Effective Mining of Protein Interactions

Rinaldi, Fabio ; Schneider, G ; Kaljurand, K ; Clematide, S

Effective Mining of Protein Interactions

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, and Simon Clematide
Institute of Computational Linguistics, University of Zurich
{rinaldi,gschneid,kalju,siclemat}@cl.uzh.ch

Abstract

The detection of mentions of protein-protein interactions in the scientific literature has recently emerged as a core task in biomedical text mining. We present effective techniques for this task, which have been developed using the IntAct database as a gold standard, and have been evaluated in two text mining competitions.

1 Introduction

As a way to cope with the constantly increasing generation of results in molecular biology, some organizations maintain various types of databases that aim at collecting the most significant information in a specific area. For example, UniProt/SwissProt (UniProt Consortium, 2007) collects information on all known proteins. IntAct (Hermjakob et al., 2004) is a database collecting protein interactions. Most of the information in these databases is derived from the primary literature by a process of manual revision known as "literature curation". Text mining solutions are increasingly requested to support the process of curation of biomedical databases.

The work presented here is part the OntoGene project[1], which aims at improving biomedical text mining through the usage of advanced natural language processing techniques. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, term recognition, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences (Rinaldi et al., 2006; Rinaldi et al., 2008). The results of the entity detection feed directly into the process of identification of protein interactions. The syntactic parser (Schneider, 2008) takes into account constituent boundaries defined by previously identified multi-word entities. Therefore the richness of the entity annotation has a direct beneficial impact on the performance of the parser, and thus leads to better recognition of interactions.

In this paper we first describe in Section 2 the process used to automatically annotate different types of entities, and ground them to reference identifiers. In Section 3 we illustrate how we collect information about the focus organisms mentioned in the articles and how we use it to disambiguate protein mentions. In Section 4 we describe our approach to the detection of interactions among entities (proteins in particular). Finally, we present some evaluation of the results in the context of two recent shared tasks (Section 5).

2 Detection and Grounding of Domain Entities

In this section, we describe our approach to the problem of detecting names of relevant domain entities in biomedical literature (we consider in particular proteins, genes, species, experimental methods, and cell lines) and grounding them to widely accepted identifiers assigned by four different knowledge bases: UniProt Knowledgebase[2], National Center for Biotechnology Information (NCBI) Taxonomy[3], Proteomics Standards Initiative Molecular Interactions Ontology (PSI-MI)[4], and Cell Line Knowledge Base (CLKB)[5].

The terms extracted from the mentioned knowledge bases are stored in a common format in a database, and mapped to a unique identifier (from the original KB). An efficient lookup procedure is used to annotate any mention of a term in the documents with the ID(s) to which it corresponds. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the known terms of the term

[1] <http://www.ontogene.org>

[2] <http://www.uniprot.org>

[3] <http://www.ncbi.nlm.nih.gov/Taxonomy/>

[4] <http://psidev.sourceforge.net/mi/psi-mi.obo>

[5] <http://clkb.ncibi.org/>

list at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings. In case the normalized strings match exactly, the input sequence is annotated with the IDs of the term list term. Finally, a disambiguation step resolves the ambiguity (i.e. multiple IDs) of the matched terms. For more details, see (Kaljurand et al., 2009a).

3 Identifying Focus Organisms

In order to disambiguate protein names, the most effective dimension is that of the organism to which they refer. We use an approach based on the identification of what we call the 'focus' organisms mentioned in the paper. This approach can be briefly summarized as (1) find all explicit mentions of organisms either by their scientific or common names; (2) count these mentions and combine the resulting numbers with a simple use of statistics to arrive at a ranked list or a simple set of organisms which can be used, among other things, to disambiguate protein names in the article under investigation.

The source of information about the organism is the NCBI taxonomy which includes entries for 319,661 different organisms. As most of these organism are unlikely to ever occur in biomedical literature, we decided to restrict our interest to the organisms for which at least one UniProt entry exists, leading to a set of 11,444 organisms.

Once all organisms mentioned in an article have been annotated, this information can be used to construct a ranked list of organisms according to the number of mentions, which in turn can serve for disambiguation purposes. A higher weight is given to mentions in the abstract, and the mention counts are further balanced using frequencies derived from manually curated databases. These balance weights play a crucial role in adapting the ranking to a particular purpose. For example, for protein disambiguation, the weights should be derived from a database which relates protein mentions to the papers in which they appear. In particular, for the task of detecting protein interactions, our weights were derived from the IntAct and MINT databases. For more details and a separate evaluation of the TX task using the IntAct dataset as a gold standard, see (Kappeler et al., 2009).

4 Detection of Protein Interactions

Using the information concerning mentions of relevant domain entities, derived as described in Section 2, and their corresponding unique identifiers obtained by the process of disambiguation described in Section 3, it is possible to create candidate interactions. In other words, the co-occurrence of two entities in a given text span (typically a sentence, or observation window) is a low-precision indication of a potential relationship among those entities. In order to obtain better precision it is necessary to take into account the syntactic structure of the sentence, and other structural information. In this section we describe the approach we have adopted for the detection of protein interactions.

