

Two approaches to principled query expansion for genomics question-answering

Helen L. Johnson William A. Baumgartner Jr. Jason Lones Lynne M. Fox
Zoya Voronovich K. Bretonnel Cohen Lawrence Hunter

`larry.hunter@uchsc.edu`

Biomedical Text Mining Group Center for Computational Pharmacology
University of Colorado School of Medicine

Abstract

The TREC 2007 Genomics track task was to provide answers to 36 questions by extracting text from a full-text document collection. The approach of the University of Colorado School of Medicine team was a concept-recognition-based effort that combined named entity and predicate expansion at the query expansion stage with concept filtering at the post-processing stage. We submitted two runs. The first run, which included systematic manual query expansion and concept filtering, outperformed the median in about 30% of the queries. The second run, which added predicate expansion to the first run methods, showed a small improvement in performance over the first run.

Introduction

The 2007 TREC Genomics track presented teams with 36 questions formulated by biomedical researchers. The document collection was the same one used for the TREC 2006 Genomics track [14]. The challenge was to return a list of answers in the form of text strings of paragraph-length or shorter from full-text articles. The semantic type of answer expected was noted in each question by a general category keyword, such as *gene*, *mutation*, or *toxicity*. There were 14 categories in all.

In order to use a concept-recognition-based ap-

proach, language processing tools that can recognize these categories in text are necessary. The biomedical language processing community has devoted much work to developing gene and protein recognizers, resulting in a number of publicly available taggers [31]. Beyond genes and proteins, there are some taggers for other biomedical concepts, such as mutations [6, 15] and UMLS concepts [2]. However, there are far fewer concept recognition tools beyond those of gene and protein taggers.

Though the TREC queries provide the semantic types expected for the answers, identifying corresponding concepts in text may not be enough. In each query there is a stated relationship with one unspecified participant. Extracting good answers from full text may require the recognition of named entities in the context of specific relationships to other named entities. Biomedical relationship recognition relies on named entity recognition to fill in the slots of relationship frames. Parallel to named entity recognition, most research in this area has focused on identifying the relationships between *genes* and *proteins* in text (e.g. [1, 13, 28] and see [9] for an overview). Some groundwork has been laid for the recognition of relationships between other biological entities, such as the relationships between *diseases* and *treatments* [27], and *proteins* and *locations* [26]. Also, an alternative way to do this is by first performing full syntactic parses on the text, then extracting the named entities as they relate to

each other across predicates, thereby getting data akin to the semantic-syntactic structure provided by PropBank[24], FrameNet [12] or PASBio [33]. However, getting full parses is a time-consuming procedure, and performing them on a large text collection may not be a reasonable method for information extraction on the fly.

Concept recognition was a major component of our TREC information extraction system, and in particular two types of concept recognition techniques characterize our first submitted run. The first is a set of automatic concept recognizers that were applied to the search engine output to filter results that did not contain the desired concepts. Concept types for which we do not have automatic taggers were identified using the second type of concept recognition technique: gazetteers of terms and synonyms denoting a particular type of concept, e.g. *diseases* or *drugs*. The gazetteers were used to expand the queries before submitting them to the search engine.

Regarding the relationships explicit in the queries, our second submitted run tests a simple procedure of adding a set of relevant predicates to the queries from the first run. Research by Bertaud et al. [5] suggests that the precision for retrieving articles that contain medical diagnosis findings is increased by using synonyms of verbs that denote *finding* and *evidence* in the search strings to retrieve articles from MEDLINE. Extending this idea, we hypothesize that the co-occurrence of a particular biomedical predicate with the appropriate entities may signal a biomedical relationship between those entities. Our second run combines the named entity expansion of the first run with synonyms of biomedical verbs in the search string as a simple way to identify the relationships between named entities, with the purpose of raising the precision of the retrieved results.

Methods

A lower-bound baseline system might measure the effectiveness of using just the keywords in the original questions. An upper-bound benchmark system might measure the effectiveness of a system that models the human behavior of iterative query building

based on search feedback. The system we present here is a benchmark model that approaches the iterative human behavior by presuming the query refinement step up front in a single step of query expansion.

