MEDSTRACT: Natural Language Tools for Mining the Biobibliome

James Pustejovsky
Computer Science
Department
Brandeis University
Waltham, MA 02254
jamesp@cs.brandeis.edu

Brent Cochran
Department of Physiology
Tufts University
136 Harrison Ave., Boston MA
University School of Medicine
cochran@opal.tufts.edu

Jose Castaño
Computer Science
Department
Brandeis University
Waltham, MA 02254
jcastano@cs.brandeis.edu

ABSTRACT

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. Here we will present demonstrations of two biological literature database servers, the *Bio-Relation Server* and the Bio-Acronym server, *Acromed*, both derived automatically from robust natural language processing of Medline text. Visualization tools for navigating the databases and exploring the results of a search are also demonstrated. The presentation consists of four distinct demos over these two databases, demonstrating the extraction results, presented in both tabular form and graphically, as well as a live illustration of how the text coherence and anaphora resolution module improves overall performance of the database results.

1. INTRODUCTION

The automatic extraction of information from Medline articles and abstracts (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 (www.medstract.org). Here we will present demonstrations of two biological literature database servers, the *Bio-Relation Server* and the Bio-Acronym server, *Acromed*, both derived automatically from robust natural language processing of Medline text. Visualization tools for navigating the databases and exploring the results of a search are also demonstrated. The presentation consists of four distinct demos over these two databases, demonstrating the extraction results, presented in both tabular form and graphically, as well as a live illustration of how the text coherence and anaphora resolution module improves overall performance of the database results. It is hoped that this set of applications will allow for organized rapid access to relevant information on individual genes by biologists over the

web. One further goal is that the automatic analysis of biological abstracts will eventually be combined with information derived from sequence databases to provide an up-to-date and comprehensive biobibliome of information regarding individual known genes and proteins.

We will have 2-3 machines with large displays running programs from remote servers.

2. BIORELATION SERVER

We have populated a bio-relational database using 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, as reported in Pustejovsky et al. (2002). 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.

3. COHERENCE AND ANAPHORA RESOLUTION IN MEDLINE

Identifying the arguments of the relations may not be enough for identifying the actual entities involved in the relation. Often anaphors, both pronominal (e.g. *it, they*) and sortal (e.g. *the protein, both enzymes*) are the actual arguments to a relation, but are not specific enough to establish a unique reference to an entity or process. We found that the use of anaphors, especially sortal anaphors, is quite prevalent in Medline abstracts. For example, out of the 62 anaphors identified in a set of 35-40 Medline documents, approximately 60% are sortal anaphors. The Coherence and Anaphora Resolution Application (CARA) labels the biologically relevant anaphors within Medline abstracts and assigns each anaphor a pointer to the sentence which contains its antecedent, while binding it to the antecedent in that sentence. Importantly, CARA focuses on the resolution of sortal anaphors as well as pronominal anaphors. Antecedent matching for an anaphor of either type

Proceedings of HLT 2002, Second International Conference on Human Language Technology Research, M. Marcus, ed., Morgan Kaufmann, San Francisco, 2002.

			Abstract
inhibitors of ERK and p38	did not inhibit	UV-induced Stat3 serine phosphorylation	Related Abstract
inhibitors of a pancreatic-like (group)	were reduced by	Tone and AA levels	Related Abstract
inhibitors of caspases	to block	etoposide-induced apoptosis	Related Abstract
inhibitors of chymotryptic proteinases	can inhibit	the degranulation of rodent mast cells	Related Abstract
inhibitors of protein synthesis RNA synthesis	completely abrogated	the EGF-induced LDH	Related Abstract
inhibitors of the proteasome such lactacystin NH(4)Cl	significantly reduced	the decrease of Na()	Related Abstract
inhibitors of two different MAPK pathways	reduce	the number of dying axotomized retinal ganglion cells	Related Abstract
inhibitors of tyrosine phosphorylation	blocked	the respiratory burst	Related Abstract

Figure 1: Relations table

is determined by finding the NP that has the highest compatibility, which is determined by matching syntactic and semantic features of the candidate NP with that of the anaphor, along with measures of textual and thematic salience. By utilizing the anaphora resolution algorithm, the performance of the relational extraction module improve by 15-20% (cf. Zhang et al, (forthcoming)).

4. BIOACRONYM DEMO

The Bio-Acronym Server (*AcroMed*) will also be demonstrated. This system is an automatically generated searchable database of over 121,000 biomedical acronyms and their associated normalized long-forms extracted from one year of Medline abstracts (2001). The process of normalization groups equivalent long forms together, which form a single meaning sense. Those acronyms (*polynoms*) which have more than one non equivalent long form are displayed with the 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.

Figure 2: Relations graph

Figure 3: Anaphor resolution

5. ADDITIONAL AUTHORS

Jason Zhang (email: jy Zhang@cs.brandeis.edu)

Roser Sauri (email: roser@cs.brandeis.edu)

Michael Morrell (email: mmorrell@argus.cs.brandeis.edu)

Wei Luo (email: weiluo@argus.cs.brandeis.edu)

Computer Science Department

Brandeis University

Waltham, MA 02254

6. REFERENCES

Pustejovsky, J., J. Castaño, B. Cochran, M. Kotecki, M. Morrell, A. Rumshisky (2001) "Extraction and Disambiguation of Acronym-Meaning Pairs in Medline." An updated version of the paper presented at Medinfo 2001, London.

Pustejovsky, J., J. Castaño, J. Zhang, B. Cochran, M. Kotecki (2002) "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations." Pacific Symposium on Biocomputing, 2002.

Zhang, J., J. Castaño, and J. Pustejovsky (forthcoming), "Coherence and Anaphora Resolution in Biomedical Literature", submission to EMNLP 2002.

Acronym	Long Form	Probable Semantic Type	
AG	Adenosine cyclase	Amino Acid, Peptide, or Protein	Pu
AG	abdominal circumference	Organism Attribute	Pu
AG	adrenal chromaffin	unknown	Pu
AG	Artisanal compliance	Quantitative Concept	Pu
AG	allergic challenge	Idea or Concept	Pu
AG	absolute configuration	Spatial Concept	Pu
AG	assimilation capacity	unknown	Pu
AG	acetylcholinesterase	unknown	Pu
AG	anterior circulation	Physiologic Function	Pu
AG	alcoholic controls	Pharmacologic Substance	Pu
AG	academic centers	Spatial Concept	Pu
AG	attraction compensation	Mental Process	Pu

Figure 4: Acronym server