Literature Data Mining for Biology

Lynette Hirschman
The MITRE Corporation

Jong C. Park
KAIST

Junichi Tsujii
University of Tokyo

Cathy Wu
Georgetown University

Limsoon Wong
Kent Ridge Digital Labs

In this introduction, we summarize the papers included in the session on Literature Data Mining for Biology. We then discuss the need for a challenge evaluation for this field, and the steps to create such an evaluation. These include creating a shared infrastructure, providing annotated data, and defining and implementing common evaluation metrics. This would enable researchers to compare differing methods, in order to accelerate progress in this field. In this context, we describe two specific applications: extraction of biological pathways from the literature and automated database curation. For each of these, we outline the task definition, the creation of an annotated corpus, and evaluation metrics.

1 Session Summary

Even though the number and the size of sequence databases are growing rapidly, most new information relevant to biology research is still recorded as free text in journal articles and in comment fields of databases like the GenBank feature table annotations. As biomedical research enters the post-genome era, new kinds of databases that contain information beyond simple sequences are needed, for example, information on cellular localization, protein-protein interactions, gene regulation and the context of these interactions. The forerunners of such databases include KEGG[1], DIP[2], BIND[3], among others. Such databases are still small in size and are largely hand curated. A factor that can accelerate their growth is the development of reliable literature data mining technologies.

This year is the third time the Pacific Symposium on Biocomputing has devoted an entire session to natural language processing and information extraction for biology. Compared to the last two years, the field has made tremendous

strides. Most of the early work on automated understanding of biomedical papers concentrated on analytical tasks such as identifying protein names [4] or relied on simple techniques such as word co-occurrence[5] and pattern matching[6]. Last year, we began to see work based on more general natural language parsers that could handle considerably more complex sentences [7,8]. This year, we see the emergence of more sophisticated natural language technologies that can handle anaphora, as well as extracting a broader range of information.

Six papers were accepted under peer-review out of a total of seventeen submissions reviewed for this session. We briefly introduce them here:

- The paper by Ding *et al.* examines an issue that is fundamental to literature data mining based on term co-occurrence methods. It systematically compares the impact on recall and precision of mining interaction information when an abstract, a sentence, or a phrase is used as the unit in which to check for term co-occurrence.

- The paper by Hahn *et al.* describes the MEDSYNDIKATE natural language processor designed for acquiring knowledge from medical reports. The system is capable of analysing co-referring sentences and is also capable of extracting new concepts given a set of grammatical constructs.

- The paper by Leroy *et al.* presents the medical parser of the GeneScene system. An interesting aspect of this parser is that it uses prepositions as entry points into phrases in the text, in contrast to earlier approaches which used verbs as entry points. It then fills in a set of basic templates of patterns of prepositions around verbs and nominalized verbs. It also has a set of rules for combining these templates to extract information from more complex sentences.

- The paper by Pustejovsky *et al.* gives us a robust parser for identifying and extracting inhibition relations from biomedical literature. The system is founded on corpus-based linguistics. A particularly interesting feature of this system is its anaphora resolution module. The results reported in this paper focus on *inhibition* relations and demonstrate that it is possible to extract biologically important information from free text with high reliability using a classical approach.

- The paper by Stapley *et al.* is an interesting combination of text processing and machine learning technologies to predict the cellular location of proteins. The performance of the classifier on a benchmark of proteins with known cellular locations is better than a support vector machine trained on amino acid composition and is comparable to an expertly hand-crafted rule-based classifier. [9]

- The paper by Wilbur formalizes the idea of a "theme" in a collection of documents as a subset of the documents and a subset of the indexing terms such that each element of the latter has a high probability of occurring in all elements of the former. An algorithm is then given to produce themes and to cluster documents according to these themes in an optimal way. Results of applying this method to over fifty thousand documents on AIDS are given as an illustration.

The response to the call for papers and the quality of the submitted papers mark this as an emerging field which combines bioinformatics and natural language processing in innovative and productive ways. We find this very encouraging, but we also feel that much research and development remains to be carried out. In the rest of this session introduction, we present some of the challenges and associated benchmarks that we feel are important to the development of the field. We have also organized an additional special session on literature data mining at this Pacific Symposium on Biocomputing to specifically discuss these challenges and benchmarks.

