A Pragmatic Information Extraction Strategy for gathering Data on Genetic Interactions

Denys Proux
Xerox Research Centre Europe
6 Chemin de Maupertuis
Meylan, France, 38000
Denys.Proux@xrce.xerox.com

François Rechenmann
INRIA Rhône-Alpes
655 avenue de l'Europe
Montbonnot, France, 38330
Francois.Rechenmann@inria.fr

Laurent Julliard
Xerox Research Centre Europe
6 Chemin de Maupertuis
Meylan, France, 38000
Laurent.Julliard@xrce.xerox.com

Key Words

Information extraction, genomics, linguistics, conceptual graph.

Abstract

We present in this paper a pragmatic strategy to perform information extraction from biologic texts. Since the emergence of the information extraction field, techniques have evolved, become more robust and proved their efficiency on specific domains. We are using a combination of existing linguistic and knowledge processing tools to automatically extract information about gene interactions in the literature. Our ultimate goal is to build a network of gene interactions. The methodologies used and the current results are discussed in this paper.

Introduction

The current electronic revolution taking place via Internet and other networked resources giving an easy on-line access to large collections of texts and data to researchers offers lots of new challenges in the field of automatic information extraction and information synthesis (Appelt 1999). In the past several projects have aimed at providing full automatic systems performing information extraction from free texts. The first prototypes were designed to work on military corpora, attempting to detect intelligence data in newspaper articles or military reports (c.f. Message Understanding Conferences). Nowadays the techniques developed for that purpose can be applied on other domains. In genomics, electronic databases are increasing rapidly, but a vast amount of knowledge still resides in large collections of scientific papers such as Medline. These data remain to be exploited. Several research projects are working in that direction. Ohta et al. (Ohta et al. 1997) describe the IFBP (Information Finding from Biological Papers) system and its application to the construction of the Transcription Factor DataBase (TFDB). Thomas et al. (Thomas et al. 2000) adapt the SRI FASTUS system to gather data on protein interactions from Medline abstracts. Different approaches are in competition, based on statistics or linguistics, using deep or shallow parsing, applying simple pattern matching or complex knowledge processing tools (Zweigenbaum et al. 1994). Attempts to use learning mechanisms are also tested to reach that goal (Craven and Kumlien 1999).

In the context of a research project in the domain of genomics that involves biology and computer science laboratories, we are developing an information extraction system using a linguistic and a knowledge processing approach. The ultimate goal of the project is to feed an object-oriented knowledge base on molecular interactions with data on several organisms. The objective is thus to automatically build a network of gene or protein interactions using information extracted from scientific papers. One interesting specificity in genomics is the redundancy of information in texts. As a matter of fact, some topics are actively studied and a same piece of information often appears in various forms in a large number of texts. The redundancy increases the chances of detection by an automated system and is taken into account by our information extraction strategy.

In the following sections we describe the architecture we have adopted combining a linguistic and a knowledge processing approach. The strategy is based on pragmatic considerations and gives priority to robustness and efficiency over large corpora. Finally we present the validation process we have engaged and discuss the results obtained.

Overall architecture and resources

Our reference corpus is a set of 1200 sentences coming from Flybase the database on *Drosophila Melanogaster*. These sentences contain two gene names and have been checked by experts to determine whether they contain gene interactions. Our architecture, however, does not rely on any domain specific feature.

We have decided to deal with the information extraction problem using a pragmatic approach based on a combination of robust technologies that have proved their efficiency. Our system has adopted a two levels architecture. The first level relies on a linguistic analysis and the second one on a knowledge-based processing of the extracted sentence structures.

The linguistic components are mainly based on the Finite State Transducer technology known to be efficient in high speed text parsing (Koskenniemi 1997). The syntactic analysis of sentences appears to be one of the most time consuming task in the information extraction process. It consists of a Part of Speech tagger that has been slightly customized for our domain specific corpus, and a Shallow Parser (Ait-Mokhtar and Chanod 1997).

