Using Dependency Parsing and Probabilistic Inference to Extract Relationships between Genes, Proteins and Malignancies Implicit Among Multiple Biomedical Research Abstracts

Ben Goertzel
Applied Research Lab for National and Homeland Security
Virginia Tech
Arlington VA 22216
ben@goertzel.org

Hugo Pinto
Novamente LLC
1405 Bernerd Place
Rockville MD 20851
hugo@vettalabs.com

Ari Heljakka
Novamente LLC
1405 Bernerd Place
Rockville MD 20851
heljakka@iki.fi

Izabela Freire Goertzel
Novamente LLC
1405 Bernerd Place
Rockville MD 20851
izabela@goertzel.org

Mike Ross
SAIC
5971 Kingstowne Village Parkway
Kingstowne, VA 22315
miross@objectsciences.com

Cassio Pennachin
Novamente LLC
1405 Bernerd Place
Rockville MD 20851
cassio@vettalabs.com

Abstract

We describe BioLiterate, a prototype software system which infers relationships involving relationships between genes, proteins and malignancies from research abstracts, and has initially been tested in the domain of the molecular genetics of oncology. The architecture uses a natural language processing module to extract entities, dependencies and simple semantic relationships from texts, and then feeds these features into a probabilistic reasoning module which combines the semantic relationships extracted by the NLP module to form new semantic relationships. One application of this system is the discovery of relationships that are not contained in any individual abstract but are implicit in the combined knowledge contained in two or more abstracts.

1 Introduction

Biomedical literature is growing at a breakneck pace, making the task of remaining current with all discoveries relevant to a given research area nearly impossible without the use of advanced NLP-based tools (Jensen et al, 2006). Two classes of tools that provide great value in this regard are those that help researchers find relevant documents and sentences in large bodies of biomedical texts (Müller, 2004; Schuler, 1996; Tanabe, 1999), and those that automatically extract knowledge from a set of documents (Smalheiser and Swanson, 1998; Rzhetsky et al, 2004). Our work falls into the latter category. We have created a prototype software system called BioLiterate, which applies dependency parsing and advanced probabilistic inference to the problem of combining semantic relationships extracted from biomedical texts, have tested this system via experimentation on research abstracts in the domain of the molecular genetics of oncology.

In order to concentrate our efforts on the inference aspect of biomedical text mining, we have built our BioLiterate system on top of a number of general NLP and specialized bioNLP components created by others. For example, we have handled entity extraction -- perhaps the most mature existing bioNLP technology (Kim, 2004) -- via incorporating a combination of existing open-source tools. And we have handled syntax parsing via integrat-

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 104–111,
New York City, June 2006.

ing a modified version of the link parser (Sleator and Temperley, 1992).

The BioLiterate system is quite general in applicability, but in our work so far we have focused on the specific task of extracting relationships regarding interactions between genes, proteins and malignancies contained in, or implicit among multiple, biomedical research abstracts. This application is critical because the extraction of protein/gene/disease relationships from text is necessary for the discovery of metabolic pathways and non-trivial disease causal chains, among other applications (Nédellec, 2005; Davulcu, 2005, Ahmed, 2005).

Systems extracting these sorts of relationships from text have been developed using a variety of technologies, including support vector machines (Donaldson et al, 2003), maximum entropy models and graph algorithms (McDonald, 2005), Markov models and first order logic (Riedel, 2005) and finite state automata (Hakenberg, 2005). However, these systems are limited in the relationships that they can extract. Most of them focus on relationships described in single sentences. The results we report here support the hypothesis that the methods embodied in BioLiterate, when developed beyond the prototype level and implemented in a scalable way, may be significantly more powerful, particularly in the extraction of relationships whose textual description exists in multiple sentences or multiple documents.

