Dependency-Based Relation Mining for Biomedical Literature

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess

Institute of Computational Linguistics,
University of Zurich, Switzerland,
{rinaldi,gschneid,kalju,hess}@cl.uzh.ch

Abstract

We describe techniques for the automatic detection of relationships among domain entities (e.g. genes, proteins, diseases) mentioned in the biomedical literature. Our approach is based on the adaptive selection of candidate interactions sentences, which are then parsed using our own dependency parser. Specific syntax-based filters are used to limit the number of possible candidate interacting pairs. The approach has been implemented as a demonstrator over a corpus of 2000 richly annotated MedLine abstracts, and later tested by participation to a text mining competition. In both cases, the results obtained have proved the adequacy of the proposed approach to the task of interaction detection.

1. Introduction

There is a considerable quantity of published literature in the Life Sciences area. The PubMed[1] repository currently contains more than 17 million references. Besides, thanks to the increased public funding and the intensifying research activities, the number of new publications is soaring. Keeping track of this vast mass of informations is becoming increasingly difficult, even for the domain experts, therefore automatic tools that can support this process are increasingly requested (Krallinger and Valencia, 2005).

One of the main obstacles in successfully extracting information from biomedical articles is the variability in the names of domain entities (proteins, genes, etc), which is a well-known problem (Jensen et al., 2006). Not only typical entity names are plagued by a high degree of polysemy and synonymy, it is also possible that very common words can be used as names of genes and proteins.

There are numerous approaches that deal with the problem of entity identification, notably the experiments performed in the Gene Normalization task of BioCreAtIvE (Morgan et al., 2007). Entity name variability is also a major focus of research (e.g. Tsuruoka et al. (2007)). A number of tools have been developed (Settles, 2005; Song et al., 2005), and some of them are freely available and can be used as components in more complex systems. The problem of relation discovery is comparably less explored; some of the existing approaches are described in section 6.

In this paper we discuss an environment which supports the process of knowledge discovery from biomedical literature, focusing in particular on the detection of interactions between biomedical entities (genes, diseases, proteins, etc.). Our approach is based on a dependency parser and modular rules which make use of a rich linguistic annotation. The results have been validated on a publicly available corpus (GENIA) and by participation to a text mining competition (BioCreAtIvE).

In the rest of this paper, we first describe the core principles of our approach, then mention the applications already developed, and finally present the results of the evaluations performed.

2. Corpus Analysis

The system that we describe includes a number of NLP tools which are organized into a pipeline (Kaljurand et al., 2006). The basic tasks that are performed by the pipeline are: sentence splitting, tokenization, PoS tagging, lemmatization, term extraction, chunking, dependency parsing. The final result of the analysis process is a set of dependency relations, which are encoded as `(sentence-id, type, head, dependent)` tuples. This is a format which is well suited for storage in a relational DB, and for delivery to other tools, either in CSV or XML format. Figure 1 shows a graphical representation of the results of the analysis.

We use a robust, deep-syntactic, broad-coverage probabilistic dependency parser (Schneider et al., 2004), which identifies grammatical relations between the heads of chunks, chunk-internal dependencies, and the majority of long-distance dependencies.

The parser expresses distinctions that are especially important for a predicate-argument based deep syntactic representation, as far as they are expressed in the Penn Treebank training data. This includes PP-attachment, most long-distance dependencies, appositions, relative clause anaphora, participles, gerunds, and argument/adjunct distinctions.

The parser is very robust and has been applied to parsing large amounts of text data, including the 100 Million word British National Corpus [2]. It does not always deliver a parse spanning the entire sentence, however it never fails completely, always delivering at least partial structures.

3. Relation Mining

Our approach to relation mining is based on cascading rules. On the first level, we exploit simple *syntactic patterns* detected in the data. On the second level we combine various patterns into a single *semantic rule*, which normalizes many possible syntactic variants (e.g. active, passive, nominalizations). On the third level we combine semantic rules with lexical and ontological constraints to obtain very specialized queries that can detect a given domain-specific relation, as specified by the user.

