Multi-way Relation Classification: Application to Protein-Protein Interactions

Barbara Rosario
SIMS
UC Berkeley
Berkeley, CA 94720
rosario@sims.berkeley.edu

Marti A. Hearst
SIMS
UC Berkeley
Berkeley, CA 94720
hearst@sims.berkeley.edu

Abstract

We address the problem of multi-way relation classification, applied to identification of the interactions between proteins in bioscience text. A major impediment to such work is the acquisition of appropriately labeled training data; for our experiments we have identified a database that serves as a proxy for training data. We use two graphical models and a neural net for the classification of the interactions, achieving an accuracy of 64% for a 10-way distinction between relation types. We also provide evidence that the exploitation of the sentences surrounding a citation to a paper can yield higher accuracy than other sentences.

1 Introduction

Identifying the interactions between proteins is one of the most important challenges in modern genomics, with applications throughout cell biology, including expression analysis, signaling, and rational drug design. Most biomedical research and new discoveries are available electronically but only in free text format, so automatic mechanisms are needed to convert text into more structured forms. The goal of this paper is to address this difficult and important task, the extraction of the interactions between proteins from free text. We use graphical models and a neural net that were found to achieve high accuracy in the related task of extracting the relation types might hold between the entities "treatment" and "disease" (Rosario and Hearst, 2004).

Labeling training and test data is time-consuming and subjective. Here we report on results using an existing curated database, the HIV-1 Human Protein Interaction Database[1], to train and test the classification system. The accuracies obtained by the classification models proposed are quite high, confirming the validity of the approach. We also find support for the hypothesis that the sentences surrounding citations are useful for extraction of key information from technical articles (Nakov et al., 2004).

In the remainder of this paper we discuss related work, describe the dataset, and show the results of the algorithm on documents and sentences.

2 Related work

There has been little work in general NLP on trying to identify different relations between entities. Many papers that claim to be doing relationship recognition in actuality address the task of role extraction: (usually two) entities are identified and the relationship is *implied* by the co-occurrence of these entities or by some linguistic expression (Agichtein and Gravano, 2000; Zelenko et al., 2002).

The ACE competition[2] has a relation recognition subtask, but assumes a particular type of relation holds between particular entity types (e.g., if the two entities in question are an EMP and an ORG, then an employment relation holds between them; which type of employment relation depends on the type of entity, e.g., staff person vs partner).

[1] www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html
[2] http://www.itl.nist.gov/iaui/894.01/tests/ace/

In the BioNLP literature there have recently been a number of attempts to automatically extract protein-protein interactions from PubMed abstracts. Some approaches simply report that a relation exists between two proteins but do not determine which relation holds (Bunescu et al., 2005; Marcotte et al., 2001; Ramani et al., 2005), while most others start with a list of interaction verbs and label only those sentences that contain these trigger words (Blaschke and Valencia, 2002; Blaschke et al., 1999; Rindflesch et al., 1999; Thomas et al., 2000; Sekimizu et al., 1998; Ahmed et al., 2005; Phuong et al., 2003; Pustejovsky et al., 2002). However, as Marcotte et al. (2001) note, "... searches for abstracts containing relevant keywords, such as interact*, poorly discriminate true hits from abstracts using the words in alternate senses and miss abstracts using different language to describe the interactions."

Most of the existing methods also suffer from low recall because they use hand-built specialized templates or patterns (Ono et al., 2001; Corney et al., 2004). Some systems use link grammars in conjunction with trigger verbs instead of templates (Ahmed et al., 2005; Phuong et al., 2003). Every paper evaluates on a different test set, and so it is quite difficult to compare systems.

In this paper, we use state-of-the-art machine learning methods to determine the interaction *types* and to extract the proteins involved. We do not use trigger words, templates, or dictionaries.

3 Data

We use the information from a domain-specific database to gather labeled data for the task of classifying the interactions between proteins in text. The manually-curated HIV-1 Human Protein Interaction Database provides a summary of documented interactions between HIV-1 proteins and host cell proteins, other HIV-1 proteins, or proteins from disease organisms associated with HIV or AIDS. We use this database also because it contains information about the *type* of interactions, as opposed to other protein interaction databases (BIND, MINT, DIP, for example[3]) that list the protein pairs interacting, without specifying the type of interactions.