On the knowledge processing side of the system we have chosen to adopt an architecture based on conceptual graphs (Sowa 1984). An information extraction system can be effective without using such a sophisticated architecture, but our objective is not only to be able to extract some very specific data but also to get a more global understanding of what is extracted. The ultimate goal is to generate a synthesis of the facts described in the corpus. This kind of architecture (see fig. 1) combined with a domain specific ontology gives the system the ability to subsume related concepts to synthesize facts and get an abstracted view of information.

In a sense this approach can be compared to the one adopted by Rassinoux et al. (Rassinoux et al. 1994) for the HELIOS project. However our work put more emphasis on the robustness and the efficiency of the linguistic analysis based on new developments in parsing techniques.

These tools and the necessary adaptation we made are described in the following paragraphs.

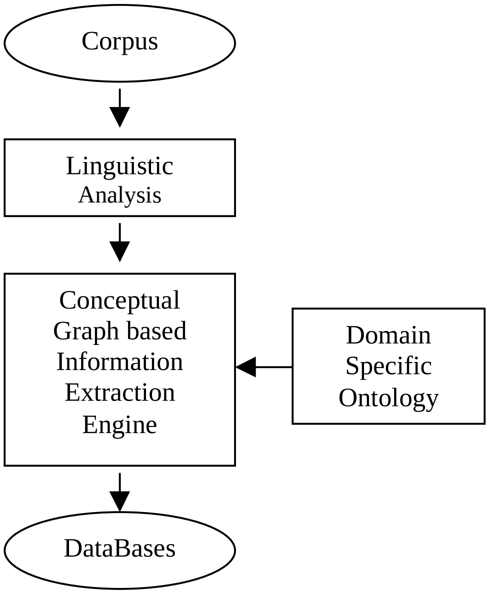

Figure 1: Overall architecture of the system displaying the information extraction process.

Linguistic analysis

The linguistic analysis of a text for information extraction purposes is a rather complex operation generally decomposed into a cascade of treatments. The first one is to get an accurate Part Of Speech (POS) tagging of every meaningful element of a sentence, which means giving a unique and reliable grammatical category to each word. In domain specific corpora this operation is complicated by the need to detect entity names, such as protein or gene names. This task has to be performed for two reasons. First of all to assign a correct POS tag to these entities (generally a proper noun), and then to help the information extraction system to establish semantic connections between these entities. As far as we are concerned we have chosen to take advantage of previous work done in POS tagging based on Finite State Transducers (FST).

The tagger we have designed operates using the following ordered process: tokenization, morphological analysis to provide possible POS tags (Schiller 1996), disambiguation using an HMM technique (Kupiec 1992), error corrections using specific modules adapted to domain specific corpora and vocabularies, and then a contextual lookup to identify gene names (Proux et al. 1998). The error recovery modules used at this level consist in a cascade of processes relying on domain specific dictionaries and on a morphologic analysis (prefix and suffix recognition). Protein names are handled in the same way. The following example (fig. 2) shows a sentence as it appears after the tagging phase.

```
Input:     "Scr is required to activate fkh expression."

Output:    #GENE# Scr    + PROP_NOUN
           is            + BE_VERB_PRES
           required      + VERB_PAST_PART
           to            + TO
           activate      + VERB_INF
           #GENE# fkh    + PROP_NOUN
           expression    + NOUN
```

Figure 2: Part of Speech tags generated for each token of a sentence.

This system has been tested on the Flybase corpus where we have reached the following results (see fig. 3).

Results on automatic detection of gene names	
Recall	94.4 %
Precision	91.4 %

Figure 3: Results obtained by our POS tagger for identification and tagging of gene names. These results have been obtained on a corpus of 750 sentences containing two gene names.

Once this first step is performed the output is processed to extract syntactic dependencies. Several strategies have been proposed to perform efficient syntactic analysis. However, the inherent complexity of natural language processing makes it almost impossible to obtain a fully accurate extraction of all these syntactic dependencies. The amount of processing needed to reach a very high level of accuracy is too heavy to be operational on very large corpora, without giving assurance of fully satisfactory results. As a consequence we opted for a shallow parsing strategy because of its robustness and its speed. The strength of Shallow Parsing is its ability to extract basic relationships such as subject-verb or verb-direct object very quickly and with high precision (see fig. 4). Its relative weakness in proper detection of prepositional phrase attachments is balanced in our system through the use of specific algorithms assigning a lower rate of confidence to the links between entities if these entities are related by a prepositional phrase connection. This kind of weak links are handled by the extraction mechanism based on the conceptual graph architecture.

