture where the NLP techniques we used were deployed using the model of the corpus analysis and development cycle described in (Pustejovsky and Hanks, July 2001). In the sections below, we describe the components of MEDSTRACT architecture and illustrate how the modules contribute to the overall goal of object and relation database construction from the biomedical literature corpus.

2 General Architecture

As mentioned above, the goal of MEDSTRACT is to provide tools and resources to biomedical researchers for better search, retrieval, and navigation of new facts and products within the biological literatures. An illustration of the relevant portion of the architecture is shown below in Figure 1. The first biomedical information servers created from this effort have been presented previously (Pustejovsky et al., 2002b)

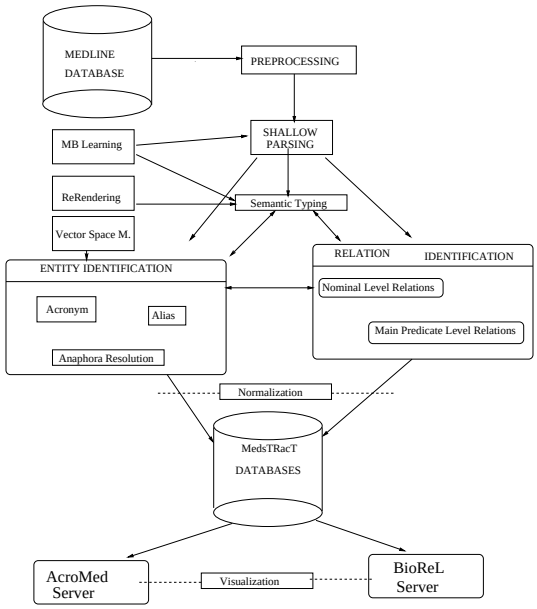

Figure 1: System Architecture

3 The Modules

3.1 Preprocessing

After identifying the corresponding XML tags of the Medline documents, titles and abstracts are tokenized. Tokens are then tagged, using a Brill-like rule-based decision procedure. A lexicon with single or multiple tags for each word is used. If the word in question has multiple tags in the lexicon, then it is tested to match a set of disambiguation rules. If it matches any, the corresponding tag is assigned. Otherwise the most probable tag is assigned. The source lexicons used were the lexicon produced by Brill's tagger and the corresponding lexicon for the UMLS Thesaurus (Humphreys et al.,

1998), with its corresponding syntactic information. The tagged elements were then stemmed with a version of the Porter stemmer. The information corresponding to the *string*, its syntactic tag, and the corresponding stem is stored in a *preterminal* object. A simple morphological analysis is performed to identify derived words that are not in the lexical databases for subsequent database inclusion.

3.2 Shallow Parsing Module

The construction of shallow parse trees involves a cascade of five separate automata, each focusing on a distinct family of grammatical constructions. This is very much in the spirit of Hindle (1983), McDonald (1992) and Pustejovsky et al. (1997). These can be distinguished as follows:
Level I: Noun chunking groups proper nouns and common nouns. It also groups some double prepositions, and compound relational terms.
Level II: Creates noun phrase chunks without prepositional phrases (including adjectives and determiners). It also creates relational terms chunks (verbal chunks, including some adjectival and adverbial terms).
Level III: Creates chunks for coordinated nouns or noun chunks, and coordinated verbal chunks or verbs.
Level IV: Creates chunks of noun phrases with *of*-prepositional phrase.
Level V: Identifies subordinate clauses chunks.

3.3 Semantic Type Identification

The role of semantic typing is to accurately determine the conceptual categories of the words and phrases encountered in Medline abstracts. Aside from direct applications, such as query reformulation or filtering of results according to type restrictions, the accurate assignment of semantic types to entities and predicates facilitates the tasks of anaphora resolution and relation extraction. The set of categories used in semantic typing needs to be adequate for these tasks, as applied to the biomedical domain. There are a number of efforts under way to develop specialized taxonomies and knowledge bases for the biomedical domain (UMLS, Gene Ontology, SWISS-PROT, OMIM, DIP). Wide coverage taxonomies combining a number of sources, such as UMLS, seem particularly fitting for the semantic typing task, especially if they can be enriched with more focused specialized knowledge bases. Our initial system for semantic typing relies on the UMLS semantic type system.

Medline abstracts are shallow-parsed, with limited prepositional attachment (only *of*-attachment) using finite-state techniques. Semantic typing of the resulting noun chunks is performed using a combination of lookup and morphology-based heuristics. The lookup is performed from UMLS. Each noun phrase is put through a cascade of hierarchically arranged type-assignment heuristics. Higher priority heuristics take absolute precedence, that is, if a semantic typing is possible, it is assigned.

