From: AAAI Technical Report WS-99-11. Compilation copyright © 1999, AAAI (www.aaai.org). All rights reserved.

Learning to Extract Relations from MEDLINE

Mark Craven
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, Pennsylvania, 15213-3891, U.S.A.
mark.craven@cs.cmu.edu

Abstract

Information in text form remains a greatly underutilized resource in biomedical applications. We have begun a research effort aimed at learning routines for automatically mapping information from biomedical text sources, such as MEDLINE, into structured representations, such as knowledge bases. We describe our application, two learning methods that we have applied to this task, and our initial experiments in learning such information-extraction routines. We also present an approach to decreasing the cost of learning information-extraction routines by learning from "weakly" labeled training data.

Introduction

The MEDLINE database is a rich source of information for the biomedical sciences, providing bibliographic information and abstracts for more than nine million articles. A fundamental limitation of MEDLINE and similar sources, however, is that the information they contain is not represented in structured format, but instead in natural language text. The goal of our research is to develop methods that can inexpensively and accurately map information in scientific text sources, such as MEDLINE, into a structured representation, such as a knowledge base or a database. Toward this end, we are investigating methods for automatically extracting key facts from scientific texts.

There are three aspects of our work that we think will be of particular interest to the *machine learning for information extraction* community. First, our application domain is novel and challenging. Second, we investigate an approach to decreasing the cost of learning information-extraction routines by learning from "weakly" labeled training data. Weakly labeled instances consist not of precisely marked up documents, but instead consist of facts to be extracted along with documents that may assert the facts. Third, we employ a learning method we have recently developed that incorporates statistical predicate invention into a relational learner.

The system we are developing is motivated by several different types of tasks that we believe would greatly benefit from the ability to extracted structured information from text:

- **Database construction and updating.** Our system could be used to help construct and update databases and knowledge bases by extracting fields from text. We are currently working with a team that is developing a knowledge base of protein localization patterns (Boland, Markey, & Murphy 1996). We are using our system to assist in developing an ontology of localization patterns and to populate the database with text-extracted facts describing the localization of individual proteins.
- **Summarization.** Another promising application of our system is to provide structured summaries of what is known about particular biological objects. For example, we are working with scientists who are studying the genetic basis of diseases by identifying gene products that are found in different concentrations in tissues in various states (e.g. healthy vs. diseased). Frequently, these scientists do time-consuming MEDLINE searches to determine if some candidate gene product is likely to be related to the disease of interest. When performing these searches, the scientists typically are trying to answer such questions as: In what types of tissues, cells and subcellular locations is the protein known to be expressed? Is the protein known to be associated with any diseases? Is the protein known to interact with any pharmacological agents? We plan to partially automate the task of extracting answers to these questions from text.
- **Discovery.** An especially compelling application of our system is its potential application to scientific discovery. The articles in MEDLINE describe a vast web of relationships among the genes, proteins, pathways, tissues and diseases of various systems and organisms of interest. Moreover, each article describes only a small piece of this web. The work of Swanson and Smalheiser (1997) has demonstrated that significant but previously unknown relationships among entities (e.g., magnesium and migraine headaches) can be discovered by automatically eliciting this information from the literature. Swanson's algorithm detects relationships among objects simply by considering the statistics of word co-occurrences in article titles. We conjecture that such relationships can be detected more accurately

by our method of analyzing sentences in the article's abstract or text. Moreover, whereas Swanson's algorithm posits only that *some* relation holds between a pair of objects, our system is designed to state what the specific relation is.

The Information Extraction Task

In the applications we are addressing, we are primarily interested in extracting instances of *relations* among objects. In particular, we want to learn extractors for the following:[1]

- subcellular-localization(Protein, Subcellular-Location): the instances of this relation represent proteins and the subcellular structures in which they are found.
- cell-localization(Protein, Cell-Type): the cell types in which a given protein is found.
- tissue-localization(Protein, Tissue): the tissue types in which a given protein is found.
- associated-diseases(Protein, Disease): the diseases with which a given protein is known to have some association.
- drug-interactions(Protein, Pharmacologic-Agent): the pharmacologic agents with which a given protein is known to interact.

In our initial experiments we are focusing on the subcellular-localization relation. As an example of the IE task, Figure 1 shows several sentences and the instances of the subcellular-localization relation that we would like to extract from them.