2 Background: State of the Art for Literature Data Mining

We start from the observation that literature data mining and natural language processing techniques have been demonstrated to work—at least in the domain of newswire. Results from various evaluations show that information extraction systems can identify and classify names (person, organization, location, times, dates) at an accuracy of greater than 90%; and they can successfully extract binary relations among these entities at over 80%, e.g, "ORGANIZATION located_at LOCATION" or "PERSON works_at ORGANIZATION" [13]. In addition, other information access and retrieval techniques have proved effective in selecting documents relevant to a specific topic or to providing answers to questions based on information located in document collections: The leading systems can provide correct answers to factual queries at 75-85% accuracy [11].

Much of the progress in the newswire domain has come about because of systematic *common evaluations* conducted in natural language processing and information retrieval. The major evaluations have been the series of Message Understanding Conferences (MUC) and the Text REtrieval Conferences (TREC). The MUC conferences were held from 1987–1998 and focused on identification of specific types of entities, relations and events from free text [14,12,13]. The TREC conferences have been held annually since 1991 [10,11]. They focus on retrieval of relevant information from large (multi-gigabyte) on-line collections, given a statement of need. Over the last several years, TREC has also included

a question answering track, where systems have been evaluated based on their ability to answer simple factual questions, such as "What do river otters eat?"

3 Challenges and Benchmarks

During one of the natural language sessions at PSB 2001, the question was raised: Would the biology community benefit from having a "challenge evaluation" for literature data mining, in the way that the CASP evaluation[a] has accelerated progress in developing computational models for protein folding. The people present at that session expressed strong interest in having such an evaluation. We view this year's PSB sessions on literature data mining as a direct follow-on to that discussion.

There is enormous potential for the application of natural language processing and literature data mining techniques to biology. These techniques can be applied to the extraction of biological pathways, to the curation of biomedical databases, or to improved access to the on-line literature. By defining a set of common evaluations, we can achieve several goals:

- Creation of sizable training and test corpora. Such corpora are necessary for the construction of high performance components, such as name taggers that identify and categorize names of relevant biological entities, e.g., genes, proteins, small molecules, or cellular locations.

- Adoption of evaluation metrics. By defining a common challenge problem, different groups can tackle the same problem using different methods. This will allow an objective comparison of these methods and their performance on a single task. Right now, it is very difficult to compare the different approaches, such as those described in this session, because they are tackling different problems, using different data sets and different evaluation metrics.

- Sharing of promising techniques. When different groups work on a single problem, it becomes much easier to exchange information, to identify the unsolved problems and to understand where techniques work or don't work. This sharing of positive and negative results accelerates progress in the field.

[a] Critical Assessment of Techniques for Protein Structure Prediction; see http://predictioncenter.llnl.gov/.

4 Organizing an Evaluation

We can identify seven ingredients for a successful evaluation:

- Challenge problem: this should be a problem of biological significance, preferably one already being addressed, such as literature search to assemble biological pathways, or creation of specialized databases to organize information and make it accessible.

- Task definition: this defines the criteria for evaluation—what constitutes a "correct" answer in the context of the challenge problem. This requires a formal specification of the target output. For the biological pathway task, this might be a formal relational language consisting of a set of predicates (activate, inhibit,...), and classes of entities that can participate in that relationship (see section 5 for examples). For the database curation task, it will consist of entries conforming to some specific ontology or nomenclature as specified by the database curators (see section 6 for examples).

- Training data: to enable developers to build systems that solve the "challenge problem," developers need annotated data. Another way to look at this is that developers need "practice tests"—instances of right answers. For a biological pathway, this might be a collection of documents and the relations extracted from each of the articles. For a database curation task, it would be the database entries and the collection of documents referenced in the curation. In addition, the training data must specify the *linkage* between the extracted information and the occurrences (phrases or sentences) in the associated article that provide the evidence for the extracted information. These linkages are the *annotations* that make it possible to create rules (by hand or using machine learning or statistical techniques) that map from the free text occurrence of information to the required target output.

- Test data: once the system is built, it must be evaluated against *blind* test data—data that neither the system nor the developers have previously seen. This makes it possible to assess the generality of the solution. Note that the test data may not need to have the linkages annotated: It is sufficient to supply the input (free text) and target output.

- Evaluation methodology and implementation: the definition of a formal evaluation methodology is a key part of creating a challenge problem. There must be a reproducible method of evaluating system performance

on the defined problem. Ideally, there would be an *automated* evaluation method and supporting software. This would allow participants to grade themselves on the training data (the "practice tests"); automated evaluation also supports system development techniques such as iterative hill climbing and machine learning.

- Evaluator: there must be a neutral group who runs the evaluation. The evaluator is responsible for providing the test data, for collecting the system runs on the test data, and for evaluating those runs.