During direct lookup, a string is assigned a given semantic type if the UMLS Metathesaurus contains a mapping of that string to one concepts so typed. If a semantic type for the whole phrase is not found in UMLS, we attempt to identify the syntactic head of the phrase using a modification of the *right hand head rule* (cf. (Pustejovsky et al., 1997)), and determine its type. For chunks with *of*-attachment, i.e., phrases of the form, `[<NP-1> of <NP-2>]`, the lookup is also attempted on NP-1 as a whole. If lookup on the given prospective head fails, it is tested for match with morphological heuristics, using a limited number of suffixes.

Direct semantic typing using the above approach confirms the need to adapt and extend the UMLS semantic type system. In particular, in must be enriched with string-to-type mappings from other sources, using more specific typing (e.g. 'protein', rather than 'Amino Acid, Peptid, or Protein). This has motivated the development of *semantic re-rendering* algorithms, designed to enrich the ontological specificity of a domain-specific semantic tag set, based on inductive techniques described in (Pustejovsky et al., 2002c).

3.4 Rerendering Semantic Ontologies

The UMLS, like many industry-standard taxonomies, contains a large number of word-concept pairings (>500K typed terms), making it potentially attractive as a resource for semantic tagging information. However, the overall type structure of UMLS is very shallow. For example, for the semantic tag "Amino Acid, Peptide, or Protein" (AAPP), there are 180,998 entries, for which there are dozens of functional subtypes that are distinguished by biologists, but not in the UMLS.

One specific example of the type system deficiencies illustrates this point very clearly: the extraction of relations and their arguments from text is greatly improved with entity and anaphora resolution capabilities. However, entity and event anaphora resolution rely on (among other things) the semantic typing of the anaphor and its potential antecedents, particularly with sortal and event anaphora, as shown in (1) below.

(1) a. [p21$_i$ inhibits the regulation of ...] ... [This inhibitor$_i$ binds to ...]
b. [A phosphorylates$_i$ B.] ... [The phosphorylation$_i$ of B ...]

Strict UMLS typing presents a problem for anaphora resolution algorithm. For example, (1a) shows a known protein (p21) being subsequently referred to as an 'inhibitor' (a functional of such class of proteins). This type does not exist in UMLS and the noun 'inhibitor' is merely typed as an AAPP. It is therefore difficult to discriminate p21 from other proteins (as potential antecedents) for the sortal anaphor "this inhibitor". A related difficulty is encountered with event anaphora cases such as (1b), where an event nominal anaphor binds to a tensed event as its antecedent, both of which are of different types in UMLS. Hence, the existing UMLS system does not allow for recognition of type-subtype relations of the kinds that are needed in order to identify anaphoric bindings in Medline texts.

Given these motivations, we are developing a set of techniques for "rerendering" an existing semantic ontology to satisfy the requirements of specific features of a (set of) application(s). For the present case (i.e., the UMLS and bio-entity and relation extraction), we will identify candidate subtypes for inclusion in the type system by two means: (a) corpus analysis on compound nominal phrases that express unique functional behavior of the compound head; (b) identification of functionally defined subtypes derived from bio-relation parsing and extraction from the corpus. The results of rerendering will be evaluated for correctness against the original type system, and against additional taxonomies, should they exist, such as the GO ontology.

Our system uses one of two resources for dynamic semantic typing of the input: (a) the UMLS Thesaurus can be exploited to assign types to nouns or noun phrases according to the UMLS type ontology; (b) the GO ontology is also available as a type resource for specific genomic data. The UMLS types were however used in the anaphora resolution task, as one of the parameters in ranking the possible antecedents list.

4 Relation Identification

The relation identification module was built independent of the specifics of any verb and associated nominals in Medline. Rather, this module was defined and designed to work on the output of the shallow parsing module to identify argument and relational chunks, independently of any specific lexical item. The extraction of a particular relation (e.g. *inhibit* or *regulate*), is accomplished by specifying stems that denote the required relation. Sentence-level parsing identifies the following constructions:

Sentence-level Relation identification

(1) Main predicate relational chunk in the sentence.

(2) Subject nominal chunk (nominal chunks at 4th level above).

(3) Object nominal chunks.

(4) Subordinate clauses (identifying also antecedents of relative clauses, and main predicates of object clauses).

