In *Proceedings of the HLT-NAACL Workshop on Linking Natural Language Processing and Biology: Towards deeper biological literature analysis (BioNLP-2006)*. pp. 49-56, New York City, NY, June 2006.

# Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline

**Razvan Bunescu, Raymond Mooney**
Department of Computer Sciences
University of Texas at Austin
1 University Station C0500
Austin, TX 78712
razvan@cs.utexas.edu
mooney@cs.utexas.edu

**Arun Ramani, Edward Marcotte**
Institute for Cellular and Molecular Biology
University of Texas at Austin
1 University Station A4800
Austin, TX 78712
arun@icmb.utexas.edu
marcotte@icmb.utexas.edu

## Abstract

The task of mining relations from collections of documents is usually approached in two different ways. One type of systems do relation extraction from individual sentences, followed by an aggregation of the results over the entire collection. Other systems follow an entirely different approach, in which co-occurrence counts are used to determine whether the mentioning together of two entities is due to more than simple chance. We show that increased extraction performance can be obtained by combining the two approaches into an integrated relation extraction model.

## 1 Introduction

Information Extraction (IE) is a natural language processing task in which text documents are analyzed with the aim of finding mentions of relevant entities and important relationships between them. In many cases, the subtask of relation extraction reduces to deciding whether a sentence asserts a particular relationship between two entities, which is still a difficult, unsolved problem. There are however cases where the decision whether the two entities are in a relationship is made relative to an entire document, or a collection of documents. In the biomedical domain, for example, one may be interested in finding the pairs of human proteins that are said to be interacting in any of the Medline abstracts, where the answer is not required to specify which abstracts are actually describing the interaction. Assembling a ranked list of interacting proteins can be very useful to biologists - based on this list, they can make more informed decisions with respect to which genes to focus on in their research.

In this paper, we investigate methods that use multiple occurrences of the same pair of entities across a collection of documents in order to boost the performance of a relation extraction system. The proposed methods are evaluated on the task of finding pairs of human proteins whose interactions are reported in Medline abstracts. The majority of known human protein interactions are derived from individual, small-scale experiments reported in Medline. Some of these interactions have already been collected in the Reactome (Joshi-Tope et al., 2005), BIND (Bader et al., 2003), DIP (Xenarios et al., 2002), and HPRD (Peri et al., 2004) databases. The amount of human effort involved in creating and updating these databases is currently no match for the continuous growth of Medline. It is therefore very useful to have a method that automatically and reliably extracts interaction pairs from Medline.

Systems that do relation extraction from a collection of documents can be divided into two major categories. In one category are IE systems that first extract information from individual sentences, and then combine the results into corpus-level results (Craven, 1999; Skounakis and Craven, 2003). The second category corresponds to approaches that do not exploit much information from the context of individual occurrences. Instead, based on co-occurrence counts, various statistical or information-theoretic tests are used to decide whether the two entities in a pair appear together more often than simple chance would predict (Lee et al., 2004; Ramani et al., 2005). We believe that a combination of the two approaches can inherit the advantages of each method and lead to improved relation extraction accuracy.

The following two sections describe the two orthogonal approaches to corpus-level relation extraction. A model that integrates the two approaches is then introduced in Section 4. This is followed by a description of the dataset used for evaluation in Section 5, and experimental results in Section 6.

## 2 Sentence-level relation extraction

Most systems that identify relations between entities mentioned in text documents consider only pair of entities that are mentioned in the same sentence (Ray and Craven, 2001; Zhao and Grishman, 2005; Bunescu and Mooney, 2005). To decide the existence and the type of a relationship, these systems generally use lexico-semantic clues inferred from the sentence context of the two entities. Much research has been focused recently on automatically identifying biologically relevant entities and their relationships such as protein-protein interactions or subcellular localizations. For example, the sentence "*TR6* specifically binds *Fas ligand*", states an interaction between the two proteins *TR6* and *Fas ligand*. One of the first systems for extracting interactions between proteins is described in (Blaschke and Valencia, 2001). There, sentences are matched deterministically against a set of manually developed patterns, where a pattern is a sequence of words or Part-of-Speech (POS) tags and two protein-name tokens. Between every two adjacent words is a number indicating the maximum number of words that can be skipped at that position. An example is: "*interaction of (3) <P> (3) with (3) <P>*". This approach is generalized in (Bunescu and Mooney, 2005), where subsequences of words (or POS tags) from the sentence are used as implicit features. Their weights are learned by training a customized subsequence kernel on a dataset of Medline abstracts annotated with proteins and their interactions.

