Molecular Cell *21*, 589–594, March 3, 2006 ©2006 Elsevier Inc. DOI 10.1016/j.molcel.2006.02.012

Biomedical Language Processing: What's Beyond PubMed?

Perspective

Lawrence Hunter[1,*] **and K. Bretonnel Cohen**[1]
[1] Center for Computational Pharmacology
University of Colorado School of Medicine
Aurora, Colorado 80045

Literature Overload

Exponential growth of the peer-reviewed literature and the breakdown of disciplinary boundaries heralded by genome-scale instruments have made it harder than ever for scientists to find and assimilate all the publications relevant to their research. The widespread adoption of title/abstract word search, primarily through the National Library of Medicine's PubMed system (http://www.ncbi.nlm.nih.gov/pubmed), was the first major change in the way bioscientists found relevant publications since the origin of *Index Medicus* in 1879. (Although it remains useful for locating pre-1966 literature (Hersh, 2003), *Index Medicus* ceased publication in 2004.) However, PubMed is only the beginning of a revolution in how scientists use the biomedical literature. Computational tools that classify documents, extract factual information, generate summaries, and generally process human language are providing powerful new tools for staying on top of the torrent of publications.

The biomedical literature is growing at a double-exponential pace; over the last 20 years, the total size of MEDLINE (the database searched by PubMed) has grown at a ~4.2% compounded annual growth rate, and the number of new entries in MEDLINE each year has grown at a compounded annual growth rate of ~3.1% (see Figure 1). There are now more than 16,000,000 publications in MEDLINE; more than three million of those were published in the last 5 years alone. The number of MEDLINE entries with a 2005 publication date was 666,029—more than 1800 per day.

Large as MEDLINE is, it captures only bibliographic information and abstracts. Electronic access to the full texts, including graphics and figures, is also on the rise, and sophisticated linkages between publications and data repositories or other supplementary materials increase the amount of information available still further. Although online full-text materials are increasingly prevalent, dramatic increases in subscription prices and decreases in library budgets have paradoxically decreased access for some researchers. Toll-free linking, where copyright owners allow free search but charge per view, is one approach to ameliorating this problem.

An alternative strategy toward this goal is the recent establishment of a movement toward a new "Open Access" model of scientific publishing. On April 11, 2003, a group of individuals interested in promoting open access to the scientific literature drafted a statement of principles that is now referred to as the Bethesda Statement on Open Access Publishing (http://www.earlham.edu/~peters/fos/bethesda.htm), later followed in Europe by The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html), ushering in the era of unrestricted use of scientific publications. Although some publishers have resisted open access, many others have responded by increasing access to archives and developing related services. In 2004, the US National Library of Medicine (NLM) created a repository, called PubMedCentral (PMC, http://www.pubmedcentral.gov/), for open access articles, which as of this writing tracks some or all of the content of 154 biomedical journals automatically and accepts individual article submissions from hundreds of others. Perhaps most important for the future was the publication in the Federal Register of a new "Policy on Enhancing Public Access to Archived Publications Resulting From NIH-Funded Research," which beginning on May 2, 2005 requests all NIH-funded investigators to submit to PMC all manuscripts resulting from research supported in whole or in part by NIH money. Less than 6 months later, more than 430,000 full-text articles (totaling more than 5TB in compressed form) are available through PMC. Furthermore, NLM is digitizing earlier print issues of many of the journals already in PMC, extending the availability of full texts back to before the implementation of the 2005 policy. Although NIH officials estimate that ~10% of the literature is NIH supported, and only about 6.5% of the MEDLINE entries for 2005 were indexed as supported by NIH extramural funding, PMC marks a significant change in the availability of full-text scientific articles in biomedicine. As stated on the NLM's web site, PMC "makes it possible to integrate the literature with a variety of other information resources such as sequence databases and other factual databases that are available to scientists, clinicians and everyone else interested in the life sciences. The intentional and serendipitous discoveries that such links might foster excite us and stimulate us to move forward." Development of novel computational tools and techniques for textual analysis are a vital prerequisite for achieving NLM's vision.