We submitted two runs to the TREC 2007 shared task. The first run was a benchmark run using manually expanded named entities in queries submitted to the Indri search engine, part of the Lemur toolkit [23]. It was designed to model the performance of a human researcher building queries to submit to a search engine, and hence was designated as a manual run. The second run was a modification of the first in that in addition to expanded named entities, an expanded set of biologically relevant predicates were added to the expanded query.

In the original queries distributed by TREC, we identified three types of keywords and then used them for distinct purposes. The first type was the named entities which were the mainsprings for query expansion. The second was biological predicates, which were used in select cases for query expansion. The third was the semantic concept types that identified the category of the expected answer, for example, the capitalized text in brackets in query #204: *What nervous system [CELL OR TISSUE TYPES] synthesize neurosteroids in the brain?* These concept types were used to select only the retrieved documents that contained the desired semantic concepts, thereby filtering out those documents that did not.

Query Creation and Expansion

For our first run, we used the named entity keywords as the starting point for query expansions. A domain expert manually expanded named entities to include synonyms and abbreviations by referencing MeSH, EntrezGene, and Wikipedia. In most cases, the domain expert found an appropriate MeSH heading for the named entity. He expanded the query to include the MeSH heading and all Entry Terms of the heading. He then traversed down the MeSH tree structure to pick up more specific terms. In a few cases MeSH was not an appropriate resource. When a term could not be found in MeSH, the domain expert consulted either EntrezGene for gene synonyms, or Wikipedia for other named entity query expansion suggestions.

This use of lexical resources was done in order to systematize the query creation and expansion procedure.

Our second run used the same named entity expansions from the first run, but added an additional verbal element. The hypothesis that motivates this method is that requiring a biomedical relation in the text would limit prodigious search results to more relevant passages. While the presence of an apparently biomedical predicate in the same text span does not guarantee a particular relation, it might raise the likelihood of the relation being asserted.

This relation requirement was imposed by expanding the biomedical verb in the query with synonymous verbs and nominalizations. This predicate requirement and expansion was performed on a subset of the queries: those queries that garnered more than 500 hits from the first submission run were examined for verbs or nominalizations that had biomedically specific meanings. For this reason this run was designated as interactive. Five queries were chosen as prime candidates for predicate expansion – #200, #211, #216, #221 and #226. Predicate expansion was performed in one of two ways. For queries #200, #211, #216, and #221, WordNet [11] term lists associated with the predicate element of the chosen query were manually mined for alternate predicates. Words were selected from WordNet’s lists of direct and full hyponyms, direct and inherited hypernyms, direct and full troponyms, sister terms, and derivationally related forms. Some verbs or nominalizations that seemed too general or that were clearly not biologically relevant were left off the expanded list of verbs. For an example of an expanded query, including the expanded predicate set, see Table 1.

We departed from this method for query #226: *What [PROTEINS] make up the murine signal recognition particle?* Although there was no biologically salient verb present in the original query, it was deemed that the underlying question was addressing protein-protein interaction concepts. We had a group of verbs already assembled for this purpose [18], and simply plugged them into the query.

Concept filtering

We dealt with the third type of keyword – the semantic type of the expected answer – in two ways. For *gene* and *protein*, *mutation*, and *biological substance* recognition we used automatic concept recognizers. The protein/gene tagger was an aggregate of the ABNER [30] and LingPipe [8] gene taggers, combined with a consensus filter to maximize precision, similar to the one used in the BioCreative II Gene Mention task [18]. To tag mutations we used MutationFinder [6]. For biological substances, we developed a concept recognizer for this task.

The recognition of biological substances poses a challenge due to the broad nature of this entity type and the impracticality of creating a list of all biological substances. Our approach focused on using machine learning to extract biological substance entities. To recognize biological substances, we trained a conditional random field model, using the ABNER system on the GENIA corpus [19]. A subset of the annotations in the GENIA corpus were selected by a domain expert as representing biological substances (amino acid monomer, inorganic, RNA molecule, carbohydrate, polynucleotide, protein family or group, protein complex, protein molecule, other organic compound, peptide, RNA family or group, and lipid).