Interaction	#Triples	Interaction	#Triples
Interacts with	1115	*Complexes with*	45
Activates	778	*Modulates*	43
Stimulates	659	*Enhances*	41
Binds	647	*Stabilizes*	34
Upregulates	316	*Myristoylated by*	34
Imported by	276	*Recruits*	32
Inhibits	194	*Ubiquitinated by*	29
Downregulates	124	*Synergizes with*	28
Regulates	86	*Co-localizes with*	27
Phosphorylates	81	*Suppresses*	24
Degrades	73	*Competes with*	23
Induces	52	*Requires*	22
Inactivates	51		

Table 1: Number of triples for the most common interactions of the HIV-1 database, after removing the distinction in directionality and the triples with more than one interaction.

In this database, the definitions of the interactions depend on the proteins involved and the articles describing the interactions; thus there are several definitions for each interaction type. For the interaction *bind* and the proteins *ANT* and *Vpr*, we find (among others) the definition "*Interaction of HIV-1 Vpr with human adenine nucleotide translocator (ANT) is presumed based on a specific binding interaction between Vpr and rat ANT.*"

The database contains 65 types of interactions and 809 proteins for which there is interaction information, for a total of 2224 pairs of interacting proteins. For each documented protein-protein interaction the database includes information about:

- A pair of proteins (PP),
- The interaction type(s) between them (I), and
- PubMed identification numbers of the journal article(s) describing the interaction(s) (A).

A protein pair PP can have multiple interactions (for example, AIP1 *binds* to HIV-1 p6 and also *is incorporated* into it) for an average of 1.9 interactions per PP and a maximum of 23 interactions for the pair CDK9 and tat p14.

We refer to the combination of a protein pair PP and an article A as a "triple." Our goal is to automatically associate to each triple an interaction

[3]DIP lists only the protein pairs, BIND has only some information about the method used to provide evidence for the interaction, and MIND does have interaction type information but the vast majority of the entries (99.9% of the 47,000 pairs) have been assigned the same type of interaction (*aggregation*). These databases are all manually curated.

type. For the example above, the triple "AIP1 HIV-1-p6 14519844" is assigned the interaction *binds* (14519844 being the PubMed number of the paper providing evidence for this interaction)[4].

Journal articles can contain evidence for multiple interactions: there are 984 journal articles in the database and on average each article is reported to contain evidence for 5.9 triples (with a maximum number of 90 triples).

In some cases the database reports multiple different interactions for a given triple. There are 5369 unique triples in the database and of these 414 (7.7%) have multiple interactions. We exclude these triples from our analysis; however, we do include articles and PPs with multiple interactions. In other words, we tackle cases such as the example above of the pair AIP1, HIV-1-p6 (that can both *bind* and *incorporate*) as long as the evidence for the different interactions is given by two different articles.

Some of the interactions differ only in the directionality (e.g., *regulates* and *regulated by*, *inhibits* and *inhibited by*, etc.); we collapsed these pairs of related interactions into one[5]. Table 1 shows the list of the 25 interactions of the HIV-1 database for which there are more than 10 triples.

For these interactions and for a random subset of the protein pairs PP (around 45% of the total pairs in the database), we downloaded the corresponding full-text papers. From these, we extracted all and only those sentences that contain both proteins from the indicated protein pair. We assigned each of these sentences the corresponding interaction I from the database ("papers").

Nakov et al. (2004) argue that the sentences surrounding citations to related work, or *citances*, are a useful resource for bioNLP. Building on that work, we use citances as an additional form of evidence to determine protein-protein interaction types. For a given database entry containing PubMed article A, protein pair PP, and interaction type I, we downloaded a subset of the papers that cite A. From these citing papers, we extracted all and only those sentences that mention A explicitly; we further filtered these to include all and only the sentences that contain PP. We labeled each of these sentences with interaction type I ("citances").

There are often many different names for the same protein. We use LocusLink[6] protein identification numbers and synonym names for each protein, and extract the sentences that contain an exact match for (some synonym of) each protein. By being conservative with protein name matching, and by not doing co-reference analysis, we miss many candidate sentences; however this method is very precise.