Biomedical Language Processing Systems

Meanwhile, over the last 5 years or so, there has been a remarkable surge of new results in biomedical language processing (BLP). BLP encompasses the many computational tools and methods that take human-generated texts as input, generally applied to tasks such as information retrieval, document classification, information extraction, plagiarism detection, or literature-based discovery. Information retrieval systems, like PubMed or Google, focus on searching large collections to find documents that are relevant to a query. They are evaluated by their sensitivity (what proportion of all of the relevant documents are found) and their specificity (what proportion of the documents found are actually relevant to the query). Document classification is another task with many biomedical applications. Such a system can be used to organize large retrieval results into meaningful categories (e.g., Tanabe et al. [1999]). Document classification can also be used to

*Correspondence: larry.hunter@uchsc.edu

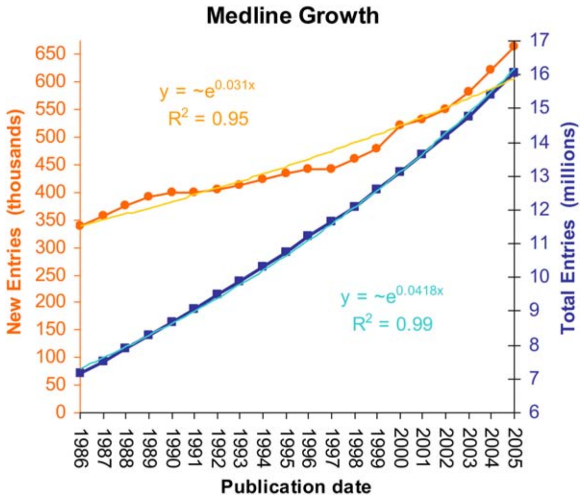

Figure 1. Growth in the Biomedical Literature, 1986–2005

Orange circles show the number of new articles indexed in MEDLINE with a publication date in each year. Blue squares show the total number of articles indexed at the end of each year. The light lines show exponential curves fitted to each series, and the corresponding formulae show growth rates and goodness of fit measures for both curves.

filter or route a flow of documents (e.g., from a service like Thompson's Current Contents or from web technologies such as RSS, the framework for providing automated "feeds" of news stories, articles, or other content types related to a specific subject matter) based on their contents. Such filtering technologies are used, for example, by several of the model organism database projects to identify publications relevant to their gene annotation efforts. Information extraction (also sometimes called text data mining) systems scan large numbers of publications to extract specific factual information, often to populate a database. The Literature Support for Alternative Transcripts (LSAT) system, discussed in more detail below, produced a database of transcript diversity in about 4000 human genes by scanning more than 14,000 MEDLINE abstracts from hundreds of different journals. Literature-based discovery is the attempt to automatically induce novel hypotheses by processing existing publications. Although a few dramatic results were obtained in the 1980s and 90s, e.g., the discovery of a linkage between magnesium and migraine (Swanson, 1988), repeated, successful, literature-based discovery remains beyond current abilities (see Weeber et al. [2005] for a good review of LBD systems).