[1] `http://www.ncbi.nlm.nih.gov/entrez`

[2] `http://www.natcorp.ox.ac.uk/`

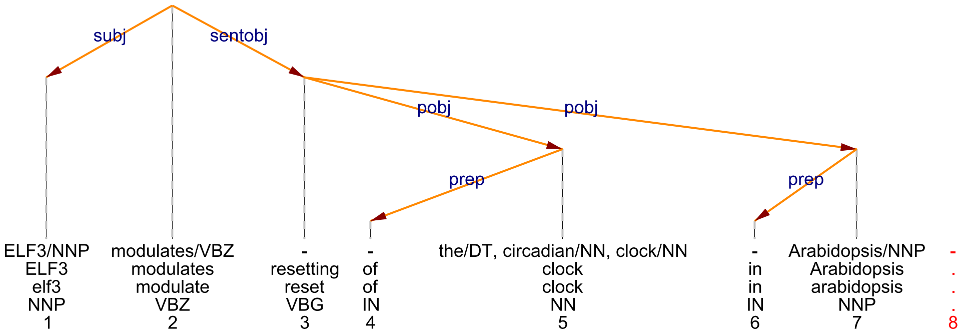


Figure 1: Tree of syntactic dependencies in the sentence *"ELF3 modulates resetting of the circadian clock in Arabidopsis"* along with other linguistic annotations

These rules are highly modular, and can therefore be reused to a large extent across different applications. For example, on the lower levels the syntactic rules will describe common structures such as passive, active or nominalized expressions. An example of a syntactic rule for the passive case is shown below:[3]

```
synRel(simple_passive, [H,B,A],
   [  dep(subj,H,B),dep(pobj,H,A), dep(prep,A,By),
      pos(H,'VBN'), lemma(By,['by','through','via']) ]).
```

Rules at the next level can combine lower-level rules into more powerful structures, which allow the users to abstract away from the syntactic level. For example, a unique query will be needed to match all sentences that describe an interaction between an `agent` A and a `target` B, regardless of how they are expressed at the surface level. The arguments of the query can either be specific entities (e.g. 'NF-KappaB') or type restrictions (e.g. protein, cell, disease).

As the set of patterns and rules is gradually enriched, so are the possible lexico-syntactic variants that can be captured. An example of a more advanced rule is the one which captures *"A triggers the H of B"*, where H represent a nominalized verb (*activation, regulation, etc.*). Similar complex rules have designed, e.g. for *"under the control of"*, *"involved in"*, *"be able to"* etc. Because such rules are built on top of the lower-level rules, they automatically capture the known syntactic variants, such as *"The H of B is triggered by A"*. We refer to relations defined at this level as *domain relations* as they rely on lexical constraints which are typical of a given domain. The user query can happen at each one of the 3 levels. So it is possible to test individual syntactic rules, semantic rules, and domain rules.

If a domain ontology is available, we can extend the interpretation of the type restriction to mean not only the objects that directly match the given type, but also those that have a type subsumed by it.

Additionally, the OntoGene Text Mining environment provides facilities for debugging and visualization. For example, each result bears the name of the rules that generated it. This allows immediate detection of problems and their quick correction. An example is shown in figure 3. We also provide a "visual diff" facility that shows in the same graphical format the matches that have been acquired or lost as a consequence of the addition of a new pattern or rule.

4. Applications

The tools that we describe have been applied to extract semantic relations from two distinct corpora: GENIA and ATCR.

GENIA (Kim et al., 2003) is a corpus of 2000 Medline abstracts which have been *manually* annotated (by domain experts) for various biological entities according to the GENIA Ontology. A detailed description of this application can be found in (Rinaldi et al., 2006).

The ATCR Corpus (Arabidopsis Thaliana Circadian Rhythms) is a corpus of 147 Medline abstracts (up to year 2004), extracted using the keywords: "Arabidopsis Thaliana" and "Circadian Rhythms". It has been *automatically* annotated using the "Biolab Experiment Assistant (BEA)"$^{\mathrm{TM}}$. This applications is described in (Rinaldi et al., 2007).