Sentence: "*ems directly regulates sc function.*"

Subject	(*ems, regulate*)
Direct-Object	(*regulate, function*)
Adverb	(*directly, regulate*)
Nominal Noun	(*sc, function*)

Figure 4: Syntactic dependencies extracted by the Shallow Parser using Part Of Speech tags generated at the first step of the linguistic analysis.

Knowledge Processing

Once sentences have been parsed the syntactic dependencies extracted at the linguistic level are used to build the semantic representation. This task is handled by the knowledge-processing module. The core of the system has been designed around a conceptual graph management tool. So far several solutions for semantic processing have been proposed to tackle this problem. Our approach with such an architecture was guided by the need for a global understanding of what is extracted. We wanted to give the system the ability to synthesize the information using the power of subsumption on concepts and unification on sub-graphs. The syntactic dependencies extracted from sentences are used to build the dependency graph. Since prepositional phrase attachments are considered less reliable, the relations between entities connected by such links are considered as weaker. This fact is taken into account by the extraction mechanism using a weighting system that assigns a quality rating to each detected relation.

The construction of the semantic representation of sentences matches the following assumptions: The verb is considered as a key element in a sentence as it generally indicates the kind of action described. Therefore it is placed at the top of the conceptual graph structure symbolizing the sentence. Nouns appearing in subject or object groups, are connected to this verb through links representing their syntactic relation (see fig. 4). The user requests are stored in the system using exactly the same structure. It provides the user the opportunity to create request scenarios in natural language to search the corpus. This feature facilitates the task of building the request scenarios giving the opportunity to query a textual database without any previous knowledge engineering background as it is the case for classical information retrieval search engine based on key words. The creation of these request scenarios is made through the definition of information patterns as they can be found in the literature. At that point this method of preliminary identification of linguistic patterns can be compared to the one used by Blaschke et al. for their system of automatic extraction of protein-protein interactions (Blaschke et al. 1999). The following example (fig. 5) shows some request samples for gene interaction detections. The patterns have been abstracted from sentences found in the literature by replacing specific nouns or verbs with more generic terms that are likely to cover a larger spectrum of linguistic expressions.

Request scenarios:

gene interacts with gene.
gene induces the expression of gene product.
gene display a strong interaction on gene.
gene product exerts an effect on gene.
gene product acts as a modifier of gene.

Which can cover a sentence like:

"*Egl protein acts as a repressor of BicD.*"

Figure 5: Gene interaction patterns used as request scenarios.

In this example, it is the fourth request pattern that matches with the sentence. *"Egl"* matches with *"gene"*, *"protein"* with *"product"*, *"repressor"* with *"modifier"* and *"BicD"* with *"gene"* (according to the specification links indicated inside the hierarchy of concepts).

The inner specificity of a conceptual graph system gives the possibility to formulate very general requests involving concepts located at the top of the lattice, or more specific ones using very specific concepts. Generic requests will cover a large scope of information and specific ones only strictly selected information. The extraction mechanism tries then to establish a projection between user request graphs and the semantic representation of sentences to detect matching patterns. According to the request graph that matches, the corresponding information are then extracted. Projection mechanism in classical conceptual graph theory generally works in a unique way. It accepts a projection between two graphs if and only if one of them contains concepts and relations that are all more abstracted than those of the second graph. It appears that this kind of projection is too restrictive for the information extraction task. We use a more adaptive tool accepting projections between heterogeneous graphs.

Sentence: "*ems directly regulates sc function*."