A brief survey of some of the existing tools illustrates the impact BLP is starting to have in molecular biology research. The LSAT system (Shah et al., 2005) surveyed mentions of transcript diversity (e.g., alternative splicing) in a large set of MEDLINE abstracts. The system was able to achieve more than 90% accuracy and identified information about the genes, tissues, species, mechanisms, number of isoforms, experimental methods, etc. regarding the transcript diversity of more than 4000 genes. LSAT is one of the tools used to populate the EBI's Alternative Splicing databases (Thanaraj et al., 2004). A similar BLP-derived database of mutations in G protein coupled receptors (http://www.gpcr.org) was created by using an information extraction system called MuteXt (Horn et al., 2004); these data are particularly important for pharmaceutical development and pharmacogenomics. Chilibot ([Chen and Sharp, 2004]; http://www.chilibot.net) is a freely available, web-based application that takes sets of gene names, and (optionally) additional keywords as input, and finds information about the relationships among them. In an initial information retrieval step, it constructs queries to send to the PubMed search engine. It uses basic language processing techniques to identify sentences that describe stimulatory, inhibitory, and other relationships between pairs of genes. For pairs of genes for which large sets of sentences are found, it then uses techniques from the field of automatic summarization (the area of natural language processing concerned with constructing shortened versions of input texts) to select the best sentences to display to the user. A graph displays the entire set of interactions amongst all of the genes that were input. By clicking on links between pairs of genes, the user can see the full set of sentences that describe relations between those genes, and clicking on the sentences themselves displays the original PubMed abstract. The Textpresso system ([Müller et al., 2004]; http://www.textpresso.org) is an example of an information extraction project that strives to create an up-to-date summary of current knowledge related to a specific model organism. Users can search for information involving any of the 33 semantic classes of things and relationships between them that the system extracts information about (e.g., genes, clones, pathways, or mutations). The original system, devoted to *C. elegans*, has scanned thousands of full-text articles, and the resulting database is accessed more than 1500 times daily by nematode biologists worldwide. Work is in progress to extend this method to a variety of other organisms, including *N. crassa*, *D. melanogaster*, and *S. cerevisiae*, and preliminary systems are now available on several model organism database sites. As an example of how document filtering can increase productivity, consider the BLP activities at the protein-protein interaction database BIND (http://www.bind.ca). A survey by the database team (Alfarano et al., 2005) estimated that ~1900 interactions are published per month, spread across about 80 journals. The database team developed the PreBIND document filtering system to assist the curation of the database. PreBIND uses a combination of statistical methods for finding relevant documents and rule-based methods for recognizing the names of biomolecules to find statements about protein interactions. BIND curators used the system to suggest candidate additions to the interaction database. An evaluation by BIND personnel estimated that the system reduced the duration of a single representative task by 70%, representing a savings of 176 person days (Donaldson et al., 2003).

Commercial language processing products for biomedicine have also recently started to become available. Bioalma's almaKnowledgeServer advertises conceptual search capabilities, document relevance ranking, and relationship detection. Ariadne Genomics markets a product called MedScanReader, which is claimed to be able to summarize the contents of large collections of documents or abstracts by identifying dozens of classes of

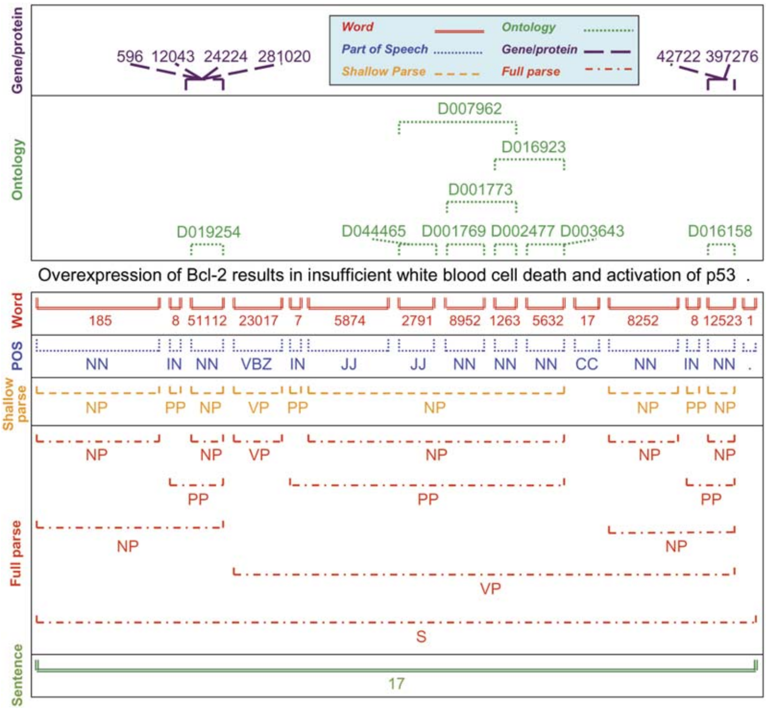

Figure 2. Some Levels of Linguistic Analysis

The gene/protein layer contains LocusLink IDs. The ontology layer contains MeSH concepts. POS is part of speech. NN is a singular noun, IN is a preposition, VBZ is a third-person singular present-tense verb, JJ is an adjective, and CC is a coordinating conjunction. Adapted from Nakov et al. (2005).

biological entities and the relationships between them. Long-time language processing software vendor L & C Global has recently oriented much of its business toward processing medical records and biomedical language.