Overall, the extraction of both entities and single-sentence-embodied inter-entity relationships has proved far more difficult in the biomedical domain than in other domains such as newspaper text (Nédellec, 2005; Jing et al, 2003; Pyysalo, 2004). One reason for this is the lack of resources, such as large tagged corpora, to allow statistical NLP systems to perform as well as in the news domain. Another is that biomedical text has many features that are quite uncommon or even non-existent in newspaper text (Pyysalo, 2004), such as numerical post-modifiers of nouns (*Serine 38*), non-capitalized entity names (*…ftsY is solely expressed during...*), hyphenated verbs (*X cross-links Y*), nominalizations, and uncommon usage of parentheses (*sigma(H)-dependent expression of spo0A*). While recognizing the critical importance of overcoming these issues more fully, we have not addressed them in any novel way in the context of our work on BioLiterate, but have rather chosen to focus attention on the other end of the pipeline: using inference to piece together relationships extracted from separate sentences, to construct new relationships implicit among multiple sentences or documents.

The BioLiterate system incorporates three main components: an NLP system that outputs entities, dependencies and basic semantic relations; a probabilistic reasoning system (PLN = Probabilistic Logic Networks); and a collection of hand-built semantic mapping rules used to mediate between the two prior components.

One of the hypotheses underlying our work is that the use of probabilistic inference in a bioNLP context may allow the capturing of relationships not covered by existing systems, particularly those that are implicit or spread among several abstracts. This application of BioLiterate is reminiscent of the Arrowsmith system (Smalheiser and Swanson, 1998), which is focused on creating novel biomedical discoveries via combining pieces of information from different research texts; however, Arrowsmith is oriented more toward guiding humans to make discoveries via well-directed literature search, rather than more fully automating the discovery process via unified NLP and inference.

Our work with the BioLiterate prototype has tentatively validated this hypothesis via the production of interesting examples, e.g. of conceptually straightforward deductions combining premises contained in different research papers.[1] Our future research will focus on providing more systematic statistical validation of this hypothesis.

2 System Overview

For the purpose of running initial experiments with the BioLiterate system, we restricted our attention to texts from the domain of molecular genetics of oncology, mostly selected from the PubMEd subset selected for the PennBioNE project (Mandel, 2006). Of course, the BioLiterate architecture in general is not restricted to any particular type or subdomain of texts.

The system is composed of a series of components arranged in a pipeline: Tokenizer →Gene,

[1] It is worth noting that inference which appear conceptually to be "straightforward deductions" often manifest themselves within BioLiterate as PLN inference chains with 1-2 dozen inferences. This is mostly because of the relatively complex way in which logical relationships emerge from semantic mapping, and also because of the need for inferences that explicitly incorporate "obvious" background knowledge.

Protein and Malignancy Tagger → Nominalization Tagger → Sentence Extractor → Dependency Extractor → Relationship Extractor → Semantic Mapper → Probabilistic Reasoning System.

Each component, excluding the semantic mapper and probabilistic reasoner, is realized as a UIMA (Götz and Suhre, 2004) annotator, with information being accumulated in each document as each phase occurs.[2]

The gene/protein and malignancy taggers collectively constitute our "entity extraction" subsystem. Our entity extraction subsystem and the tokenizer were adapted from PennBioTagger (McDonald et al, 2005; Jin et al, 2005; Lerman et al, 2006). The tokenizer uses a maximum entropy model trained upon biomedical texts, mostly in the oncology domain. Both the protein and malignancy taggers were built using conditional random fields.

The nominalization tagger detects nominalizations that represent possible relationships that would otherwise go unnoticed. For instance, in the sentence excerpt "… intracellular signal transduction leading to transcriptional activation…" both "transduction" and "activation" are tagged. The nominalization tagger uses a set of rules based on word morphology and immediate context.

Before a sentence passes from these early processing stages into the dependency extractor, which carries out syntax parsing, a substitution process is carried out in which its tagged entities are replaced with simple unique identifiers. This way, many text features that often impact parser performance are left out, such as entity names that have numbers or parenthesis as post-modifiers.

The dependency extractor component carries out dependency grammar parsing via a customized version of the open-source Sleator and Temperley link parser (1993). The link parser outputs several parses, and the dependencies of the best one are taken.[3]

The relationship extractor component is composed of a number of template matching algorithms that act upon the link parser's output to produce a semantic interpretation of the parse. This component detects implied quantities, normalizes passive and active forms into the same representation and assigns tense and number to the sentence parts. Another way of conceptualizing this component is as a system that translates link parser dependencies into a graph of semantic primitives (Wierzbicka, 1996), using a natural semantic metalanguage (Goddard, 2002).