These taggers were applied to the returned results of the query engine to filter results that did not have a named entity of the respective semantic type.

For those concept types for which we did not have concept recognizers, gazetteers were constructed by automatically parsing resources for lists of terms, synonyms, and abbreviations. See Table 2 for a list of the resources used for each semantic concept type. Many terms mined from those resources contained nomenclature that would rarely if ever be found in free text, such as terms in parentheses – *Dry Heaves (Nausea)* [22] – and terms with adjectives following punctuation – *Abrasion, localized* [16]. These terms were programmatically retrieved, their order reversed, and/or their punctuation removed in order to maximize the coverage of the gazetteers. Most of the time this text processing step works to increase the viable synonyms in the expanded query, as in the

Table 1: Example of an expanded query (#216) including expanded predicate. The “expanded terms” for each “keyword” were concatenated using Boolean *OR* operators. The “expanded term” sets that resulted from that were concatenated with Boolean *AND* operators. Rows labeled “1” and “2” stand for uchsc1 and uchsc2 runs, respectively.

original query		What [GENES] regulate puberty in humans?	
keyword		expanded terms	
2	1	puberty	puberty, sexually mature, sexual maturity, adolescent, adolescence
		human	human, homo sapien, patient, man, woman, child, children, girl, boy, teen, teenager
		regulate	initiate, initiation, origin, originate, origination, start, create, bring about, bring forth, generate, generation, invoke, invocation, evoke, evocation, establish, produce, production, give rise, gave rise, originate in, originative, starter, delay, detain, hold up, stall, decelerate, slow, slow down, slow up, retard, retardation, decrease, diminish, lessen, fall behind, fell behind, change, postpone, hold over, set back, defer, deferment, remit, remittance, put off, stimulate, stimulation, bring on, brought on, develop, development, induce, induction, inducement, inducing, encourage, encouragement, lead to, instigate, instigation, compel, effectuate, precipitate, rush, hasten

examples above. However, it is not a perfect technique – for instance, reversing the order of the disease term *hand, foot & mouth disease* [16] gives *foot & mouth disease hand*.

Finally, the lists of gazetteer terms were added to the queries in the query expansion stage.

Information Retrieval

The documents from the TREC collection had been previously preprocessed and indexed using Lemur [23] for our TREC 2006 effort, (c.f. [7] for a full description). Documents were divided into eligible paragraphs and Porter stemming [25] and stop word removal were applied prior to document indexing.

Expanded terms in sets of synonyms were concatenated using Boolean *OR*, and the sets were submitted to the Indri search engine using the *#band* operator which enforces a true Boolean *AND* operation. Like indexing, Porter stemming and stop word removal were used on queries before submitting them to the search engine. A cap of 1000 returned documents was enforced.

Results

Our first run, uchsc1, was the manual run that filtered the potential passage hits for the presence of the correct concept in terms of the expected answer. As described in the methods section, the filtering was performed in one of two ways: either by an automatic entity recognition tagger, or by matching words in the document to a gazetteer list of words.

This run outperformed the median in about 30% of the queries across all four performance metrics (document-level, passage-level, passage2-level, and in diversity of aspects). See Figures 1 – 4 for a comparison of uchsc1 run results against the median and max MAP scores for all teams. The system’s five best performances were on queries #201, #220, #224, #229, and either #214 or #215 depending on metric. Of those six queries, five were handled with the automatic named entity taggers and one used the gazetteer filtering method, see Table 3. Both the automatic named entity tagger and the gazetteer methods contributed to results after the top five.