In the case of the GENIA corpus, users can create complex queries which make use of the ontological relations, because the entity annotations have been created according to the types defined in the GENIA Ontology [4].

Additionally, we have used an approach based on the rules described above as one of the key components within a text mining system aimed at extracting specific protein-protein interactions from biomedical literature. This system was used in our participation to the BioCreative text mining challenge (Krallinger et al., 2007), obtaining competitive results (for details see Rinaldi et al. (2008)).

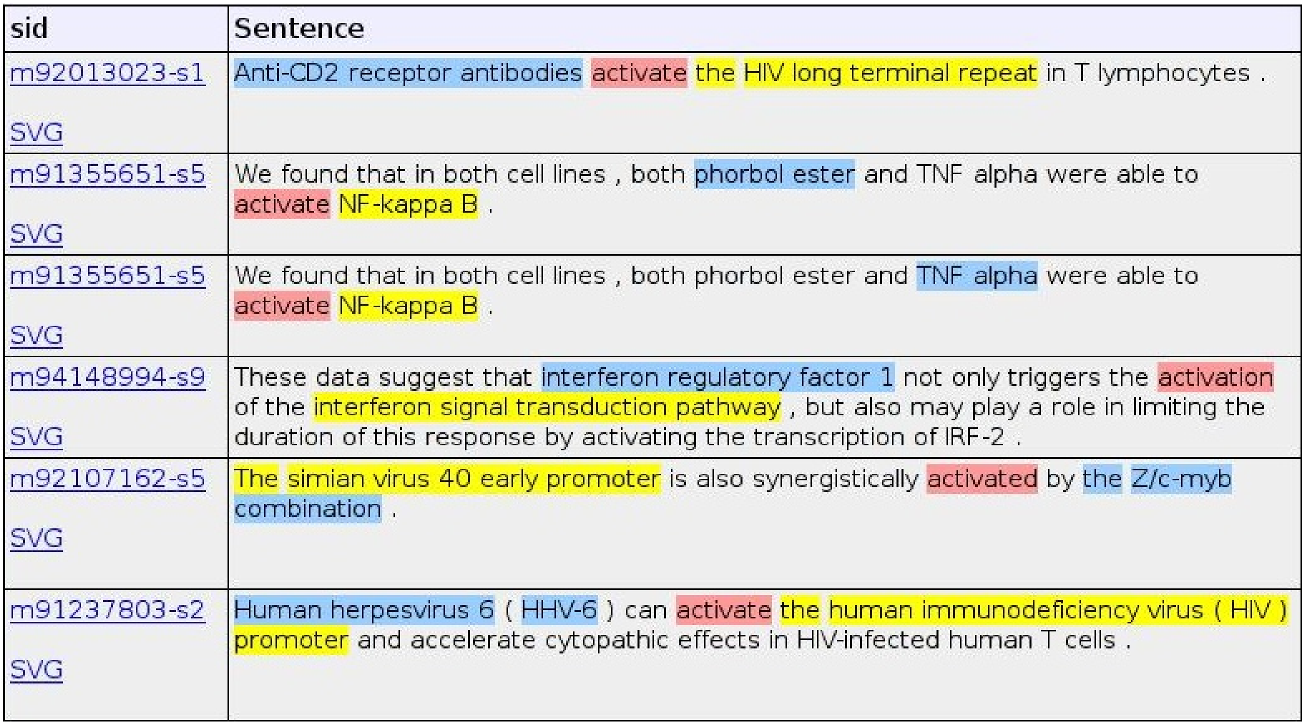
BioCreAtIvE provides a framework for testing and evaluating text mining tools over a number of shared tasks of biological significance. One of them is the protein-protein interaction task, which consists of detecting from articles the main protein interactions reported by the authors.