(5) Sentential coordination.

It has also the capability of identifying:

(6) Preverbal adjuncts.

(7) Post Object target adjuncts (ambiguous between adjuncts and nominal modifiers, PP attachment ambiguity).

In the nominal domain, head nouns may typically carry relational semantics; for example the noun *inhibitor* can refer to both the relation as well as the biological entity itself. The parsing decisions involved for these forms are distinct from the verbal form. The constructions and relations identified by the nominal-level module are given below:

Nominal-level Relation Identification

(1) Nominal chunks of Level IV.

(2) Prepositional relational chunks.

Note that relations inside Level IV are decomposed first, i.e., *of*-prepositional relations. Our next step will be to add reduced relative clauses and gerundive relations to this parser module.

As mentioned briefly above, the initial corpus analysis distinguished the verbal forms from the nominal forms. Because of their distinct argument binding and complementation behaviors, we decided to develop separate automata for each form, and then merge the results in a subsequent database population phase. In fact, however, there is reason to believe that keeping the results extracted from the two modules separate is actually desirable for database purposes as well; this is due in large part to the degree of relevance associated with 'given' versus 'new' information as presented in documents (cf. (Pustejovsky et al., in preparation)).

5 Entity Recognition

5.1 Acronym Sense Disambiguation

Acronyms are widely used in biomedical and other technical texts. Understanding their meaning constitutes an important problem in the automatic extraction and mining of information from text. Moreover, an even harder problem is sense disambiguation of acronyms; that is, where a single acronym has a multiplicity of meanings (polynyms), a common occurrence in the biomedical literature. In such cases, it is necessary to identify the correct corresponding sense for the polynym, which is often not directly specified in the text. For example, the acronym AC appeared associated to the following different long forms in a search through different documents:

atrioventricular connection

anterior colporrhaphy procedure

auditory cortex

atypical carcinoid

abdominal circumference

acalculous cholecystitis

adenylyl cyclase

adenyl cyclase

Adenylyl cyclase

adenylate cyclase

Here we present a system called AcroMed which finds acronym-meaning pairs as part of a set of information extraction tools designed for processing and extracting data from abstracts in the Medline database. Our strategy for finding acronym-meaning pairs differs from previous automated acronym extraction methods by incorporating shallow parsing of the text into the acronym recognition algorithm. The performance of our system has been tested with a highly diverse set of Medline texts, giving 98 % precision. We then present an initial approach for disambiguating polynyms, using a vector space model. Our disambiguation tests produced 97.62 % accuracy in one test for acronyms and 86.6 % accuracy in another for aliases.

A second task here is that of grouping together equivalent long forms which constitute a single sense of a given acronym. To that end we have used two formal criteria. Two long forms are considered to mean the same when they have a common normalized form (the corresponding

version of a long form in lower case and not containing either hyphens or a plural 's') or when an 80% of the longest form is also present in the shortest one. Some examples of grouping are:

PDA:	*patency of the ductus arteriosus*
	patency of ductus arteriosus
	patent ductus arteriosus
AC:	*adenylyl cyclase*
	Adenylyl cyclase
	adenyl cyclase
	adenylate cyclase

Currently we are also working on the identification and correlation of acronym aliases (i.e., different terms that refer to the same entity, or synonyms). Our final aim is to be able to point to a concept from different aliases and know what other denominations the concept receives.

5.2 Coherence and Anaphora Resolution

Identifying the arguments of the relations may not be enough for identifying the actual entities involved in the relation. Quite often anaphors (e.g., *it, they*) and sortal anaphoric noun phrases (e.g. *the protein*, *both enzymes*) are the actual arguments to a relation, but unfortunately are not specific enough to establish a unique reference to an entity or process. Although the use of anaphoric terms seems to be relatively infrequent in Medline abstracts, the use of sortal anaphors is quite prevalent. This module focuses on the resolution of biologically relevant sortal terms (i.e., proteins, genes, and bioprocesses), as well as pronominal anaphors, including third person pronouns and reflexive pronouns. The initial data source for this resolution algorithm is the preprocessed Medline text (shallow parsed), where each noun phrase (NP) has been identified and annotated with a syntactic tag and semantic tag(s). The anaphora resolution algorithm examines the text sequentially and represents each sentence as a "frame environment". Every NP within a sentence is a potential referent and is made into an entity with a unique ID and syntactic/semantic tags, and added to the sentence environment in which it occurs. If an NP is identified as an anaphor, then the resolution algorithm will attempt to resolve it by traversing through the sentence environments from the most recent (which contains the anaphor), back to the first sentence of the abstract, and selecting the NP among the sentences that has the highest compatibility with the anaphor as the antecedent (cf. Kennedy and Boguraev (1996)). The choice of antecedent is determined by matching syntactic and semantic features of the candidate NP with that of the anaphor, which includes person/number agreement, semantic type, as well as physical string comparisons. In the case that more than one NP is found to be equally compatible, preference is given to the one that is most adjacent to the anaphor in the text. If an anaphor requires multiple antecedents (e.g., the anaphor *both enzymes*) then the resolution algorithm will continue in the sentence environment where the first antecedent is found, and then select the subsequent antecedent which is most compatible with both the anaphor and the first antecedent.