On average, for "papers," we extracted 0.5 sentences per triple (maximum of 79) and 50.6 sentences per interaction (maximum of 119); for "citances" we extracted 0.4 sentences per triple (with a maximum of 105) and 49.2 sentences per interaction (162 maximum). We required a minimum number (40) of sentences for each interaction type for both "papers" and "citances"; the 10 interactions of Table 2 met this requirement. We used these sentences to train and test the models described below[7].

Since all the sentences extracted from one triple are assigned the same interaction, we ensured that sentences from the same triple did not appear in both the testing and the training sets. Roughly 75% of the data were used for training and the rest for testing.

As mentioned above the goal is to automatically associate to each triple an interaction type. The task tackled here is actually slightly more difficult: given some sentences extracted from article A, assign to A an interaction type I and extract the proteins PP involved. In other words, for the purpose of classification, we act as if we do not have information about the proteins that interact. However, given the way the sentence extraction was done, all the sentences extracted from A contain the PP.

[4]To be precise, there are for this PP (as there are often) multiple articles (three in this case) describing the interaction *binds*, thus we have the following three triples to which we associate *binds*: "AIP1 HIV-1-p6 14519844," "AIP1 HIV-1-p6 14505570" and "AIP1 HIV-1-p6 14505569."

[5]We collapsed these pairs because the directionality of the interactions was not always reliable in the database. This implies that for some interactions, we are not able to infer the *different* roles of the two proteins; we considered only the pair "prot1 prot2" or "prot2 prot1," not both. However, our algorithm can detect *which* proteins are involved in the interactions.

[6]LocusLink was recently integrated into Entrez Gene, a unified query environment for genes (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene).

[7]We also looked at larger chunks of text, in particular, we extracted the sentence containing the PP along with the previous and the following sentences, and the three consecutive sentences that contained the PP (the proteins could appear in any of the sentences). However, the results obtained by using these larger chunks were consistently worse.

Interaction	Papers	Citances
Degrades	60	63
Synergizes with	86	101
Stimulates	103	64
Binds	98	324
Inactivates	68	92
Interacts with	62	100
Requires	96	297
Upregulates	119	98
Inhibits	78	84
Suppresses	51	99
Total	821	1322

Table 2: Number of interaction sentences extracted.

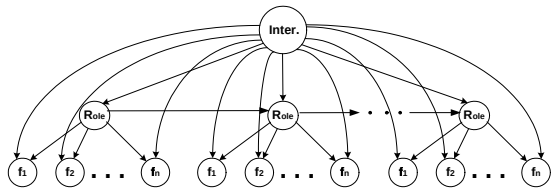

Figure 1: Dynamic graphical model (DM) for protein interaction classification (and role extraction).

A hand-assessment of the individual sentences shows that not every sentence that mentions the target proteins PP actually describes the interaction I (see Section 5.4). Thus the evaluation on the test set is done at the document level (to determine if the algorithm can predict the interaction that a curator would assign to a document as a whole given the protein pair).

Note that we assume here that the papers that provide the evidence for the interactions are given – an assumption not usually true in practice.

4 Models

For assigning interactions, we used two generative graphical models and a discriminative model. Figure 1 shows the generative dynamic model, based on previous work on role and relation extraction (Rosario and Hearst, 2004) where the task was to extract the entities TREATMENT and DISEASE and the relationships between them. The nodes labeled "Role" represent the entities (in this case the choices are PROTEIN and NULL); the children of the role nodes are the words (which act as features), thus there are as many role states as there are words in the sentence; this model consists of a Markov sequence of states where each state generates one or multiple observations. This model makes the additional assumption that there is an interaction present in the sentence (represented by the node "Inter.") that generates the role sequence and the observations. (We assume here that there is a single interaction for each sentence.) The "Role" nodes can be observed or hidden. The results reported here were obtained using only the words as features (i.e., in the dynamic model of Figure 1 there is only one feature node per role) and with the "Role" nodes hidden (i.e., we had no information regarding which proteins were involved). Inference is performed with the junction tree algorithm[8].