5. Evaluation

The applications over the GENIA corpus and the ATRC corpus were evaluated by asking domain experts to use the system and build rules to match the information of interest. During this process, they can make use of visual feedback showing the coverage of each rule. The evaluation was based on a random selection of a limited number of sentences from the corpora, which were then manually assessed for the detection of relations and arguments of inter-

[3] A prolog-based syntax is adopted.

[4] `http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html`



sid	Sentence
m92013023-s1 SVG	Anti-CD2 receptor antibodies activate the HIV long terminal repeat in T lymphocytes .
m91355651-s5 SVG	We found that in both cell lines , both phorbol ester and TNF alpha were able to activate NF-kappa B .
m91355651-s5 SVG	We found that in both cell lines , both phorbol ester and TNF alpha were able to activate NF-kappa B .
m94148994-s9 SVG	These data suggest that interferon regulatory factor 1 not only triggers the activation of the interferon signal transduction pathway , but also may play a role in limiting the duration of this response by activating the transcription of IRF-2 .
m92107162-s5 SVG	The simian virus 40 early promoter is also synergistically activated by the Z/c-myb combination .
m91237803-s2 SVG	Human herpesvirus 6 (HHV-6) can activate the human immunodeficiency virus (HIV) promoter and accelerate cytopathic effects in HIV-infected human T cells .

Figure 2: Sample output for the ‘activate’ relation

	agent				target			
	Y	A	P	N	Y	A	P	N
activate	72	64	5	8	77	54	8	10
bind	36	18	1	8	39	18	1	5
block	3	0	0	0	1	1	0	1
TOTAL	111 52%	82 38%	6 3%	16 7%	117 55%	73 34%	9 4%	16 7%
	correct 90%		incorrect10%		correct 89%		incorrect11%	

Table 1: Analysis of precision for selected relations over GENIA

est. Table 1 summarizes the values of precision. The column identified by [Y] shows cases where the relation (and the corresponding argument) was considered by the biologist as correct and relevant. The [A] column refers to argument which were only partially correct (but sufficient to understand the nature of the relationship). The other columns consider different types of errors.

In the absence of a gold standard, only approximative recall values can be reported. However, in the case of the ATCR application, we have been able to measure a “worst-case” recall value of 40% , which basically implies that our actual recall is at least as good as this value. On a smaller subset of the corpus we actually measured a recall value of 60% (for details see Rinaldi et al. (2007)).

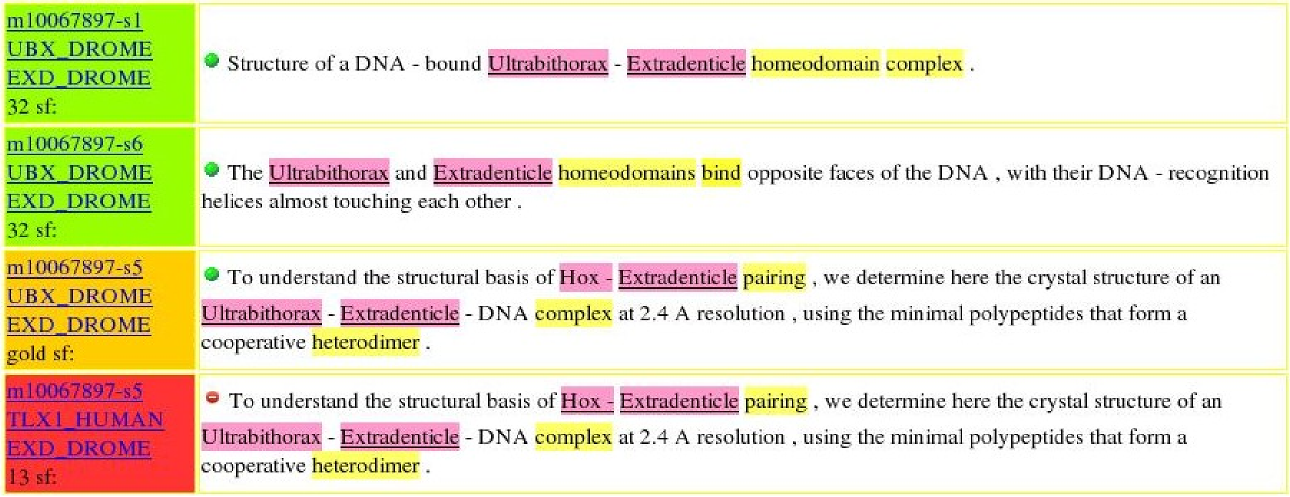
In the context of the BioCreative competition, the task of detecting protein-protein interactions is additionally complicated by the need of locating protein names in the abstracts provided by the organizers, and normalizing them to unique identifiers. We developed techniques capable of locating protein names and their variants starting from a set of seed terms obtained from UniProt (UniProt Consortium, 2007). Figure 4 shows an example of automatically annotated abstract.