Why BLP Is Hard

Language processing applications have been built–or attempted—since the earliest days of computer science. Nonetheless, the goal of natural language understanding—full, human-like computational comprehension of language—remains unrealized and is likely to remain so for some time to come. The primary source of difficulty in NLP comes from ambiguity: the possibility of multiple interpretations for strings that represent language. As a very basic example, consider periods in PubMed abstracts. An obvious role of periods is to mark the boundaries of sentences. However, they also serve a variety of other functions, such as marking abbreviations, indicating decimal points, and demarcating levels in hierarchical identifiers. Consequently, the apparently simple task of locating sentence boundaries is problematic. It is even more so in biomedical texts, where sentences can begin with lower-case letters, e.g., "lush mutants are also defective for pheromone-evoked behavior" (PMID 15664171).

Ambiguity is a pervasive phenomenon throughout language. It exists at every level of linguistic analysis (see Figure 2). It complicates almost every linguistic processing task: determining a word's part of speech (is "sense" a noun or a verb?); deciding which meaning a given word has (is the noun "group" a chemical entity or an assemblage?); and determining the intended groupings of words and phrases (is "regulation of cell migration and proliferation" the regulation of some unspecified sort of proliferation and also regulation of cell migration or the regulation of cell migration and the regulation of cell proliferation?). Ambiguities at more basic levels of linguistic analysis contribute to an error rate that compounds from one level of analysis to the next. Additional problems of reference arise in the context of the biomedical domain. Ambiguity of acronym and abbreviation definitions (does "PDA" mean "patent ductus arteriosus," "posterior descending artery," or "phorbol 12,13 diacetate"?) is quite common, with almost 22% of abbreviations in one sample of biomedical text having more than one possible expansion and an average of 4.61 possible definitions for abbreviations six or fewer characters long (Chang et al., 2002). The molecular biology domain in particular presents very specific, and very thorny, examples of this kind of problem related to gene symbols. The general phenomenon of gene symbol polysemy refers to the fact that a single name or symbol can refer to more than one gene, both within a single species and in disparate organisms. Within a species, a single symbol may name multiple genes—Hirschman et al. (2005b) report that 3.6% of all *Drosophila* identifiers map to multiple FlyBase entries. If species is uncertain for a gene reference, the problem is greatly compounded; the Entrez Gene database contains more than 800 distinct genes that have been called P60. Sometimes a shared name across species indicates homology, but not always: e.g., *TRP1* refers to several apparently nonhomologous gene families. The situation is especially difficult when dealing with *Drosophila* literature, because many *Drosophila* symbols and names are the same as common English words—*a*, *to*, and *And* are all symbols of *Drosophila* genes (Entrez Gene IDs 43852, 43036, and 44913, respectively). Besides this ambiguity, with respect to which gene a name or symbol refers to, gene names are also subject to a type of ambiguity related to the phenomenon of metonymy, or referring to something by an entity that is related to it (Lakoff and Johnson, 1980): it is often not clear to a computer (or, indeed, to a human reader) whether a string like *p53* refers to the gene of that name, to the protein that it codes for, or to its mRNA (Hatzivassiloglou et al., 2001).

The consequence of this ubiquitous ambiguity is that many of the most basic tasks in language processing are frustratingly difficult to program. It is difficult to accurately recognize all gene names mentioned in texts and harder still to identify specifically which gene is being referenced (e.g., via a GenBank accession number). Ambiguity is thought to be an innate characteristic of all human language, and it is almost certainly impossible to legislate it out of published texts, the mostly successful efforts of the yeast community to standardize references to particular genes in their publications notwithstanding. Fully and accurately capturing the subtle and complex relationships among structures and functions described in molecular biology publications is well beyond the current state of the art. Nevertheless, as illustrated above, many successful applications are now in wide use in the biomedical research community.