Our second run, uchsc2, was the interactive run in which we manually expanded the biomedical predicates in five queries with the most hits returned from

Table 2: Resources used to construct gazetteers.

concept type	resource	types of terms extracted
antibodies	AbMiner [21] MeSH [20]	all entries mn=D12.776.124.486.485.114.* mh, entry, print entry, fx
cell or tissue types	Cell Type Ontology [4]	name, exact_synonym, broad_synonym, narrow_synonym, related_synonym
diseases	Brenda Tissue Type Ontology [29] Human Disease Ontology [16]	name, synonym
drugs	FDA Orange Book [10] MeSH [20]	drugname, activeingredient mn=D26.* mh, entry, print entry, fx
molecular functions	Gene Ontology (Molecular Function) [3]	name, exact_synonym, broad_synonym, narrow_synonym, related_synonym
signaling pathways/pathways	Event INOH pathway ontology [17] Pathway Ontology [32]	name, synonym name, synonym
strains	—	—
signs and symptoms	MedicineNet.com [22]	all entries
toxicities	hand-built list of keywords	—
tumor types	MeSH [20]	mn=C04.* mh, entry, print entry, fx

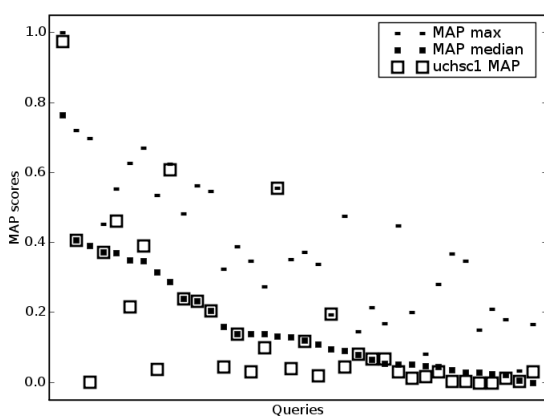


Figure 1: Document MAP, ordered by descending median MAP.

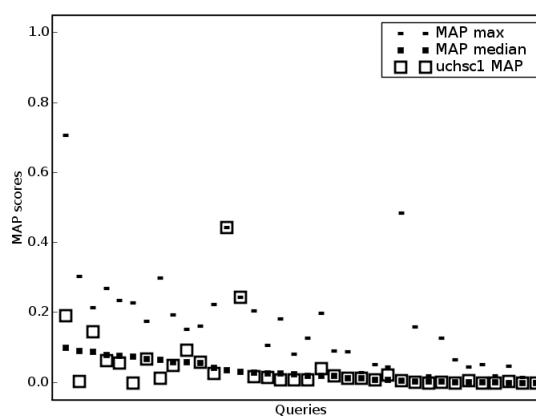


Figure 2: Passage MAP, ordered by descending median MAP.

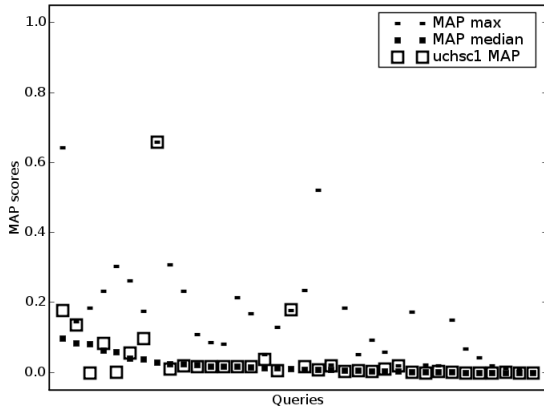


Figure 3: Passage2 MAP, ordered by descending median MAP.

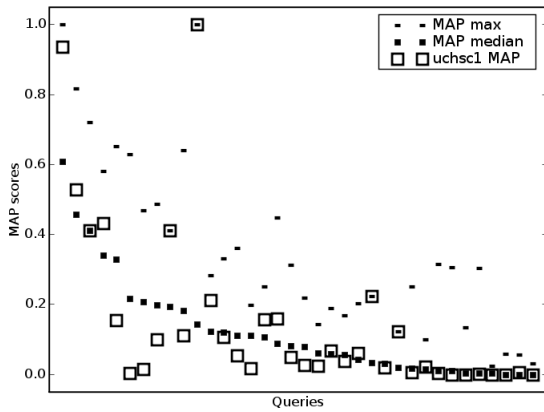


Figure 4: Aspect MAP, ordered by descending median MAP.