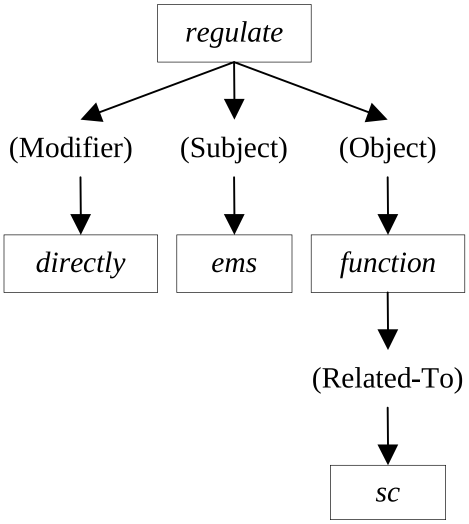

Figure 6: Conceptual graph generated using the syntactic dependencies extracted by the Shallow Parser (fig. 4). Expressions within brackets are semantic relations between concepts.

As for the results returned by the extraction mechanism, the quality of these results can be customized. The system may return only information matching strictly with terms of the user scenarios, or try to enlarge the requests to capture information that might also have some interest. This customization is made possible by the inherent capabilities of our conceptual graph management system and to several mechanisms that degrade requests when they failed to match. These mechanisms can extend the vocabulary, using reverse morphology, to provide from a single word other related words generated from its root (e.g. from *"repressor"*, we can generate *"repress"*, *"repressive*", …). Applying such linguistic techniques can significantly broaden the request coverage. Other functionalities automatically capture the semantic of multiple word expressions, or accept projection on concepts linked together by a common ancestor in the hierarchy (e.g. *"protein"* and *"gene"* can be seen as related by a common ancestor *"biological entity"*).

Conceptual graph theory presents an inherent problem for real life application, which is the creation and the maintenance of a hierarchy of concepts. This difficulty is addressed on our system by the ability to work with a two level ontology. The first one is limited and very domain specific in order to increase the quality of information detection and to reduce the cost of design. The second one is more general such as Wordnet and is used only to provide a more global vision of the text if needed.

First experiment

A first evaluation of the extraction mechanism has been conducted on a small set of 200 sentences from the Flybase corpus. The sentences have been divided in three categories: one containing gene interaction descriptions, a set of sentences containing no interactions (but at least two gene names in it) and a smaller set with sentences where experts have not been able to determine whether there were interactions or not. The average length was 18 words per sentence. The search engine was asked to extract information when two gene entities were connected to an interaction class verb.

On the set of sentences containing interactions, the system has been able to correctly detect it in 34% of the cases (fig. 7). This first result appears to be low in recall but according to the state of development of the system at the time this test was performed it was perfectly aligned with our expectations. As for the sentences where no interactions have been detected, we can classified them in the following categories.

In 32% of the cases (fig. 7), the main verb was too general to be classified as an interaction verb (e.g. "*Kr **is** a strong **repressor** of gt in the embryo*"). For this test the request scenarios were not designed to consider a "be" verb with a qualifier such as "repressor" or "activator" as relevant. Therefore the extraction mechanism was not supposed to retrieve it. This fact conducted us to design and implement for the new version a sentence simplification mechanism that automatically transform sentence like "*Kr **is** a strong **repressor** of gt …*" into "*Kr strongly **represses** gt…*" which modify a general verb into a specific verb using its qualifier to perform this operation. The goal of this functionality is to reduce the complexity of the request scenarios, decreasing the number of linguistic configurations needed to cover the different expressions of a same fact (e.g. *"is a repressor"*, *"represses"*, *"has a repressive impact"*, …).

In 6% of the cases, a Part of Speech tagging error has occurred on a critical word of the sentence, such as a verb (fig. 7). This error induced an incorrect syntactic dependency extraction and therefore bad semantic connections between entities. The POS tagger has therefore been modified to better take into account the specificity of the corpus to correct those problems.

In 17% of the cases, interaction descriptions were not clear enough for a non-specialist (fig. 7). Sentences like "*Lethality of three doses of Tpl can be rescued by dosage of the Is locus*." are not easy to handle by an automatic system. Those cases are problematic to solve, as it is even hard for a human being to decide.