- Participants: any evaluation is only as good as the groups (and systems) that participate in it. Therefore, it is critical to identify beforehand a core set of groups who would be willing to perform such an evaluation, if the rest of the infrastructure (as listed above) were provided.

- Funding: To create a successful challenge evaluation, there must be funding for the infrastructure. The evaluation itself must be funded (in particular, the designated evaluator group), and finally, participants are more likely to participate if there is funding associated with the evaluation. For the participants, the association with funding may be indirect, e.g., it may be sufficient that there are funded programs (government or private) that might directly or indirectly reward good results in such an evaluation.

In the remaining sections of the introduction, we look at two examples of challenge problems: the extraction of biological pathways from the literature and techniques for automating database curation.

5 Extraction of Biological Pathways

To a biologist, a biological pathway is generally a chain of events and decision points that pertain to specific biological functions, such as the production of a desaturated fatty acid. In contrast to the situation with genes, where a detailed ontology for their classification and annotation has been established[17], there is no widely accepted ontology for biological pathways.

Ideally, given an established ontology for annotating biological pathways, a benchmark natural language-based extraction of biological pathways can also be established. First, a database structured according to the ontology can be adopted. Second, a set of scientific texts and abstracts can be chosen. Third, the database can be populated from these texts and abstracts by domain experts. Then a set of tests and evaluation criteria with respect to this database

can be set up as a benchmark for evaluating technologies for extracting biological pathway information.

However, no such ontology has been widely accepted yet. Moreover, the ideal ontology that most closely reflects biological pathways as a biologist sees them may not be suitable for information extraction tasks. Here, we adopt a more relaxed view instead and consider biological pathways as a network of interactions and events between proteins, drugs, and other molecules. We propose three layers of challenges with respect to this more relaxed view:

- At the first layer, the task is to recognize names of proteins, drugs, and other molecules.
- At the second layer, the task is to recognize basic interaction events between molecules.
- At the third layer, the task is to recognize the relationships between the basic interaction events.

Before we describe the three tasks above in more detail, let us first set up the framework for benchmarking these tasks. The framework is oriented towards information extraction rather than deep natural language understanding. That is, we see each task as filling in a set of prescribed templates for each problem, as opposed to obtaining detailed parse trees and complete semantic representations of each sentence. We have three primary reasons for this orientation. First, we still do not have grammars that provide sufficiently broad coverage for the language found in the biomedical literature. This is both because the language is complex and because the articles or abstracts may not always be written in (or translated into) grammatical English. Second, filling in a template is closer to the application scenario of filling in a database table. Third, information extraction may not necessarily be natural language-based, and hence the present choice allows us to assess a broader range of techniques.

The framework is as follows. A number of test databases are constructed. Each test database is organized as a set of records. Each record should have a piece of text to be tested and a list of expected facts. The text can be at the sentence level, the abstract level, or the entire article level. The list of expected facts should contain everything that a "perfect" information extractor for the task on hand can extract and nothing else. For convenience, each fact can be thought of as a short sentence in a highly standardized form such as "P_1 activate P_2". More abstractly, we see a test database db as a set $\{(t_1, F_1)$, ..., $(t_m, F_m)\}$, where t_i are the texts and $F_i = \{f_{i,1}$, ..., $f_{i,n_i}\}$ are expected facts. There are two primary levels of evaluation. The first is at the level of individual records. The second is at the level of the entire test database.

The traditional performance evaluation of information retrieval systems calls for the following. At either level, we evaluate the sensitivity (or recall) and specificity (or precision) of an information extractor E against the list of expected facts, where

$$\mathit{recall}(E) = \frac{TP(E)}{TP(E) + FN(E)} \qquad \mathit{precision}(E) = \frac{TP(E)}{TP(E) + FP(E)}$$

The definitions for $TP(E)$ (ie. true positives), $FN(E)$ (ie. false negatives), and $FP(E)$ (ie. false positives) depend on whether we are evaluating at the record level or at the database level. Note that it is not possible to define the usual notion of true negatives in our context because there is no theoretical bound on the number of "facts" that can be generated from a sentence and because it is not reasonable to use the closed world assumption in biology. At the record level, each expected fact in a separate record is counted as a separate instance. If $E(t)$ is the set of facts that E extracts from a text t, then

$$TP(E) = \sum_{(t,F)\in db} |E(t) \cap F| \qquad FN(E) = \sum_{(t,F)\in db} |F| - TP(E)$$