Table 3: Characteristics of the five best query performances. The “Method” column refers to the automatic vs gazetteer method of concept recognition.

Query-ID	Concept	Method
#224	genes	auto
#201	mutations	auto
#220	proteins	auto
#229	signs/sympt.	gazette
#215	proteins	auto
#214	genes	auto

the uchsc1. These five queries had low MAP scores in the uchsc1 run across all metrics because of the large amounts of hits they returned. In three of the five queries, (#216, #221, #226), expanding the predicate improved the score. In queries #200 and #211, the performance stayed the same or deteriorated. See Figures 5 – 8 for a comparison of uchsc2 to uchsc1 results on the five queries. Of particular note is the vast improvement in aspect diversity for query #226 when the predicates were expanded (see Figure 8).

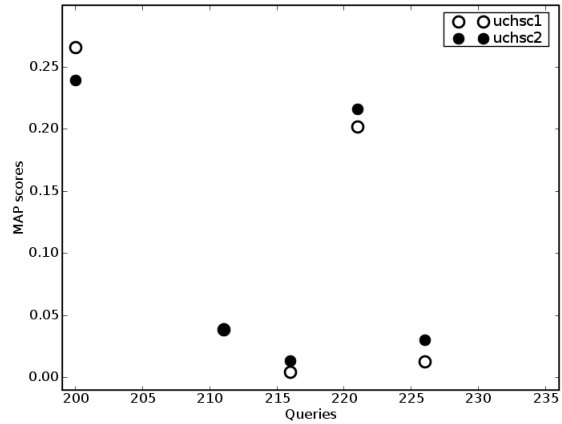


Figure 5: Comparison of document MAP: uchsc1 vs uchsc2 predicate expansions. Ordered by query number.

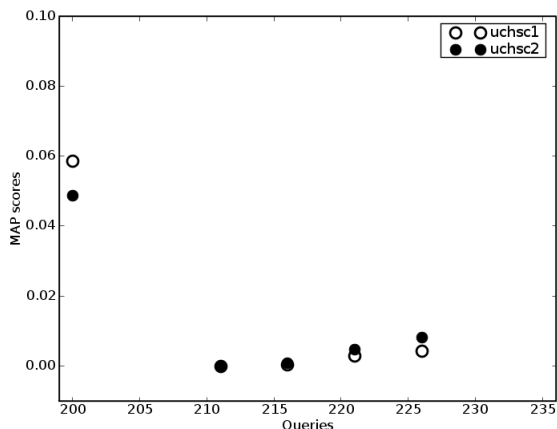


Figure 6: Comparison of passage MAP: uchsc1 vs uchsc2 predicate expansions. Ordered by query number.

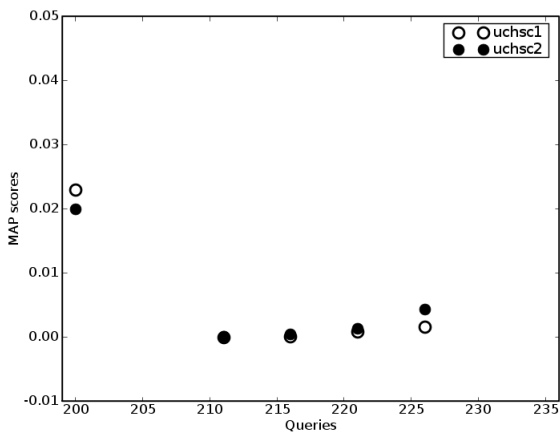


Figure 7: Comparison of passage2 MAP: uchsc1 vs uchsc2 predicate expansions. Ordered by query number.

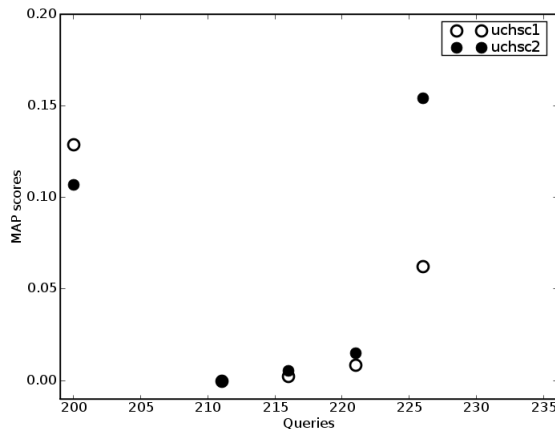


Figure 8: Comparison of aspect MAP: uchsc1 vs uchsc2 predicate expansions. Ordered by query number.