We use the GENIA corpus (Kim et al., 2003), augmented by manual decisions, as training corpus for the interaction detection task, based on an approach described in (Schneider et al., 2009). We use more features, however, and have set the feature scores to optimise on the BioCreative training data. The GENIA corpus has been parsed with our state-of-the art dependency parser which has been adapted to and evaluated on the biomedical domain (Schneider, 2008). After parsing, we collect all syntactic connections that exist between all the terms as follows. For each term-coocurrence, i.e. two terms appearing in the same sentence, a collector traverses the tree from one term up to the lowest common mother node, and down the second term, recording all intervening nodes. We record the head lemma of the top node, and the grammatical labels plus prepositions connecting all intervening nodes.

Only a minority of the paths extracted by the method just introduced actually express a biomedical interaction. The decision to classify the paths observed in the training data as positive or negative is taken manually. Among the observed paths, 309 were positively classified. We noticed that in many sentences one of the interactors is embedded in a way that the whole sentence, albeit not directly expressing the interaction, implies it, or can be paraphrased as to imply the interaction. For example, the sentence *"A activates groups of B"* typically implies that *"A activates B"*, or *"A blocks activation of B"* implies that *"A blocks B"*, whereas *"A activates C, which has a binding site for B"* does not express that *"A activates B"*. There is a large set of words like *group* and *activation*, for which we have adopted the

term *transparent words*. All the words intervening inside a path are candidates for being transparent words. For each word appearing inside a path, we calculate a score which divides its frequency inside a path by its total frequency. Words above a threshold are treated as transparent in the application phase. Depending on the threshold, our transparent words resource contains between 100 and 800 words. For a related approach, see (Pyysalo et al., 2009)

The paths that are extracted from GENIA can directly be used for PPI detection. For example, in the sentence shown in Figure 1, a pattern with the decision 'yes' exists for the relation between *Tim18* and *Tim12*, i.e. the pattern with top node *coimmunoprecipitate*, left path `[subj]` and right path `[pobj]`. Since using the full paths would lead to data sparseness problems, various backoff stages are used to ensure their usability, gradually relaxing some constraints. Dependencies with no semantic content (e.g. conjunctions, appositions) are first cut from the path, followed if necessary by dependencies containing transparent words ('portion of' in figure 1), which has the effect that the pattern directly reporting the relation between Tim18p and Tim12p also finds the relation between Tim18p and Tim54p in Figure 1. Surface patterns are then used as an ultima ratio backoff step. A related approach is discussed in (Buyko et al., 2009).

5 Evaluation

5.1 Evaluation: BioNLP Shared Task

We participated in the BioNLP shared task (Kaljurand et al., 2009b), where we achieved a recall of 26% and a precision of 44% in the official test run, which placed us at rank 8 of 24 teams. The task was difficult due to the fact that events needed to be labelled, and complex events (e.g. interactions between interactions) were included. Our performance on non-complex events was 57% precision at 40% recall, which is comparable to our previous experimental results. We included more term resources in order to increase recall, at a certain cost for precision. It turned out that transparent words were valuable indicators for the event type label, so that we could not use the transparent words resource to reduce sparseness.

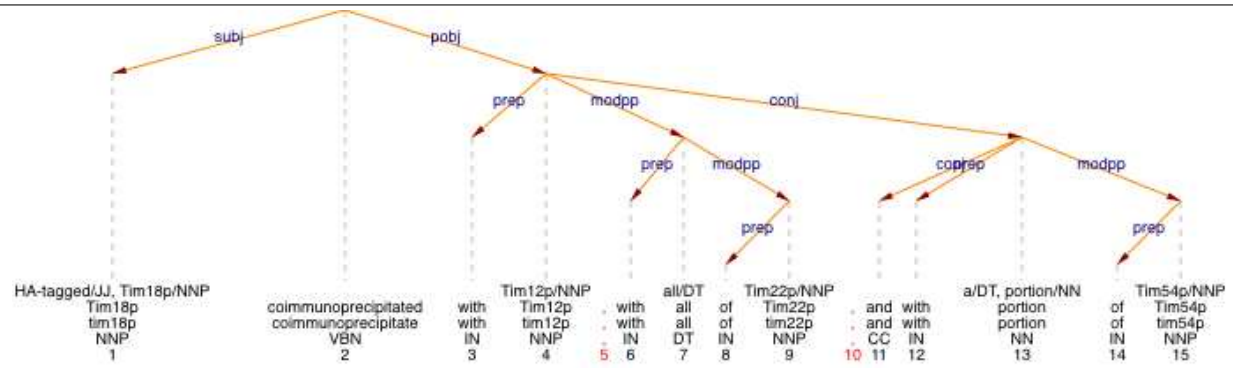
5.2 Evaluation: BioCreative II.5

More recently, we participated to the BioCreative II.5 competition, which focused on the extraction of protein-protein interactions. In this competition, which was based on unlabelled events, the transparent words resource proved to be beneficial. We added further term resources to boost recall, and extended the backoff chain, including WordNet synsets and training data from the BioNLP shared task. We noticed that precision is quite low because the task includes the difficult distinction between novel and background interactions: the only interaction which are considered relevant are those reported in the target article as results of the experiments performed by the authors. Inspecting the test data showed us that, where both terms are grounded correctly, recall is quite high, and that most of the remaining missed interactions cannot be found without several logical inference steps, or they involve several sentences, which means that they are beyond the scope of our current approach. In order to increase precision, we have added further features and re-weighted their scores, as follows.

