Text Empirics-Based Mining of Biomolecular Interactions from Texts: Theory and Application

Daniel Berleant[1], Lifeng Zhang[2], Jing Ding[3], Tuan Cao[2], Daniel Nettleton[2], Jun Xu[4], Andy Fulmer[4], and Eve Wurtele[3]

[1]Dept. of Information Science, University of Arkansas at Little Rock, jdberleant@ualr.edu, 501-569-3448; [2]Iowa State University, Ames; [3]Ohio State University Medical Center; [4]Procter & Gamble Co., Miami, Ohio

Abstract

Vast quantities of biomedical data exist, not only in well-structured databases, but in texts as well. Two of these that have driven considerable text mining research in biomedicine are PubMed, which serves mostly the 18-million record MEDLINE resource of abstracts and other bibliographic information, and PubMedCentral which is increasingly influential due to the US's recent policy requiring online availability of papers on research funded by its NIH funding agency. Textual information, regardless of language or national origin, presents a fundamental problem: it is in a form friendly to humans but not to computers, yet it is desirable to automatically extract and use knowledge in it nevertheless.

We describe several empirically determined facts about biomedical text passages and how they can support extracting protein interactions from biomedical texts. A system, PathBinder, built using text empirics is also presented.

Introduction

Full automated natural language understanding (NLU) remains a long-term dream. Until this dream is achieved at some undetermined time in the future, extracting knowledge automatically from texts must rely on shallower methods. Development of shallow methods is thus critical for the foreseeable future. Even should a comprehensive NLU vision ultimately be achieved, shallow methods can still make an important contribution: they comprise a source of evidence about meaning that can is computationally simpler to process and thus intrinsically faster than NLU. This means that in principle, integrating results from shallow methods run concurrently with NLU could be used to speed up NLU in real time, for example by trimming parse search spaces. The principle has been successfully applied (e.g. Frank et al. 2003). Consequently shallow methods may be expected to both be essential now, and to continue to be relevant indefinitely.

One shallow approach that has generated an extensive body of literature is statistical, corpus-based analyses. For example, machine learning can determine facts about language from tagged examples and use those facts implicitly to reach conclusions about text meaning. An alternative way to determine such facts is to investigate *text empirics*. The text empirics approach is to *manually analyze texts to*

empirically determine easily computed, useful properties, and use those properties to automatically extract knowledge from the texts. One advantage of text empirics is that the empirically determined properties can stand on their own as explicit facts about texts. Once in the public domain, such facts can be used by anyone to develop systems that extract knowledge from texts.

The best-known of the classical work on statistical properties of words and their frequencies is that of George Zipf (1935, 1945), the source of what is now called Zipf's Law (Fedorowicz 1982; Li 2003). We developed Figure 1 to test whether protein occurrences in MEDLINE are typical in conforming to Zipf's empirically derived exponential law.

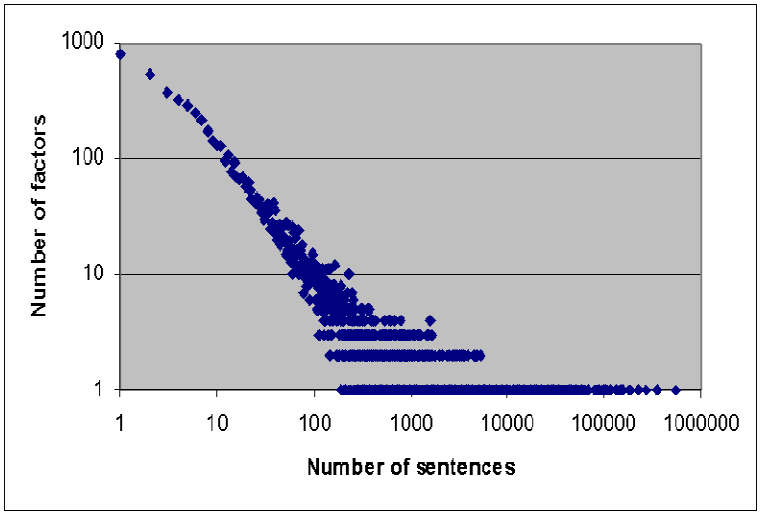

Figure 1. Analysis of MEDLINE showing the approximately exponential drop off (linear on a log-log plot in biomolecule names vs. sentences containing them. Every *x*-axis value appears at most once, but plotting may seem to make points overlap and re-occur for different *y* values.