A relation extraction system that works at the sentence-level and which outputs normalized confidence values for each extracted pair of entities can also be used for corpus-level relation extraction. A straightforward way to do this is to apply an aggregation operator over the confidence values inferred for all occurrences of a given pair of entities. More exactly, if $p_1$ and $p_2$ are two entities that occur in a total of $n$ sentences $s_1$, $s_2$, ..., $s_n$ in the entire corpus $C$, then the confidence $P(R(p_1, p_2)|C)$ that they are in a particular relationship $R$ is defined as:

$$P(R(p_1, p_2)|C) = \Gamma(\{P(R(p_1, p_2)|s_i)|i=1:n\})$$

Table 1 shows only four of the many possible choices for the aggregation operator $\Gamma$.

| | |
|---|---|
| *max* | $\Gamma_{max} = \max_i P(R(p_1, p_2)\|s_i)$ |
| *noisy-or* | $\Gamma_{nor} = 1 - \prod_i (1 - P(R(p_1, p_2)\|s_i))$ |
| *avg* | $\Gamma_{avg} = \sum_i \frac{P(R(p_1, p_2)\|s_i)}{n}$ |
| *and* | $\Gamma_{and} = \prod_i P(R(p_1, p_2)\|s_i)^{1/n}$ |

Table 1: Aggregation Operators.

Out of the four operators in Table 1, we believe that the *max* operator is the most appropriate for aggregating confidence values at the corpus-level. The question that needs to be answered is whether there is a sentence somewhere in the corpus that asserts the relationship $R$ between entities $p_1$ and $p_2$. Using *avg* instead would answer a different question - whether $R(p_1, p_2)$ is true in most of the sentences containing $p_1$ and $p_2$. Also, the *and* operator would be most appropriate for finding whether $R(p_1, p_2)$ is true in all corresponding sentences in the corpus. The value of the *noisy-or* operator (Pearl, 1986) is too dependent on the number of occurrences, therefore it is less appropriate for a corpus where the occurrence counts vary from one entity pair to another (as confirmed in our experiments from Section 6). For examples, if the confidence threshold is set at 0.5, and the entity pair $(p_1, p_2)$ occurs in 6 sentences or less, each with confidence 0.1, then $R(p_1, p_2)$ is false, according to the noisy-or operator. However, if $(p_1, p_2)$ occur in more than 6 sentences, with the same confidence value of 0.1, then the corresponding noisy-or value exceeds 0.5, making $R(p_1, p_2)$ true.

## 3 Co-occurrence statistics

Given two entities with multiple mentions in a large corpus, another approach to detect whether a relationship holds between them is to use statistics over their occurrences in textual patterns that are indicative for that relation. Various measures such as pointwise mutual information (PMI) , chi-square ($\chi^2$) or log-likelihood ratio (LLR) (Manning and Schütze, 1999) use the two entities' occurrence statistics to detect whether their co-occurrence is due to chance, or to an underlying relationship.

A recent example is the *co-citation* approach from (Ramani et al., 2005), which does not try to find specific assertions of interactions in text, but rather exploits the idea that if many different abstracts reference both protein $p_1$ and protein $p_2$, then $p_1$ and $p_2$ are likely to interact. Particularly, if the two proteins are co-cited significantly more often than one would expect if they were cited independently at random, then it is likely that they interact. The model used to compute the probability of random co-citation is based on the hypergeometric distribution (Lee et al., 2004; Jenssen et al., 2001). Thus, if $N$ is the total number of abstracts, $n$ of which cite the first protein, $m$ cite the second protein, and $k$ cite both, then the probability of co-citation under a random model is:

$$P(k|N,m,n) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}} \qquad (1)$$

The approach that we take in this paper is to constrain the two proteins to be mentioned in the same sentence, based on the assumption that if there is a reason for two protein names to co-occur in the same sentence, then in most cases that is caused by their interaction. To compute the "degree of interaction" between two proteins $p_1$ and $p_2$, we use the information-theoretic measure of pointwise mutual information (Church and Hanks, 1990; Manning and Schütze, 1999), which is computed based on the following quantities:

1. $N$ : the total number of protein pairs co-occurring in the same sentence in the corpus.

