Learning Language in Logic - Genic Interaction Extraction Challenge

C. Nédellec CLAIRE.NEDELLEC@JOUY.INRA.FR
Laboratoire Mathématique, Informatique et Génome (MIG), INRA,
Domaine de Vilvert, 78352 F- Jouy-en-Josas cedex.

Abstract

We describe here the context of the LLL challenge of Genic Interaction extraction, the background of its organization and the data sets. We discuss then the results of the participating systems.

1. Introduction

The Learning Language in Logic **(**LLL05) challenge is part of the 2005 LLL workshop. The LLL05 challenge task is to learn rules to extract protein/gene interactions in the form of relations from biology abstracts from the Medline bibliography database. The goal of the challenge is to test the ability of the participating ML systems to learn rules for identifying the gene/proteins that interact and their roles, agent or target. The training data contains the following information:

- The Agent and Target of the genic interactions.
- A dictionary of named entities (including typographic variants and synonyms)
- Linguistic information: word segmentation, lemmatization and syntactic dependencies.

The participants have tested their Information Extraction (IE) rules on a separate test set in a limited amount of time. The challenge organizers have provided the facilities for computing the scores of the results. Six different teams have participated and reported their results in the papers in this volume. This paper aims at summarizing the motivation for the challenge, the presentation of the training and test data and comparing the participant results.

2. Motivation

2.1 Biological motivation

Developments in biology and biomedicine are reported in large bibliographical databases either focused on a specific species (*e.g.* Flybase, specialized on *Drosophila Melanogaster*) or not (*e.g.* Medline). These types of information sources are crucial for biologists, but there is a lack of tools to explore them and extract relevant information.

While recent named entity recognition tools have gained a certain success on these domains, event-based Information Extraction (IE) is still challenging. Biologists can search bibliographic databases via the Internet, using keyword queries that retrieve a large set of relevant papers. To extract the requisite knowledge from the retrieved papers, they must identify the relevant paragraphs or sentences. Such manual processing is time consuming and repetitive, because of the bibliography size, the relevant data sparseness, and because the database is continually updated. For example, from the Medline database, the focused query "*Bacillus subtilis* and transcription", which returned 2,209 abstracts in 2002 retrieves more than 2,693 today. We chose this example because *Bacillus subtilis* is a model bacterium and because transcription is both a central phenomenon in functional genomics involved in gene interaction and a popular IE problem.

Example:
GerE *stimulates* ***cotD*** *transcription and inhibits* ***cotA*** *transcription in vitro by* ***sigma K*** *RNA polymerase, as expected from in vivo studies, and, unexpectedly, profoundly inhibits in vitro transcription of the gene (****sigK****) that encode* ***sigma K****.*

In this example, there are 6 genes and proteins mentioned and among the 30 potential ordered couples, 5 couples actually interact: (GerE, cotD), (GerE, cotA), (sigma K, cotA), (GerE, SigK) and (sigK, sigma K). The precision of the baseline method that extracts gene/protein cocitations is then 20 % for 100 % recall. In gene interactions, the agent is distinguished from the target of the interaction. Such interactions are central in functional genomics because they form regulation networks that are very useful for determining the function of the genes. The description of such gene interactions is not available in structured databases but only in scientific papers. Figure 1 gives an example of such a regulation network.

