HAL
archives-ouvertes.fr

Machine learning for information extraction in genomics: state of the art and perspectives

Claire Nédellec

▶ To cite this version:

Claire Nédellec. Machine learning for information extraction in genomics: state of the art and perspectives. 1. International workshop on text mining and its application, Apr 2003, Patras, Greece. hal-02764351

HAL Id: hal-02764351
https://hal.inrae.fr/hal-02764351

Submitted on 4 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Machine Learning for Information Extraction in Genomics – State of the art and perspectives

C. NÉDELLEC

Laboratoire Mathématique, Informatique et Génome (MIG), INRA, Domaine de Vilvert, 78352 F-Jouy-en-Josas

1. Introduction

The considerable development of multimedia communication goes along with an exponentially increasing volume of textual information. Information Retrieval (IR) technology provides information at a document collection level and thus it is not able to answer requests for specific pieces of information when needed. The development of intelligent tools and methods that give access to document content and extract relevant information, is more than ever a key issue for knowledge and information management. Information Extraction is one of the main research fields that attempt to fulfill this need. The IE field has been initiated by the DARPA's MUC program (Message Understanding Conference in 1987 (MUC Proceedings). MUC has originally defined IE as the task of (1) extracting specific, well-defined pieces of information from homogeneous sets of textual documents in restricted domains (2) in order to fill the slots of pre-defined form or templates. MUC has also brought about a new evaluation paradigm: the comparison of machine-extracted information to human-produced results. MUC inspired a large amount of work in IE and has become a major reference in the text-mining field. Even in the above restrictive definition, the design of an efficient IE system with good recall (coverage) and precision (correctness) rates remains a challenging task.

Building IE systems is time-consuming because even in the simplest case, they rely on manually encoded vocabularies and on extraction rules or patterns that are specific to the domains and the tasks at hand and therefore not easily reusable. In the more complex cases, they require linguistic analysis that involves lexical, syntactic and semantic resources proper to the domain. Therefore, the automated learning of resources and extraction rules for IE has appeared as very attractive since the early nineties (Riloff, 1993). In this area, the main research effort in machine learning (ML) has been devoted to named entity recognition and IE rules.

In biomedical domains as well as in many technical and scientific domains, researchers are looking to IE for tools that will enable them to deal with information overflow. In genomics, the demand for automating the access to the content of texts in electronic form, and for automated identification and interpretation of the relevant information in these texts, grew with the evolution of the research scope. Earlier approaches focused on a given specie metabolism and a limited set of genes; recent genome research applies experimental approaches, such as DNA chips, at the level of whole organisms. Access to many previous results in the form of textual information is essential to select promising subjects of study and to interpret the experimental results.

After sequencing, one of the next main challenges in genomics, is to identify the role of genes and proteins in regulation networks and metabolism. Unfortunately, most of the knowledge in functional genomics is not directly and easily retrievable from databanks; it is only available in scientific abstracts and articles written in natural language. However, most of the literature is available in large, open, online datatbases. For instance, the main generalist bibliographic database, MedLine, contains approximately 12 millions entries. Therefore, the capability to explor bibliographies and to extract useful knowledge from the literature would be a major advance toward developing functional models. Most of the few applications of IE to genomics are devoted to gene interaction, protein localization and function discovery. They have met with considerable interest in the bioinformatics community as demonstrated by the success of the text sessions at PSB and ISMB, the main bioinformatics conferences. Up to now, most of the IE methods applied to genomics rely on manually encoded resources. ML is mainly applied to named entity recognition. Some isolated but encouraging results have been obtained in learning lexical, syntactic and ontological knowledge for semantic labeling and in IE rule learning.

The specificity of the sublanguages of genomics makes existing dictionaries and lexicons of little use. However, as shown by (Harris *et al.*, 1989) in immunology, the variability of the sublanguages in specific research domains is limited: the vocabulary, the polysemy, the syntactic forms, the variety of concepts represented are restricted compared to wider domains. Therefore, the acquisition of linguistic resources and IE rules can be usefully based on the observation of lexical and linguistic regularities in selected documents from a specific domain. This idea is now being popularized in Machine Learning (ML) papers in the IE field and its application to genomics is starting.

The future directions of the domain are difficult to foresee, the domain being very new - the first papers were published in 1998; the research competencies required for developing such applications are diverse and the IE tasks require much more investment and expertise to be fruitful in this application domain than in MUC competitions.
The trend in current projects is towards the involvement of linguistic text processing and semantic knowledge, rather than shallow processing and simple IE patterns: segmentation into words, morpho-syntactic tagging (the part-of-speech categories of words are identified), syntactic analysis (sentence constituents such as noun or verb phrases are identified and the structure of complex sentences is analyzed) and sometimes additional processing such as lexical disambiguation, semantic tagging and anaphora resolution.
As in MUC, statistics provide good methods for low level tasks such as named entity recognition while more knowledge-intensive ML systems are applied to higher level tasks such as IE rule learning where more expressive representations and background knowledge are needed.
The field of genomics, like all quickly evolving research domains, raises problems that did not appear so crucial in MUC domains, such as the problem of feature selection and combination among the huge amount of candidate text features, the integration of existing resources with learned knowledge and the lack of standard corpora and expertise.