BLP Methods and Resources

Two broad classes of approaches have been taken to addressing these problems. The first is rule- or pattern-based approaches. These use varying sources of background knowledge (such as dictionaries, grammars, thesauri, or ontologies) to create methods to disambiguate among different possible interpretations (or analyses) of a text. The alternative approach involves building statistical discriminators (such as support vector machines, e.g., Lee et al., [2004]), generally trained on the surrounding words in the text. Each approach has challenges: the expert labor involved in building rule- or pattern-based systems can be prohibitive, and performance can be disappointing. Statistical systems require large quantities of training data, texts that have been manually marked up with correct classifications by human annotators. Limits on the ability to generate the huge quantities of text reliably annotated with the desired information have constrained the statistical approach. A variety of hybrid approaches are possible as well. This work is surveyed in Cohen and Hunter (2004).

A critical aspect of the information extraction task (and sometimes other BLP tasks as well) is the definition of the desired output—that is, a formal representation for the to-be-extracted knowledge. An emerging consensus has developed around the idea that ontologies (such as the Gene Ontology; [Ashburner et al., 2000]) are the most effective representational mechanism, at least for noun-like concepts (as opposed to verbs and relations, which can beneficially be represented by richer, frame-like or event-related structures [Fillmore et al., 2001; Kipper and Palmer, 2000; Wattarujeekrit et al., 2004; Cohen and Hunter, unpublished data]). Bodenreider (2006) reviews the primary ontological resources in the biomedical domain, and Spasic et al. (2005) reviews the use of such resources in various BLP applications. A growing body of recent work explores the concept-based approaches to language processing extrinsically, i.e., by quantifying the contributions of concept recognition to other tasks, such as information retrieval and text classification (Cohen and Hersh, 2005; Caporaso et al., 2005). (Extrinsic evaluation contrasts with intrinsic evaluation: testing a technique against a gold standard, independently of whether or not it actually contributes significantly to the performance of a system on some larger task [Sparck-Jones and Galliers, 1996].) Even apparently dissenting viewpoints (Tsujii and Ananiadou, 2005) actually agree on most points regarding the overall utility of such approaches, differing more with respect to their theoretical underpinnings than with respect to implications for implementation.

Competitive Evaluations

As in the area of protein structure prediction, competitive evaluations have played a role in pushing the field of biomedical natural language processing forward. Four competitive evaluations have been held in recent years; Hirschman and Blaschke (2006) reviews their structure and results. Perhaps the most influential of these evaluations has been the TREC Genomics Track (Hersh and Bhupatiraju, 2003; Hersh et al., 2004, 2005). Supported by the National Institute of Standards and Technology, TREC provides a forum for evaluation of information retrieval systems. The Genomics Track, begun in 2003 and now an annual event, focuses on document retrieval and classification tasks related to molecular biology. The 2004 competition had 29 participating research groups, and the 2005 competition was the largest track in TREC, with 41 participants. The BioNLP competition (Kim et al., 2004) focused on entity identification as a prerequisite step in information extraction. BioCreative (Hirschman et al., 2005a) was a 2004 competition involving a number of facets of the information extraction task. Twenty-seven teams participated in one or more of three related tasks: location of gene mentions in text, normalization of gene mentions to Entrez Gene entries, and extraction of gene/function associations. The earliest biomedical text data mining competition was an information extraction task sponsored by the Knowledge Discovery and Data mining (KDD) Cup in which participants built systems to aid in the FlyBase curation process (Yeh et al., 2003). In the aggregate, these events have demonstrated that competitive evaluation of biomedical text data mining is practical, that it scales well to realistic problems, and that there are potential applications, especially in the area of model organism database curation, for concept-based text data mining from biomedical documents.