Extraction via Text Classification

Our first approach to learning information extractors uses a statistical text classification method. Without loss of generality, assume that we are addressing the task of extracting instances of a binary relation, r(X, Y). This approach assumes that for the variables of the relation, X and Y, we are given semantic lexicons, L(X) and L(Y), of the possible words that could be used in instances of r. For example, the second constant of each instance of the relation subcellular-localization, described in the previous section, is in the semantic class Subcellular-Structure. Our semantic lexicon for this class consists of words like *nucleus, mitochondrion*[2], etc. Given such lexicons, the first step in this approach is to identify the *instances* in a document that could possibly express the relation. In the work reported here, we make the assumption that these instances consist of individual sentences. Thus, we can frame the information-extraction task as one of sentence classification. We extract a relation instance r(x, y) from the sentence if (i) the sentence contains a word x ∈ L(X) and a word y ∈ L(Y), and (ii) the sentence is classified as a positive instance by a statistical model. Otherwise, we consider the sentence to be a negative instance and we do not extract anything from it. We can learn the statistical model used for classification from labeled positive and negative instances (i.e. sentences that describe instances of the relation, and sentences that do not).

As stated above, we make the assumption that instances consist of individual sentences. It would be possible, however, to define instances to be larger chunks of text (e.g. paragraphs) or smaller ones (e.g. sentence clauses) instead. One limitation of this approach is that it forces us to assign a single class label to each instance. This limitation provides an argument for setting up the task so that instances are relatively small.

In order to learn models for classifying sentences, we use a statistical text-classification method. Specifically, we use a Naive Bayes classifier with a *bag-of-words* representation. This approach involves representing each document (i.e. sentence) as a bag of words. Given a document, d of n words $(w_1, w_2, \ldots, w_n)$, Naive Bayes estimates the probability that the document belongs to each possible class $c_j \in C$ as follows:

$$\Pr(c_j|d) = \frac{\Pr(c_j)\Pr(d|c_j)}{\Pr(d)} \approx \frac{\Pr(c_j)\prod_{i=1}^{n}\Pr(w_i|c_j)}{\Pr(d)}.$$

The prior probability of the document, $\Pr(d)$ does not need to be estimated directly. Instead we can get the denominator by normalizing over all of the classes. The conditional probability, $\Pr(w_i|c_j)$, of seeing word w_i given class c_j is estimated from the training data using a Laplace estimate.

To evaluate our approach, we assembled a corpus of abstracts from the MEDLINE database. This corpus, consisting of 2,889 abstracts, was collected by querying on the names of six proteins and then downloading the first 500 articles returned for each query protein, discarding entries that did not include an abstract. We selected the six proteins for their diversity and for their relevance to the research of one of our collaborators. We created a labeled data set for our IE experiments as follows. A collaborator (Johan Kumlien), who is trained in medicine and clinical chemistry, hand-annotated each abstract in the corpus with instances of the target relation subcellular-localization. To determine if an abstract should be annotated with a given instance, subcellular-localization(x, y), the abstract had to clearly indicate that protein x is found in location y. This labeling process resulted in a total of thirty-three instances of the subcellular-localization relation. Individual instances were found in from one to thirty different abstracts. For example, the fact that calcium channels are found in the sarcoplasmic reticulum was indicated in eight different abstracts.

The goal of the information-extraction task is to correctly identify the instances of the target relation

[1] We use the following notation to describe relations: constants, such as the names of specific relations and the objects they characterize, start with lowercase letters; the names of variables begin with uppercase letters.

[2] Our lexicons also include adjectives and the plural forms of nouns.

Immunoprecipitation of biotinylated type XIII collagen from surface-labeled HT-1080 cells, subcellular fractionation, and immunofluorescence staining were used to demonstrate that type XIII collagen molecules are indeed located in the plasma membranes of these cells.	subcellular-localization(collagen, plasma-membranes)
HSP47 is a collagen-binding stress protein and is thought to be a collagen-specific molecular chaperone, which plays a pivotal role during the biosynthesis and secretion of collagen molecules in the endoplasmic reticulum.	subcellular-localization(collagen, endoplasmic-reticulum)

Figure 1: An illustration of the IE task. On the left are sentences from MEDLINE abstracts. On the right are instances of the subcellular-localization relation that we might extract from these sentences.

that are represented in the corpus, without predicting spurious instances. Furthermore, although each instance of the target relation, such as subcellular-localization(calcium-channels, sarcoplasmic-reticulum), may be represented multiple times in the corpus, we consider the information-extraction method to be correct as long it extracts this instance from *one* of its occurrences. We estimate the accuracy of our learned sentence classifiers using leave-one-out cross validation. Thus, for every instance in the data set, we induce a classifier using the other instances as training data, and then treat the held-out instance as a test case. We compare our learned information extractors against a baseline method that we refer to as the *sentence co-occurrence* predictor. This method predicts that a relation holds if a protein and a sub-cellular location occur in the same sentence.