2. $P(p_1, p_2) \simeq n_{12}/N$ : the probability that $p_1$ and $p_2$ co-occur in the same sentence; $n_{12}$ = the number of sentences mentioning both $p_1$ and $p_2$.

3. $P(p_1, p) \simeq n_1/N$ : the probability that $p_1$ co-occurs with any other protein in the same sentence; $n_1$ = the number of sentences mentioning $p_1$ and $p$.

4. $P(p_2, p) \simeq n_2/N$ : the probability that $p_2$ co-occurs with any other protein in the same sentence; $n_2$ = the number of sentences mentioning $p_2$ and $p$.

The PMI is then defined as in Equation 2 below:

$$\begin{aligned} PMI(p_1, p_2) &= \log \frac{P(p_1, p_2)}{P(p_1, p) \cdot P(p_2, p)} \\ &\simeq \log N \frac{n_{12}}{n_1 \cdot n_2} \end{aligned} \qquad (2)$$

Given that the PMI will be used only for ranking pairs of potentially interacting proteins, the constant factor $N$ and the $log$ operator can be ignored. For sake of simplicity, we use the simpler formula from Equation 3.

$$sPMI(p_1, p_2) = \frac{n_{12}}{n_1 \cdot n_2} \qquad (3)$$

## 4 Integrated model

The $sPMI(p_1, p_2)$ formula can be rewritten as:

$$sPMI(p_1, p_2) = \frac{1}{n_1 \cdot n_2} \cdot \sum_{i=1}^{n_{12}} 1 \qquad (4)$$

Let $s_1$, $s_2$, ..., $s_{n_{12}}$ be the sentence contexts corresponding to the $n_{12}$ co-occurrences of $p_1$ and $p_2$, and assume that a sentence-level relation extractor is available, with the capability of computing normalized confidence values for all extractions. Then one way of using the extraction confidence is to have each co-occurrence weighted by its confidence, i.e. replace the constant 1 with the normalized scores $P(R(p_1, p_2)|s_i)$, as illustrated in Equation 5. This results in a new formula $wPMI$ (*weighted PMI*), which is equal with the product between $sPMI$ and the average aggregation operator $\Gamma_{avg}$.

$$\begin{aligned} wPMI(p_1, p_2) &= \frac{1}{n_1 \cdot n_2} \cdot \sum_{i=1}^{n_{12}} P(R(p_1, p_2)|s_i) \\ &= \frac{n_{12}}{n_1 \cdot n_2} \cdot \Gamma_{avg} \end{aligned} \qquad (5)$$

The operator $\Gamma_{avg}$ can be replaced with any other aggregation operator from Table 1. As argued in Section 2, we consider $max$ to be the most appropriate operator for our task, therefore the integrated model is based on the weighted PMI product illustrated in Equation 6.

$$\begin{aligned} wPMI(p_1, p_2) &= \frac{n_{12}}{n_1 \cdot n_2} \cdot \Gamma_{max} \qquad (6) \\ &= \frac{n_{12}}{n_1 \cdot n_2} \cdot \max_i P(R(p_1, p_2)|s_i) \end{aligned}$$

If a pair of entities $p_1$ and $p_2$ is ranked by $wPMI$ among the top pairs, this means that it is unlikely that $p_1$ and $p_2$ have co-occurred together in the entire corpus by chance, and at the same time there is at least one mention where the relation extractor decides with high confidence that $R(p_1, p_2) = 1$.

## 5 Evaluation Corpus

Contrasting the performance of the integrated model against the sentence-level extractor or the PMI-based ranking requires an evaluation dataset that provides two types of annotations:

1. The complete *list of interactions* reported in the corpus (Section 5.1).
2. Annotation of *mentions* of genes and proteins, together with their corresponding *gene identifiers* (Section 5.2).

We do not differentiate between genes and their protein products, mapping them to the same gene identifiers. Also, even though proteins may participate in different types of interactions, we are concerned only with detecting whether they interact in the general sense of the word.