A number of reports involve mining of protein-protein interactions from text (Blaschke et al. 1999, Donaldson et al. 2003, Humphreys et al. 2000, Marcotte et al. 2001, Ng and Wong 1999, Ono et al. 2001, Park et al. 2001, Thomas et al. 2000, Wong 2001). Available empirical information on text attributes as evidence that a passage describes a protein-protein interaction is not extensive. However a number of works with related foci describe some findings. Craven and Kumlien (1999) give a list 20 word stems and the ability of each to predict that a sentence describes the subcellular location of a protein, given that it contains the stem, a protein name, and a subcellular location term. However list content and order are noisy due to limited training data. Marcotte et al. (2001) provide a ranked list of the 20 words found most useful in identifying abstracts describing protein interactions. Results were derived from yeast-related abstracts and therefore may be yeast-specific, and the list includes words like *from* and *required* with little comment. Ono et al. (2001) assessed the abilities of four common interaction-indicating terms, each associated with a custom set of templates, to detect descriptions of protein-protein interactions. The quantitative performances of the four are hard to interpret because each used a different template set, but it is interesting that when ordered by precision their order was the same for both the yeast and *E. coli* domains, suggesting domain independence for precision. Thomas et al. (2000) proposed four categories of passages using a rule-based scoring strategy, and gave the IR performance of each category. However the set of rules is vaguely described and apparently complex, making it unclear how the results might be applied by others.

Other reports have focused on text properties with the potential for more concrete guidance in system design. Sekimizu et al. (1998) measured the IR performances of 8 interaction-indicating verbs in the context of a shallow parser. The IR capabilities of the verbs could be meaningfully compared, although the extent to which these results would apply to other passage analysis techniques or specifically to protein-protein interactions is not clear. Ding et al. (2002) found that, as vehicles for describing protein-protein interactions, sentences had slightly higher IR effectiveness than phrases despite lower precision, and considerably higher IR effectiveness than whole abstracts.

Methods

We used a body of 303 MEDLINE abstracts chosen because they matched one of ten representative queries to PubMed. Each query consisted of two protein names, and was elicited from biologists to be typical of the kinds of queries biologists are likely to make. Some further details about the corpus appear in Ding et al. (2002). Each sentence in the corpus mentioning both terms or their synonyms in the query that retrieved its containing abstract was analyzed to determine empirical facts likely to be useful for extracting protein interactions automatically.

We investigated several properties of passages (sentences and phrases) containing co-occurrences of biomolecule names. These analyses are given next.

Interaction-indicating terms. One basic characteristic is whether the passage contains an interaction-indicating term (e.g. 'regulates,' 'inhibits,' etc.) along with the biomolecules. We found that for the great majority of co-occurrences at least one interaction-indicating term was also present in the phrase (Figure 2) or sentence containing the co-occurrence. The ones that were not had low precisions: 0% of the co-occurrences were described as interacting for co-occurrences in phrases, and 8% for co-occurrences in sentences (but not a single phrase in the sentence). This was significant ($p<.001$, χ^2 test, for both phrase and sentence co-occurrences). Thus, we conclude that as a source of interactions to be mined, co-occurrences not associated with interaction-indicating terms have comparatively little to offer.

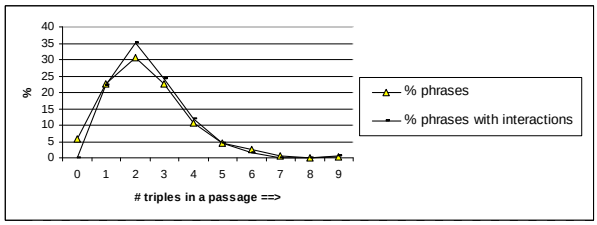

Figure 2. Percentages of the 285 phrases containing containing 1 co-occurrence and 0, 1, 2,… interaction-indicating terms. 199 described interactions.

Order of terms. A co-occurrence associated with an interaction-indicating term may have this term intervening between the co-occurring biomolecules, or in some other part of the phrase or sentence. Table 1 shows that when the interaction-indicating term intervenes, the recall is relatively high. This means that a system design would be well advised to handle such passages. On the other hand, when the interactor does not

intervene, the recall is much lower, indicating less benefit in analyzing those passages. The precision of these passages is also relatively low, so their overall effectiveness is also low. Consequently, while it would make little sense for a system design to ignore passages with co-occurrences and an intervening interaction-indicating term, ignoring passages with non-intervening interaction-indicating terms is a reasonable option.

