Determining causal and non-causal relationships in biomedical text by classifying verbs using a Naive Bayesian Classifier

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

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, but in this paper we will focus on subject-verb-object constructions. If a certain protein is mentioned in the subject of a sentence, and another protein in the object, we assume in this paper that some interaction is described between those proteins. The verb phrase that links the subject and object together plays an important role in this. However, there are a great many different verbs in the English language that can be used in a description of a protein-protein interaction. Since it is practically impossible to manually determine the specific biomedical meanings for all of these verbs, we try to determine these meanings automatically. We define two classes of protein-protein interactions, *causal* and *non-causal*, and using a Naive Bayesian Classifier, we predict for a given verb in which class it belongs. This process is a first step in automatically creating a useful network of interacting proteins out of information from biomedical journals.

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 each other.

3 Classification

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

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

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.

Since we leave out the information of the states of the proteins in this work, the first class covers positive, negative and neutral causal relations. The second class includes not just verbs that describe a correlation (*interact*), but also verbs such as *require*

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

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 112–113,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

and *bind* that describe a biologically important relationship, but not specifically a causal one.

We use a Naive Bayesian Classifier to estimate the probability $P(c_i|V)$ that a given verb belongs to a certain class. In the retrieved subject-verb-object constructions, such 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. Each pair pp_j independently contributes to the estimation of $P(c_i|V)$.

$$V = \{pp_1, pp_2, ..., pp_n\} \quad (1)$$

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

4 Experimental results

To test our approach, we retrieved a set of subject-verb-object relations from PubMed. We choose to test our approach on yeast proteins rather than e.g. human proteins to avoid Named Entity Recognition problems.

To get rid of 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. 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. They represent the different types within each class, and of these it is clear they belong in that specific class. By using WordNet[3] we can augment this set. Table 1 shows the results of the different tests, using different parameter settings in WordNet to augment the training set ('l1' means recursive level 1, 's2' means WordNet senses 1 to 2, 'sa' means all WordNet senses are taken). It contains the number of verbs classified in the leave-one-out cross validation (*V*), the number of verbs that were correctly classified (*C*), the precision ($P = \frac{C}{V}$) and the probability Q that a random classifier would perform as good or better than this classifier, given by Equation 3

[3]http://wordnet.princeton.edu/

	V	*C*	*P*	*Q*
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: Results for different settings.

$$Q = \sum_{i=C}^{V} \binom{V}{i} p^i \cdot (1-p)^{V-i} \quad (3)$$

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 used two classes of verbs, making the distinction between relations that describe proteins *affecting* other proteins (*causal relation*) and any other relation (*non-causal relation*). Verbs like *require* and *bind* describe biologically distinct interactions however, and preferably should be put into classes separate from general correlations. We chose to use a two-way distinction as a first step however, which was still biologically relevant. 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 syntactical constructions, other than just subject and object.