Extracting Protein-Protein interactions using simple contextual features

Leif Arda Nielsen
School of Informatics
University of Edinburgh
leif.nielsen@gmail.com

1 Introduction

There has been much interest in recent years on the topic of extracting Protein-Protein Interaction (PPI) information automatically from scientific publications. This is due to the need that has emerged to organise the large body of literature that is generated through research, and collected at sites such as PubMed. Easy access to the information contained in published work is vital for facilitating new research, but the rate of publication makes manual collection of all such data unfeasible. Information Extraction approaches based on Natural Language Processing can be, and are already being used, to facilitate this process.

The dominant approach so far has been the use of hand-built, knowledge-based systems, working at levels ranging from surface syntax to full parses (Blaschke and Valencia, 2002; Huang et al., 2004; Plake et al., 2005; Rebholz-Schuhmann et al., 2005; Yakushiji et al., 2005). A similar work to the one presented here is by (Sugiyama et al., 2003), but it is not possible to compare results due to differing datasets and the limited information available about their methods.

2 Data

A gene-interaction corpus derived from the BioCreAtIvE task-1A data will be used for the experiments. This data was kindly made available by Jörg Hakenberg[1] and is described in (Plake et al., 2005). The data consists of 1000 sentences marked up for POS tags, genes (both genes and proteins are marked as 'gene'; the terms will be used interchangeably in this paper) and iWords. The corpus contains 255 relations, all of which are intra-sentential, and the "interaction word" (iWord)[2] for each relation is also marked up.

I utilise the annotated entities, and focus only on relation extraction. The data contains directionality information for each relation, denoting which entity is the 'agent' and which the 'target', or denoting that this distinction cannot be made. This information will not be used for the current experiments, as my main aim is simply to identify relations between entities, and the derivation of this information will be left for future work.

I will be using the Naive Bayes, KStar, and JRip classifiers from the Weka toolkit, Zhang Le's Maximum Entropy classifier (Maxent), TiMBL, and LibSVM to test performance. All experiments are done using 10-fold cross-validation. Performance will be measured using Recall, Precision and F1.

3 Experiments

Each possible combination of proteins and iWords in a sentence was generated as a possible relation 'triple', which combines the relation extraction task with the additional task of finding the iWord to describe each relation. 3400 such triples occur in the data. After each instance is given a probability by the classifiers, the highest scoring instance for each protein pairing is compared to a threshold to decide

[1] See http://www.informatik.hu-berlin.de/ hakenber/publ/suppl/sac05/

[2] A limited set of words that have been determined to be informative of when a PPI occurs, such as *interact, bind, inhibit, phosphorylation*. See footnote 1 for complete list.

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 120–121,
New York City, June 2006.

the outcome. Correct triples are those that match the iWord assigned to a PPI by the annotators.

For each instance, a list of features were used to construct a 'generic' model :

interindices The combination of the indices of the proteins of the interaction; "P1-position:P2-position"

interwords The combination of the lexical forms of the proteins of the interaction; "P1:P2"

p1prevword, p1currword, p1nextword The lexical form of P1, and the two words surrounding it

p2prevword, p2currword, p2nextword The lexical form of P2, and the two words surrounding it

p2pdistance The distance, in tokens, between the two proteins

inbetween The number of other identified proteins between the two proteins

iWord The lexical form of the iWord

iWordPosTag The POS tag of the iWord

iWordPlacement Whether the iWord is between, before or after the proteins

iWord2ProteinDistance The distance, in words, between the iWord and the protein nearest to it

A second model incorporates greater domain-specific features, in addition to those of the 'generic' model :

patterns The 22 syntactic patterns used in (Plake et al., 2005) are each used as boolean features[3].

lemmas and stems Lemma and stem information was used instead of surface forms, using a system developed for the biomedical domain.

4 Results

Tables 1 and 2 show the results for the two models described above. The system achieves a peak performance of 59.2% F1, which represents a noticeable improvement over previous results on the same dataset (52% F1 (Plake et al., 2005)), and demonstrates the feasibility of the approach adopted.

It is seen that simple contextual features are quite informative for the task, but that a significant gains can be made using more elaborate methods.

Algorithm	Recall	Precision	F1
Naive Bayes	61.3	35.6	45.1
KStar	65.2	41.6	50.8
Jrip	**66.0**	**45.4**	**53.8**
Maxent	58.5	48.2	52.9
TiMBL	49.0	41.1	44.7
LibSVM	49.4	56.8	52.9

Table 1: Results using 'generic' model

Algorithm	Recall	Precision	F1
Naive Bayes	64.8	44.1	52.5
KStar	60.9	45.0	51.8
Jrip	44.3	45.7	45.0
Maxent	57.7	56.6	57.1
TiMBL	42.7	74.0	54.1
LibSVM	**54.5**	**64.8**	**59.2**

Table 2: Results using extended model

[3]These patterns are in regular expression form, i.e. "P1 word{0,n} Iverb word{0,m} P2". This particular pattern matches sentences where a protein is followed by an iWord that is a verb, with a maximum of *n* words between them, and following this by *m* words maximum is another protein. In their paper, (Plake et al., 2005) optimise the values for *n* and *m* using Genetic Algorithms, but I will simply set them all to 5, which is what they report as being the best unoptimized setting.

References

C. Blaschke and A. Valencia. 2002. The frame-based module of the suiseki information extraction system. *IEEE Intelligent Systems*, (17):14–20.

Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G. Payan, Kunbin Qu 2, and Ming Li. 2004. Discovering patterns to extract proteinprotein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

Conrad Plake, Jörg Hakenberg, and Ulf Leser. 2005. Optimizing syntax-patterns for discovering protein-protein-interactions. In *Proc ACM Symposium on Applied Computing, SAC, Bioinformatics Track*, volume 1, pages 195–201, Santa Fe, USA, March.

D. Rebholz-Schuhmann, H. Kirsch, and F. Couto. 2005. Facts from text–is text mining ready to deliver? *PLoS Biol*, 3(2).

Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa, and Shunsuke Uemura. 2003. Extracting information on protein-protein interactions from biological literature based on machine learning approaches. *Genome Informatics*, 14:699–700.

Akane Yakushiji, Yusuke Miyao, Yuka Tateisi, and Jun'ichi Tsujii. 2005. Biomedical information extraction with predicate-argument structure patterns. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine*, pages 60–69.