	Interaction-indicating term intervenes	**Interaction-indicating term elsewhere**	**Interaction-indicating term anywhere**
Phrase co-occurrences	196 out of 270 *r*=0.55 *p*=0.73	63 out of 231 *r*=0.18 *p*=0.27	**259 out of 501** ***r*=0.73 *p*=0.52**
Sentence co-occurrences	77 out of 210 *r*=0.22 *p*=0.37	21 out of 219 *r*=0.059 *p*=0.096	**98 out of 429** ***r*=0.27 *p*=0.23**
All co-occurrences	**273 out of 480** ***r*=0.76 *p*=0.57**	**84 out of 450** ***r*=0.24 *p*=0.19**	**357 out of 930** ***r*=1 *p*=0.38**

Table 1. Analysis of co-occurrences with respect to recall (*r*) and precision (*p*).

Separation of co-occurring terms. Co-occurring terms can be separated by any number of intervening words, from zero up. The separation can be zero when the terms are next to each other or when intervening material is hyphen-connected to a term. For example, "…A-induced B…" is considered to have zero full words between A and B. Different separations are associated with different recalls and precisions. These recalls and precisions can impact system design in different ways.

Recall. Figure 3 shows percentages of co-occurrences (vertical axis) that have *x* words or fewer between the co-occurring terms, for each of four disjoint categories. The two curves for phrase co-occurrences level off sooner than the two for sentence co-occurrences, probably because phrases tend to be shorter than sentences.

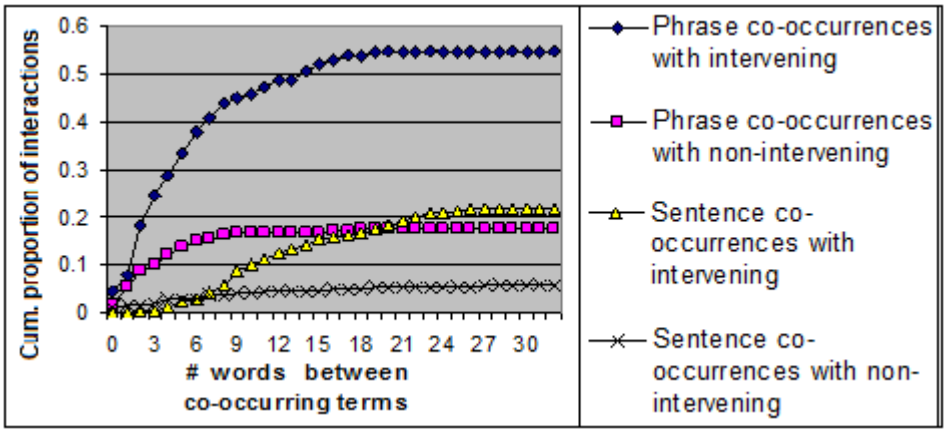

Figure 3. Cumulative interactions for four disjoint categories of co-occurrences.

Precision. We found that different separations are associated with different precisions. Thus text mining systems could use the number of intervening words to help estimate the likelihood that an interaction is described. Figure 4 shows the precisions for different separations for each of the four co-occurrence categories of Figure 3. The number of data for any given separation was often small. Therefore statistical analysis was employed to see the underlying tendency while characterizing the noise. We used

the logistic regression model to describe precision *p* as a function of separation *x*. Maximum likelihood methods (Agresti 1990 pp. 112-117) were used to obtain estimates and error bars representing plus and minus one standard error.

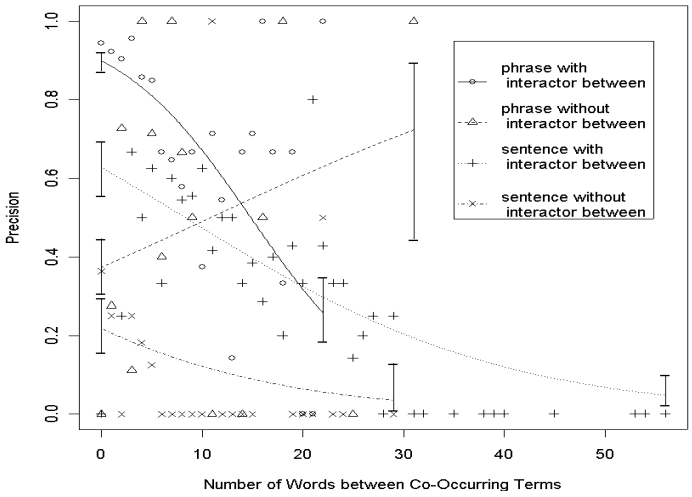

Figure 4. Precision data for four categories of co-occurrences and various separations.