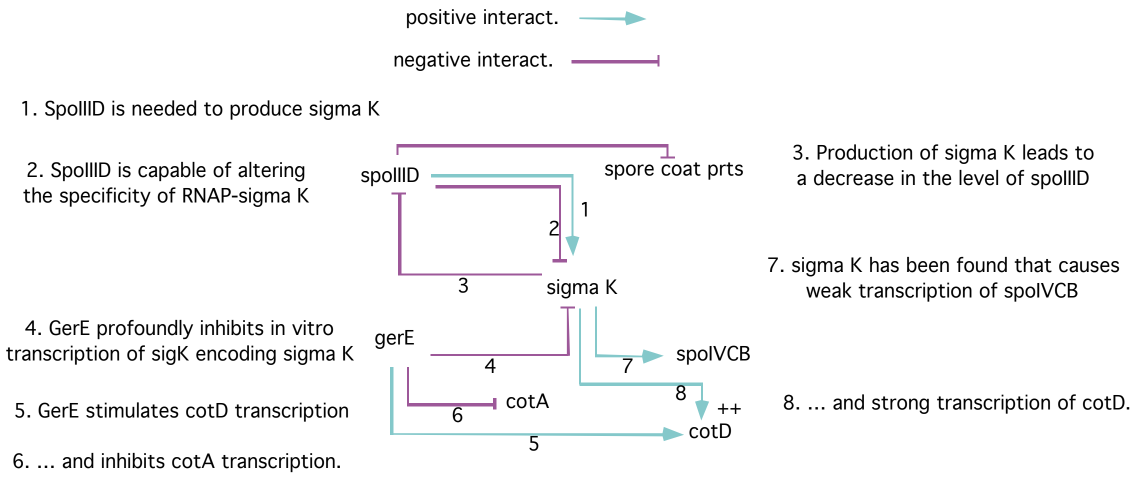
positive interact.
negative interact.
1. SpoIIID is needed to produce sigma K
2. SpoIIID is capable of altering the specificity of RNAP-sigma K
3. Production of sigma K leads to a decrease in the level of spoIIID
4. GerE profoundly inhibits in vitro transcription of sigK encoding sigma K
5. GerE stimulates cotD transcription
6. ... and inhibits cotA transcription.
7. sigma K has been found that causes weak transcription of spoIVCB
8. ... and strong transcription of cotD.
spoIIID
spore coat prts
sigma K
gerE
spoIVCB
cotA
cotD
++
1
2
3
4
5
6
7
8

Figure 1. Example of a regulation network

The arrows in Figure 1. represent the interactions between proteins and genes of *Bacillus subtilis* involved into the sporulation process. The numbered textual annotations around represent the fragments of MedLine abstracts the interactions have been extracted from.

2.2 Learning Language in Logic motivation

Applying IE to genomics and more generally to biology is not an easy task because IE systems require deep analysis methods to extract the relevant pieces of information. As shown in the example, retrieving that GerE is the agent of the inhibition of the transcription of the gene sigK requires at least coordination processing and syntactic dependency analysis *(e.g.* GerE is the subject of inhibits and cotA transcription is the object of inhibits). Such a relational representation of the text motivates relational learning to be applied to automatically acquire the information extraction rules.

For instance:
genic_interaction(X,Z):-
is-a(**X**,protein), **subject(X,Y)**, verb(**Y**), is-a(**Y**,interaction_action), **Obj**(**Z**,**Y**), is-a(**Z**,gene-expression).

Interpretation of the rule
If the **subject X** of an interaction action **verb Y**, is a protein name, and the **object Z** is a gene name or a gene expression, **then**, X is the *agent* and Z is the *target* of the interaction.

2.3 Expected impact on Machine Learning research and field of interest

Information Extraction has been a ML application area since the beginning of the nineties. However, most of the work focuses on the named-entity recognition problem with mainly statistics-based methods applied on shallow text representations. There were few attempts to develop ML methods for extracting relations from text although the development of relational methods and inductive learning yield excellent results in other application areas. The main reason for the lack of relational learning development in IE is due to the lack of dataset in IE that ML researchers could use without any investment in natural language processing (NLP). Indeed, relational event extraction requires that the text is deeply processed by syntactic parsing including syntactic dependencies. Most of the ML research groups do not have the NLP competencies and tools for performing this processing in specific domains with a good quality level. As a consequence, the training data set has been prepared so that ML researchers only could perform basic format change to be able to apply their methods.

The LLL challenge data set meets this requirement. Its use does not need any investment in biology neither in NLP. All the needed information is provided at a good quality level. The syntactic dependencies, which are critical here, have been automatically produced by LinkParser (Sleator and Temperley, 1993) and manually crosschecked by specialists of syntactic analysis of MIG and LIPN laboratories.