Discussion

The results of the uchsc1 run suggest that the tasks of concept recognition and filtering are better performed by the automatic named entity recognizers than by the gazetteers. This is not surprising, since it has been reported repeatedly that gazetteer approaches are less effective than automatic named entity taggers. This suggests that energy spent on developing named entity recognizers for the wide range of biology concepts would be of value. Years of effort has been invested in creating automatic named entity recognizers for genes and proteins [9]. The result is rule-based and machine learning approaches using a variety of features that have yielded good results. The named entity recognizer for mutations, MutationFinder, relies on the consistent nature of mutation mentions in text. Other concepts, such as diseases or symptoms, may not be so quite so easy to characterize in free text. The success of concept recognizers may rest on the development of ontologies that characterize such concepts in terms of their relationships to other concepts.

The small improvement by using expanded pred-

icates is less than we hypothesized. This could be for several reasons: 1) Named entity recognition is sufficient to extract the relevant relations. Requiring relationship predicates to co-occur in paragraphs is redundant. 2) The predicate expansion used in uchsc2 was too permissive. Restricting the required co-occurring predicates to only the most synonymous words of the desired relation is a more appropriate solution. 3) The biomedical verbs included in the expansion are too polysemous to narrow the search results as desired. Further research in both this method and in extracting relations from text is necessary to answer this open question.

Acknowledgments

We would like to thank Greg Caporaso, Dave Farell and Trevor Pincock for technical support during the competition.

References

- [1] S. Ahmed, D. Chidambaram, H. Davulcu, and C. Baral. Intex: A syntactic role driven protein-protein interaction extractor for bio-medical text. In *Proceedings ISMB/ACL Biolink 2005*, 2005.
- [2] A. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proc AMIA 2001*, pages 17–21, 2001.
- [3] Michael Ashburner and et al. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000.
- [4] J. Bard, S. Y. Rhee, and M. Ashburner. An ontology for cell types. *Genome Biol*, 6(2), 2005. Retrieved file cell.obo (version 1.24 25:05:2007 09:56) from <http://obofoundry.org/cgi-bin/detail.cgi?cell> on June 14, 2007.
- [5] V. Bertaud. The value of using verbs in Medline searches. *Medical Informatics & The Internet in Medicine*, 32(2):117–122, 2007.
- [6] J. Gregory Caporaso, William A. Baumgartner Jr., David A. Randolph, K. Bretonnel Cohen, and Lawrence Hunter. Mutationfinder: A high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23:1862–1865, 2007.
- [7] J.G. Caporaso, W.A. Baumgartner, H. Kim, Z. Lu, H.L. Johnson, O. Medvedeva, A. Lindemann, L. Fox, E. White, K.B. Cohen, and L. Hunter. Concept recognition, information retrieval, and machine learning in genomics question answering. In *TREC 2006 Proceedings*, 2006.
- [8] B. Carpenter. Phrasal queries with lingpipe and lucene. *13th Text REtrieval Conference*, 2004.
- [9] Aaron M. Cohen and William Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005.
- [10] Drugs@FDA. url=<http://www.fda.gov/cder/drugsatfda/datafiles/>. Retrieved file drugsatfda.zip on June 14, 2007.
- [11] C. Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, Massachusetts, May 1998a.
- [12] Charles J. Fillmore and Collin F. Baker. Frame semantics for text understanding. In *Proceedings of NAACL WordNet and Other Lexical Resources Workshop*, Pittsburg, Pennsylvania, 2001. ACL.
- [13] Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl. 1):S74–S82, 2001.
- [14] William Hersh, Aaron M. Cohen, Phoebe Roberts, and Hari K. Rekapalli. Trec 2006 genomics track overview. In *TREC Notebook*. NIST, 2006.