2. Information Extraction

A typical IE task is illustrated by Fig. 1 from a CMU corpus of seminar announcements (Freitag, 1998). IE process recognizes a name (*John Skvoretz*) and classifies it as a person name. It also recognizes a seminar event and creates a seminar event form (John Skvoretz is the seminar speaker whose presentation is entitled "Embedded commitment"). Even in such a simple example, IE should not be considered as a mere keyword filtering method. Filling a form with some extracted words and textual fragments involves a part of interpretation with respect to the "context" (*i.e.* domain knowledge or other pieces of information extracted from the same document) and according to its "type" (*i.e.* the information is the value of an attribute / feature / role represented by a slot of the form). In the document of Fig. 1, "4-5:30" is understood as a time interval and background knowledge about seminars is necessary to interpret "4" as "4 pm" and as the seminar starting time.

Document: Professor John Skvoretz, U. of South Carolina, Columbia, will present a seminar entitled "Embedded commitment", on Thursday, May 4th from 4-5:30 in PH 223D.
Filled form (partial)
place: *PH 223D*
starting time: *4 pm*
title: *Embedded commitment*
speaker: *Professor John Skvoretz* [...]

Fig 1. A seminar announcement event example.

IE overall process

Operationally, IE relies on document preprocessing and extraction rules (or equivalently extraction patterns) to identify and interpret the information to be extracted. The rules specify the conditions that the preprocessed text must verify and how the relevant textual fragments can be interpreted to fill the forms. In the simplest case, the textual fragment and the coded information are the same and there are neither text preprocessing nor interpretation.
More precisely, in a typical IE system, three processing steps can be identified (Hobbs *et al.* 1997; Cowie and Wilks, 2000):

1. *text preprocessing,* whose level ranges from mere text segmentation into sentences and sentences into tokens to a full linguistic analysis;
2. *rule selection*: the extraction rules are associated with triggers (*e.g.* keywords), the text is scanned to identify the triggering items and the corresponding rules are selected;
3. *rule application,* which checks the conditions of the selected rules and fills the forms according to the conclusions of the matching rules.

The rules are usually declarative. The conditions are expressed in a Logics-based formalism (Fig. 2), in the form of regular expressions, patterns or transducers. The conclusion explains how to identify in the text the value that should fill a slot of the form. The result may be a filled form, as in Fig. 1, or equivalently, a labeled text as in Fig. 2. The more explicit (*i.e.* the more semantic and conceptual) the IE rule, the more powerful, concise and understandable it is. However, it requires the input text being parsed and semantically tagged.
Extraction usually proceeds by filling forms of increasing complexity (Wilks, 1997):

- *Filling entity forms* aims at identifying the items representing the domain referential entities. These items are called "named entities" (*e.g. Analysis & Technology Inc.*) and assimilated to proper names (company, person, gene names) but they can be any kind of word or expression that refers to a domain entity.

- *Filling domain event forms*: The information about the events extracted by the rules is then encoded into forms in which a specific event of a given type and its role fillers are described. An entity form may fill an event role.
- *Merging forms* that are issued from different parts of the text but provide information about a same entity or event.
- *Assembling scenario forms*: Ideally, various event and entity forms can be further organized into a larger scenario form describing a temporal or logical sequence of actions/events.

As shown in Fig. 2, the condition part of the extraction rules may check the presence of a given lexical item (*e.g.* the verb *named*), the syntactic category of words and their syntactic dependencies (*e.g.* object and subject relations). Different clues such as typographical characteristics, relative position of words, semantic types or even coreference relations can also be exploited.

Sentence: "NORTH STONINGTON, Connecticut (Business Wire) - 12/2/94 - Joseph M. Marino and Richard P. Mitchell have been named senior vice president of Analysis & Technology Inc. (NASDAQ NMS: AATI), Gary P. Bennett, president and CEO, has announced."
Rule
Conditions:
noun-phrase (PNP, head(isa(person-name))), noun-phrase (TNP, head(isa(title))),
noun-phrase (CNP, head(isa(company-name))), verb-phrase (VP, type(passive),head(named or elected)),
preposition (PREP, head(of or at or by)),
subject (PNP, VP), object (VP, TNP), post_nominal_prep (TNG,PREP), prep_object (PREP, CNP)
Conclusion: management_appointment (M, person(PNP), title (TNP), company (CNP)).
Comment:
IF there is a noun phrase (NP) whose head is a person name (PNP), an NP whose head is a title name (TNP), an NP whose head is a company name (CNP), a verb phrase whose head is a passive verb (named or elected or appointed), a preposition of, at or by. If PNP and TNP are respectively subject and object of the verb, and if CNP modifies TNP,
THEN it can be stated that the person “PNP” is named "TNP" of the company “CNP”.
Labeled document
NORTH STONINGTON, Connecticut (Business Wire) - 12/2/94 - <Person>Joseph M. Marino and Richard P. Mitchell</Person> have been named <Title>senior vice president</Title> of <Company>Analysis & Technology Inc</Company>. (NASDAQ NMS: AATI), Gary P. Bennett, president and CEO, has announced.