We consider using our learned Naive Bayes models in two ways. In the first method, we use them as classifiers: given an instance, the model either classifies it as positive and returns an extracted relation instance, or the model classifies it as negative and extracts nothing. In the second method, the model returns its estimated posterior probability that the instance is positive. With this method, we do not strictly accept or reject sentences.

For each method, we rank its predictions by a confidence measure. For a given relation instance, r(x, y), we first collect the set of sentences that would assert this relation if classified into the positive class (i.e. those sentences that contain both the term x and the term y). For the sentence co-occurrence predictor, we rank a predicted relation instance by the size of this set. When we use the Naive Bayes models as classifiers, we rank a predicted relation instance by the number of sentences in this set that are classified as belonging to the positive class. When we use the probabilities produced by Naive Bayes, we estimate the posterior probability that each sentence is in the positive class and combine the class probabilities using the *noisy or* function (Pearl 1988):

$$\text{confidence} = 1 - \prod_{k}^{N} [1 - \Pr(c = \text{pos } | s_k)] .$$

Here, $\Pr(c = \text{pos } | s_k)$ is the probability estimated by Naive Bayes for the kth element of our set of sentences. This combination function assumes that each sentence in the set provides independent evidence for the truth of the asserted relation.

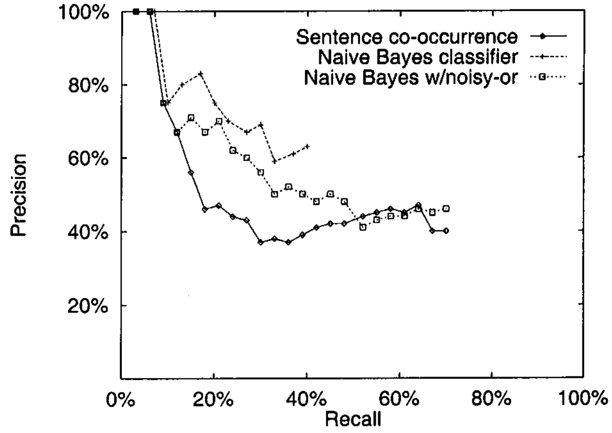

Figure 2: Precision vs. recall for the co-occurrence predictors and the Naive Bayes model.

Figure 2 plots *precision* versus *recall* for the three methods on the task of extracting instances of the subcellular-localization relation. The figure illustrates several interesting results. The most significant result is that both versions of the Naive Bayes predictor generally achieve higher levels of precision than the sentence co-occurrence predictor. This result indicates that the learning algorithm has captured in its models some of the statistical regularities that arise in how authors describe the subcellular localization of a protein. None of the methods is able to achieve 100% recall since some positive relation instances are not represented by individual sentences. In the limit, the recall of the Naive Bayes classifiers is not as high as it is for the baseline predictor because the former incorrectly classifies as negative some sentences representing positive instances. Since the Naive Bayes models with noisy-or do not reject any sentences in this way, their recall is the same as the baseline method. Their precision is lower than the Naive Bayes classifier, however, indicating that even when Naive Bayes makes accurate classifications, it often does not estimate probabilities well. An interesting possibility would be to combine these predictors to get the high precision of the Naive Bayes classifiers along with the high recall of the Naive Bayes moles using noisy-or. Provost and Fawcett (1998) have developed a method especially well suited to this type of combination.

Exploiting Existing Databases for Training Data

Although machine learning offers a promising alternative to hand-coding information extraction routines, providing labeled training data to the learner is still quite time-consuming and tedious. In fact, labeling the corpus used in the previous section required approximately 35 hours of an expert's time. In this section, we present an approach to learning information extractors that relies on existing databases to provide something akin to labeled training instances.

Our approach is motivated by the observation that, for many IE tasks, there are existing information sources (knowledge bases, databases, or even simple lists or tables) that can be coupled with documents to provide what we term "weakly" labeled training examples. We call this form of training data weakly labeled because each instance consists not of a precisely marked document, but instead it consists of a fact to be extracted along with a document that *may* assert the fact. To make this concept more concrete, consider the Yeast Protein Database (YPD) (Hodges, Payne, & Garrels 1998), which includes a *subcellular localization* field for many proteins. Moreover, in some cases the entry for this field has a reference (and a hyperlink to the MEDLINE entry for the reference) to the article that established the subcellular localization fact. Thus, each of these entries along with its reference could be used as a weakly labeled instance for learning our subcellular-localization information extractors.