Table 1 below shows some of the primitive semantic relationships used, and their associated link parser links:

subj	Subject	S, R, RS
Obj	Direct object	O, Pv, B
Obj-2	Indirect object	O, B
that	Clausal Complement	TH, C
to-do	Subject Raising Complement (do)	I, TO, Pg

Table 1. Semantic Primitives and Link Parser Links

For a concrete example, suppose we have the sentences:

a) Kim kissed Pat.
b) Pat was kissed by Kim.

Both would lead to the extracted relationships:

```
subj(kiss, Kim), obj(kiss, Pat)
```

For a more interesting case consider:

c) Kim likes to laugh.
d) Kim likes laughing.

Both will have a `to-do (like, laugh)` semantic relation.

Next, this semantic representation, together with entity information, is feed into the Semantic Mapper component, which applies a series of hand-created rules whose purpose is to transform the output of the Relationship Extractor into logical relationships that are fully abstracted from their syntactic origin and suitable for abstract inference. The need for this additional layer may not be apparent a priori, but arises from the fact that the output of the Relationship Extractor is still in a sense "too close to the syntax." The rules used within the Relationship Extractor are crisp rules with little context-dependency, and could fairly easily be built into a dependency parser (though the link parser is not architected in such a way as to make this pragmatically feasible); on the other

[2] The semantic mapper will be incorporated into the UIMA framework in a later revision of the software.

[3] We have experimented with using other techniques for selecting dependencies, such as getting the most frequent ones, but variations in this aspect did not impact our results significantly.

hand, the rules used in the Semantic Mapper are often dependent upon semantic information about the words being interrelated, and would be more challenging to integrate into the parsing process.

As an example, the semantic mapping rule

```
by($X,$Y) & Inh($X, transitive_event) →
subj ($X,$Y)
```

maps the relationship `by(prevention, inhibition)`, which is output by the Relationship Extractor, into the relationship subj(prevention, inhibition), which is an abstract conceptual relationship suitable for semantic inference by PLN. It performs this mapping because it has knowledge that "prevention" inherits (`Inh`) from the semantic category `transitive_event`, which lets it guess what the appropriate sense of "by" might be.

Finally, the last stage in the BioLiterate pipeline is probabilistic inference, which is carried out by the Probabilistic Logic Networks[4] (PLN) system (Goertzel et al, in preparation) implemented within the Novamente AI Engine integrated AI architecture (Goertzel and Pennachin, 2005; Looks et al, 2004). PLN is a comprehensive uncertain inference framework that combines probabilistic and heuristic truth value estimation formulas within a knowledge representation framework capable of expressing general logical information, and possesses flexible inference control heuristics including forward-chaining, backward-chaining and reinforcement-learning-guided approaches.

Among the notable aspects of PLN is its use of two-valued truth values: each PLN statement is tagged with a truth value containing at least two components, one a probability estimate and the other a "weight of evidence" indicating the amount of evidence that the probability estimate is based on. PLN contains a number of different inference rules, each of which maps a premise-set of a certain logical form into a conclusion of a certain logical form, using an associated truth-value formula to map the truth values of the premises into the truth value of the conclusion.

The PLN component receives the logical relationships output by the semantic mapper, and performs reasoning operations on them, with the aim at arriving at new conclusions implicit in the set of relationships fed to it. Some of these conclusions may be implicit in a single text fed into the system; others may emerge from the combination of multiple texts.

In some cases the derivation of useful conclusions from the semantic relationships fed to PLN requires "background knowledge" relationships not contained in the input texts. Some of these background knowledge relationships represent specific biological or medical knowledge, and others represent generic "commonsense knowledge." The more background knowledge is fed into PLN, the broader the scope of inferences it can draw.