We use the following five features for the interaction detection task. (1) *Syntactic path*: if the path between the two proteins is equivalent to one of the paths previously seen in GENIA and classified as positive, the candidate interaction receives a higher score. Various backoff stages (as discussed in Section 4) allow to deal gracefully with data sparseness. (2) *Known interaction*: Interactions that are already reported in the IntAct and MINT databases receive a low score. The older the entry data in the database, the lower the score. (3) *Novelty score*: on the basis of linguistic clues (e.g. *"Here we report that..."*) we attempt to distinguish between sentences that report the results detected by the authors from sentences that report background results. Interactions in 'novelty' sentences are scored higher than interactions in 'background' sentences. (4) *Zoning*: The abstract and the conclusions are the typical places for mentioning novel interactions, the introduction and methods section are less likely and get lower scores. (5) *Pair salience*: Proteins that are mentioned frequently in an article are more likely to participate in a relevant interaction than proteins that are mentioned only once. We use the following simple calculation to assign a value to this feature: $sal(p_1, p_2) = \frac{f(p_1) * f(p_2)}{f(proteins\ in\ article)}$

The weights of each feature are currently set heuristically, we intend to explore ways to optimize them on the basis of a training collection. The scores of each feature are multiplied, and the

Figure 1 Dependency parser output example.



total score of a protein-protein interaction is the sum of its occurrences. The result is then normalized to the range `[0,1]` with the following formula: $log(score)/log(maxscore)$. The value of this score is then used for ranking the candidate interactions. A low threshold can then be used to remove the least promising candidates, leading to an increase in precision at the cost of a minimal loss of recall. The organizers of the BioCreative II.5 competition have adopted as official scoring criteria the AUC of the iP/R graph, which is a indication of the quality of the ranking of the results. In many applications, what is most important to the end user is to be able to get quickly at the relevant information. A good ranking of the results is therefore of more practical relevance than optimal P/F/R scores. Our AUC over the BioCreative training set is 22%.

6 Conclusions

In this paper we have discussed applications of a dependency parser to advanced text mining tasks, such as the extraction of protein-protein interactions or of more complex events. We have shown the effectiveness of the selected approach through participation to a number of shared competitive evaluations.

Acknowledgments This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002. Fabio Rinaldi is currently supported by the SNF fellowship PBZHP2-125509.

References

[Buyko et al.2009] Ekaterina Buyko, Erik Faessler, Joachim Wermter, and Udo Hahn. 2009. Event extraction from trimmed dependency graphs. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Share d Task*, pages 19–27, Boulder, Colorado, June. Association for Computational Linguistics.

[Hermjakob et al.2004] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David Sherman, and Rolf Apweiler. 2004. IntAct: an open source molecular interaction database. *Nucl. Acids Res.*, 32(suppl 1):D452–455.

[Kaljurand et al.2009a] Kaarel Kaljurand, Fabio Rinaldi, Thomas Kappeler, and Gerold Schneider. 2009a. Using existing biomedical resources to detect and ground terms in biomedical literature. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME09)*.

[Kaljurand et al.2009b] Kaarel Kaljurand, Gerold Schneider, and Fabio Rinaldi. 2009b. UZurich in the BioNLP 2009 Shared Task. In *Proceedings of the BioNLP workshop, Boulder, Colorado*.

[Kappeler et al.2009] Thomas Kappeler, Kaarel Kaljurand, and Fabio Rinaldi. 2009. TX Task: Automatic Detection of Focus Organisms in Biomedical Publications. In *Proceedings of the BioNLP workshop, Boulder, Colorado*.

[Kim et al.2003] J.D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus — a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):180–182.

[Pyysalo et al.2009] Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*, pages 1–9.

[Rinaldi et al.2006] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. 2006. An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3.

[Rinaldi et al.2008] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. 2008. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13.

[Schneider et al.2009] Gerold Schneider, Kaarel Kaljurand, Thomas Kappeler, and Fabio Rinaldi. 2009. Detecting protein-protein interactions in biomedical texts using a parser and linguistic resources. In *Proceedings of CICLING 2009*.

[Schneider2008] Gerold Schneider. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

[UniProt Consortium2007] UniProt Consortium. 2007. The universal protein resource (uniprot). *Nucleic Acids Research*, 35:D193–7.