### 5.1 Medline Abstracts and Interactions

In order to compile an evaluation corpus and an associated comprehensive list of interactions, we exploited information contained in the HPRD (Peri et al., 2004) database. Every interaction listed in HPRD is linked to a set of Medline articles where the corresponding experiment is reported. More exactly, each interaction is specified in the database as a tuple that contains the LocusLink (now EntrezGene) identifiers of all genes involved and the PubMed identifiers of the corresponding articles (as illustrated in Table 2).

| Interaction (XML) (HPRD) |
|---|
| <interaction> |
| <gene>**2318**</gene> |
| <gene>**58529**</gene> |
| <pubmed>**10984498 11171996**</pubmed> |
| </interaction> |
| **Participant Genes (XML) (NCBI)** |
| <gene id="**2318**"> |
| <name>**FLNC**</name> |
| <description>**filamin C, gamma**</description> |
| <synonyms> |
| <synonym>**ABPA**</synonym> |
| <synonym>**ABPL**</synonym> |
| <synonym>**FLN2**</synonym> |
| <synonym>**ABP-280**</synonym> |
| <synonym>**ABP280A**</synonym> |
| </synonyms> |
| <proteins> |
| <protein>**gamma filamin**</protein> |
| <protein>**filamin 2**</protein> |
| <protein>**gamma-filamin**</protein> |
| <protein>**ABP-L, gamma filamin**</protein> |
| <protein>**actin-binding protein 280**</protein> |
| <protein>**gamma actin-binding protein**</protein> |
| <protein>**filamin C, gamma</protein**> |
| </proteins> |
| </gene> |
| <gene id="**58529**"> |
| <name>**MYOZ1**</name> |
| <description>**myozenin 1**</description> |
| <synonyms> ... </synonyms> |
| <proteins> ... </proteins> |
| </gene> |
| **Medline Abstract (XML) (NCBI)** |
| <PMID>**10984498**</PMID> |
| <AbstractText> |
| We found that this protein binds to three other Z-disc proteins; therefore, we have named it **FATZ**, **gamma-filamin**, alpha-actinin and telethonin binding protein of the Z-disc. |
| </AbstractText> |

Table 2: Interactions, Genes and Abstracts.

The evaluation corpus (henceforth referred to as the *HPRD corpus*) is created by collecting the Medline abstracts corresponding to interactions between human proteins, as specified in HPRD. In total, 5,617 abstracts are included in this corpus, with an associated list of 7,785 interactions. This list is comprehensive - the HPRD database is based on an annotation process in which the human annotators report all interactions described in a Medline article. On the other hand, the fact that only abstracts are included in the corpus (as opposed to including the full article) means that the list may contain interactions that are not actually reported in the HPRD corpus. Nevertheless, if the abstracts were annotated with gene mentions and corresponding GIDs, then a "quasi-exact" interaction list could be computed based on the following heuristic:

**[H]** *If two genes with identifiers $gid_1$ and $gid_2$ are mentioned in the same sentence in an abstract with PubMed identifier $pmid$, and if $gid_1$ and $gid_2$ are participants in an interaction that is linked to $pmid$ in HPRD, then consider that the abstract (and consequently the entire HPRD corpus) reports the interaction between $gid_1$ and $gid_2$.* ∎

An application of the above heuristic is shown at the bottom of Table 2. The HPRD record at the top of the table specifies that the Medline article with ID 10984498 reports an interaction between the proteins *FATZ* (with ID 58529) and *gamma-filamin* (with ID 2318). The two protein names are mentioned in a sentence in the abstract for 10984498, therefore, by **[H]**, we consider that the HPRD corpus reports this interaction.

This is very similar to the procedure used in (Craven, 1999) for creating a "weakly-labeled" dataset of *subcellular-localization* relations. **[H]** is a strong heuristic – it is already known that the full article reports an interaction between the two genes. Finding the two genes collocated in the same sentence in the abstract is very likely to be due to the fact that the abstract discusses their interaction. The heuristic can be made even more accurate if a pair of genes is considered as interacting only if they co-occur in a (predefined) minimum number of sentences in the entire corpus – with the evaluation modified accordingly, as described later in Section 6.