Further, not all potential interactions are requested, but only those that the authors present as their main results. Interactions that are mentioned only as ‘background’ information should be ignored. The combination of these requirements makes the task extremely challenging, and no participating system was capable of achieving results above 50%.

In our own system, after generating candidate interactions on the basis of co-occurrence of protein names within the same sentence, we have applied the methodology described above as a ‘syntactic filter’ in order to separate syntactically meaningful interactions from accidental ones. These filters proved to be extremely efficient: on the BioCreative training set they allowed us to increase precision from 20% to almost 50% with only a minimal loss in recall. More details can be found in (Rinaldi et al., 2008).

6. Related Work

The field of text mining from biomedical literature has been flourishing in the past few years (Cohen and Hunter,



m10067897-s1 UBX_DROME EXD_DROME 32 sf:	● Structure of a DNA - bound Ultrabithorax - Extradenticle homeodomain complex .
m10067897-s6 UBX_DROME EXD_DROME 32 sf:	● The Ultrabithorax and Extradenticle homeodomains bind opposite faces of the DNA , with their DNA - recognition helices almost touching each other .
m10067897-s5 UBX_DROME EXD_DROME gold sf:	● To understand the structural basis of Hox - Extradenticle pairing , we determine here the crystal structure of an Ultrabithorax - Extradenticle - DNA complex at 2.4 A resolution , using the minimal polypeptides that form a cooperative heterodimer .
m10067897-s5 TLX1_HUMAN EXD_DROME 13 sf:	⊖ To understand the structural basis of Hox - Extradenticle pairing , we determine here the crystal structure of an Ultrabithorax - Extradenticle - DNA complex at 2.4 A resolution , using the minimal polypeptides that form a cooperative heterodimer .

Figure 3: Debugging facilities

2004; Ananiadou et al., 2006), with important contributions from the computational linguistics field (Miyao et al., 2006; Pyysalo et al., 2006; Daraselia et al., 2004).

In particular, the task of relation extraction has seen a number of different approaches, which primarily can be classified into three groups according to the amount of NLP involved.

Surface-based approaches (Hakenberg et al., 2007; Ehrler et al., 2007) make use of lexical or PoS patterns without attempting any deeper understanding of the nature of the interaction. The advantage is that such patterns can in many cases be automatically learnt from annotated corpora.

Shallow parsing approaches typically detect the main constituents of the sentences, without building a complete syntactic analysis. An example is (Corney et al., 2004).

Approaches based on full parsing attempt to build a complete syntactic structure for each sentence in the corpus, which is then used to extract or confirm candidate interactions. For example, (Gonzalez et al., 2007) use the Link Grammar parser, (Daraselia et al., 2004) makes use of an LFG parser, specifically adapted to Medline, (Saetre et al., 2007) use an HPSG parser, (Erkan et al., 2007) uses the Stanford dependency parser.

7. Conclusion

We have developed various techniques that can support the process of relation discovery from biomedical literature. These techniques have been evaluated on a corpus of 2000 medline abstracts and by participation to a competitive evaluation for text mining systems. In both cases, the results prove the effectiveness of the proposed approach.

Acknowledgments

This research is partially supported by the Swiss National Science Foundation (grant 100014-118396/1).