The expected impact on ML is a growing interest for IE and more generally for semantic knowledge learning from textual data. It is a great opportunity for ILP to evaluate, compare, adapt and develop methods on a large application domain that is critical from both a research and economic point of view. For instance, automatically producing meta data for the semantic Web from textual Web pages is strongly related to this ML and IE domain.

Moreover, the biologist expectations are very high and the particular task proposed here is not artificial but is critical in functional genomics. Even a partial automatization of the information extraction would be a considerable progress. We also expect a high impact of the availability of this data on the development of ML in bioinformatics for the access to textual content.

3. Description of the data

The challenge focuses on information extraction of gene interactions in *Bacillus subtilis*. Extracting gene interaction is the most popular *event* extraction task in biology. *Bacillus subtilis* (*Bs*) is a model bacterium and many papers have been published on direct gene interactions involved in sporulation, as opposed to what happens for eukaryotes. The gene interactions are generally mentioned in the abstract and the full text of the paper is not needed here. The relevant abstracts have been selected by querying MedLine on Bacillus subtilis transcription and sporulation. The relevant information is mostly local to single sentences (Ding *et al.*, 2002). The main exception comes from coreferences. For instance, the gene/protein name is mentioned in a sentence and referred to in the form of a pronoun or an hyperonym in the next sentence. We do not consider this case here. The abstracts have been segmented into sentences. Sentences have been automatically filtered by the STFilter system in order to retain those that contain at least two gene/protein names and are most probable to denote interactions (Nedellec *et al.*, 2000). MIG-INRA expert biologists have annotated with the XML editor CADIXE[1] hundreds of the interactions and the experimental conditions. For this challenge, a simple subset of them is provided as training and test data. The protein/gene names that can play the roles of agent and target of the gene interaction in the data sets are also recorded in a named-entity dictionary in the form of lists of canonical forms and variants. There could be more than one interaction per sentence and a given protein / gene may be involved in several interactions in different roles, agent or target.

3.1 Biological typology

The data has been selected on the following basis, the gene interaction is expressed,

- By an **explicit action** such as, ***GerE** stimulates **cotD** transcription*
- Or by a **binding of the protein** on the promoter of the target gene, *Therefore, **ftsY** is solely expressed during sporulation from a **sigma(K)**- and **GerE**-controlled promoter that is located immediately upstream of **ftsY** inside the smc gene.*
- Or by **membership to a regulon** family, ***yvyD** gene product, being a member of the **sigmaB** regulon* [..]

The sentences relying on other biological models have not been considered. For instance, a very frequent case involves gene mutants where the role of the genes in the interactions can be derived from the comparison with the normal experimental conditions. Other biological models are less represented. Then, the three selected categories are well representative of the interaction distribution excluding the mutant category.

3.2 Linguistic typology

The data set is decomposed into two subsets of increasing difficulties. The first subset does not include coreferences neither ellipsis, as opposed to the second subset. The coreferences selected are kept very simple. Most of them are just appositions.

For example,

> *Transcription of the **cotD** gene is activated by a protein called **GerE**, [..]*
>
> ***GerE** binds to a site on one of this promoter, **cotX** [..]*

Notice that when the absence of interaction between two genes is explicitly stated, it is represented as interaction information.

For example,

> *There likely exists another **comK**-independent mechanism of **hag** transcription.*

3.3 Linguistic information

These two subsets are available with two kinds of linguistic information,

1. **The Basic data set** includes sentences, word segmentation and biological target information: agents, targets and genic interactions
2. **The Enriched data set** includes also lemmas and syntactic dependencies manually checked.

The corpora and the information extraction task are the same in both cases. The two sets differ only by the nature of the linguistic information available. The participants to the challenge were free to use or not this linguistic information or to apply their own linguistic tools. When publishing their results, the participants had to be clear about the kind of information that has been used for training the learning methods.