Biognostic Systems

In the postgenomic era, finding and integrating all of the information relevant to a particular hypothesis requires transcending many disciplinary boundaries. Different research communities may "speak different languages," even when making reference to the same underlying concepts. BLP tools for information retrieval, organization, and extraction are already helping to bridge the gaps between these communities, but even when they are effective, the sheer amount of relevant information can be overwhelming, and new computational tools for effectively assimilating all this information are needed. Recently, several somewhat speculative approaches have been suggested for changing the nature of scientists' interaction with the literature, interposing biognostic ("life-knowing") systems as mediators. For example, the HyBrow system (Racunas et al., 2004) provides graphical and textual tools for a researcher to specify a detailed molecular hypothesis, and then the system identifies statements in the literature that support or contradict aspects of the hypothesis. The program can also suggest small modifications to the stated hypothesis that would increase support and/or decrease conflict with existing knowledge. So far, its knowledge base has been restricted to a small, manually constructed data set related to yeast galactose metabolism; an example related to this system can be explored at http://www.hybrow.org. Providing a text-mining back-end to HyBrow would result in the creation of a knowledge discovery system possessing a flexibility and hypothesis space larger than any text-based knowledge discovery system to date. Although major challenges remain in devising information extraction, knowledge representation, and user interface schemes adequate to achieving the vision of a biognostic machine, the potential for richer forms of computer mediation between researchers and the primary literature is significant.

Further Reading

Three recent books provide in-depth treatment of biomedical text data mining. Hersh (2003) focuses on information retrieval. Ananiadou and McNaught (2006) covers the various BLP task types, as well as describing resources and evaluation methods. Shatkay and Craven (2007) discusses biomedical text mining from a primarily machine-learning-based perspective. A variety of review papers and tutorial materials with varying foci are available. Many of them, along with a wide variety of other publications on BLP, can be accessed through the BLIMP website (http://blimp.cs.queensu.ca/); http://www.BioNLP.org is another useful resource for biologists who are interested in language processing. "General-domain" NLP is an active area of research in computer science and linguistics, and a number of industry and government groups view it as having great strategic importance (e.g., Google, Microsoft, DARPA, and ARDA). For general-domain NLP, Jackson and Moulinier (2002) is an excellent overview of the field. Jurafsky and Martin (2000) is the standard general text, and Manning and Schütze (1999) is the standard text for statistical approaches.

Acknowledgments

The authors thank Hao Chen, Hans-Michael Müller, and Parantu Shah for usage data; Lynne Fox for information on PubMed; MITRE's Biomedical Information Processing group for data on Drosophila name ambiguity; and Preslav Nakov and Marti Hearst for adapting Figure 2.

References

Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., et al. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. *33(Database issue)*, D418–D424.

Ananiadou, S., and McNaught, J. (2006). Text Mining for Biology and Biomedicine (Norwood, MA: Artech).

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Bodenreider, O. (2006). Lexical, terminological and ontological resources for biological text mining. In Text Mining for Biology and Biomedicine, S. Ananiadou and J. McNaught, eds. (Norwood, MA: Artech), pp. 43–66.

Caporaso, J.G., Baumgartner, Jr., W.A., Johnson, H.L., Paquette, J., and Hunter, L. (2005). Concept recognition improves performance on the TREC Genomics tasks. The Fourteenth Text Retrieval Conference (TREC 2005). National Institute of Standards and Technology.

Chang, J.T., Schütze, H., and Altman, R.B. (2002). Creating an online dictionary of abbreviations from MEDLINE. J. Am. Med. Inform. Assoc. *9*, 612–620.

Chen, H., and Sharp, B.M. (2004). Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics *5*, 147.

Cohen, A., and Hersh, W. (2005). A survey of current work in biomedical text mining. Brief. Bioinform. *6*, 57–71.

Cohen, K.B., and Hunter, L. (2004). Natural language processing and systems biology. In Artificial Intelligence Methods and Tools for Systems Biology, W. Dubitzky and F. Azuaje, eds. (Norwell, MA: Springer), pp. 147–173.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., et al. (2003). PreBIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics *4*, 11.

Fillmore, C.J., Wooters, C., and Baker, C.F. (2001). Building a large lexical databank which provides deep semantics. Proceedings of the Pacific Asian Conference on Language, Information and Computation.

Hatzivassiloglou, V., Duboué, P.A., and Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics *17* (*Suppl. 1*), S97–S106.

Hersh, W.R. (2003). Information Retrieval: A Health and Biomedical Perspective (New York: Springer).

Hersh, W.R., and Bhupatiraju, R.T. (2003). TREC Genomics track overview. The Twelfth Text Retrieval Conference (TREC 2003), National Institute of Standards and Technology.