In this section we evaluate the utility of learning from weakly labeled training instances. From the YPD Web site, we collected 1,213 instances of the subcellular-localization relation that are asserted in the YPD database, and from MEDLINE we collected the abstracts from 924 articles that are pointed to by these entries in YPD. For many of the relation instances, the associated abstracts do not say anything about the subcellular localization of the reference protein, and thus they are not helpful to us. However, if we select the relation instances for which an associated abstract contains a sentence mentioning both the protein and a subcellular location, then we get 336 relation instances described in 633 sentences.

As in the previous section, we treat individual sentences as instances to be processed by a Naive Bayes text classifier. Moreover, we make the assumption that every one of the 633 sentences mentioned above represents a positive training example for our text classifier. In other words, we assume that if we know that relation subcellular-localization(x, y) holds, then any sentence in the abstract(s) associated with subcellular-localization(x, y) that references both x and y is effectively stating that x is located in y. Of course this assumption is not always valid in practice, and we are currently investigating approaches that allow us to relax it. We take the remaining sentences in the YPD corpus as negative training examples.

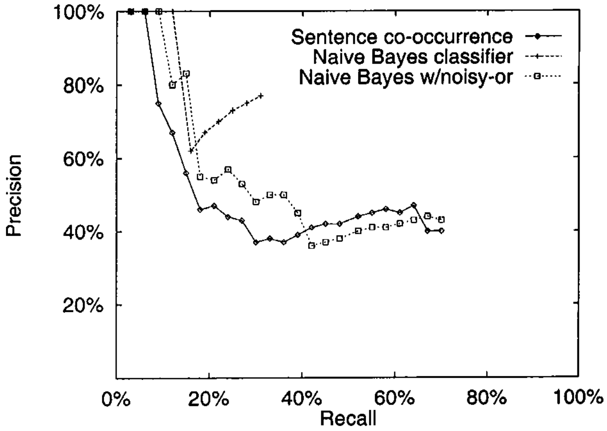

Figure 3: Precision vs. recall for the Naive Bayes model trained on the YPD data set.

The hypothesis that we consider in this section is that it is possible to learn accurate information-extraction routines using weakly labeled training data, such as that we gathered from YPD. To test this hypothesis we train a Naive Bayes model using the YPD data as a training set, and then we evaluate it using our hand-labeled corpus as a test set. We train our statistical text classifier in the same manner as described in the previous section.

Figure 3 shows the precision vs. recall curves for the YPD-trained model and for the baseline sentence co-occurrence predictor described in the previous section. From this figure we can see that the curve for the Naive Bayes model learned from the YPD data is comparable to the curve for the models learned from the hand-labeled data. Whereas the Naive Bayes classifiers from the previous section achieved 69% precision at 30% recall, the Naive Bayes classifier trained on the YPD data reaches 77% precision at 30% recall. Moreover, the YPD model achieves better precision at comparable levels of recall than the sentence co-occurrence classifier.

These two results support our hypothesis. It should be emphasized that the result of this experiment was not a foregone conclusion. Although the YPD data set contains many more positive instances than our hand-labeled data set, it represents a very different distribution than our test set. The YPD data set has a particular focus on the localization of yeast proteins. The test set, in contrast does not concentrate on protein localization and barely mentions yeast. We argue that the result of this experiment is very significant result because it suggests that effective information-extraction routines can be learned without an expensive hand-coding or hand-labeling process.

Extraction via Relational Learning

The primary limitation of the statistical classification approach to IE presented in the preceding sections is that it does not represent the linguistic structure of the text being analyzed. In this section, we describe an approach that involves parsing sentences, and learning relational rules in terms of these parses. Our approach uses a sentence analyzer called Sundance (Riloff 1998) that assigns part-of-speech tags to words, and then builds a shallow parse tree that segments sentences into clauses and noun, verb, or prepositional phrases. Given these parses, we learn IE rules using a relational learning algorithm that is similar to FOIL (Quinlan 1990).