We used a second type of graphical model, a simple Naive Bayes, in which the node representing the interaction generates the observable features (all the words in the sentence). We did not include role information in this model.

We defined joint probability distributions over these models, estimated using maximum likelihood on the training set with a simple absolute discounting smoothing method. We performed 10-fold cross validation on the training set and we chose the smoothing parameters for which we obtained the best classification accuracies (averaged over the ten runs) on the training data; the results reported here were obtained using these parameters on the held-out test sets[9].

In addition to these two generative models, we also used a discriminative model, a neural network. We used the Matlab package to train a feed-forward network with conjugate gradient descent. The network has one hidden layer, with a hyperbolic tangent function, and an output layer representing the relationships. A logistic sigmoid function is used in the output layer. The network was trained for several choices of numbers of hidden units; we chose the best-performing networks based on training set error. We then tested these networks on held-out testing data. The features were words, the same as those used for the graphical models.

[8]Using Kevin Murphy's BNT package: http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html.

[9]We did not have enough data to require that the sentences in the training and test sets *of the cross validation procedure* originate from disjoint triples (they do originate from disjoint triple in the final held out data). This may result in a less than optimal choice of the parameters for the aggregate measures described below.

	All		Papers		Citances	
	Mj	**Cf**	**Mj**	**Cf**	**Mj**	**Cf**
DM	60.5	59.7	**57.8**	55.6	53.4	54.5
NB	58.1	61.3	**57.8**	55.6	55.7	54.5
NN	**63.7**	–	44.4	–	**55.8**	–
Key	20.1	–	24.4	–	20.4	–
KeyB	25.8	–	40.0	–	26.1	–
Base.	21.8		11.1		26.1	

Table 3: Accuracies for classification of the 10 protein-protein interactions of Table 2. DM: dynamic model, NB: Naive Bayes, NN: neural network. Baselines: Key: trigger word approach, KeyB: trigger word with backoff, Base: the accuracy of choosing the most frequent interaction.

The task is the following: given a triple consisting of a PP and an article, extract the sentences from the article that contain both proteins. Then, predict for the entire document one of the interactions of Table 2 given the sentences extracted for that triple. This is a 10-way classification problem, which is significantly more complex than much of the related work in which the task is to make the binary prediction (see Section 2).

5 Results

The evaluation was done on a document-by-document basis. During testing, we choose the interaction using the following aggregate measures that use the constraint that all sentences coming from the same triple are assigned the same interaction.

- **Mj**: For each triple, for each sentence of the triple, find the interaction that maximizes the posterior probability of the interaction given the features; then assign to *all* sentences of this triple the most frequent interaction among those predicted for the individual sentences.
- **Cf**: Retain all the conditional probabilities (do not choose an interaction per sentence), then, for each triple, choose the interaction that maximizes the sum over all the triple's sentences.

Table 3 reports the results in terms of classification accuracies averaged across all interactions, for the cases "all" (sentences from "papers" and "citances" together), only "papers" and only "citances". The accuracies are quite high; the dynamic model achieves around 60% for "all," 58% for "papers" and 54% for "citances." The neural net achieves the best results for "all" with around 64% accuracy. From these results we can make the following observations: all models greatly outperform the baselines; the performances of the dynamic model DM, the Naive Bayes NB and the NN are very similar; for "papers" the best results were obtained with the graphical models; for "all" and "citances" the neural net did best. The use of "citances" allowed the gathering of additional data (and therefore a larger training set) that lead to higher accuracies (see "papers" versus "all").

In the confusion matrix in Table 5 we can see the accuracies for the individual interactions for the dynamic model DM, using "all" and "Mj." For three interactions this model achieves perfect accuracy.

5.1 Hiding the protein names

In order to ensure that the algorithm was not overfitting on the protein names, we ran an experiment in which we replaced the protein names in all sentences with the token "PROT_NAME." For example, the sentence: *"Selective CXCR4 antagonism by Tat"* became: *"Selective PROT_NAME2 antagonism by PROT_NAME1."*

Table 5.1 shows the results of running the models on this data. For "papers" and "citances" there is always a decrease in the classification accuracy when we remove the protein names, showing that the protein names do help the classification. The differences in accuracy in the two cases using "citances" are much smaller than the differences using "papers" at least for the graphical models. This suggests that citation sentences may be more robust for some language processing tasks and that the models that use "citances" learn better the linguistic context of the interactions. Note how in this case the graphical models always outperform the neural network.