Interation-indicating term category. We empirically investigated the ability of five different syntactic categories of interaction-indicating terms to predict that a passage containing such a term along with a biomolecule name co-occurrence describes an interaction. These categories were adjective, simple form verb, verb ending in –ing, past or perfect tense verb, and noun. Likelihoods for each were determined for the case of phrase co-occurrences (the biomolecules and the interaction-indicating term all in one phrase) and sentence co-occurrences (all in the same sentence but not in the same phrase). Detailed figures will be given later in Tables 2 and 3 when they are used.

Using text empirics to assess phrases and sentences. We expanded the analysis above, quantifying as needed to get, for a given phrase or sentence containing two given biomolecules, the probability that it describes an interaction between the biomolecules based on its characteristics. This was a two stage process.

- First, we identified *specific characteristics* and, for each, the conditional probability that an interaction is described based on just that characteristic: $p(\text{interaction} \mid f_i)$ for text feature i. For example, f_i might be some number of words intervening between the two biomolecules, 0 if the biomolecule names were next to each other, 1 if there was one word between the biomolecule names, and so on.
- Second, we combined the evidence provided by the various characteristics whose conditional probabilities were separately determined. The combination method assumed the different sources of evidence were independent. In the text mining domain this assumption is commonly made. While universally acknowledged to only approximate the true situation, it has been found

frequently useful in practice. This was verified in the present work, as discussed later.

The Semi-Naïve Evidence Combination Model (see Appendix for a more complete discussion if desired) is the basis of our algorithm for evaluating the probability that a sentence describes an interaction between two given biomolecules. The formula in the Semi-Naïve evidence combination model is $o(interaction \mid f_1,...,f_n) = o_1 o_2 ... o_n / o_p^{n-1}$, where $o(interaction \mid f_1,...,f_n)$ describes the odds that a passage describes an interaction if it has features f_1 through f_n, o_k are the odds that a passage with feature k describes an interaction, and the *prior odds* (i.e. over all cases irrespective of their features) are o_p. The use of odds here instead of probabilities makes the formula simpler in appearance, but is otherwise unimportant because odds and probabilities are easily converted: $p=o/(o+1)$, and $o=p(1-p)$. The appendix explains Semi-Naïve Evidence Combination.

In detail, given a sentence containing a co-occurrence of two biomolecules, here is how the probability that it describes an interaction between them is estimated. (Special cases are explained following.)

1) Determine the odds that an interaction is described using Semi-Naïve Evidence Combination on evidence based on the locations of given biomolecule pair in the sentence. Independently, determine the odds based on the presence and morphological form of an interaction-indicating term using Semi-Naïve Evidence Combination.
2) Convert the two odds o to probabilities p according to $p=o/(1+o)$.
3) Combine the probabilities according to the standard formula p (interaction)$=1 - (1- p_1)(1-p_2)$.

Special cases. If a sentence contains multiple co-occurrences (example: biomolecule A occurs once and B occurs twice, so there are two AB co-occurrences), calculate the probability for each and use the highest. We do not combine the probabilities, because we have observed that multiple co-occurrences do not necessarily improve the probability that a sentence describes an interaction, as the co-occurrences do not provide independent evidence that an interaction is described.

Analogous to the possibility of multiple co-occurrences, multiple interaction-indicating words may be present. For the same reasons as for multiple co-occurrences, multiple interaction-indicating terms are handled by calculating the evidence provided by each, using the one that is best, and discarding the rest.

Using Semi-Naïve Evidence Combination.

1) As noted, we treated evidence related to co-occurring term location in a sentence separately from evidence related to interaction-indicating terms. Other system builders may wish to treat them together as it is not a requirement that it be done one way or the other.
2) *Odds from co-occurring term location*.
    a) If the co-occurrence is within a phrase:
        i. If no interaction-indication term is in the phrase, estimate p=0.1.

ii. If an interaction-indicating term is in the phrase:
    1. If there is 1 co-occurrence in the phrase, estimate o_{2aii1}=0.7/0.3=2.33.
    2. If there is >1 co-occurrences in the phrase, estimate o_{2aii2}=0.86/0.14=6.1.
    3. If there is not an interaction-indicating term between the co-occurring terms, estimate o_{2aii3}=0.24/0.76=0.316, except if separation=0. In that case, o_{2aii3}=1/9 (as an estimate of 0+ε, for small but unknown ε).
    4. If there is an interaction-indicating term between the co-occurring terms, and separation>0, estimate o_{2aii4}=(-0.03*k*+0.9)/(1-(-0.03*k*+0.9) =(-0.03*k*+0.9)/(0.1+0.03*k*)), where *k* is the number of words between the co-occurring terms. However, if this is below 0, set o_{2aii4}=0. If separation=0 then o_{2aii4}=17.
    5. Let prior odds o_{phrase}=0.68.
    6. Compute the product of all the $o_{2aii\_}$ that apply.
    7. Divide by o_{phrase}^{n-1} where *n* is the number of $o_{2aii\_}$ that apply.

