Classifying Verbs in Biomedical Text Using Subject-Verb-Object Relationships

Pieter van der Horn and **Bart Bakker** and **Gijs Geleijnse** and **Jan Korst**
Philips Research Laboratories
High Tech Campus 12a, 5656 AE Eindhoven, The Netherlands
{pieter.van.der.horn,bart.bakker,gijs.geleijnse,jan.korst}@philips.com

Sergei Kurkin
BioFocus DPI, a Galápagos company
Darwinweg 24, P.O. Box 127, 2300 AC Leiden, The Netherlands
sergei.kurkin@glpg.com

Abstract

A protein-protein interaction in a biomedical text is often described using a wide range of verbs, e.g. *activate, bind, interact*. In order to determine the specific type of interaction described, we must first determine the meaning of the verb used. In biomedical context, however, some verbs can be considered synonyms, yet may not be so in standard lexical databases, like WordNet. Furthermore, some verbs will not be mentioned at all in such a dictionary, since they are too area specific. We propose a simple classification scheme to predict the correct class (meaning) of the verb. With this, one can identify the types of protein-protein interactions described in subject-verb-object constructions in PubMed abstracts.

1 Introduction

Since scientific journals are still the most important means of documenting biological findings, biomedical articles are the best source of information we have on protein-protein interactions. The mining of this information will provide us with specific knowledge of the presence and types of interactions, and the circumstances in which they occur.

There are various linguistic constructions that can describe a protein-protein interaction. (Tateisi et al., 2004) use predicate-argument structures in the mining of protein-protein interactions. These are used to identify the specific roles of encountered proteins in an interaction, but not to determine the biomedical meaning of the verb itself. In (Wattarujeekrit et al., 2004), an extended model based on PropBank (Palmer et al., 2005) is used to group verbs according to their differences and similarities in sense, structure and number of arguments between their use in biomedical and regular text. Their main focus is domain-specific verbs that are used to describe molecular events in biology. We share their considerations about the fact that verbs are used in a different way in biomedical text compared to regular text. Our goal, however, is to group general verbs according to their meanings in biomedical text. This will allow us to identify the type of interaction indicated by any possible verb encountered, instead of having to rely on a limited number of predefined domain-specific verbs. Following (Jensen et al., 2006), we focus on causality to create a biologically meaningful distinction. We use two classes of verbs, making the distinction between relations that describe proteins *affecting* other proteins (*causal relation*) and any other relation (*non-causal relation*). Future work will incorporate more classes in order to be able to make a more specific distinction between different meanings of verbs.

2 Preprocessing

The protein-protein interactions we are interested in are described in the subject, the object and the interlinking verb phrase of a sentence. To determine which parts of the sentence make up this construction, we need to preprocess the sentence. For this, we use the Genia Chunker[1] to break the sentence into different chunks (in particular we are interested in noun phrases and verb phrases). We combine this information with the result of the Stanford Dependency Parser[2] to determine how these different chunks (phrases) are connected to

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/
[2] http://nlp.stanford.edu/downloads/lex-parser.shtml

each other.

Using WordNet (Fellbaum, 1998), we can increase the number of verbs for which we know (or can reasonably assume) the right class. WordNet identifies synonyms for each verb, grouped by different senses (meanings) that are ordered by frequency (most common meaning first). We can choose how many senses we use (at least one), and how many recursive levels of synonyms we want (synonyms, synonyms of those synonyms, etc). However, this can create noise, since WordNet is a lexicon for the general use of words, and not specifically for biomedical context (Poprat et al., 2008). Lacking a proper biomedical lexicon, we will make limited use of WordNet in order to test our approach.

3 Classification

The subject-verb-object construction can be schematically represented as follows:

[(state of) protein] [verb] [(state of) protein]

We assume the interaction between the two proteins to be determined by the combination of the states in the noun phrases and the relevant verb in the verb phrase. Such states can be described by single words (e.g. *activation*, *suppression*, *overexpression*) or far more complicated descriptions. However, detection of these descriptions of states of proteins can be difficult and is a separate research topic. Since the focus of this paper is on the meanings of verbs, we will leave this detection of protein states for future work.

We make a distinction between two classes of verbs. One class describes a strict *causal relation* and the other covers all other types of meanings (*non-causal*). Table 1 shows some example verbs for the two classes.

Class	**Examples**
causal	*activate, inhibit, cause*
non-causal	*interact, require, bind*

Table 1: Two classes of verbs.

The second class includes not just verbs that describe a correlation (e.g. *interact*), but also verbs such as *require* and *bind*. One could argue that these latter verbs also describe a directed action from agent to target, like a strict causal relation does. However, they do not describe a direct change of the state of the target protein, and therefore we choose not to put them in the first class. The three verbs in the *causal* class represent the positive, negative and general causal relations. The three verbs in the *non-causal* class represent three different types of relations that occur very often in the text. Since these relations are not synonymous to each other, each of them has to be represented by a separate verb. Having labeled these six verbs manually, we will use this to attempt to automatically predict the right class for the possibly many unknown verbs that can occur in subject-verb-object constructions in biomedical text.