Here we are concerned with 2 types of anaphors:

- Pronominal Anaphors: including third person pronouns and reflexive pronouns. E.g., *Human growth hormone (hGH) binds to its receptor (hGHr) in a three-body interaction: one molecule of* **it** *and two identical monomers of the receptor form a trimer.*
- Sortal Anaphors as Epithets. E.g., *we quantitatively analyzed the relationship between the structure and inhibiting activity of these substances toward acetylcholinesterase and butyryl cholinesterase. Hydrophobic interactions were found to be important for the inhibition of* **both enzymes** *but are more pronounced in the case of butyryl cholinesterase.*

From our analysis of the Medline corpus, we found that the use of both types of anaphors, especially sortal anaphors, is quite prevalent in Medline abstracts. For instance, out of 100 distinct anaphors derived from a set of 60-70

Medline documents, approximately 60% are sortal anaphors. Our initial tests (Castaño et al., 2002) resulted in 77% precision and 71% recall.

6 Information Servers

6.1 BioRelation Server

We have populated a bio-relational database using a a robust parser for identifying and extracting biomolecular relations from the biomedical literature (Medline) (e.g., *A inhibits B*, *X regulates Y*). We have measured the system at a performance of 90% precision, 59% recall, and 22% partial recall, results which were published previously (Pustejovsky et al., 2002a). The Bio-Relation server provides the ability to search the Medstract relational database, which contains various regulatory and inhibitory relations of bio-entities (proteins, cellular processes, etc.) extracted from six months of Medline data (250,000 abstracts). The user can select the type of relation to be searched (i.e.,*inhibit*, *regulate*, etc.) as well as the UMLS type (e.g. *gene*, *amino acid*) or name of the bio-entities which are the arguments of the relation. As a result of the search, all biological relations relevant to the specified bio-entities are returned in either the form of a database table or a navigable hyperbolic graph. Both forms link directly to the citations from the abstracts. Figure 2 is the screenshot of the database table.

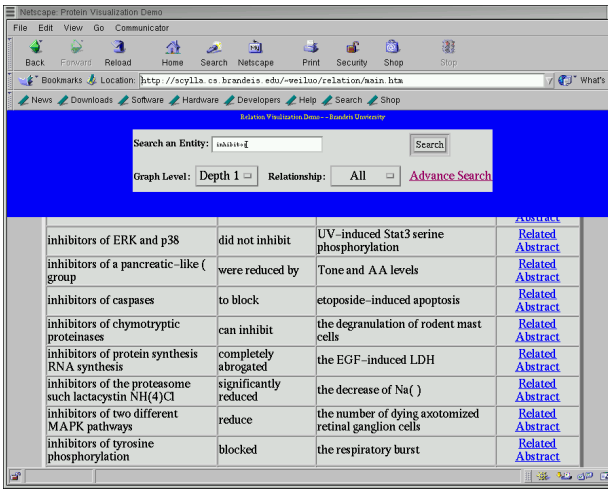

Figure 2: Relations table

6.2 BioAcronym Server

The Bio-Acronym Server (*AcroMed*) is an automatically generated searchable database of over 121,000 biomedical acronyms and their associated normalized long-forms extracted from one year and a half of Medline abstracts (release 2001, 4.5G approx.). Every acronym is displayed with its corresponding set of senses. Each acronym-long form pair in the database is linked to the abstracts in which it was discovered, and the set of equivalent long forms corresponding to a single sense can be submitted directly to PubMed as searches by a single click as a query reformulation. Furthermore, Acromed also attempts to classify each acronym-long form pair by its semantic type, using an ontology comprised of both UMLS and Go taxonomic terms. Currently we are incorporating into the acronym server, which include aliases of named entities (e.g.: *WAF1* as alias of *p21*). The system is live and currently web viewable.