5.2 Using a "trigger word" approach

As mentioned above, much of the related work in this field makes use of "trigger words" or "interaction words" (see Section 2). In order to (roughly) compare our work and to build a more realistic baseline, we created a list of 70 keywords that are repre-

	Prediction										Acc.
Truth	***D***	***SyW***	***St***	***B***	***Ina***	***IW***	***R***	***Up***	***Inh***	***Su***	**(%)**
Degrades (D)	5	0	0	0	0	0	0	0	0	0	100.0
Synergizes with (SyW)	0	1	0	0	0	1	0	3	3	0	12.5
Stimulates (St)	0	0	4	0	0	0	6	0	1	0	36.4
Binds (B)	0	0	0	18	0	4	1	1	3	0	66.7
Inactivates (Ina)	0	0	0	0	9	0	0	0	0	0	100.0
Interacts with (IW)	0	0	4	3	0	5	1	0	1	2	31.2
Requires (R)	0	0	0	0	0	3	3	0	1	1	37.5
Upregulates (Up)	0	0	0	2	1	0	0	12	2	0	70.6
Inhibits (Inh)	0	0	0	3	0	0	1	1	12	0	70.6
Suppresses (Su)	0	0	0	0	0	0	0	0	0	6	100.0

Table 4: Confusion matrix for the dynamic model DM for "all," "Mj." The overall accuracy is 60.5%. The numbers indicate the number of articles A (each paper has several relevant sentences).

	All			Papers			Citances		
	Mj	**Cf**	**Diff**	**Mj**	**Cf**	**Diff**	**Mj**	**Cf**	**Diff**
DM	**60.5**	**60.5**	0.7%	44.4	40.0	-25.6%	52.3	**53.4**	-2.0%
NB	59.7	59.7	0.1%	46.7	**51.1**	-11.7%	**53.4**	**53.4**	-3.1%
NN	51.6		-18.9%	44.4		0%	50.0		-10.4%

Table 5: Accuracies for the classification of the 10 protein-protein interactions of Table 2 with the *protein names removed*. Columns marked Diff show the difference in accuracy (in percentages) with respect to the original case of Table 3, averaged over all evaluation methods.

sentative of the 10 interactions. For example, for the interaction *degrade* some of the keywords are "degradation," "degrade," for *inhibit* we have "inhibited," "inhibitor," "inhibitory" and others. We then checked whether a sentence contained such keywords. If it did, we assigned to the sentence the corresponding interaction. If it contained more than one keyword corresponding to multiple interactions consisting of the generic *interact with* plus a more specific one, we assigned the more specific interaction; if the two predicted interactions did not include *interact with* but two more specific interactions, we did not assign an interaction, since we wouldn't know how to choose between them. Similarly, we assigned no interaction if there were more than two predicted interactions or no keywords present in the sentence. The results are shown in the rows labeled "Key" and "KeyB" in Table 3. Case "KeyB" is the "Key" method with back-off: when no interaction was predicted, we assigned to the sentence the most frequent interaction in the training data. As before, we calculated the accuracy when we force all the sentences from one triple to be assign to the most frequent interaction among those predicted for the individual sentences.

KeyB is more accurate than Key and although the KeyB accuracies are higher than the other baselines, they are significantly lower than those obtained with the trained models. The low accuracies of the trigger-word based methods show that the relation classification task is nontrivial, in the sense that not all the sentences contain the most obvious word for the interactions, and suggests that the trigger word approach is insufficient.

5.3 Protein extraction

The dynamic model of Figure 1 has the appealing property of simultaneously performing interaction recognition and protein name tagging (also known as role extraction): the task consists of identifying all the proteins present in the sentence, given a sequence of words. We assessed a slightly different task: the identification of all (and only) the proteins present in the sentence *that are involved in the interaction*.