The objective of the learning algorithm is to learn a definition for the predicate: localization-sentence(Sentence-ID, Phrase-ID, Phrase-ID). Each instance of this relation consists of (i) an identifier corresponding to the sentence represented by the instance, (ii) an identifier representing the phrase in the sentence that contains an entry in the protein lexicon, and (iii) and identifier representing the phrase in the sentence that contains an entry in the subcellular location lexicon. Thus, the learning task is to recognize pairs of phrases that correspond to positive instances of the target relation. The models learned by the relational learner consist of logical rules constructed from the following *background relations*:

- phrase-type(Phrase-ID, Phrase-Type): This relation allows a particular phrase to be characterized as a noun phrase, verb phrase, or prepositional phrase.
- next-phrase(Phrase-ID, Phrase-ID): This relation specifies the order of phrases in a sentence. Each instance of the relation indicates the successor of one particular phrase.
- constituent-phrase(Phrase-ID, Phrase-ID): This relation indicates cases in which one phrase is a constituent of another phrase (e.g., prepositional phrases usually have constituent noun phrases).
- subject-verb(Phrase-ID, Phrase-ID), verb-direct-object(Phrase-ID,Phrase-ID): These relations enable the learner to link subject noun phrases to their corresponding verb phrases, and verb phrases to their direct object phrases.
- same-clause(Phrase-ID, Phrase-ID): This relation links phrases that occur in the same sentence clause.

This set of background relations enables the learner to characterize the relations among phrases in sentences. Our learner also has the ability to describe the words occurring in sentences and phrases by using a statistical predicate-invention method (Slattery & Craven 1998). This predicate-invention method devises Naive Bayes models on the fly to characterize phrases or sentences as a whole, and the learner considers using these Naive Bayes models as Boolean predicates in rules.

Using a procedure similar to *relational pathfinding* (Richards & Mooney 1992), our learning algorithm initializes each rule by trying to find the combination of next-phrase, constituent-phrase, subject-verb, verb-direct-object, and same-clause literals that link the phrases of the greatest number of uncovered positive instances. After the rule is initialized with these literals, the learning algorithm uses a hill-climbing search to add additional literals.

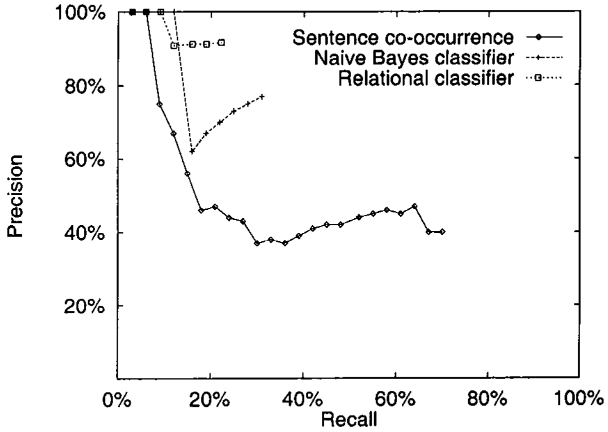

Figure 4: Precision vs. recall for the relational classifier trained on the YPD data set.

To evaluate our relational IE approach, we learned a set of rules using the YPD data set as a training set, and tested the rules on the hand-labeled data set. Our relational method learned a total of 26 rules covering the positive instances in the training set.

Figure 4 shows the precision vs. recall curve for the learned relational rules. The confidence measure for a given example is the estimated accuracy of the first rule that the example satisfies. For comparison, Figure 4 also shows the precision vs. recall curves for the YPD-trained Naive Bayes classifier discussed in the previous section, and for the sentence co-occurrence baseline. As this figure illustrates, although the recall of the relational rule set is rather low (21%), the precision is quite high (92%). In fact, this precision value is considerably higher than the precision of the Naive Bayes classifier at the corresponding level of recall. This result indicates the value of representing grammatical structure when learning information extractors. We believe that the recall level of our relational learner can be improved by tuning the set of background relations it employs, and we are investigating this issue in our current research.

Discussion and Conclusions

One may ask whether the learned classifiers we described in this paper are accurate enough to be of use. We argue that, for many tasks, they are. As discussed in Section , two of the motivating applications for our work are (i) providing structured summaries of particular biological objects, and (ii) supporting discovery by eliciting connections among biological objects. As demonstrated by the work of Swanson *et al.* (1997), even word co-occurrence predictors can be quite useful for these tasks. Therefore, any method that can provide a boost in predictive power over these baselines is of practical value. For tasks such as automatic genome annotation, where the predictions made by the information extractors would be put directly into a database, the standard for accuracy is higher. For this type of task, we believe that extraction routines like those described in this paper can be of value either by (i) making only high-confidence predictions, thereby sacrificing recall for precision, or (ii) operating in a semi-automated mode in which a person reviews (some) of the predictions made by the information extractors.