### 5.2 Gene Name Annotation and Normalization

For the annotation of gene names and their normalization, we use a dictionary-based approach similar to (Cohen, 2005). NCBI[1] provides a comprehensive dictionary of human genes, where each gene is specified by its unique identifier, and qualified with an official name, a description, synonym names and one or more protein names, as illustrated in Table 2. All of these names, including the description, are considered as potential referential expressions for the gene entity. Each name string is reduced to a normal form by: replacing dashes with spaces, introducing spaces between sequences of letters and sequences of digits, replacing Greek letters with their Latin counterparts (capitalized), substituting Roman numerals with Arabic numerals, decapitalizing the first word if capitalized. All names are further tokenized, and checked against a dictionary of close to 100K English nouns. Names that are found in this dictionary are simply filtered out. We also ignore all ambiguous names (i.e. names corresponding to more than one gene identifier). The remaining non-ambiguous names are added to the final gene dictionary, which is implemented as a trie-like structure in order to allow a fast lookup of gene IDs based on the associated normalized sequences of tokens.

Each abstract from the HPRD corpus is tokenized and segmented in sentences using the OpenNLP[2] package. The resulting sentences are then annotated by traversing them from left to right and finding the longest token sequences whose normal forms match entries from the gene dictionary.

## 6 Experimental Evaluation

The main purpose of the experiments in this section is to compare the performance of the following four methods on the task of corpus-level relation extraction:

1. Sentence-level relation extraction followed by the application of an aggregation operator that assembles corpus-level results (**SSK.Max**).
2. Pointwise Mutual Information (**PMI**).
3. The integrated model, a product of the two base models (**PMI.SSK.Max**).
4. The hypergeometric co-citation method (**HG**).

## 7 Experimental Methodology

All abstracts, either from the HPRD corpus, or from the entire Medline, are annotated using the dictionary-based approach described in Section 5.2. The sentence-level extraction is done with the subsequence kernel (SSK) approach from (Bunescu and Mooney, 2005), which was shown to give good results on extracting interactions from biomedical abstracts. The subsequence kernel was trained on a set of 225 Medline abstracts which were manually

[1] URL: http://www.ncbi.nih.gov

[2] URL: http://opennlp.sourceforge.net

annotated with protein names and their interactions. It is known that PMI gives undue importance to low frequency events (Dunning, 1993), therefore the evaluation considers only pairs of genes that occur at least 5 times in the whole corpus.

When evaluating corpus-level extraction on HPRD, because the "quasi-exact" list of interactions is known, we report the precision-recall (PR) graphs, where the precision (P) and recall (R) are computed as follows:

$$P = \frac{\#\textit{true interactions extracted}}{\#\textit{total interaction extracted}}$$
$$R = \frac{\#\textit{true interactions extracted}}{\#\textit{true interactions}}$$

All pairs of proteins are ranked based on each scoring method, and precision recall points are computed by considering the top $N$ pairs, where $N$ varies from 1 to the total number of pairs.

When evaluating on the entire Medline, we used the shared protein function benchmark described in (Ramani et al., 2005). Given the set of interacting pairs recovered at each recall level, this benchmark calculates the extent to which interaction partners in a data set share functional annotation, a measure previously shown to correlate with the accuracy of functional genomics data sets (Lee et al., 2004). The KEGG (Kanehisa et al., 2004) and Gene Ontology (Ashburner et al., 2000) databases provide specific pathway and biological process annotations for approximately 7,500 human genes, assigning human genes into 155 KEGG pathways (at the lowest level of KEGG) and 1,356 GO pathways (at level 8 of the GO biological process annotation).

The scoring scheme for measuring interaction set accuracy is in the form of a log odds ratio of gene pairs sharing functional annotations. To evaluate a data set, a log likelihood ratio (LLR) is calculated as follows:

$$LLR = ln\frac{P(D|I)}{P(D|\neg I)} = ln\frac{P(I|D)P(\neg I)}{P(\neg I|D)P(I)} \quad (7)$$

where $P(D|I)$ and $P(D|\neg I)$ are the probability of observing the data $D$ conditioned on the genes sharing benchmark associations ($I$) and not sharing benchmark associations ($\neg I$). In its expanded form (obtained by Bayes theorem), $P(I|D)$ and $P(\neg I|D)$ are estimated using the frequencies of interactions observed in the given data set $D$ between annotated genes sharing benchmark associations and not sharing associations, respectively, while the priors $P(I)$ and $P(\neg I)$ are estimated based on the total frequencies of all benchmark genes sharing the same associations and not sharing associations, respectively. A score of zero indicates interaction partners in the data set being tested are no more likely than random to belong to the same pathway or to interact; higher scores indicate a more accurate data set.