b) If the co-occurrence is within a sentence but not on one phrase:
    1. If the sentence has 1 co-occurrence, estimate o_{2b1}=0.4/0.6=0.67.
    2. If the sentence has >1 co-occurrences, estimate o_{2b2}=0.32/0.68=0.47.
    3. If there is an interaction-indicating term between the co-occurring terms, and separation>0, estimate o_{2b3}=(-0.01*k*+0.6)/(1+0.01*k*-0.6) =(-0.01*k*+0.6)/(0.4+0.01*k*) where *k* is the number of words between the co-occurring terms. However, if this is below 0, set o_{2b3}=0). If separation=0, then o_{2b3}=1/9 (this is an estimate of 0+ε for small but unknown ε).
    4. If there is not an interaction-indicating term between the co-occurring terms, and separation>0, estimate o_{2b4}=(-0.0033*k*+0.2)/(0.8+0.0033*k*) =(-0.0033*k*+0.2)/(0.8+0.0033*k*) where *k* is the number of words between the co-occurring terms. However, if this is below 0, set o_{2b4}=0). If separation=0, then o_{2b4}=4/7.
    5. Let prior odds $o_{sentence}$=0.33.
    6. Compute the product of all the $o_{2b\_}$ that apply.
    7. Divide by $o_{sentence}^{n-1}$ where *n* is the number of $o_{2b\_}$ that apply.

3) To calculate the odds *o*(co-occurrence *i* is part of an interaction description) from interaction-indicating term evidence, do the following.

a. If the co-occurrence is within a phrase:
   i. If no interaction-indicating term is in the phrase, return no value.
   ii. If there are interaction-indicating term(s) in the phrase, find odds o_{3aii} based on Table 2 for each one, and return the highest.

b. If the co-occurrence is within a sentence (but not the same phrase):
   i. If no interaction-indicating term is in the sentence, return no value.
   ii. If there are interaction-indicating term(s) in the sentence, for each term, find its odds O_{3bii} based on Table 3. Then return the highest one found.

Form	Odds
noun	1.902
adj	0.75
simple	2.818
-ing	1.231
past/perfect	1.867

Table 2. Interaction-indicating term form odds for phrases.

Form	Odds
noun	1.469
adj	0.818
simple	1.923
-ing	1.029
past/perfect	1.203

Table 3. Interaction-indicating term form odds in sentence.

4) *Odds normalization*. Our corpus was made of sentences containing pairs of biomolecules known to interact, whether or not a given sentence describes them as interacting. Thus the odds obtained from analysis of these sentences and there values used in steps 2) and 3) above deviate from actual odds that should be used by systems that automatically assess sentences containing arbitrary co-occurring biomolecules. The reason is that arbitrary co-occurring biomolecules might not interact, in which case the passage containing them is unlikely to say they do regardless of its characteristics.

To better understand this, we created another corpus made of 300 sentences containing at least two biomolecules other from those in the corpus analyzed earlier. The default odds for these sentences of describing an interaction were $o_{background} = 0.723$, equivalent to probability $p_{background} = 0.419$.

Theory indicates (and preliminary testing confirmed) that results are better if the odds given in steps 2) and 3) above are normalized. Based on normalizing their corresponding probabilities by

$p_{normalized}=p_{unnormalized} \times p_{background}$, and because

$o_{normalized}= p_{normalized} /(1- p_{normalized})$,

algebra gives the corresponding formula for normalizing an odds o_m as

$o_{normalized}= o_{normalized} \times o_{background} / (1+ o_{normalized} + o_{background})$.

Results

This interaction network has been applied in our software system, PathBinder, which serves as a query gateway to users. If users provide a biomolecule, PathBinder can

find other biomolecules likely to interact with it. Users can also choose a biomolecule pair as a query, as illustrated in Figure 5. Much of the database was populated by processing MEDLINE. Once PathBinder gets a query, it searches the database for sentences satisfying the query and displays them to users in a new window (Figure 6). It can rank the result sentences by their assessed probability of describing an interaction, or by PubMed's PMID number. Users can click the PMID and the browser will go directly to PubMed to display the relevant record.