3.1 Naive Bayesian Classifier

Using a Naive Bayesian Classifier, we estimate the probability that a given verb belongs to a certain class. Bayes' Theorem describes this probability.

$$P(c_i|V) = \frac{P(c_i) \cdot P(V|c_i)}{P(V)} \tag{1}$$

In the retrieved subject-verb-object constructions, a verb V will occur a number of times, each time in combination with a specific ordered pair of proteins pp_j, one in the subject and one in the object.

$$V \equiv \{pp_1, pp_2, ..., pp_n\}$$

These pairs of proteins are the different features of this verb. In Naive Bayesian Classification, these features are assumed to *independently* contribute to the estimate of the posterior probability.

$$\begin{aligned} P(c_i|V) &= P(c_i|pp_1, pp_2, ..., pp_n) \\ &= \frac{P(c_i) \cdot \prod_{j=1}^{n} P(pp_j|c_i)}{P(pp_1, pp_2, ..., pp_n)} \end{aligned} \tag{2}$$

Given a test set of instances (in Section 4 we elaborate on how we get those instances), we define the following variables:

$f_{j,i}$	*number of occurrences of protein pair pp_j around verbs of class c_i (frequency)*
$Q_i = \sum_j f_{j,i}$	*number of protein pairs around verbs of class c_i*
$Q = \sum_i Q_i$	*total number of protein pairs encountered*
U	*number of unique protein pairs encountered in the training set*

With these, we can estimate the necessary probabilities:

$P(c_i) \cong \frac{Q_i}{Q}$ *prior probability of class* c_i

$P(pp_j|c_i) \cong \frac{f_{j,i}+1}{Q_i+U}$ *conditional probability of pair* pp_j *given class* c_i

For the conditional probability, we use Laplace estimates. That is, we add 1 to the numerator and U to the denominator, in order to compensate for pairs for which $f_{j,i} = 0$. If we would use $P(pp_j|c_i) \cong \frac{f_{j,i}}{Q_i}$ instead, the conditional probability would become equal to 0 if $f_{j,i} = 0$. This would cause the posterior probability $P(c_i|pp_1, pp_2, ..., pp_n)$ to be equal to 0 as well (Equation 2), leaving us without a reasonable estimate of this posterior probability. The probability $P(V)$ is the factor with which we normalize the numerator of Equation 2 for each class c_i. This gives us $P(c_i|V)$ for each class. A verb V is then classified to be in the class c_i for which the posterior probability $P(c_i|V)$ is highest.

$$C(V) = \underset{c_i}{\operatorname{argmax}}\ P(c_i|V)$$

4 Experiments

4.1 Setup

In order to test our approach, we retrieved a set of subject-verb-object relations from abstracts stored in PubMed. We chose to test our approach on yeast proteins rather than e.g. human proteins to avoid Named Entity Recognition problems. We used a predefined data set of names to detect yeast proteins in text.

To remove any excess information, the verb phrases are normalized. We assume the last verb in the phrase to be the relevant verb and check the direction of the relation (active or passive form of that verb). Finally, the verb is stemmed using the Porter stemmer (Porter, 1980). For those verbs that are in the passive form, the order of the protein pairs around it was reversed, and, for simplification, verb phrases that describe a negation were removed. More than one protein can occur in the subject and/or object, so we count each possible pair as an occurrence around the particular verb.

We used the 6 verbs as shown in Table 1 as a starting set to test the classifier. The training set is then augmented using WordNet. For the resulting verbs in the classes, we run a leave-one-out cross validation. That means, we classify each of these verbs by training the Naive Bayesian Classifier on the frequencies of the occurring pairs of proteins around the other known verbs. Some verbs we retrieved from WordNet may not occur at all in the subject-verb-object instances we have. These verbs are ignored in the leave-one-out cross-validation.

4.2 Results

	V	*C*	*A*	*P*
no WN	6	3	0.50	0.66
l1/s1	13	7	0.54	0.50
l1/s2	18	13	0.72	0.05
l1/sa	19	14	0.74	0.03
l2/s1	19	12	0.63	0.18
l2/s2	27	21	0.78	2.96E-3
l2/sa	55	32	0.58	0.14
l3/s1	26	20	0.77	4.68E-3
l3/s2	42	35	0.83	7.55E-6
l3/sa	73	43	0.59	0.08

Table 2: Leave-one-out cross-validation results.

Table 2 shows the results of the different tests, using different parameter settings in WordNet to augment the training set. They contain the number of verbs classified in the leave-one-out cross-validation (*V*), the number of verbs that were correctly classified (*C*), the accuracy ($A = \frac{C}{V}$) and the probability P that a random classifier would perform as good or better than this classifier, given by

$$P = \sum_{i=C}^{V} \binom{V}{i} p^i \cdot (1-p)^{V-i}$$

in which $p = \frac{1}{2}$ (determined by the number of classes). We have run the program with different settings for WordNet ('l1' means recursive level 1, 's2' means WordNet senses 1 to 2, 'sa' means all WordNet senses are taken).