Hersh, W.R., Bhupatiraju, R.T., Ross, L., Johnson, P., Cohen, A.M., and Kraemer, D.F. (2004). TREC 2004 Genomics track overview. The Thirteenth Text Retrieval Conference (TREC 2004), National Institute of Standards and Technology.

Hersh, W.R., Cohen, A.M., Yang, J., Bhupatiraju, R.T., Roberts, P., and Hearst, M. (2005). TREC 2005 Genomics track overview. The Fourteenth Text Retrieval Conference (TREC 2005), National Institute of Standards and Technology.

Hirschman, L., and Blaschke, C. (2006). Evaluation of text mining in biology. In Text Mining for Biology and Biomedicine, S. Ananiadou and J. McNaught, eds. (Norwood, MA: Artech), pp. 213–235.

Hirschman, L., Colosimo, M., Morgan, A., and Yeh, A. (2005a). Overview of BioCreAtIvE Task 1B: normalized gene lists. BMC Bioinformatics *6* (*Suppl. 1*), S11.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005b). Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics *6* (*Suppl. 1*), S1.

Horn, F., Lau, A.L., and Cohen, F.E. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics *20*, 557–568.

Jackson, P., and Moulinier, I. (2002). Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization (Amsterdam: John Benjamins Publishing Company).

Jurafsky, D., and Martin, J.H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Upper Saddle River, NJ: Prentice Hall).

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA.

Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04).

Kipper, K., and Palmer, M.S. (2000) Representation of actions as an interlingua. Proceedings of the Third Workshop on Applied Interlinguas, ANLP-NAACL 2000.

Lakoff, G., and Johnson, M. (1980). Metaphors We Live By (Chicago, IL: University of Chicago Press).

Lee, K.J., Hwang, Y.S., Kim, S., and Rim, H.C. (2004). Biomedical named entity recognition using two-phase model based on SVMs. J. Biomed. Inform. *37*, 436–447.

Manning, C.D., and Schütze, H. (1999). Foundations of Statistical Natural Language Processing (Cambridge, MA: The MIT Press).

Müller, H.-M., Kenny, E.E., and Sternberg, P.W. (2004). Textpresso: an ontology-baesd information retrieval and extraction system for biological literature. PLoS Biol. *2*, e309 10.1371/journal.pbio.0020309.

Nakov, P., Schwartz, A., Wolf, B., and Hearst, M. (2005). Supporting annotation layers for natural language processing. Proceedings of the ACL interactive poster and demonstration sessions, Association for Computational Linguistics.

Racunas, S.A., Shah, N.H., Albert, I., and Fedoroff, N.V. (2004). HyBrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics *20* (*Suppl. 1*), i257–i264.

Shah, P., Jensen, J.J., Boué, S., and Bork, P. (2005). Extraction of transcript diversity from scientific literature. PLoS Comput. Biol. *1*, e10 10.1371/journal.pcbi.0010010.

Shatkay, H., and Craven, M. (2007). Biomedical Text Mining (Cambridge, MA: MIT Press), in press.

Sparck-Jones, K., and Galliers, J.R. (1996). Evaluating Natural Language Processing Systems: An Analysis and Review (Berlin: Springer).

Spasic, I., Ananiadou, S., McNaught, J., and Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. Brief. Bioinform. *6*, 239–251.

Swanson, D.R. (1988). Migraine and magnesium: eleven neglected connections. Perspect. Biol. Med. *31*, 526–557.

Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Michaels, G.S., Hunter, L., and Weinstein, J.N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques *27*, 1210–1217.

Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.-J., Le Texier, V., and Muilu, J. (2004). ASD: the Alternative Splicing Database. Nucleic Acids Res. *32(Database issue)*, D64–D69.

Tsujii, J., and Ananiadou, S. (2005). Thesaurus or logical ontology, which one do we need for text mining? Language Resources and Evaluation *39*, 77–90.

Wattarujeekrit, T., Shah, P.K., and Collier, N. (2004). PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics *5*, 155.

Weeber, M., Kors, J.A., and Mons, B. (2005). Online tools to support literature-based discovery in the life sciences. Brief. Bioinform. *6*, 277–286.

Yeh, A.S., Hirschman, L., and Morgan, A.A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. Bioinformatics *19* (*Suppl. 1*), i331–i339.