8. References

Sophia Ananiadou, Douglas B Kell, and Junichi Tsujii. 2006. Text mining and its potential applications in systems biology. *Trends Biotechnol*, 24(12):571–579.

K. Bretonnel Cohen and Lawrence Hunter. 2004. Natural language processing and systems biology. In Werner Dubitzky and Francisco Azuaje, editors, *Artificial Intelligence Methods and Tools for Systems Biology*, pages 147–173. Springer Netherlands.

D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D.T. Jones. 2004. BioRAT: Extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–13.

Nikolai Daraselia, Anton Yuryev, Sergei Egorov, Svetalana Novichkova, Alexander Nikitin, and Ilya Mazo. 2004. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611.

Frederic Ehrler, Julien Gobeill, Imad Tbahriti, and Patrick Ruch. 2007. GeneTeam site report for BioCreAtIvE ii: Customizing a simple toolkit for text mining in molecular biology. In (Krallinger and Hirschman, 2007), pages 199–208.

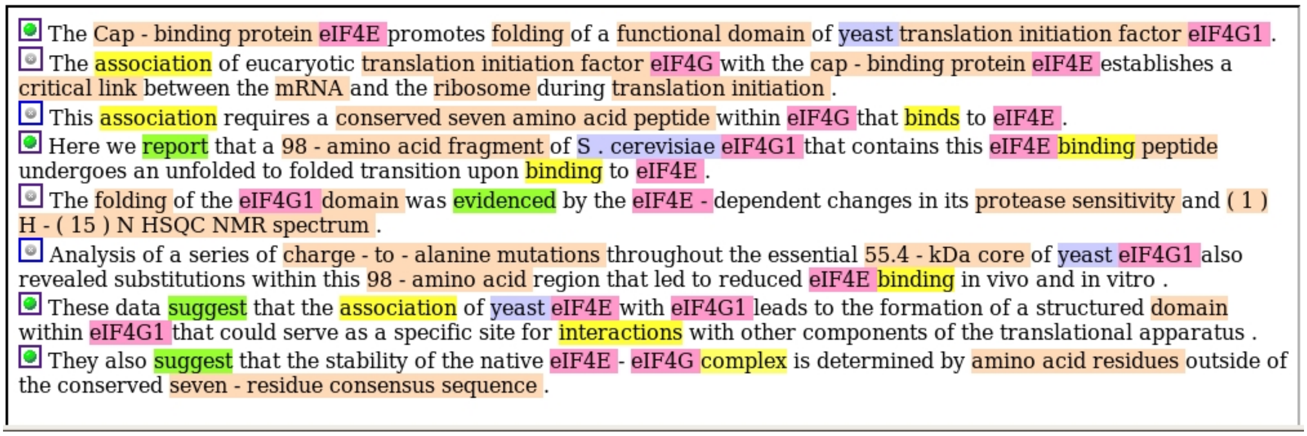
Güneş Erkan, Arzucan Ozgur, and Dragomir R. Radev. 2007. Extracting interacting protein pairs and evidence sentences by using dependency parsing and machine learning techniques. In (Krallinger and Hirschman, 2007), pages 287–292.

Graciela Gonzalez, Luis Tari, Anthony Gitter, Robert Leaman, Shawn Nikkila, Ryan Wendt, Amanda Zeigler, and Chitta Baral. 2007. Integrating knowledge extracted from biomedical literature: normalization and evidence statements for interactions. In (Krallinger and Hirschman, 2007), pages 227–236.

Jörg Hakenberg, Michael Schröder, and Ulf Leser. 2007. Consensus pattern alignment to find protein-protein interactions in text. In (Krallinger and Hirschman, 2007), pages 213–216.

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

Kaarel Kaljurand, Fabio Rinaldi, and Gerold Schneider. 2006. Prolog-based query interface to syntactic depen-



● The Cap - binding protein eIF4E promotes folding of a functional domain of yeast translation initiation factor eIF4G1 .