One of the major unknowns regarding the current approach is how much background knowledge will need to be supplied to the system in order to enable truly impressive performance across the full range of biomedical research abstracts. There are multiple approaches to getting this knowledge into the system, including hand-coding (the approach we have taken in our BioLiterate work so far) and automated extraction of relationships from relevant texts beyond research abstracts, such as databases, ontologies and textbooks. While this is an extremely challenging problem, we feel that due to the relatively delimited nature of the domain, the knowledge engineering issues faced here are far less severe than those confronting projects such as Cyc (Lenat, 1986; Guha, 1990; Guha, 1994) and SUMO (Niles, 2001) which seek to encode commonsense knowledge in a broader, non-domain-specific way.

3 A Practical Example

We have not yet conducted a rigorous statistical evaluation of the performance of the BioLiterate system. This is part of our research plan, but will involve considerable effort, due to the lack of any existing evaluation corpus for the tasks that BioLiterate performs. For the time being, we have explored BioLiterate's performance anecdotally via observing its behavior on various example "inference problems" implicit in groups of biomedical abstracts. This section presents one such example in moderate detail (full detail being infeasible due to space limitations).

Table 2 shows two sentences drawn from different PubMed abstracts, and then shows the conclusions that BioLiterate draws from the combination of these two sentences. The table shows the conclusions in natural language format, but the system

[4] Previously named Probabilistic Term Logic

actually outputs conclusions in logical relationship form as detailed below.

Premise 1	Importantly, bone loss was almost completely prevented by p38 MAPK inhibition. (PID 16447221)
Premise 2	Thus, our results identify DLC as a novel inhibitor of the p38 pathway and provide a molecular mechanism by which cAMP suppresses p38 activation and promotes apoptosis. (PID 16449637)
(Uncertain) Conclusions	DLC prevents bone loss. cAMP prevents bone loss.

Table 2. An example conclusion drawn by BioLiterate via combining relationships extracted from sentences contained in different PubMed abstracts. The PID shown by each premise sentence is the PubMed ID of the abstract from which it was drawn.

Tables 3-4 explore this example in more detail. Table 3 shows the relationship extractor output, and then the semantic mapper output, for the two premise sentences.

Premise 1 Rel Ex. Output	_subj-n(bone, loss) _obj(prevention, loss) _subj-r(almost, completely) _subj-r(completely, prevention) by(prevention, inhibition) _subj-n(p38 MAPK, inhibition)
Premise 2 Sem Map Output	subj (prevention, inhibition) obj (prevention, loss) obj (inhibition, p38_MAPK) obj (loss, bone)
Premise 1 Rel Ex Output	_subj(identify, results) as(identify, inhibitor) _obj(identify, DLC) _subj-a(novel, inhibitor) of(inhibitor, pathway) _subj-n(p38, pathway)
Premise 2 Sem Map Output	subj (inhibition, DLC) obj (inhibition, pathway) inh(pathway, p38)

Table 3. Intermediary processing stages for the two premise sentences in the example in Table 2.

Table 4 shows a detailed "inference trail" constituting part of the reasoning done by PLN to draw the inference "DLC prevents bone loss" from these extracted semantic relationships, invoking background knowledge from its knowledge base as appropriate.

The notation used in Table 4 is so that, for instance, `Inh inhib1 inhib2` is synonymous with `inh(inhib1, inhib2)` and denotes an Inheritance relationship between the terms inhibition$_1$ and inhibition$_2$ (the textual shorthands used in the table are described in the caption). The logical relationships used are Inheritance, Implication, AND (conjunction) and Evaluation. Evaluation is the relation between a predicate and its arguments; e.g. *Eval subj(inhib2, DLC)* means that the `subj` predicate holds when applied to the list `(inhib2, DLC)`. These particular logical relationships are reviewed in more depth in (Goertzel and Pennachin, 2005; Looks et al, 2004). Finally, indent notation is used to denote argument structure, so that e.g.

```
R
    A
    B
```

is synonymous with `R(A,B)`.