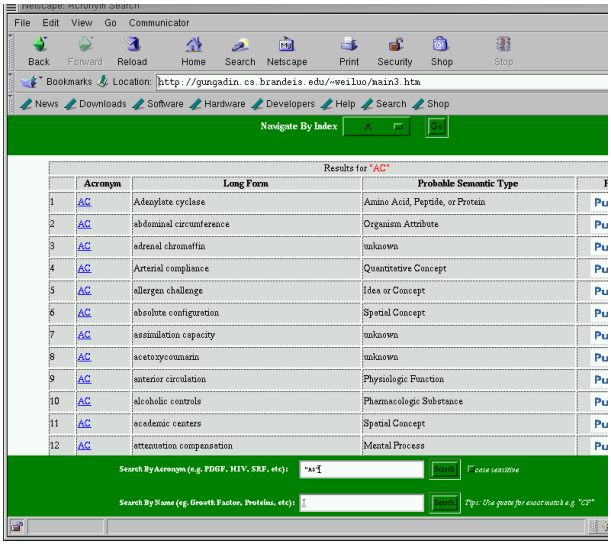

Figure 3: Acronym server

7 Conclusions

In this paper we have described an approach to architecting domain specific information servers, where the target Data Bases are constructed in incremental and independent layers. This is a methodology that we would like to promote generally within the CL community. From an engineering perspective this approach embodies elements of best practices, where the end-user interacts in the development cycle providing feedback and enabling the community to supply the Quality Analysis (QA). We find this is a promising perspective on the deployment of large NL systems.

References

C. Blasche, M. Andrade, C. Ouzounis, and A. Valencia. 1999. Automatic extraction of biological information from scientific text: protein-preotein interactions. In *AAAI*.

J. Castaño, J. Zhang, and Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution in NLP, Alicante, Spain*.

M. Craven and J. Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*.

D Hindle. 1983. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*.

T. Hishiki, N. Collier, C. Nobata, T. Okazaki-Ohta, N. Ogata, T. Sekimizu, R. Steiner, H.S. Park, and J. Tsujii. 1998. Developing nlp tools for genome informatics: An information extraction perspective. In *Proc. of Genome Informatics, Tokyo, Japan*, pages 81–90.

B. L. Humphreys, D. A. B. Lindberg, Schoolman H. M., and Barnett G. O. 1998. The unified medical language system: An informatics research collaboration. *Journal of the American Medical Informatics Association*, (5).

C. Kennedy and B. Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING), Kopenhagen*, volume Vol. I, pages 113–118.

A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and J. R. 3rd. Yates. 1999. Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, (17):676–82.

D. D. McDonald. 1992. Robust partial parsing through incremental multi-algorithm processing. In P. Jacobs, editor, *Text-based Intelligent Systems*.

T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. 2001. Automatic extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, pages 155–161.

J. Pustejovsky and P. Hanks. July, 2001. Very large lexical databasees: A tutorial primer. In *Association for Computational Linguistics, Toulouse*.

J. Pustejovsky, B. Boguraev, M. Verhagen, P. Buitelaar, and M. Johnston. 1997. Semantic indexing and typed hyperlinking. In *AAAI Symposium on Language and the Web, Stanford, CA*.

J. Pustejovsky, J. Castaño, J. Zhang, and B. Cochran. 2002a. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Pacific Symposium on Biocomputing*.

J. Pustejovsky, B. Cochran, J. Castaño, , M. Morrell, J. Zhang, and R. Saurí. 2002b. Medstract: Natural language tools for mining the biobibliome. In *Demo presented at HLT-2002, San Diego, CA*.

J. Pustejovsky, A. Rumshinsky, and J. Castaño. 2002c. Automatic extensions to UMLS through corpus analytics. In *Ontolex 2002*.

J. Pustejovsky, J. Castaño, and B. Cochran. in preparation. Exploiting given versus new information for information extraction tasks.

T. C. Rindflesch, Jayant V. Rajan, and Lawrence Hunter. 2000. Extracting molecular binding relationships from biomedical text. In *Proceedings of the ANLP-NAACL*, pages 188–195. Association for Computational Linguistics.

T. Sekimizu, H. S. Park, and J. Tsujii. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. In *Proc. of Genome Informatics,Tokyo,Japan,*, pages 62–71.

B. J. Stapley and G. Benoit. 2000. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. In *Pacific Symposium on Biocomputing*, pages 529–40.