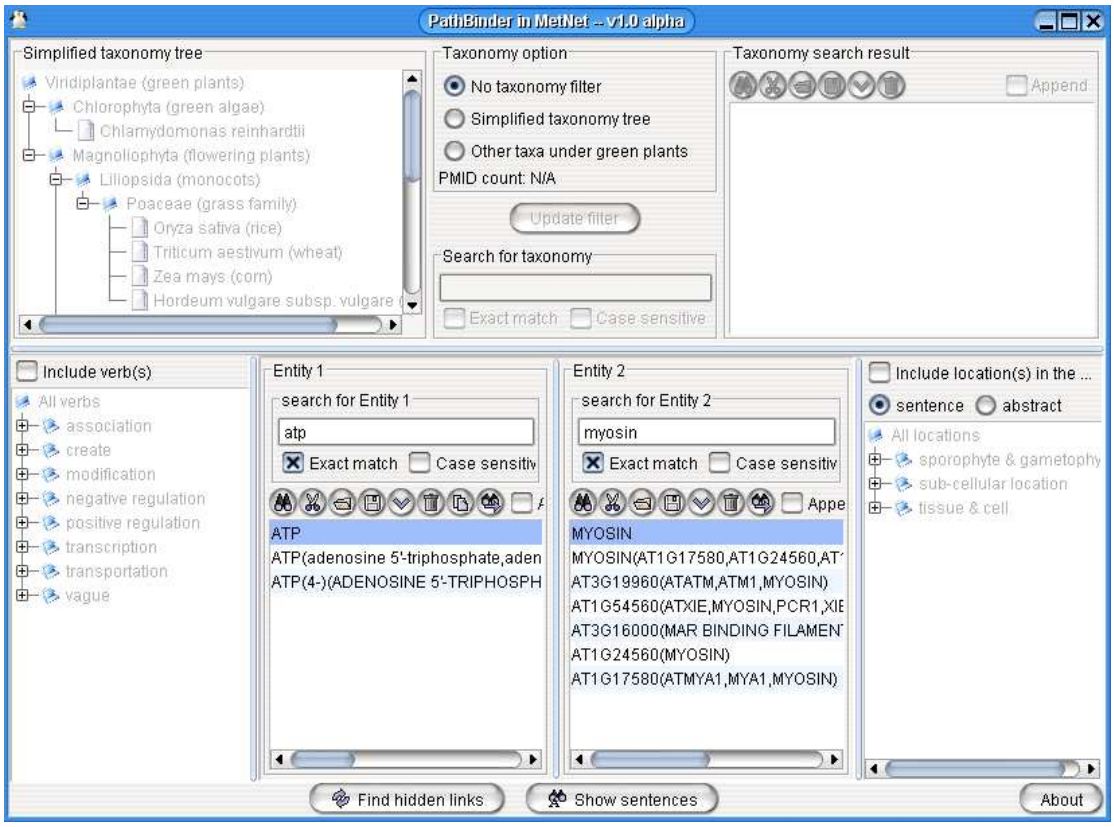

Figure 5. PathBinder Main Interface

Training set analysis. We tested the approach by taking sentences that were assessed a probability of close to 0.5 and manually tagging them as in fact describing an interaction or not. (If half of them described an interaction and half did not, then the assessed probability of 0.5 would be accurate.) We did this for sets of sentences from the training set assessed at other probabilities also, covering probabilities 0, 0.1, 0.2,…0.7 (see Figure 7). There were no sentences assessed at 0.8 or above. A linear regression analysis produced the top curve in Figure 7. It clearly diverges, though only modestly, from the theoretical ideal of $y = x$, also shown in Figure 7 for comparison. This reflects sources of error in the Semi-Naïve Evidence Combination approach to assessing the probabilities, such as the independence assumption it uses when combining evidence. We then shifted the modestly divergent curve arithmetically with a basic algebraic transformation, by incorporating adjustment factors for slope and y-axis intercept to move it (and therefore every point on it) so that it exactly matched the ideal $y = x$ curve. This shifting process adjusts the probability assessment for any

given sentence, which now is derived from Semi-Naïve Evidence Combination followed by application of the adjustment factors.