The remaining 6% was due to misspellings, or unknown words, or miscellaneous problems (fig. 7).

Sentences with Interactions	
34 %	Good detection
32 %	Main verb too general
6 %	POS tag error
17 %	Ambiguous formulation
6 %	Miscellaneous

Figure 7: Results obtained from the set of sentences with interaction descriptions.

As for the sentences with no interaction in it (fig. 8), the system has extracted nothing in 80% of the cases, and detected something, for the remaining 20%. These wrong extractions were due in part to the specificity of the corpus. All these sentences have been selected because they contain two gene names. This unusual configuration increases the probability that the information displayed inside sentences look like one of the interaction scenarios provided to the search engine. The precision would rise with sentences taken blindly from a full text.

For the last set where the experts were not able to confirm an interaction, the system has detected nothing. Those cases can be combined to the set of sentences with no interaction, raising therefore the level of accuracy.

Sentences without Interactions	
80 %	No detection
20 %	Detection of something

Figure 8: Results obtained from the set of sentences without interaction descriptions.

Second experiment and validation

After this first experiment rich in learning, two major improvements were done to the system. The first one has been to increase the performance of the POS tagger to avoid cascading errors, and the second one to introduce a sentence simplification mechanism. A new series of tests has then been conducted on the Flybase corpus with the new system. It has been performed on 294 sentences with interactions, 288 sentences without interaction (but with two gene names in it), and 52 ambiguous sentences (with two gene names in it).

For the first set (fig. 9), the extraction engine has been able to detect a strong interaction (which means two gene entities related explicitly with an interaction class verb) in 44% of these sentences. A weak interaction (which means only a link between one gene entity and an interaction class verb, but without the other link to the second gene entity) in 26%, and nothing in 30%.

In the set of sentences with no interaction (fig. 9), the extraction engine has detected nothing (which was the objective) in 81% of the cases. For the remaining sentences it has detected a strong interaction in 7% of them, and a weak interaction in 12%. This level of errors can appears a little high but one should remember that these sentences have been selected because they contain two gene names which increases the risk of incorrect detection. The precision can be raised or decreased by rejecting or accepting the weak interaction detection.

As for the remaining set of ambiguous sentences (fig. 9), the results for no detection, and for a detection of a strong and a weak interaction are respectively, 83%, 13% and 4%.

Sentence sets	Detection	Results
With interaction	Strong	44 %
	Weak	26 %
	Nothing	30 %
Without interaction	Strong	7 %
	Weak	12 %
	Nothing	81 %
Ambiguous	Strong	13 %
	Weak	4 %
	Nothing	83 %

Figure 9: Results obtained with the new version of the system on sentences from the Flybase corpus

The next figure (fig. 10) display an example of results obtained by the system on a sentence from the Flybase corpus. This sentence has been identified by experts as “containing an interaction”. The request made to the system was to detect gene entities connected by interaction class verbs. The orientation of the interaction is obtained thanks to the recognition of subject-verb and verb-object dependencies.

The user can adjust the quality of recall and precision by allowing the capture of weak and strong detection increasing therefore recall, or just strong detection increasing then precision. Extraction of information in specific domains like genomics can also be helped by the redundancy of information in texts. Therefore the recall can be addressed by the following assumption: if a specific information occurs once in a document it is likely to appear again elsewhere in the document. So if the system did not detect it the first time because of an unusual formulation, we can consider that it might detect it in another sentence or at in another document (according to the information redundancy assumption). Based on that strategy we can assume that if a specific information is detected a great number of times, it is likely to be valid, at the information extraction level.

A new series of tests is planned on Medline abstracts to validate both the quality of results of the extraction mechanism and the assumption that recall can be improved using the redundancy of information.