## 8 Experimental Results

The results for the HPRD corpus-level extraction are shown in Figure 1. Overall, the integrated model has a more consistent performance, with a gain in precision mostly at recall levels past $40\%$. The SSK.Max and HG models both exhibit a sudden decrease in precision at around $5\%$ recall level. While SSK.Max goes back to a higher precision level, the HG model begins to recover only late at $70\%$ recall.

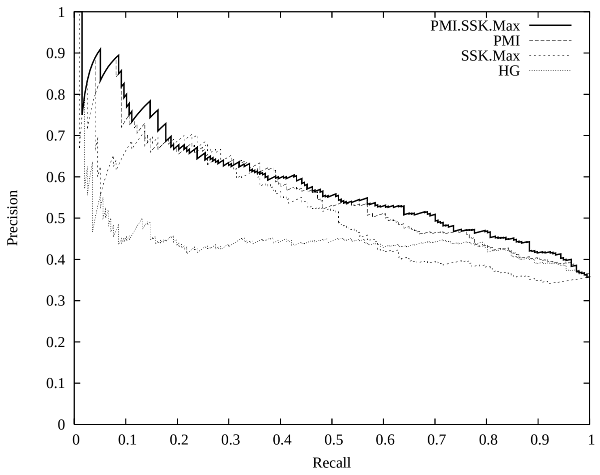


Figure 1: PR curves for corpus-level extraction.

A surprising result in this experiment is the behavior of the HG model, which is significantly outperformed by PMI, and which does only marginally better than a simple baseline that considers all pairs to be interacting.

We also compared the two methods on corpus-level extraction from the entire Medline, using the shared protein function benchmark. As before, we considered only protein pairs occurring in the same sentence, with a minimum frequency count of 5. The resulting 47,436 protein pairs were ranked according to their PMI and HG scores, with pairs that are most likely to be interacting being placed at the top. For each ranking, the LLR score was computed for the top N proteins, where N varied in increments of 1,000.

The comparative results for PMI and HG are shown in Figure 2, together with the scores for three human curated databases: HPRD, BIND and Reactome. On the top 18,000 protein pairs, PMI outperforms HG substantially, after which both converge to the same value for all the remaining pairs.

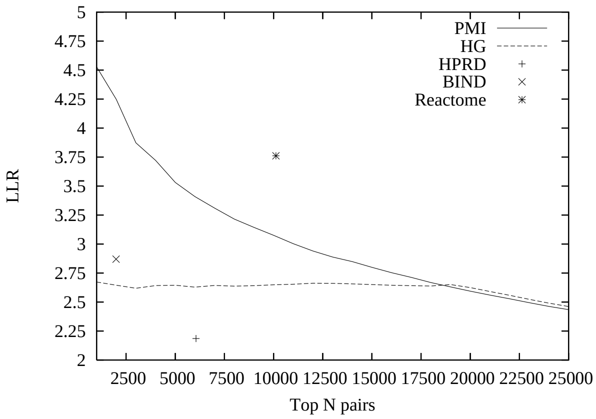


Figure 2: Functional annotation benchmark.

Figure 3 shows a comparison of the four aggregation operators on the same HPRD corpus, which confirms that, overall, *max* is most appropriate for integrating corpus-level results.

## 9 Future Work

The piece of related work that is closest to the aim of this paper is the Bayesian approach from (Skounakis and Craven, 2003). In their probabilistic model, co-occurrence statistics are taken into account by using a prior probability that a pair of proteins are interacting, given the number of co-occurrences in the corpus. However, they do not use the confidences of the sentence-level extractions. The GeneWays system from (Rzhetsky et al., 2004) takes a different approach, in which co-occurrence frequencies are simply used to re-rank the ouput from the relation extractor.