$$FP(E) = \sum_{(t,F)\in db} |E(t)| - TP(E)$$

At the database level, all different instances of an expected fact are counted as one. Then we have instead

$$TP(E) = \left| \bigcup_{(t,F)\in db} E(t) \cap F \right| \qquad FN(E) = \left| \bigcup_{(t,F)\in db} F \right| - TP(E)$$

$$FP(E) = \left| \bigcup_{(t,F)\in db} E(t) \right| - TP(E)$$

However, it is not straightforward to compare two information extractors each characterized by a pair of numbers. The usual mechanism in diagnostic systems is to generate a range of precision numbers over a range of recall numbers to derive a single number called the area under the relative operating characteristic curve (the aROC number) [18] and compare the aROC numbers of two systems. Unfortunately, it is not always possible to obtain aROC numbers of the information extractors we are considering because they are typically not based on a continuous decision threshold. In order to choose an alternative, two conditions should be imposed [19]. The first condition is that it must be able

to distinguish the ideal information extractor from the worst information extractor. The second condition is that it shows a gradual and strictly monotonic change in value when the information extractor is changed from the worst to the best one. Note that neither recall nor precision alone satisfies these two conditions.

Many choices that satisfy these two conditions are possible[19,20]. However, many of them depend on the definition for "true negatives", which is not available in our context. So we propose a variation of the simple matching coefficient (SMC) which simply measures the probability of the information extractor correctly extracting a fact.[b] It is defined as follows and is easily verified to satisfy the two conditions above:

$$SMC(E) = \frac{TP(E)}{TP(E) + FP(E) + FN(E)},\ 0 \leq SMC(E)\ \leq 1$$

Now we return to our three information extraction tasks. The first task is obvious. We want to recognize proper names of proteins, drugs, and other molecules mentioned in texts. We do not want to recognize names of authors, processes, and any other entities mentioned in these texts.

The second task is slightly more complicated. We want to recognize interaction events between proteins, drugs, or other molecules. These events should include events at transcription, translation, post translational modification, complexing, dissociation, and other interactions. As we are viewing each fact as a highly standardized short sentence, we propose here a grammar for them.

[b] A related metric has been proposed in the spoken language processing community for both transcribing audio input and for identifying entities and relations among entities. This is the *slot error rate*, which is the ratio of insertions, deletions and substitution errors divided by the true positives. In our context, we can interpret insertions as false positives and deletions as false negatives; substitutions are not directly relevant [21]. Another related measure is the F-measure, defined as the harmonic mean of recall and precision. That is, $F(E) = (2 \times recall(E) \times precision(E))/(recall(E) + precision(E))$. After substituting the definitions for recall and precision, this reduces to $F(E) = (2 \times TP(E))/(2 \times TP(E) + FN(E) + FP(E))$. There is no intuitive statistical reason for having the multiplicative factor of 2 on $TP(E)$. However, if we drop this multiplicative factor, the result is precisely $SMC(E)$.

PositiveEvent	::=	P phosphorylate P [on T] [at L]
	|	P dephosphorylate P [on T] [at L]
	|	P ubiquinate P
	|	P acetylate P
	|	...
	|	P interact-with P [to-produce P]
	|	P [at L] bind-to P [at L] [to-produce P]
	|	P dissociate [to-produce P+]
	|	P degrade P
	|	P activate-transcription P [to-produce P]
	|	P inhibit-transcription P
	|	P activate [F activity-of] P
	|	P inhibit [F activity-of] P
	|	P transport P [from C] [to C]
Event	::=	PositiveEvent [mediated-by P+] [independent-of P+]
	|	not PositiveEvent [mediated-by P+] [independent-of P+]

In the grammar above, P denotes proteins, drugs, or other molecules; T denotes amino acids; L denotes positions; F denotes biological function; and C denotes cellular locations. In evaluating an information extractor for this task, we can further consider its performance with or without extracting the optional components in the grammar. In the few clauses where a plurality of P's are expected (ie. the $P+$'s), we can consider the situation of a complete or an incomplete match.

It is important to understand that this grammar is not intended as a grammar for parsing scientific texts. Rather, it is more appropriately treated as a standardized grammar to convert pertinent parts of scientific texts into. As such, an information extractor should convert or normalize different expressions of the same fact into the semantically closest standard form in the grammar. It should not make fine distinctions between different sentence forms. For example, it should convert "camptothecin, an inhibitor of human TOP1" to "camptothecin inhibit TOP1". It should also not make fine distinctions between shades of meanings. For example, "caspase-8 was also stimulated by NB-506" to "NB-506 activate caspase-8".