The F-measure[10] achieved by this model for this task is 0.79 for "all," 0.67 for "papers" and 0.79 for "citances"; again, the model parameters were chosen with cross validation on the training set, and "ci-

[10]The F-measure is a weighted combination of precision and recall. Here, precision and recall are given equal weight, that is, F-measure = $(2 * PRE * REC)/(PRE + REC)$.

tances" had superior performance. Note that we did not use a dictionary: the system learned to recognize the protein names using only the training data. Moreover, our role evaluation is quite strict: every token is assessed and we do not assign partial credit for constituents for which only some of the words are correctly labeled. We also did not use the information that all the sentences extracted from one triple contain the same proteins.

Given these strong results (both F-measure and classification accuracies), we believe that the dynamic model of Figure 1 is a good model for performing both name tagging and interaction classification simultaneously, or either of these task alone.

5.4 Sentence-level evaluation

In addition to assigning interactions to protein pairs, we are interested in sentence-level semantics, that is, in determining the interactions that are actually expressed in the sentence. To test whether the information assigned to the entire document by the HIV-1 database record can be used to infer information at the sentence level, an annotator with biological expertise hand-annotated the sentences from the experiments. The annotator was instructed to assign to each sentence one of the interactions of Table 2, "not interacting," or "other" (if the interaction between the two proteins was not one of Table 2).

Of the 2114 sentences that were hand-labeled, 68.3% of them disagreed with the HIV-1 database label, 28.4% agreed with the database label, and 3.3% were found to contain multiple interactions between the proteins. Among the 68.3% of the sentences for which the labels did not agree, 17.4% had the vague *interact with* relation, 7.4% did not contain any interaction and 43.5% had an interaction different from that specified by the triple[11]. In Table 6 we report the mismatch between the two sets of labels. The total accuracy of 38.9%[12] provides a useful baseline for using a database for the labeling at the sentence level. It may be the case that certain interactions tend to be biologically related and thus tend to co-occur (*upregulate* and *stimulate* or *inactivate* and *inhibit*, for example).

	All	Papers	Citan.
DM	48.9	28.9	47.9
NB	47.1	33.3	53.4
NN	**52.9**	**36.7**	**63.2**
Key	30.5	18.9	38.3
KeyB	46.2	36.3	52.6
Base	36.3	34.4	37.6

Table 7: Classification accuracies when the models are trained and tested on the hand labeled sentences.

We investigated a few of the cases in which the labels were "suspiciously" different, for example a case in which the database interaction was *stimulate* but the annotator found the same proteins to be related by *inhibit* as well. It turned out that the authors of the article assigned *stimulate* found little evidence for this interaction (in favor of *inhibit*), suggesting an error in the database. In another case the database interaction was *require* but the authors of the article, while supporting this, found that under certain conditions (when a protein is too abundant) the interaction changes to one of *inhibit*. Thus we were able to find controversial facts about protein interactions just by looking at the confusion matrix of Table 6.

We trained the models using these hand-labeled sentences in order to determine the interaction expressed *for each sentence* (as opposed to for each document). This is a difficult task; for some sentences it took the annotator several minutes to understand them and decide which interaction applied. Table 7 shows the results on running the classification models on the six interactions for which there were more than 40 examples in the training sets. Again, the sentences from "papers" are especially difficult to classify; the best result for "papers" is 36.7% accuracy versus 63.2% accuracy for "citances." In this case the difference in performance of "papers" and "citances" is larger than for the previous task of document-level relation classification.

6 Conclusions

We tackled an important and difficult task, the classification of different interaction types between proteins in text. A solution to this problem would have an impact on a variety of important challenges in modern biology. We used a protein-interaction

[11] For 28% of the triples, none of the sentences extracted from the target paper were found by the annotator to contain the interaction given by the database. We read four of these papers and found sentences containing that interaction, but our system had failed to extract them.

[12] The accuracy without the vague *interact with* is 49.4%.