An interesting direction for future research is to

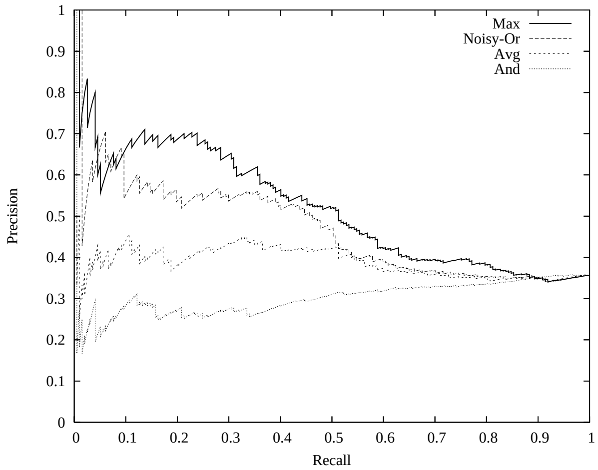


Figure 3: PR curves for aggregation operators.

design a model that takes into account both the extraction confidences and the co-occurrence statistics, without losing the probabilistic (or information-theoretic) interpretation. One could investigate ways of integrating the two orthogonal approaches to corpus-level extraction based on other statistical tests, such as chi-square and log-likelihood ratio.

The sentence-level extractor used in this paper was trained to recognize relation mentions *in isolation*. However, the trained model is later used, through the *max* aggregation operator, to recognize whether *multiple mentions* of the same pair of proteins indicate a relationship between them. This points to a fundamental mismatch between the training and testing phases of the model. We expect that better accuracy can be obtained by designing an approach that is using information from multiple occurrences of the same pair in both training and testing.

## 10 Conclusion

Extracting relations from a collection of documents can be approached in two fundamentally different ways. In one approach, an IE system extracts relation instances from corpus sentences, and then aggregates the local extractions into corpus-level results. In the second approach, statistical tests based on co-occurrence counts are used for deciding if a given pair of entities are mentioned together more often than chance would predict. We have described a method to integrate the two approaches, and given experimental results that confirmed our intuition that an integrated model would have a better performance.

## 11 Acknowledgements

This work was supported by grants from the N.S.F. (IIS-0325116, EIA-0219061), N.I.H. (GM06779-01), Welch (F1515), and a Packard Fellowship (E.M.M.).

## References

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. et al. Eppig. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29.

G. D. Bader, D. Betel, and C. W. Hogue. 2003. Bind: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250.

C. Blaschke and A. Valencia. 2001. Can bibliographic pointers for known biological data be found automatically? protein interactions as a case study. *Comparative and Functional Genomics*, 2:196–206.

Razvan C. Bunescu and Raymond J. Mooney. 2005. Subsequence kernels for relation extraction. In *Proceedings of the Conference on Neural Information Processing Systems*, Vancouver, BC.

Kenneth W. Church and Patrick W. Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Aaron M. Cohen. 2005. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Minining Biological Semantics*, pages 17–24, Detroit, MI.

Mark Craven. 1999. Learning to extract relations from MEDLINE. In *Papers from the Sixteenth National Conference on Artificial Intelligence (AAAI-99) Workshop on Machine Learning for Information Extraction*, pages 25–30, July.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28.

G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, and et al. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33 Database Issue:D428–432.

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32 Database issue:D277–280.

I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Judea Pearl. 1986. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288.

S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, and et al. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, 32 Database issue:D497–501.

A. K. Ramani, R. C. Bunescu, R. J. Mooney, and E. M. Marcotte. 2005. Consolidating the set of know human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5):r40.

Soumya Ray and Mark Craven. 2001. Representing sentence structure in hidden Markov models for information extraction. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, pages 1273–1279, Seattle, WA.

A. Rzhetsky, T. Iossifov, I. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P.A. Duboue, W. Weng, W.J. Wilbur, V. Hatzivassiloglou, and C. Friedman. 2004. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37:43–53.

Marios Skounakis and Mark Craven. 2003. Evidence combination in biomedical natural-language processing. In *Proceedings of the 3nd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD 2003)*, pages 25–32, Washington, DC.

I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. 2002. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305.

Shubin Zhao and Ralph Grishman. 2005. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 419–426, Ann Arbor, Michigan, June. Association for Computational Linguistics.