The third task is to recognize relationships between the basic events already outlined above. In contrast to the basic events which focus on interactions between molecules, this task is focused on the causal and temporal relationships between two such events. The grammar we propose for them is given below.

Relationship	::=	Event [is-caused-by Event+] [provided Event+]
	|	Event [is-independent-of Event+] [provided Event+]
	|	Event [is-inhibited-by Event+] [provided Event+]

The intention of a relationship such as "E_1 is-caused-by E_2 provided E_3" is as follows. The event E_3 is assumed to have taken place some time in the past and its resultant conditions have remained true. This allows event E_2 to take place and as a result the event E_1 will also take place at the completion of E_2. Again, an information extractor should convert or normalize different expressions of the same event relationships into the semantically closest standard form in the grammar.

Having now described the three tasks, we would also like to propose some candidates for forming the benchmark databases for these tasks. We would like to suggest that the appendix of the paper by Kohn [22] as one of the candidates. This appendix lists about 200 statements of interaction events and has sentences of a fairly complex form. Another candidate is the set of MEDLINE abstracts matching the term "Topoisomerase inhibitors." Presently this set includes just over 200 abstracts. A preliminary analysis shows that they contain less than 1000 names and less than 200 interaction events. These numbers are small enough for a small team of experts to construct a benchmark database manually.

6 Automated Database Curation

Automated database curation represents a second challenge application. It is important because the rate of published experiments is outstripping the ability of database curators to keep up with the relevant literature. In addition, automated curation techniques could allow curators of databases to check the consistency and completeness of their databases.

Database curation is interesting for another reason: curated databases represent a repository of "gold standard" data. A database entry is typically associated with the literature reference(s) from which it is derived – this means that the human curator has already done the extraction from the literature. Craven and Kumlien[23] report an experiment in which they were able to use the subcellular localization field of the Yeast Protein Database [16]. They collected instances of this relation from the database, traced the references associated with each database entry back to the PubMed abstract, and then within each abstract, identified, where possible, the sentence within the abstract that gave rise to the annotation. This gave them a set of extracted relations (from the database) and the underlying text sources (sentences from the abstracts).

They were then able to train and compare several classifiers that extracted the desired localization information.

This experiment is suggestive of the ways in which curated databases can be exploited to create "cheap" annotated corpora. It is relatively straightforward to associate the entry in a database field with the underlying article from which it is derived. The harder part is to provide an explicit linkage from the database entry to the phrases and sentences from which it is derived. When the database uses a controlled vocabulary or an ontology to define legal entries for each field, the phrases appearing in the journal article or abstract may not correspond to the actual entry in the database.

In the examples below, we see some of the possible relations between the mention in the literature and its representation in the fields of the database. The example is from the Flybase gene expression database;[c] the first list shows three fields from the polypeptide Appl+P130kD entry. Note that for each of these entries, the first field is the article from which this information was derived[24] (in Flybase, this is a hotlink to the abstract).

```
1   Protein size (kD):      Luo et al, 1990   130
2.  Cell location:          Luo et al, 1990   axon
3.  Expression pattern
    Luo et al, 1990
        Stage         Tissue/Position
        Embryo        Embryonic Central Nervous System
        Embryo        Peripheral Nervous System
```

The next list contains sentences from the abstract of the Luo article[24] cited above; the phrases in boldface show the specific source of information within each sentence.

1. APPL is synthesized as a 145-kDa membrane-associated precursor that is converted to a **130-kDa** secreted for ...
2. APPL proteins are first detected in developing neurons concomitant with **axonogenesis** ...
3. In the **embryo**, APPL proteins are expressed exclusively in the **CNS and PNS** neurons ...

Even in this very small sample, we see that simple pattern matching suffices in example 1 to find *130-kDa*, complex morphology is needed in example

[c] Available at http://flybase.bio.indiana.edu.

2 to associate *axonogenesis* with "cell location: axon", and we must decode abbreviations (*CNS* = central nervous system, *PNS* = peripheral nervous system) as well as using information derived from multiple parts of the sentence in example 3. A larger sample would contain many more complex mappings between database fields and the underlying literature reference, including entries that require resolution of coreference across sentences or entries that require an analysis of the underlying syntactic relations among entities.