	Annotator											
Database	***D***	***SyW***	***St***	***B***	***Ina***	***R***	***Up***	***Inh***	***Su***	***IW***	***Ot***	***No***
Degrades (D)	44	0	2	5	6	5	2	0	23	9	11	6
Synergizes with (SyW)	0	78	3	14	0	13	8	0	0	26	31	11
Stimulates (St)	0	5	23	12	0	8	7	5	1	26	60	18
Binds (B)	0	6	9	118	0	25	8	10	1	129	77	22
Inactivates (Ina)	0	0	4	25	0	2	4	33	6	14	27	11
Requires (R)	0	5	29	20	0	63	8	54	0	85	80	33
Upregulates (Up)	0	4	24	0	0	0	124	2	0	21	32	4
Inhibits (Inh)	0	8	4	8	2	2	2	43	9	24	37	19
Suppresses (Su)	3	0	0	1	5	0	0	42	34	33	24	4
Interacts with (IW)	0	1	5	28	1	12	6	1	1	49	27	28
Accuracy	93.6	72.9	22.3	51.1	0	48.5	73.4	22.7	45.3	11.8		

Table 6: Confusion matrix comparing the hand-assigned interactions and those extracted from the HIV-1 database. Ot: sentences for which the annotator found an interaction different from those in Table 2. No: sentences for which the annotator found no interaction. The bottom row shows the accuracy of using the database to label the individual sentences.

database to automatically gather labeled data for this task, and implemented graphical models that can simultaneously perform protein name tagging and relation identification, achieving high accuracy on both problems. We also found evidence supporting the hypothesis that citation sentences are a good source of training data, most likely because they provide a concise and precise way of summarizing facts in the bioscience literature.

Acknowledgments. We thank Janice Hamer for her help in labeling examples and other biological insights. This research was supported by a grant from NSF DBI-0317510 and a gift from Genentech.

References

E. Agichtein and L. Gravano. 2000. Snowball: Extracting relations from large plain-text collections. *Proc. of DL '00*.

S. Ahmed, D. Chidambaram, H. Davulcu, and C. Baral. 2005. Intex: A syntactic role driven protein-protein interaction extractor for bio-medical text. In *Proceedings ISMB/ACL Biolink 2005*.

C. Blaschke and A. Valencia. 2002. The frame-based module of the suiseki information extraction system. *IEEE Intelligent Systems*, 17(2).

C. Blaschke, M.A. Andrade, C. Ouzounis, and A. Valencia. 1999. Automatic extraction of biological information from scientific text: Protein-protein interactions. *Proc. of ISMB*.

R. Bunescu, R. Ge, R. Kate, E. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong. 2005. Comparative experiments on learning information extractors for protiens and their interactions. *Artificial Intelligence in Medicine*, 33(2).

D. Corney, B. Buxton, W. Langdon, and D. Jones. 2004. Biorat: extracting biological information from full-length papers. *Bioinformatics*, 20(17).

E. Marcotte, I. Xenarios, and D. Eisenberg. 2001. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4).

P. Nakov, A. Schwartz, and M. Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*.

T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. 2001. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(1).

T. Phuong, D. Lee, and K-H. Lee. 2003. Learning rules to extract protein interactions from biomedical text. In *PAKDD*.

J. Pustejovsky, J. Castano, and J. Zhang. 2002. Robust relational parsing over biomedical literature: Extracting inhibit relations. *Proc. of Pac Symp Biocomputing*.

C. Ramani, E. Marcotte, R. Bunescu, and R. Mooney. 2005. Using biomedical literature mining to consolidate the set of known human protein-protein interactions. In *Proceedings ISMB/ACL Biolink 2005*.

T. Rindflesch, L. Hunter, and L. Aronson. 1999. Mining molecular binding terminology from biomedical text. *Proceedings of the AMIA Symposium*.

Barbara Rosario and Marti A. Hearst. 2004. Classifying semantic relations in bioscience texts. In *Proc. of ACL 2004*.

T. Sekimizu, H.S. Park, and J. Tsujii. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Gen. Informat.*, 9.

J. Thomas, D. Milward, C. Ouzounis, and S. Pulman. 2000. Automatic extraction of protein interactions from scientific abstracts. *Proceedings of the Pac Symp Biocomput*.

D. Zelenko, C. Aone, and A. Richardella. 2002. Kernel methods for relation extraction. *Proceedings of EMNLP 2002*.