PLN is an uncertain inference system, which means that each of the terms and relationships used as premises, conclusions or intermediaries in PLN inference come along with uncertain truth values. In this case the truth value of the conclusion at the end of Table 4 comes out to <.8,.07>, which indicates that the system guesses the conclusion is true with probability .8, and that its confidence that this probability assessment is roughly correct is .07. Confidence values are scaled between 0 and 1: .07 is a relatively low confidence, which is appropriate given the speculative nature of the inference. Note that this is far higher than the confidence that would be attached to a randomly generated relationship, however.

The only deep piece of background knowledge utilized by PLN in the course of this inference is the knowledge that:

```
Implication
      AND
            Inh X1 causal_event
            Inh X2 causal_event
            subj(X1, X3)
            subj(X2, X1)
      subj(X2,X3)
```

which encodes the transitivity of causation in terms of the `subj` relationship. The other knowledge

used consisted of simple facts such as the inheritance of inhibition and prevention from the category `causal_event`.

Rule	Premises Conclusion
Abduction	Inh $inhib_1$, inhib Inh $inhib_2$, inhib Inh $inhib_1$, $inhib_2$ <.19, .99>
Similarity Substitution	Eval subj ($prev_1$, $inhib_1$) Inh $inhib_1$, $inhib_2$ Eval subj ($prev_1$ $inhib_2$) <1, .07>
Deduction	Inh $inhib_2$, inhib Inh inhib, causal_event Inh $inhib_2$, causal_event <1,1>
AND	Inh $inhib_2$, causal_event Inh $prev_1$, causal_event Eval subj ($prev_1$, $inhib_2$) Eval subj ($inhib_2$, DLC) AND <1, .07> Inh $inhib_2$, causal_event Inh $prev_1$, causal_event Eval subj ($prev_1$, $inhib_2$) Eval subj ($inhib_2$, DLC)
Unification	ForAll (X_0, X_1, X_2) Imp AND Inh X_0, causal_event Inh X_1, causal_event Eval subj (X_1, X_0) Eval subj (X_0, X_2) Eval subj (X_1, X_2) AND Inh $inhib_2$, causal_event Inh $prev_1$, causal_event Eval subj ($prev_1$, $inhib_2$) Eval subj ($inhib_2$, DLC) Eval subj ($prev_1$, $inhib_2$) <1,.07>
Implication Breakdown (Modus Ponens)	Imp AND Inh $inhib_2$, causal_event Inh $prev_1$, causal_event Eval subj ($prev_1$, $inhib_2$) Eval subj ($inhib_2$, DLC) Eval subj ($prev_1$, DLC) Eval subj ($prev_1$, DLC) <.8, .07>

Table 4. Part of the PLN inference trail underlying Example 1. This shows the series of inferences leading up to the conclusion that the prevention act prev1 is carried out by the subject DLC. A shorthand notation is used here: Eval = Evaluation, Imp = Implication, Inh = Inheritance, inhib = inhibition, prev = prevention. For instance, $prev_1$ and $prev_2$ denote terms that are particular instances of the general concept of prevention. Relationships used in premises along the trail, but not produced as conclusions along the trail, were introduced into the trail via the system looking in its knowledge base to obtain the previously computed truth value of a relationship, which was found via prior knowledge or a prior inference trail.

4 Discussion

We have described a prototype bioNLP system, BioLiterate, aimed at demonstrating the viability of using probabilistic inference to draw conclusions based on logical relationships extracted from multiple biomedical research abstracts using NLP technology. The preliminary results we have obtained via applying BioLiterate in the domain of the genetics of oncology suggest that the approach is potentially viable for the extraction of hypothetical interactions between genes, proteins and malignancies from sets of sentences spanning multiple abstracts. One of our foci in future research will be the rigorous validation of the performance of the BioLiterate system in this domain, via construction of an appropriate evaluation corpus.

In our work with BioLiterate so far, we have identified a number of examples where PLN is able to draw biological conclusions by combining simple semantic relationships extracted from different biological research abstracts. Above we reviewed one of these examples. This sort of application is particularly interesting because it involves software potentially creating relationships that may not have been explicitly known by any human, because they existed only implicitly in the connections between many different human-written documents. In this sense, the BioLiterate approach blurs the boundary between NLP information extraction and automated scientific discovery.