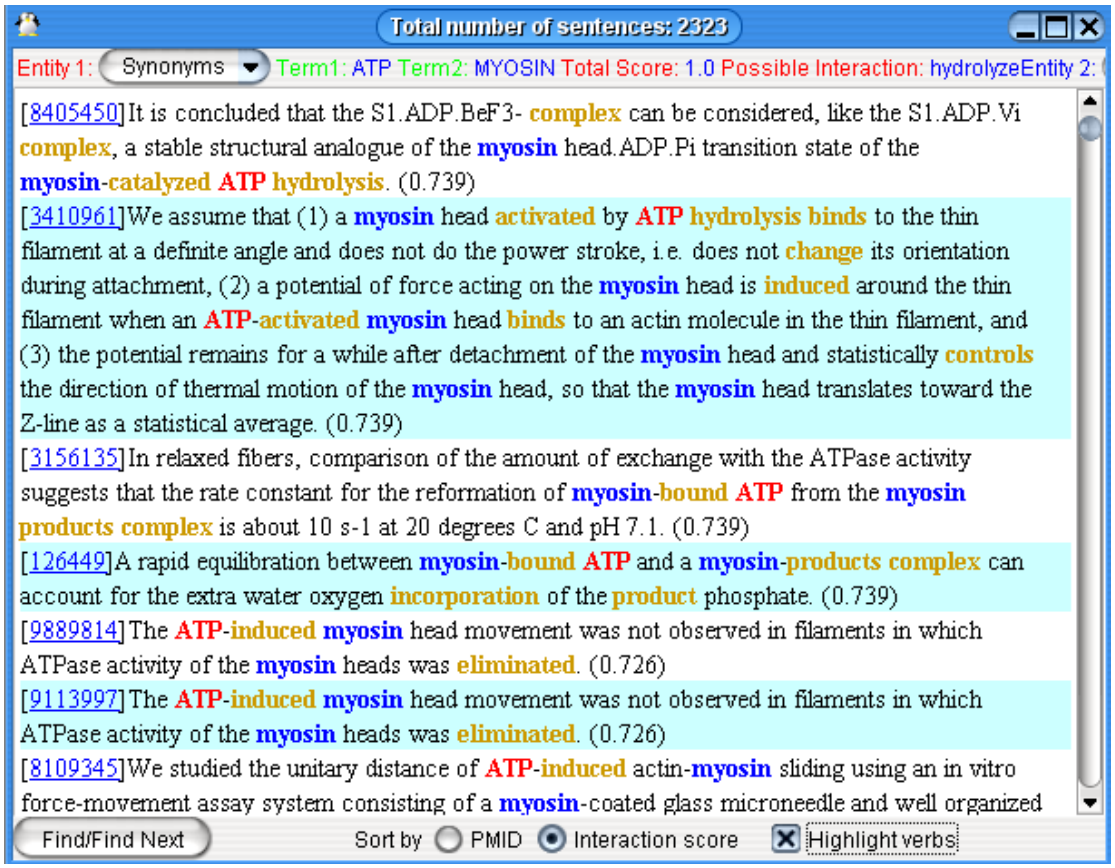

Figure 6. PathBinder Search Results Window.

Test set analysis. 600 new sentences were obtained. 123 of them contained the same biomolecule name co-occurrences as were in the training set, while the other 477 only contained other biomolecule co-occurrences. The assessed probabilities, adjusted as in the previous paragraph, were chosen pseudorandomly to be distributed over the range of interest (0 to 0.7). Then these sentences were manually analyzed to see if they really did describe the interaction or not, the regression line for this data obtained, and plotted in Figure 7. Note that the regression curve almost exactly matches $y = x$, indicating successful application of the method to the newly obtained test set.