3.4 Data representation

The data representation is detailed on the Web site: http://genome.jouy.inra.fr/texte/LLLchallenge/ The training data includes the target information to be extracted, the agent and target of the interaction.

---

[1] It has been developed by the National inter-EPST Caderige project and mainly involves LEIBNIZ-IMAG, MIG-INRA, LIPN-CNRS and ENSAR-INRA. It is available on demand.

Example from the Basic data set:

```
ID      11011148-1

sentence     ykuD was transcribed by
SigK  RNA  polymerase  from  T4  of
sporulation.

words   word(0,'ykuD',0,3)
        word(1,'was',5,7)
        word(2,'transcribed',9,19)
        word(3,'by',21,22)
        word(4,'SigK',24,27)
        word(5,'RNA',29,31)
        word(6,'polymerase',33,42)
        word(7,'from',44,47)
        word(8,'T4',49,50)
        word(9,'of',52,53)
        word(10,'sporulation',55,65)

agents        agent(4)

targets       target(0)

genic_interactions
        genic_interaction(4,0)
```

There is one genic interaction involving one agent and target here. The arguments of the agent, target and genic-interaction literals refer to the unique identifier of the word.

Example from the enriched data set:

```
ID      10747015-5

sentence      Localization  of  SpoIIE
was  shown  to  be  dependent  on  the
essential cell division protein FtsZ.

words   word(0,'Localization',0,11)
        word(1,'of',13,14)
word(2,'SpoIIE',16,21)

lemmas lemma(0,'localization')
        lemma(1,'of') lemma(2,'SpoIIE')

syntactic_relations
        relation('comp_of:N-N',0,2)
        relation('mod_att:NADJ',13,10)
        relation('mod_pred:N-ADJ',0,7)
        relation('mod_att:N-N',14,13)

agents        agent(14)

targets       target(2)

genic_interactions
        genic_interaction(14,2)
```

The lemma of named-entities is the canonical form as defined in the associated named-entity dictionary. For instance, the canonical form of kinD is ykvD according to the dictionary. The syntactic relations are defined in the Syntactic Analysis Guidelines document. For instance, `relation('comp_of:N-N',0,2)` means that word 0 and 2, namely, 'Localization' and 'SpoIIE' are two nouns and SpoIIE is a modifier of Localization which is the head of the relation introduced by the preposition 'of'.

Participants were free to use all external information that they find useful, annotated Medline abstracts included. However, for this latter resource, they had to select abstracts later than year 2000 in order to avoid overlapping with the test data.

3.5 Training data set

The training set without coreferences includes 57 sentences describing *106 positive examples* of genic interactions:

- 70 examples of action
- 30 examples of binding and promoter
- 6 examples of regulon

The training set with coreferences includes 23 sentences describing *165 positive examples* of interactions with coreferences

- 42 examples of action
- 10 examples of binding and promoter
- 7 examples of regulon

There are then *271 training examples* in 80 sentences. The training data does not explicitly describe negative examples. A straightforward way for generating negative examples is to use the Closed-World Assumption: if no interaction is specified between two given biological objects A and B, then they do not interact and form a negative example. This way, they could be easily derived from the training data and the dictionary as near-miss examples.

3.6 Test set

The test data are examples from sentences following the same biological typology as the training data. The distribution of the positive examples among the biological categories (action, binding, promoter and regulon) and with / without coreferences is the same as in the training data. The test set also includes negative examples, namely sentences without any genic interaction. This set follows the same distribution as in the initial corpus selected by MedLine query and containing at least two gene names, i.e. 50 % of the sentences are negative. The test set includes 87 sentences describing *106 positive examples* of genic interactions:

- 55 examples of action
- 23 examples of binding and promoter
- 5 examples of regulon

There is no sentence in the test data with no clear separation between the agent and the target (*e.g.*, "gene products x and y are known to interact").

The distinction between the sentences, with and without coreferences is not done in the test set and is not known by the participants because the test data set also contains sentences without any interaction. Marking "coreferences" sentences in the test set would bias the test task by giving hints for identifying the sentences without any interaction. However, the distinction is taken into account by the score computation.