Finally, by experimenting with the BioLiterate prototype we have come to some empirical conclusions regarding the difficulty of several parts of the pipeline. First, entity extraction remains a challenge, but not a prohibitively difficult one. Our system definitely missed some important relationships because of imperfect entity extraction but this was not the most problematic component.

Sentence parsing was a more serious issue for BioLiterate performance. The link parser in its pure form had very severe shortcomings, but we were able to introduce enough small modifications to obtain adequate performance. Substituting un-

common and multi-word entity names with simple noun identifiers (a suggestion we drew from Pyysalo, 2004) reduced the error rate significantly, via bypassing problems related to wrong guessing of unknown words, improper handling of parentheses, and excessive possible-parse production. Other improvements we may incorporate in future include augmenting the parser's dictionary to include biomedical terms (Slozovits, 2003), pre-processing so as to split long and complex sentences into shorter, simpler ones (Ding et al, 2003), modifying the grammar to handle with unknown constructs, and changing the link parser's ranking system (Pysalo, 2004).

The inferences involved in our BioLiterate work so far have been relatively straightforward for PLN once the premises have been created. More complex inferences may certainly be drawn in the biomedical domain, but the weak link inference-wise seems to be the provision of inference with the appropriate premises, rather than the inference process itself.

The most challenging aspects of the work involved semantic mapping and the supplying of relevant background knowledge. The creation of appropriate semantic mapping rules can be subtle because these rules sometimes rely on the semantic categories of the words involved in the relationships they transform. The execution of even commonsensically simple biomedical inferences often requires the combination of abstract and concrete background knowledge. These are areas we will focus on in our future work, as achieving a scalable approach will be critical in transforming the current BioLiterate prototype into a production-quality system capable of assisting biomedical researchers to find appropriate information, and of drawing original and interesting conclusions by combining pieces of information scattered across the research literature.

Acknowledgements

This research was partially supported by a contract with the NIH Clinical Center in September-November 2005, arranged by Jim DeLeo.

References

Chan-Goo Kang and Jong C. Park. 2005. *Generation of Coherent Gene Summary with Concept-Linking Sentences*. Proceedings of the International Symposium on Languages in Biology and Medicine (LBM), pages 41-45, Daejeon, Korea, November, 2005.

Claire Nédellec. 2005. *Learning Language in Logic - Genic Interaction Extraction Challenge*. Proceedings of The 22nd International Conference on Machine Learning, Bonn, Germany.

Cliff Goddard. 2002. *The On-going Development of the NSM Research Program*. Ch 5 (pp. 301-321) of *Meaning and Universal Grammar - Theory and Empirical Findings. Volume II*. Amsterdam: John Benjamins.

Davulcu, H et Al. 2005. *IntEx?: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text*. Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. Detroit.

Donaldson, Ian, Joel Martin, Berry de Bruijn, Cheryl Wolting et al. 2003. *PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine*. BMC Bioinformatics, 4:11,

Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky. 2001. *A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles*. Bioinformatics Jun;17 Suppl 1:S74-82.

Goertzel, Ben and Cassio Pennachin. 2005. *Artificial General Intelligence*. Springer-Verlag.

Goertzel, Ben, Matt Ikle', Izabela Goertzel and Ari Heljakka. 2006. *Probabilistic Logic Networks*. In preparation.

Götz, T and Suhre, O. 2004. *Design and implementation of the UIMA Common Analysis System*. IBM Systems Journal. V 43, number 3. pages 476-489 .

Guha, R. V., & Lenat, D. B. 1994. *Enabling agents to work together*. Communications of the ACM, 37(7), 127-142.

Guha, R.V. and Lenat,D.B. 1990. *Cyc: A Midterm Report*. AI Magazine 11(3):32-59.

Hakenberg, . et al. 2205. *LLL'05 Challenge: Genic Interaction Extraction -- Identification of Language Patterns Based on Alignment and Finite State Automata*. Proceedings of The 22nd International Conference on Machine Learning, Bonn, Germany. 2005.