○ The association of eucaryotic translation initiation factor eIF4G with the cap - binding protein eIF4E establishes a critical link between the mRNA and the ribosome during translation initiation .

○ This association requires a conserved seven amino acid peptide within eIF4G that binds to eIF4E .

● Here we report that a 98 - amino acid fragment of S . cerevisiae eIF4G1 that contains this eIF4E binding peptide undergoes an unfolded to folded transition upon binding to eIF4E .

○ The folding of the eIF4G1 domain was evidenced by the eIF4E - dependent changes in its protease sensitivity and (1) H - (15) N HSQC NMR spectrum .

○ Analysis of a series of charge - to - alanine mutations throughout the essential 55.4 - kDa core of yeast eIF4G1 also revealed substitutions within this 98 - amino acid region that led to reduced eIF4E binding in vivo and in vitro .

● These data suggest that the association of yeast eIF4E with eIF4G1 leads to the formation of a structured domain within eIF4G1 that could serve as a specific site for interactions with other components of the translational apparatus .

● They also suggest that the stability of the native eIF4E - eIF4G complex is determined by amino acid residues outside of the conserved seven - residue consensus sequence .

Figure 4: Example of an annotated abstract. The terms marked in violet are those identified by the system as protein names, the terms marked in blue are those identified as organism names, while those marked in orange are other classes of terms. Words marked in yellow are indicators for a relation, words marked in green might suggest the presence of a curatable relation. The green dot on the left of a sentence indicates that the system considers that sentence as potentially containing a “curatable” relation.

dencies extracted from biomedical literature. Technical report, IFI, University of Zurich. Technical Report IFI-2006.04.

J.D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(1):i180–i182.

M Krallinger and L Hirschman, editors. 2007. *Proc. of Second BioCreative Challenge Evaluation Workshop 2007*. CNIO, Madrid.

Martin Krallinger and Alfonso Valencia. 2005. Text-mining and information-retrieval services for molecular biology. *Genome Biology*, 6(7):224.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2007. Overview of the protein-protein interaction annotation extraction task of biocreative ii. *Genome Biology*.

Y Miyao, T Ohta, K Masuda, Y Tsuruoka, K Yoshida, T Ninomiya, and J Tsujii. 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of COLING-ACL 2006, Sydney, Australia*, pages 1017–1024.

Alexander A. Morgan, Zhiyong Lu, Xinglong Wang, Aaron M. Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, Chengjie Sun, Heng hui Liu, Rafael Torres, Michael Krauthammer, William W. Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, and Lynette Hirschman. 2007. Overview of biocreative ii gene normalization. *Genome Biology*.

Sampo Pyysalo, Tapio Salakoski, Sophie Aubin, and Adeline Nazarenko. 2006. Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. *BMC Bioinformatics*, 7(Suppl 3):S2.

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.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, Christos Andronis, Ourania Konstanti, and Andreas Persidis. 2007. Mining of Functional Relations between Genes and Proteins over Biomedical Scientific Literature using a Deep-Linguistic Approach. *Journal of Artificial Intelligence in Medicine*, 39:127–136.

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*. (to appear).

Rune Saetre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. 2007. AKANE system: Protein-protein interaction pairs in the BioCreAtIvE2 challenge, PPI-IPS subtask. In (Krallinger and Hirschman, 2007), pages 209–212.

Gerold Schneider, Fabio Rinaldi, and James Dowdall. 2004. Fast, Deep-Linguistic Statistical Minimalist Dependency Parsing. In G. Kruijff and D. Duchier, editors, *COLING-2004 workshop on Recent Advances in Dependency Grammars, August 2004, Geneva, Switzerland*, pages 33–40.

Burr Settles. 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Yu Song, Eunju Kim, Gary Geunbae Lee, and Byoung-Kee Yi. 2005. POSBIOTM-NER: a trainable biomedical named-entity recognition system. *Bioinformatics*, 21(11):2794–2796.

Yoshimasa Tsuruoka, John McNaught, Jun’ichi Tsujii, and Sophia Ananiadou. 2007. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23(20):2768–2774.

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