To create an annotated corpus from a curated database, we need to map from entries in database fields back to the text. To do automated database curation, we need the inverse mapping from free text to database entry. We believe that we can create a reversible set of tools that can be used in either direction: mapping from database field to literature or mapping from literature to database. By providing collections of linked pairs of database entry and associated text for use as training and evaluation sets, we would enable many researchers to participate in building tools to automate the database curation process. Although the database curation application is different in its structure from the biological pathway detection experiment outlined in section 5, it is amenable to the same kinds of automated evaluation techniques outlined there.

7 Conclusion

We have outlined how we might go about creating a challenge evaluation for literature data mining in biology. The papers in this session illustrate both the promise of literature data mining and the need for challenge evaluations. They show how current language processing approaches can be successfully used to extract and organize information from the literature. They also illustrate the diversity of applications and evaluation metrics. By defining several biologically important challenge problems and by providing the associated infrastructure (annotated data and a common evaluation framework), we can accelerate progress in this field. This will allow us to compare approaches, to scale up the technology to tackle important problems, and to learn what works and what areas still need work.

References

1. H. Ogata et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 27(1):29–34, January 1999.
2. I. Xenarios, D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, and D. Eisenberg. DIP: The database of interacting proteins. *Nucleic Acid Res*, 28(1):289–291, January 2000.

3. G.D. Bader, I. Donaldson, C. Wolting, B.F. Ouellette, T. Pawson, and C.W. Hogue. BIND–the biomolecular interaction network database. *Nucleic Acids Res*, 29(1):242–245, January 2001.
4. K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Proceedings of Pacific Symposium on Biocomputing'98*, pages 707–718, Maui, Hawaii, January 1998.
5. B.J. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. In *Proceedings of Pacific Symposium on Biocomputing*, pages 529–540, 2000.
6. S.-K. Ng and M. Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*, 10:104–112, December 1999.
7. J.C. Park, H.S. Kim, and J.J. Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *Proceedings of Pacific Symposium on Biocomputing*, pages 396–407, 2001.
8. A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomedical papers using a full parser. In *Proceedings of Pacific Symposium on Biocomputing*, pages 408–419, 2001.
9. F. Eisenhaber and P. Bork. Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, 15(528–535), 1999.
10. E. M. Voorhees and D. K. Harman, eds. The Eighth Text REtrieval Conference (TREC-8). *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, NIST (National Institute of Standards and Technology) Special Publication 500-246, 2000.
11. E. M. Voorhees and D. K. Harman, eds. The Ninth Text REtrieval Conference (TREC-9). *Proceedings of the Eighth Text REtrieval Conference (TREC-9)*, NIST (National Institute of Standards and Technology) Special Publication 500-XXX, 2001. Available at http://trec.nist.gov/pubs/trec9/t9_proceedings.html
12. DARPA (Defense Advanced Research Projects Agency). Sixth Message Understanding Conference (MUC-6). *Proceedings of MUC-6*, Columbia, Maryland, Morgan Kaufmann, 1995.
13. DARPA (Defense Advanced Research Projects Agency). Seventh Message Understanding Conference (MUC-7), available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/, 1998.
14. L. Hirschman. The Evolution of Evaluation: Lessons from the Message

Understanding Conferences. *Computer Speech and Language*, 12(281–305), 1998.
15. E. M. Voorhees and D. M.Tice. The TREC-8 question answering track evaluation. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, NIST (National Institute of Standards and Technology) Special Publication 500-246, pages 83–105, 2000.
16. P. E. Hodges, W. E. Payne and J. I. Garrels. Yeast Protein Database (YPD): A database for the complete proteome of the saccharomyces cerevisiae. *Nucleic Acids Research* 26:68–72, 1998.
17. M. Ashburner et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.*, 25(1):25–29, 2000.
18. J. A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, June 1988.
19. V. B. Bajic. Comparing the success of different prediction software in sequence analysis: A review. *Briefings in Bioinformatics*, 1(3):214–228, 2000.
20. M.R. Anderberg. *Cluster Analysis for Applications.* Academic Press, New York, 1973.
21. J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance Measures for Information Extraction. *Proc. of the DARPA Broadcast News Workshop*, 249–254, Herndon, VA, Morgan Kaufmann, 1999.
22. K. W. Kohn. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10:2703–2734, August 1999.
23. M. Craven and J. Kumlien. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. *Proc. of the 7th International Conference on Intelligent Systems in Molecular Biology (ISMB-99)*, 1999.
24. L.Q. Luo, L. Martin-Morris, and K. White. Identification, secretion, and neural expression of APPL, a Drosophila protein similar to human amyloid protein precursor. *J. Neuroscience* 10(12):3849–3861, 1990.