Perhaps the most significant contribution of our work is the approach to using "weakly" labeled training data. Most previous work in learning information extractors has relied on training examples consisting of documents precisely marked with the facts that should be extracted along with their locations within the document. Our approach involves (i) identifying existing databases that contain instances of the target relation, (ii) associating these instances with documents so that they may be used as training data, (iii) dividing the documents into training instances and weakly labeling these instances (e.g. by assuming that all sentences that mention a protein and a subcellular location represent instances of the subcellular-localization relation). Currently, we are investigating objective functions for learning that take into account the nature of the weakly labeled instances. We believe that this approach has great promise because it vastly reduces the time and effort involved in assembling training sets.

Several other research groups have addressed the task of information extraction from biomedical texts. Our research differs considerably, however, in the type of knowledge we are trying to extract and in our approach to the problem. A number of groups have developed systems for extracting *keywords* from text sources. Andrade and Valencia (1997) describe a method for extracting keywords characterizing functional characteristics of protein families. In similar work, Ohta *et al.* (1997) extract keywords using an information-theoretic measure to identify those words that carry the most information about a given document. Weeber and Vos (1998) have developed a system for extracting information about adverse drug reactions from medical abstracts. Their system isolates words that occur near the phrase "side effect" and then uses statistical techniques to identify words that possibly describe adverse drug reactions. Fukuda *et al.* (1998) consider the task of recognizing protein names in biological articles. Their system uses both orthographic and part-of-speech features to recognize and extract protein names. The prior research most similar to ours is that of Leek (1997), who investigated using hidden Markov models (HMMs) to extract facts from text fields a biomedical database. The task addressed by Leek, like our task, involved extracting instances of a binary relation pertaining to location. His location relation, however, referred to the positions of genes on chromosomes. The principal difference between Leek's approach and our approach is that his HMMs involved a fair amount of domain-specific human engineering.

In summary, we believe that the work presented herein represents a significant step toward making textual sources of biological knowledge as accessible and interoperable as structured databases.

References

Andrade, M. A., and Valencia, A. 1997. Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. In *Proc. of the 5th International Conf. on Intelligent Systems for Molecular Biology*, 25–32. Halkidiki, Greece: AAAI Press.

Boland, M. V.; Markey, M. K.; and Murphy, R. F. 1996. Automated classification of protein localization patterns. *Molecular Biology of the Cell* 8(346a).

Fukuda, K.; Tsunoda, T.; Tamura, A.; and Takagi, T. 1998. Toward information extraction: Identifying protein names from biological papers. In *Pacific Symposium on Biocomputing*, 707–718.

Hodges, P. E.; Payne, W. E.; and Garrels, J. I. 1998. Yeast protein database (YPD): A database for the complete proteome of saccharomyces cerevisiae. *Nucleic Acids Research* 26:68–72.

Leek, T. 1997. Information extraction using hidden markov models. Master's thesis, Department of Comp. Sci. and Eng., Univ. of California, San Diego, CA.

Ohta, Y.; Yamamoto, Y.; Okazaki, T.; Uchiyama, I.; and Takagi, T. 1997. Automatic construction of knowledge base from biological papers. In *Proc. of the 5th International Conf. on Intelligent Systems for Molecular Biology*, 218–225. Halkidiki, Greece: AAAI Press.

Pearl, J. 1988. *Probabalistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.

Provost, F., and Fawcett, T. 1998. Robust classification systems for imprecise environments. In *Proc. of the 15th National Conf. on Artificial Intelligence*, 706-713. Madison, WI: AAAI Press.

Quinlan, J. R. 1990. Learning logical definitions from relations. *Machine Learning* 5:239–2666.

Richards, B. L., and Mooney, R. J. 1992. Learning relations by pathfinding. In *Proc. of the 10th National Conf. on Artificial Intelligence*, 50–55. San Jose, CA: AAAI/MIT Press.

Riloff, E. 1998. The sundance sentence analyzer. http://www.cs.utah.edu/projects/nlp/.

Slattery, S., and Craven, M. 1998. Combining statistical and relational methods for learning in hypertext domains. In *Proc. of the 8th International Conf. on Inductive Logic Programming.* Springer Verlag.

Swanson, D. R., and Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence* 91:183–203.

Weeber, M., and Vos, R. 1998. Extracting expert medical knowledge from texts. In *Working Notes of the Intelligent Data Analysis in Medicine and Pharmacology Workshop*, 23–28.