- [15] Florence Horn, Anthony L. Lau, and Fred E. Cohen. Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, 20(4):557–568, Mar 2004.
- [16] Human Disease Ontology. url=<http://diseaseontology.sourceforge.net/#communication>. Retrieved file human_disease.obo (version 0.8 beta) from http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology on June 14, 2007.
- [17] Integrated Network Objects with Hierarchies: Event Pathway Ontology. url=<http://www.inoh.org/>. Retrieved file EventOntology.obo (version 1.5 28:03:2007 18:22) from <http://www.obofoundry.org/cgi-bin/detail.cgi?id=event> on June 14, 2007.
- [18] William A. Baumgartner Jr., Zhiyong Lu, Helen L. Johnson, J. Gregory Caporaso, Jesse Paquette, Anna Lindemann, Elizabeth K. White, Olga Medvedeva, K. Bretonnel Cohen, and Lawrence Hunter. An integrated approach to concept recognition in biomedical text. In *Proceedings of BioCreative 2006*, 2007.
- [19] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl. 1):180–182, 2003.
- [20] DA Lindberg, BL Humphreys, and AT McCray. The unified medical language system. *Methods Inf Med*, 32(4):281–291, 1993. Retrieved file d2007.bin from <http://www.nlm.nih.gov/cgi/request.meshdata> on June 20, 2007.
- [21] S.M. Major, S. Nishizuka, D. Morita, R. Rowland, M. Sunshine, U. Shankavaram, F. Washburn, D. Asin, H. Kouros-Mehr, D. Kane, and J. Weinstein. AbMiner: a bioinformatic resource on available monoclonal antibodies and corresponding gene identifiers for genomic, proteomic, and immunologic studies. *BMC Bioinformatics*, 7(1):192, 2006. Retrieved terms from website <http://discover.nci.nih.gov/abminer/searchantibody.do> on June 22, 2007.
- [22] MedicineNet.com. url=http://www.medicinenet.com/symptoms_and_signs/article.htm. Retrieved terms from website on June 15, 2007.
- [23] Paul Ogilvie and James P. Callan. Experiments using the Lemur Toolkit. In *Text REtrieval Conference*, 2001.
- [24] Martha Palmer, Paul Kingsbury, and Daniel Gildea. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.
- [25] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [26] S. Ray and M. Craven. Representing sentence structure in hidden Markov models for information extraction. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 273–1. Washington, US: Morgan Kaufmann Publishers, 2001.
- [27] Barbara Rosario and Marti A. Hearst. Classifying semantic relations in bioscience texts. In *Proceedings of ACL 2004*, pages 430–437, 2004.
- [28] Jasmin Saric, Lars Juhl J. Jensen, Rossitza Ouzounova, Isabel Rojas, and Peer Bork. Extraction of regulatory gene/protein networks from medline. *Bioinformatics*, July 2005.
- [29] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*, 32(Database issue), 2004. Retrieved file BrendaTissue.obo from <http://obofoundry.org/cgi-bin/detail.cgi?brenda> on June 14, 2007.
- [30] Burr Settles. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

- [31] Larry Smith, Lorraine Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christof Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig Struble, Richard Povinelli, Andreas Vlachos, William Baumgartner, Jr., Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres Perez, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Mana, Jacinto Mata-Vazquez, and w. John Wilber. Overview of BioCreative II gene mention recognition. *Genome Biology*, to appear.
- [32] S.N. Twigger, M. Shimoyama, S. Bromberg, A.E. Kwitek, and H.J. Jacob. The Rat Genome Database, update 2007—Easing the path from disease to data and back again. *Nucleic Acids Research*, 35(Database issue):D658, 2007. Retrieved file pathway.obo (version 07:06:2007 16:25) from <http://www.obofoundry.org/cgi-bin/detail.cgi?id=pathway> on June 14, 2007.
- [33] Tuangthong Wattarujeekrit, Parantu K. Shah, and Nigel Collier. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(155), 2004.