“ *There is a distinct* ***signaling pathway activated*** *by*
egfr *that* ***interacts*** *with* ***ras85D signal transduction***

***cascade** to induce crossvein formation in the wing that might be used for signaling processes elsewhere in the developing fly .* **”**

Creation of the information scenario:

Key Entities wanted: *gene*
Key Verbs wanted: *interact*

Information extracted:

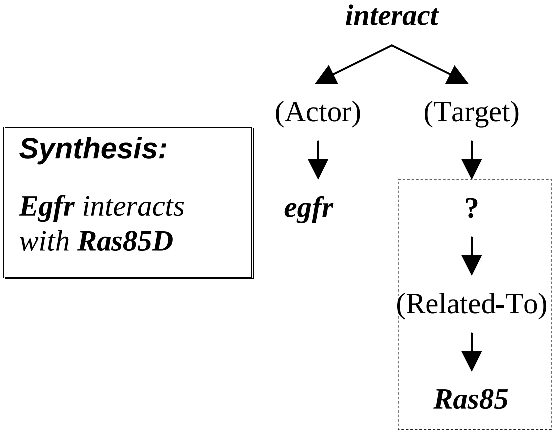

Figure 10: Information extracted on a real life example. The sub-graph detected in the sentence matches a graph generated according to the user request specifications.

Result analysis

According to the results produced by the second experiment it appears that in 30 % of the sentences expressing an interaction this one has not been properly detected. After close examination it appears that these sentences are often complex or ambiguous. The key entities were detected but the system was unable to identify the semantic links between these entities. This fact confirms the assumption we have that it is important to reduce the complexity of sentences before attempting to extract the semantic relations. The other possibility to reduce this percentage is to increase the number of linguistic formulae introduced in the request scenarios, but this would also increase the pain of building these scenarios. So an automatic way of reducing the complexity of sentences seems to be a much better approach. Next figure (fig. 11) shows two sample sentences where no detection has been performed.

> *“abd-A has no effect on the ability of scr to direct the formation of salivary glands .”*
>
> *“DNA sequence analysis reveals four E box binding site, for the binding of hetero-oligomeric complexes composed of da or AS-C proteins, in the first 877bp of the ac upstream region.”*

Figure 11: Sentences with interaction that has not been detected by the system.

The second experiment results also showed that the system has incorrectly detected an interaction in 7 % of the sentences where no interaction was expressed This can be explained by the fact that all these sentences contained at least two gene names. This linguistic configuration combined with the use of an interaction class vocabulary and with a possible bad syntactic dependency recognition can lead the system to extract wrong semantic relations. This consequence is often related to a complex sentence construction. The following samples shows two sentences where wrong interactions were detected.

> *“ubx and abd-A are required for the expression of the abdominal variant of the NB1-1 lineage.”*
>
> *“In the embryo, ac and sc are expressed coincidentally, at reproducible anterior-posterior and dorso-ventral coordinates, in clusters from which neuroblasts will arise.”*

Figure 12: Sentences without interaction but where the system has detected something between *“ubx”* and *“abd-A”* in the first one, and between *“ac”* and *“sc”* in the second one.

The weakness of the system is therefore related to the complexity of the input sentences. This is due to the linguistic approach chosen for the system. Speed and robustness are obtained at the expense of a deep syntactic analysis. A way to help the system is to reduce the complexity of these input sentences. An alternative is to focus on the precision, assuming that the redundancy of information throughout the literature will lead the system to reach a satisfactory recall value through the analysis of not one sentence but a full document or corpus.

Conclusions

We presented an information extraction system currently under development and evaluation that relies on a linguistic and on a knowledge processing approach. The linguistic tools in use are based on the Finite State Transducer technology, combining a Part Of Speech tagger and a Shallow Parser. The extraction mechanism is build around a conceptual graph architecture, adopting a domain specific ontology to improve its efficiency. This system embeds features such as automatic capture of semantics for compound nouns, customization of the level of relevance of extracted information using request graph degradations or automatic enlargement of the relevant vocabulary. Our information extraction strategy put the emphasis on the precision of extracted data, relying on the number of occurrences for a same data description in various papers to increase the recall.