4. Information extraction task

Given the description of the test examples and the named-entity dictionary, the task consists in automatically extracting the agent and the target of all genic interactions.

In order to avoid ambiguous interpretations, the agents and targets have to be identified by the canonical forms of their names as they are defined in the dictionary and by lemmas in the enriched version of the data. Thus there are two ways of retrieving the canonical name, given the actual name.

The agent and target roles should not be exchanged. If the sentence mentions different occurrences of an interaction between a given agent and target, then the answer should include all of them. For instance, in *A low level of GerE activated transcription of cotD by sigmaK RNA polymerase in vitro, but a higher level of GerE repressed cotD transcription*. There are two interactions to extract between GerE and cotD.

5. Computation of the score

The evaluation is based on the usual counting of false positive and false negative examples and on recall and precision. Partially correct answers will be considered as wrong answers. By partially correct answer we mean answers where the roles are exchanged, or only one of the two arguments (agent or target) of the genic interaction is correct. The score computation has been measured by the organizers on the results provided by the participants by applying the score computation program available to download as well as the check format program. These official scores are compared in section 6. The details on how scores are computed can be found on line in the user's manual of the score computation program.

The learning methods have been trained either on the file without coreferences or with coreferences, or on both of them (union). The participants have to specify which data set they compete for, so that the score computation program takes it into account for computing the scores.

The organizers also provide Web facilities to the participants for automatically uploading result files and compute the scores on the test data after the result submission deadline. These results have been further improved by the participants after the deadline. These "non official" results are not considered here for comparison because of the risk of over-fitting on the test data. However, they are interpreted and analyzed in the participant papers in this volume.

6. Result interpretation and comparison

Six research groups have participated in the challenge by submitting the results of the test set. The papers reporting their method and results are included in this volume. This section compares the official results among the participants.

6.1 Participating systems

Group 1 (KMB, Univ. Berlin and EBI) has applied alignment and finite-state automata technology for generating IE patterns from the LLL data set and an additional corpus of 256 positive examples manually annotated. The corpus has been enriched by POS tags and a list of words denoting interactions.

Group 2 (CS, Univ. Sheffield) method generates candidate patterns from examples parsed by MiniPar and semantically tagged by WordNet and PASBio. The candidates are manually filtered and then generalized with respect to a similarity criterion with already learned patterns. The training set has been augmented by weakly labeled training examples (cocitations of genes and proteins from positive examples, occurring in new sentences).

Group 3 (HCS Lab, Univ. Amsterdam) has applied the rule induction method Ripper to lexical-semantic-syntactic subtrees obtained by unification of the enriched form of the training examples. The semantics is given by an *ad'hoc* ontology designed for the challenge purpose.

Group 4 (KDLab, Univ. Brno) has applied the ILP method Aleph on the enriched data set without coreferences. Two features have been added, POS tags by the Brill tagger and WordNet hyperonyms.

Group 5 (Biostats and CS, Univ. Madison) has applied the ILP method Aleph on the enriched data set with and without coreferences wrapped into Gleaner that selects the best point on recall-precision curves. The data sets have been preprocessed and enriched by 215 new predicates including position, neighborhood, typographic, syntactic, semantic (belonging to MesH) and counting features.

Group 6 (ICCS, Univ. Edinburgh) has applied ILP and Markov Logic methods on the data parsed by the CCG and CCG2sem parsers that build syntactic and semantic paths. The best results are obtained without such preprocessing.

6.2 Results

Most of the results were obtained from the test set without coreferences (Table 1). The ML method of

Group 1. and 6. have achieved the best F-measures with balanced recall and precision around 50 %, which is high compared to other challenges on event or relation extraction such as the Succession Management MUC competition. Both systems are based on the representation of the examples as sequences. It would be interesting to study the role of the semantic tagging of word denoting interaction as done by Group 1. The other methods achieved a high recall but a poor precision. The reasons for such an overgeneralization could be explained by the fact that the training data did not include sentences without any interaction, as opposed to test data. The systems trained without such sentences or on weakly labeled additional data could have been thus handicapped. The results obtained with and without linguistic information cannot be easily compared here, since only Group 5. has provided results on both data sets. The role played by the syntactic dependencies cannot then be analyzed.

