Benchmarking machine learning techniques for the extraction of protein-protein interactions from text

Sofie Van Landeghem SOFIE.VANLANDEGHEM@PSB.UGENT.BE
Yvan Saeys YVAN.SAEYS@PSB.UGENT.BE
Yves Van de Peer YVES.VANDEPEER@PSB.UGENT.BE
Department of Plant Systems Biology, VIB, 9052 Gent, Belgium
Department of Molecular Genetics, University of Ghent, 9052 Gent, Belgium

Bernard De Baets BERNARD.DEBAETS@UGENT.BE
Department of Applied Mathematics, Biometrics and Process Control, University of Ghent, 9000 Gent, Belgium

Abstract

Accurately extracting information from text is a challenging discipline because of the complexity of natural language. We have studied state-of-the art systems that extract biological relations from research articles. It has become clear that this field is still struggling with a heterogeneous collection of data sets, data formats and evaluation methods. While recent developments look promising, there is still plenty of room for improvement.

1. Introduction

In the field of life sciences it is vital to automatically link experimental results to data already published in online literature resources. Fully automated systems that extract biological knowledge from text have thus become a necessity. We have studied the feasibility of applying machine learning approaches for the extraction of protein-protein interactions (PPIs). During our comparative study, it became clear that there is a great need for the standardization of evaluation procedures.

2. Corpora

Over the past few years, different methods have been proposed to extract biological relations from text. The development of standard benchmarking data sets is a step forward towards meaningful comparison between these systems. Such corpora include LLL, AImed and BioInfer, which have all been published in different dataformats. Only recently, software has been introduced to convert these and two smaller data sets into a common dataformat (Pyysalo et al., 2008), which facilitates comparison between different methods.

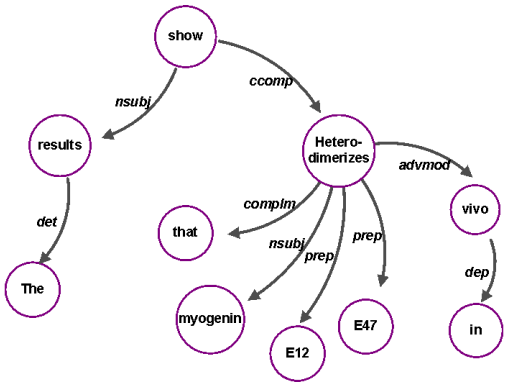

Figure 1. Dependency parse for 'The results show that myogenin heterodimerizes with E12 and E47 in vivo.'

3. PPI extraction

Sentences selected from biomedical text usually contain complex structures with multiple subordinate clauses. Interacting proteins often occur in a sentence with some distance between them. Therefore, pattern-based approaches and algorithms using word order suffer from low recall. On the other hand, techniques solely based on co-occurrence of named entities exhibit low precision. To better capture the semantics of a sentence, recent systems make use of information derived from dependency trees (see Fig 1).

By extracting properties from dependency trees, explicit features can be obtained for each pair of proteins. These feature vectors are used by classifiers such as decision trees, BayesNet and SVM to identify sentences which express a protein-protein interaction. Useful features relate to lexical and syntactic information about the children and ancestors of the proteins in the tree, the presence of common interaction words and depth of the named entities in the tree.

4. Ideas to improve benchmarking

4.1. Common set of benchmark data

A comparative study between different PPI extraction systems is a non trivial task as different studies often benchmark on different data sets. The RelEx system of Fundel et al. (2006) has been reimplemented with the goal of evaluating it on different corpora (Pyysalo et al., 2008). An F score of 0.77 was obtained when benchmarking on LLL, and a score between 0.41 and 0.44 when evaluated on AImed and Bioinfer. We obtain similar results when applying the walk kernel of Kim et al. (2008) to the AImed data set, which results in an F score of 0.44. In contrast, the original paper reports a score of 0.77 for the evaluation on LLL. This shows that for the same extraction method, performance can differ up to 36% depending on the choice of the corpus. It is therefore meaningful to evaluate new algorithms on a collection of different data sets.

4.2. Instance extraction

When benchmarking on the same corpus, different preprocessing steps can yield different instances. Homodimers, which are self-interacting proteins, are sometimes simply discarded. A similar issue is raised by annotations which are nested. The ability of the preprocessing techniques to deal with such annotations influences the final number of instances in the data set and ultimately the performance of the system.

Most corpora do not deal with the construction of negative training data. It has become common practice to adapt the closed world assumption, stating that no interaction exists between two entities when there is no annotated evidence. Even though AImed provides an explicit set of abstracts with no annotated interactions, these are not always used, resulting in different numbers of negative instances in the training set.

Ideally, abstracts for the testing phase should be completely hidden during training. Saetre et al. (2008) pointed out that some evaluations suffer from an artificial boost of performance by using features from the same sentence in both training and testing steps of the machine learning algorithm. This boost of performance has been estimated between 10 and 20%.

4.3. Counting true positives

The definition of true positives varies between different evaluation approaches. Most approaches consider every protein pair as an individual instance and evaluate whether an interaction is stated between these two particular entities. Some however state that an interaction between two proteins may be expressed in the same corpus by more than one instance. To extract a true interaction, retrieving one such instance suffices. The latter evaluation technique exhibits higher recall. Even though this technique may be useful for the evaluation of complete information retrieval systems, we feel the first is more representative for the subtask of extracting interactions between named entities from individual sentences.

4.4. Directed interactions

Finally, the definition of PPI extraction task is not unambiguously defined across corpora. The LLL data set and Bioinfer both consider the role of the different proteins in their interaction and discriminate between effectors and effectees. In AImed however, protein-protein interactions are considered to be symmetrical. This has led to the common practice of treating LLL annotations as symmetrical as well, resulting in artificially higher precision rates.

5. Conclusions

The comparison of different PPI extraction methods is hindered by the lack of standard evaluation procedures. We have pointed out the main problems for such a comparative study and indicated some practical guidelines for setting up a meaningful evaluation.

Acknowledgments

SVL would like to thank the Special Research Fund (BOF) for funding her research. YS would like to thank the Research Foundation Flanders (FWO) for funding his research.

References

Fundel, K., Küffner, R., & Zimmer, R. (2006). Relex–relation extraction using dependency parse trees. *Bioinformatics*, *23*, 365–371.

Kim, S., Yoon, J., & Yang, J. (2008). Kernel approaches for genic interaction extraction. *Bioinformatics*, *24*, 118–126.

Pyysalo, S., Airola, A., Heimonen, J., Bjrne, J., Ginter, F., & Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, *9*.

Saetre, R., Sagae, K., & Tsujii, J. (2008). Syntactic features for protein-protein interaction extraction. *Proceedings of the Second International Symposium on Languages in Biology and Medicine (LBM2007)*.