Hoffmann, R., Valencia, A. 2005. *Implementing the iHOP concept for navigation of biomedical literature*. Bioinformatics 21(suppl. 2), ii252-ii258 (2005).

Ian Niles and Adam Pease. 2001. *Towards a Standard Upper Ontology*. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine, October 2001

Jensen, L.J., Saric, J and Bork, P. 2006. *Literature Mining for the biologist: from information retrieval to biological discovery*. Nature Reviews. Vol 7. pages 119-129. Natura Publishing Group. 2006.

Jing Ding. 2003. *Extracting biomedical interactions with from medline using a link grammar parser*. Proceedings of 15th IEEE international Conference on Tools With Artificial Intelligence.

Kim, Jim-Dong et al. 2004. *Introduction to the Bio-NLP Entity Task at JNLPBA 2004*. In Proceedings of JNLPBA 2004.

Lenat, D., Prakash, M., & Shepard, M. 1986. *CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks*. AI Magazine, 6(4), 65-85

Lerman, K , McDonal, R., Jin, Y. and Pancoast, E. University of Pennsylvania *BioTagger.* 2006. http://www.seas.upenn.edu/~ryantm/software/BioTagger/

Looks, Moshe, Ben Goertzel and Cassio Pennachin. 2004. Novamente: An Integrative Approach to Artificial General Intelligence. *AAAI Symposium on Achieving Human-Level Intelligence Through Integrated Systems and Research*, Washington DC, October 2004

Mandel, Mark. 2006. *Mining the Bibliome*. February, 2006 http://bioie.ldc.upenn.edu

Mark A. Greenwood, Mark Stevenson, Yikun Guo, Henk Harkema, and Angus Roberts. 2005. *Automatically Acquiring a Linguistically Motivated Genic Interaction Extraction System*. In Proceedings of the 4th Learning Language in Logic Workshop (LLL05), Bonn, Germany.

McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin and P. White. 2005. Simple Algorithms for Complex Relation Extraction with Applications to Biomedical IE. R. 43rd Annual Meeting of the Association for Computational Linguistics, 2005.

Müller, H. M., Kenny, E. E. and Sternberg, P. W. 2004. *Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature*. PLoS Biol 2(11): e309

Pyysalo, S. et al. 2004. *Analisys of link Grammar on Biomedical Dependency Corpus Targeted at Protein-Protein Interactions*. In Proceedings of JNLPBA 2004.

Riedel, et al. 2005. *Genic Interaction Extraction with Semantic and Syntactic Chains*. Proceedings of The 22nd International Conference on Machine Learning, Bonn, Germany.

Ryan McDonald and Fernando Pereira. 2005. *Identifying gene and protein mentions in text using conditional random fields*. BMC Bioinformatics 2005, 6(Suppl 1):S6

Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbur WJ, Hatzivassiloglou V, Friedman C. 2004. *GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data*. Journal of Biomedical Informatics 37(1):43-53.

Sleator, Daniel and Dave Temperley. 1993. *Parsing English with a Link Grammar*. Third International Workshop on Parsing Technologies, Tilburg, The Netherlands.

Smalheiser, N. L and Swanson D. R. 1996. *Linking estrogen to Alzheimer's disease: an informatics approach*. Neurology 47(3):809-10.

Smalheiser, N. L and Swanson, D. R. 1998. *Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses*. Comput Methods Programs Biomed. 57(3):149-53.

Syed Ahmed et al. 2005. *IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text*. Proc. of BioLink '2005, Detroit, Michigan, June 24, 2005

Szolovits, Peter. 2003. *Adding a medical lexicon to an English parser*. Proceedings of 2003 AMIA Annual Symposium. Bethesda. MD.

Tanabe, L. U. Scherf, L. H. Smith, J. K. Lee, L. Hunter and J. N. Weinstein. 1999. *MedMiner: an Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling*. BioTechniques 27:1210-1217.

Wierzbicka, Anna. 1996. *Semantics, Primes and Universals*. Oxford University Press.