Table 1. Results on the test set without coreferences

Gr. #	Basic test set			Enriched test set		
	prec.	rec.	F	prec.	rec.	F
1.	50,0	53,8	51,8			
2.	10,6	98,1	19,1			
4.				37,9	55,5	45,1
5.	25,0	81,4	38,2	20,5	90,7	33,4
6.				60,9	46,2	52,6

Table 2. Results on the test set with coreferences

Gr. #	Basic test set			Enriched test set		
	prec.	rec.	F	prec.	rec.	F
5.	14,0	82,7	24,0	14,0	93,1	24,4

Table 3. Results on the test set with and without coreferences

Gr. #	Enriched test set		
	prec.	rec.	F
3.	51,8	16,8	25,4
6.	55,6	53,0	54,3

Table 2 presents the results as obtained on the test data with coreferences while Table 3 presents the results as obtained on the union of the test data with, and without coreferences. As shown by Table 2, the F-measure of Group 5 on the basic and linguistically enriched data set is not significantly different, as it is the case in Table 1. In all cases, the precision is poor, the recall high and the recall improved by the linguistic information.

Only the two groups 3. and 6. have provided results on the union of both test sets with and without coreferences. In both cases, the linguistic information has been exploited. Surprisingly, despite the difficulty of dealing with coreferences, the scores obtained on the set without coreferences (Table 1.) are similar: 52,6 against 54,3. Note that most of the coreferences in the test set were denoted by simple appositions and represented by explicit syntactic dependencies.

7. Conclusion

The high scores (more than 50 %) yields by the best system as well as by further experiments done by the other participants are very encouraging. As described in section 3., the data have been carefully selected in order to keep the underlying biological models simple. The parsing results as computed by LinkParser have been corrected by hand. The next challenges now consist in extending the data sets so that it becomes more representative of the real data as it can be found in MedLine abstracts and leave the syntactic parsing partially incorrect as it is when produced by automatic methods. The influence of the domain knowledge such as for instance, semantic classes of actions and their role in interactions has not been fully explored here but only through *ad'hoc* lists or patterns. It would certainly worthwhile to explore this direction.

Acknowledgements

MIG-INRA laboratory and LIPN-RCLN group have been deeply involved in the data preparation, biological annotation and syntactic dependency checking. The research and software development have been funded by the Caderige Project, Inter-EPST bioinformatics (1999-2003), ExtraPloDocs, RNTL (2000-2005) and Alvis, FP6-IST-STREP (2004-2007). The data preparation has been partially founded by Pascal FP6-IST-NoE (2004-2007).

References

Alphonse E., Aubin S., Bessières P., Bisson G., Hamon T., Lagarrigue S., Nazarenko A, Manine A.-P., Nédellec C., Ould Abdel Vetah M., Poibeau T. and Weissenbacher D. (2004). Event-Based Information Extraction for the biomedical domain: the Caderige project. Proceedings of the *Workshop BioNLP (Biology and Natural language Processing), Conférence Computational Linguistics* (Coling 2004).

Ciravegna F. (2000). Learning to Tag for Information Extraction from Text. Proceedings of the ECAI-2000 Workshop on Machine Learning for Information Extraction, F. Ciravegna *et al.* (eds), Berlin.

Cohen A. M, Hersh W. R. (2005). A survey of current work in biomedical text mining. Brief Bioinform. Mar;6(1):57-71..

Collier N., Ruch P. and Nazarenko A. (2004). Proceedings of the Joint Coling workshop on *Natural Language Processing in Biomedicine and its Applications*.

Daraselia N., Yuryev A., Egorov S., Novichkova S., Nikitin A., Mazo I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*. 22;20(5):604-11.