We have started a first validation process on sentences from the Flybase corpus describing gene interactions. The first results were encouraging and gave us useful lessons for system improvements. Two main lines of modifications have therefore been implemented in the new version of the extraction system. These modifications intend to increase the Part Of Speech tagger accuracy on domain specific corpus as it remains one of the critical step of the analysis process. The other big improvement introduces an automatic sentence simplification mechanism using reverse morphology to help handling very general verbs such as "*be*", or "*have*", therefore transforming expressions like "*is a repressor*" into "*repress*".

A new series of tests have confirmed our expectations on the system performance and new tests are planned on abstracts from Medline to validate the approach on a much larger scale.

Referencing to our initial goal, our gene interaction detection system is still under development. The kernel and information extraction core functionalities are operational even if some necessary improvements are still needed. The module generating extracted networks is still under development.

Although the system is currently evaluated on genomics, it can be tuned to perform analysis on other domains provided that the domain specific ontology and the corresponding request scenarios are made available.

Acknowledgements

This work has been supported in part by a grant from CNRS (Centre National de la Recherche Scientifique) in the context of its Genome research program, and by ANRT (Association Nationale de la Recherche Scientifique) through a CIFRE grant for Denys Proux. We are grateful to Bernard Jacq from the LGPD (Laboratoire de Génétique et Physiologie du Développement) and to Violain Pillet from the CRRM (Centre de Recherches Rétrospectives de Marseille) for helpful discussions and for providing the first Flybase corpus. We also want to thanks Claude Roux from XRCE (Xerox Research Centre Europe) for its help on the conceptual graph aspects of the problem.

References

Aït-Mokhtar S., Chanod J.P., 1997. *Incremental Finite-State Parsing.* In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97), 72-79. Washington, USA, March 31st to April 3rd.

Appelt D. E. 1999. *Introduction to information extraction.* In Artificial Intelligence Communications. 12(3):161-172, ISSN 0921-7126.

Blaschke C., Andrade M., Ouzounis C. Valencia A. 1999. *Automatic extraction of biological information from scientific text: protein-protein interactions.* In Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB 99), 77-86. Heidelberg, Germany: AAAI Press.

Craven M., Kumlien J. 1999. *Constructing Biological Knowledge Bases by Extracting Information from Text Sources.* In Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB 99), 77-86. Heidelberg, Germany: AAAI Press.

Koskenniemi K., 1997. *Representations and Finite-State Components in Natural Language*. In Finate-State language Processing. Roches & Schabes Eds., 99-116, MIT Press.

Kupiec J. 1992. *Robust Part-of-speech Tagging Using a Hidden Markov Model.* In journal of Computer Speech and Language. Vol. 6.

Ohta Y., Yamamoto Y., Okazaki T., Uchiyama I., and Takagi, T. 1997. *Automatic Constructing of Knowledge Base from Biological Papers*. In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB'97), 218-225. Halkidiki, Greece: AAAI Press.

Proux D., Rechenmann F., Julliard L., Pillet V., Jacq B. 1998. *Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction.* In Proceedings of the Eight Workshop on Genome Informatics (GIW'98). 72-80. Tokyo Japan: Universal Academy Press, Inc.

Rassinoux A.M., Michel P.A, Juge C., Baud R., Scherrer J.R., 1994. *Natural Language Processing of Medical Texts within the HELIOS Environment.* In Computer Methods and Programs in Biomedicine, (45)79-96.

Schiller A. 1996. *Multilingual Part-of-Speech Tagging and Noun Phrase Mark-up.* In Proceedings of the 15th European Conference on Grammar and Lexicon of Romance Languages. Munich, Germany, September 19th to 21st.

Sowa, J.F. 1984. *Conceptual Structures.* Information Processing in Mind and Machine. Reading, Mass.: Addison-Wesley.

Thomas J., Milward D., Ouzounis C., Pulman S and Carroll M. 2000. *Automatic Extraction of Protein Interactions from Scientific Abstracts*. In Proceedings of Pacific Symposium on Biocomputing (PSB 2000), 5:538-549. Honolulu, USA: World Scientific Press.

Zweigenbaum P. and Consortium MENELAS. 1994. *MENELAS: an access system for medical records using natural language.* In Computer Methods and Programs in Biomedicine, (45)117-120.