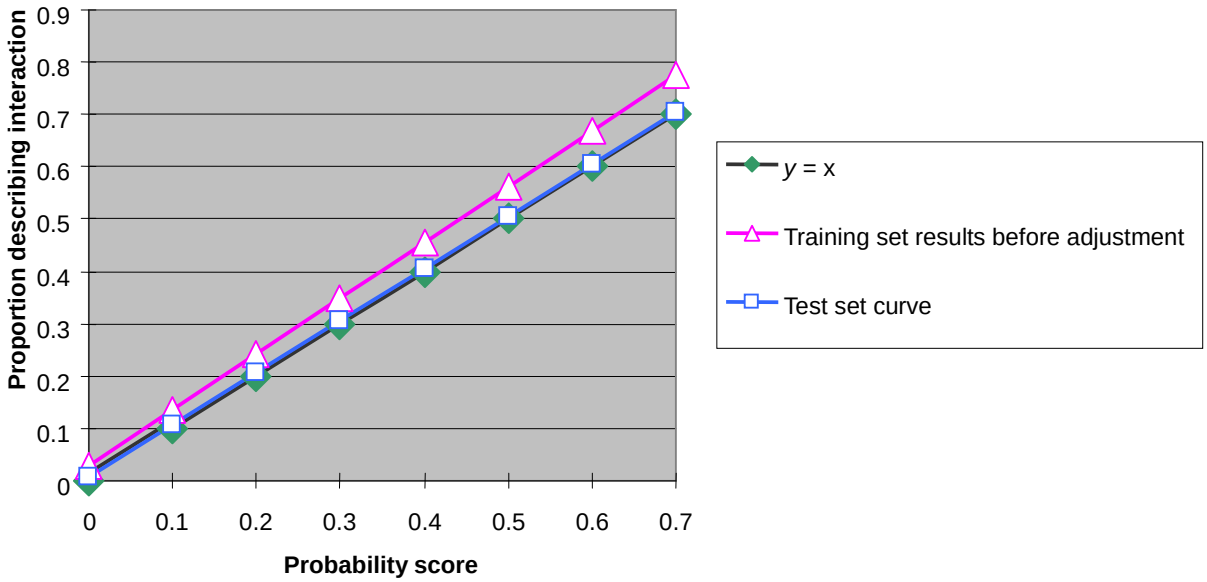

Figure 7. The comparison among (i) $y = x$, (ii) the unadjusted regression line (not from the test set), and (iii) the regression line for the test set after it was adjusted based on the discrepancy between (i) and (ii).

References

Agresti A (1990), *Categorical Data Analysis*, Wiley, New York.

Blaschke C, M Andrade, C Ouzounis, and A Valencia (1999), Automatic extraction of biological information from scientific text: protein-protein interactions, *AAAI Conference on Intelligent Systems in Molecular Biology*, 60-67.

Craven M and J Kumlien (1999), Constructing biological knowledge bases by extracting information from text sources, *Proc. 7th Int. Conf. on Intelligent Systems for Molecular Biology* (*ISMB*), 77-86.

Ding J, D Berleant, D Nettleton, and E Wurtele (2002), Mining MEDLINE: abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing* **7** (*PSB*), Kaua'i, Hawaii, 326-337.

Donaldson I, J Martin, B de Bruijn, C Wolting, V Lay, B Tuekam, S Zhang, B Baskin, GD Bader, K Michalickova, T Pawson, and CW Hogue (2003), PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine, *BMC Bioinformatics* **4** (11), www.biomedcentral.com/1471-2105/4/11.

Fedorowicz J (1982), A Zipfian model of an automatic bibliographic system: an application to MEDLINE, *J Am Soc Inf Sci.*, **33** (4):223-232.

Frank M, B Becker, B Crysmann, B Kiefer, and U Schäfer (2003), Integrated shallow and deep parsing: TopP meets HPSG, *Proceedings of the ACL 2003*, Sapporo, Japan.

Humphreys K, G Demetriou, and R Gaizauskas (2000), Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures, *Pacific Symposium on Biocomputing* **5**, 502-513.

Lewis, D (1998), Naïve Bayes at forty: the independence assumption in information retrieval. in *Conf. Proc. European Conference on Machine Learning*, Chemnitz, Germany.

Li W, Zipf's Law, http://www.nslij-genetics.org/wli/zipf/.

Marcotte EM, I Xenarios, and D Eisenberg (2001), Mining literature for protein-protein interactions, *Bioinformatics* **17** (4):359-363.

Ng S-K and M Wong (1999), Toward routine automatic pathway discovery from on-line scientific text abstracts, *Proc. 10th Workshop on Genome Informatics*, 104-112.

Ono T, H Hishigaki, A Tanigami, and T Takagi (2001), Automated extraction of information on protein-protein interactions from the biological literature, *Bioinformatics* **17** (2):155-161.

Park JC, HS Kim, and JJ Kim (2001), Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar, *Pacific Symposium on Biocomputing* **6**, 396-407.

Sekimizu T, HS Park, and J Tsujii (1998), Identifying the interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts, *Proc. 9th Workshop on Genome Informatics*, 62-71.

Thomas J, D Milward, C Ouzounis, S Pulman, and M Carroll (2000), Automatic extraction of protein interactions from scientific abstracts, *Pacific Symposium on Biocomputing* **5**, 538-549.

Wong L, (2001), A protein interaction extraction system, *Pacific Symposium on Biocomputing* (*PSB*) **6**.

Zipf GK (1935), *The Psychobiology of Language*, Houghton Mifflin.

Zipf GK (1945), The repetition of words, time-perspective, and semantic balance, *The Journal of General Psychology* **32**: 127-148.

Appendix: Semi-Naïve Evidence Combination

[A detailed discussion of this model was submitted as a supplementary document.]