Ding J., Berleant D., Nettleton D., Wurtele E. (2002). Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomputing* pp. 326-37.

Ding J., Berleant D., Xu J., and Fulmer A. W. (2003). Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser. In *15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)*.

Freitag D. (1998). Toward General-Purpose Learning for Information Extraction. *Proceedings of COLING-ACL-98*.

Grover C., Lapata M., and Lascarides A. (2004). A Comparison of Parsing Technologies for the Biomedical Domain. *Journal of Natural Language Engineering*.

Hishiki T., Collier N., Nobata C., Ohta T., Ogata N., Sekimizu T., Steiner R., Park H. S., Tsujii J. (1998). Developping NLP tools for Genome Informatics: An Information Extraction Perspective. *Genome Informatics*. Universal Academy Press Inc., Tokyo, Japan.

Huang M., Zhu X., Hao Y., Payan D. G., Qu K., Li M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*. 12;20(18):3604-12.

Leroy G., Chen H., Martinez J. D. (2003). A shallow parser based on closed-class words to capture relations in biomedical text. *J Biomed Inform*. Jun;36(3):145-58.

McDonald D. M., Chen H., Su H., Marshall B. B. (2004). Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser., *Bioinformatics*. 12;20(18):3370-8..

Nédellec C. (2004). Machine Learning for Information Extraction in Genomics - State of the Art and Perspectives, *Text Mining and its Applications: Results of the NEMIS Launch Conference Series: Studies in Fuzziness and Soft Computing,* Sirmakessis, Spiros (Ed.), Springer Verlag.

Nédellec C., Ould Abdel Vetah M. and Bessières P. (2001). Sentence Filtering for Information Extraction in Genomics: A Classification Problem. In *Proceedings of the International Conference on Practical Knowledge Discovery in Databases* (PKDD'2001), pp. 326–338. Springer Verlag, LNAI 2167, Freiburg.

Ng S., Wong M. (2004). Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*. 10:104-112.

Ono T., Hishigaki H., Tanigami A., Takagi T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*. 17(2): 155-161.

Park J. C., Kim H. S., Kim J. J. (2001). Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In proceedings of *PSB'2001*.

Pyysalo S., Ginter F., Pahikkala T., Boberg J., Järvinen J., Salakoski T. and Koivula J. (2004). Analysis of Link grammar on Biomedical Dependency Corpus Targeted at Protein-Protein Interactions. Proceedings of the *Workshop BioNLP (Biology and Natural language Processing), Conférence Computational Linguistics* (Coling 2004).

Rindflesch T. C., Tanabe L., Weinstein J. N., Hunter L. (2000). EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. *Proceedings of PSB'2000*, vol 5:514-525.

Roux C., Proux D., Rechenmann F., Julliard L. (2000). An Ontology Enrichment Method for a Pragmatic Information Extraction System gathering Data on Genetic Interactions. Proceedings of the *ECAI'2000 Ontology Learning Workshop*, S. Staab *et al.* (eds.).

Sasaki Y., Matsuo Y. (2000). Learning Semantic-Level Information Extraction Rules by Type-Oriented ILP. *Proceedings of COLING-2000*, Kay M. (ed), Saarbrücken.

Sleator D. and Temperley D. (1993). Parsing English with a Link Grammar. In *Third International Workshop on Parsing Technologies*. Tilburg. Netherlands.

Soderland S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning Journal*, vol 34.

Temkin J. M., Gilder M. R. (2003). Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*. Nov 1;19(16):2046-53.

Thomas J., Milward D., Ouzounis C., Pulman S., Carroll M. (2000). Automatic extraction of protein interactions from scientific abstracts. *PSB'2000*, pp 541-52.

Valencia A. and Blaschke C., (2004). Proceedings of the workshop *A critical assessment of text mining methods in molecular biology*, Spain.

Yakushiji A., Tateisi Y., Miyao Y., Tsujii J.-I., (2001). Extraction from biomedical papers using a full parser. *Proceedings of PSB'2001*.