From the cross-validations, we can see that the algorithm performs reasonably well. There are multiple settings that obtain an accuracy higher than 0.70, and one setting in particular ('l3/s2') reached an accuracy of 0.83. The probability that a random classifier would perform as good or better than this is $7.55 \cdot 10^{-6}$. In Table 3, the results of the cross-validation for this setting are shown for each of the 42 verbs, highest P_1 first (P_1 is

verb	P_1	error	verb	P_1	error
suppress	1.00	0	turn	0.68	0
have	1.00	0	position	0.67	1
lead	1.00	0	perform	0.59	0
activate	1.00	0	make	0.52	0
stimulate	1.00	0	comprise	0.51	0
reduce	1.00	0	investigate	0.49	0
cause	1.00	0	see	0.49	0
contain	1.00	0	incorporate	0.49	1
induce	1.00	0	do	0.49	1
repress	1.00	0	impact	0.49	0
allow	0.99	0	pull	0.49	0
control	0.99	0	situate	0.49	0
inhibit	0.97	0	displace	0.49	1
trigger	0.95	0	hold	0.49	1
give	0.85	0	attach	0.35	0
carry	0.81	0	occupy	0.32	0
keep	0.80	0	need	0.18	0
expect	0.79	1	involve	0.06	0
maintain	0.78	0	bind	2.64E-15	0
bear	0.75	0	interact	5.74E-30	0
affect	0.75	1	require	1.12E-54	0

Table 3: Cross-validation of 42 verbs.

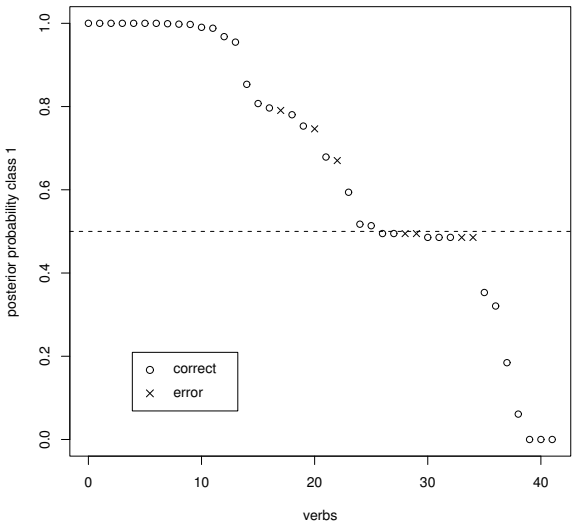

Figure 1: Results of leave-one-out cross-validation.

the posterior probability that a verb belongs to class 1, the *causal* class). Figure 1 visualizes the distances of the classifications from the decision boundary. The crosses indicate the errors made. The confidence of the classification is defined by distance of the posterior probability to the decision boundary. This confidence clearly differs for each verb. We can see that there is a group of 11 verbs for which the confidence is very low (*make* to *hold*). This group accounts for four of the errors made, out of a total of seven. For these 11 verbs it is unclear, even for humans, which class they belong to. Some of these verbs, however, may not describe any interaction at all. One could use a confidence threshold to discard the verbs of which the classifications are very uncertain.

5 Conclusions and Future Work

Given an appropriate set of known verbs, we can predict the meanings of unknown verbs with reasonable confidence. This automatic prediction is very useful, since it is infeasible to manually determine the meanings of all possible verbs. We chose to use a two-way distinction as a first step. Verbs like *require* and *bind* describe biologically distinct interactions however, and preferably should be put into classes separate from general correlations. In order to create a more detailed network of interacting proteins, one can take these other types into account as well.

Furthermore, it would be useful to separate the causal relationship into positive and negative relations. This specific distinction however is not just described in the connecting verb, but also in possible state descriptions in the noun phrases. Further research is necessary to extract these descriptions from the text. Finally, it would be useful to look at different syntactic constructions, other than just subject and object.

References

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Lars J. Jensen, Jasmin Saric, and Peer Bork. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2):119–129.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics*, (31):71–105.

M. Poprat, E. Beisswanger, and U. Hahn. 2008. Building a biowordnet using wordnet data structures and wordnet's software infrastructure - a failure story. In *ACL 2008 workshop "Software Engineering, Testing, and Quality Assurance for Natural Language Processing"*.

M.F. Porter. 1980. An algorithm for suffix stripping. In *Program*, volume 14, pages 130–137.

Y. Tateisi, T. Ohta, and J. Tsujii. 2004. Annotation of predicate-argument structure on molecular biology text. In *IJCNLP-04*.

T. Wattarujeekrit, P. K. Shah, and N. Collier. 2004. Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5, October.