Fig. 2. Example from MUC-6, a newswire about management succession

3. Machine Learning for Information Extraction

Among all IE tasks, most of the effort in Machine Learning has been devoted to named entity recognition and IE rule acquisition.

Named entity recognition (NER)

Recognizing and classifying named entities in texts require knowledge on the domain entities. Specialized lexical or keyword lists are commonly used to identify the referential entities in documents. Usual manual approaches also combine pattern matching with manually constructed dictionary in order to associate abbreviations, typographic and morpholological variations to the appropriate references. Semantic tagging by the type of the entities (company name, place, date) is quite straightforward in this case. The patterns may include constraints on the context of the entity to disambiguate the type if needed.

Hidden Markov Models (HMM) based on sequences of bigrams (pairs of tokens) has become a popular method for learning named entity recognition patterns from annotated corpora since Nymble (Bikel *et al.,* 1997). Simple bigrams appear as sufficient for learning efficient rules. In this framework, the learning problem comes to associate category tags, *(i.e.* the *entity* types and the *other* type) to the text words, according to the only previous word in the sentence. Named entities can be represented by compound nouns and not only simple nouns, then type categories can be associated to *type* "*beginning" tag* and *type* "*in" tag* while the rest of the words are tagged by the *other tag*. The HMMs differ in their ability to learn the model structure or not, in the way they estimate the transition probabilities (from training data or models built by hand) and in their reusability in different domains according to (Collier *et al.* 2000).

More recently, approaches based on the Maximum Entropy (ME) appear as very powerful and relevant (Mikheev *et al.*, 1998; Borthwick, 1999; Chieu and Ng, 2002). As in HMM, the method computes the probability to output a given label, given the word to tag. In this model, dependencies between word labels are easier to represent and the role of useful text features (simple words, case, length, POS tags, semantic categories, numbers, specific symbols, prefix, suffix, context) is coded is a more explicit way and easier to take into account.

Classical ML discriminant classification methods such as SVMs (Takeuchi and Collier, 2002; Isozaki and Kazawa, 2002), k-KNN, Neural Networks have also been applied. As for HMM, the learning task is coded as a classification problem where each term/word is associated to a tag.
Manually encoded patterns are generally more efficient but also more time-consuming. Then depending on the tasks and the type of entities, SVMs, ME and HMM yield more or less similar results.

Learning IE rules

In the classical framework, a ML system is fed with pairs of filled forms and annotated texts, where substrings in the text are associated to the filled slots in the form.
Learning can be then viewed either as a classification task (Freitag, 1998) (as illustrated by Fig. 3), where the extraction rules to be learned represent the conditions for filling a given slot, or equivalently, as pattern learning where the patterns are regular expression to be matched to text substrings.

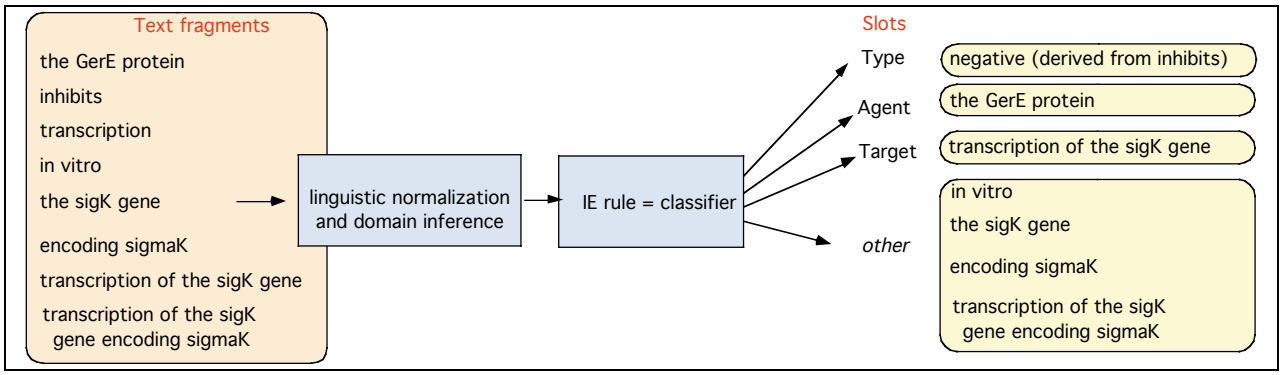


Fig. 3. IE rule learning viewed as a classification task

The learning methods then differ in:
The type of text: free, semi-structured, structured text, more or less domain restricted, (physician discharges, gene interactions, newswires about company joint ventures and terrorist attacks, job or seminar announcements).

- The type of slots to fill, (symbolic / numeric, text substring or more abstract);
- The role of the context of the relevant fragment in the text (size of the context);
- The type of features for describing the documents, which can be relational (relative position of two words, word neighborhood, syntactic relation, thematic role) or not (exact word, lemma, word position, part-of-speech tag, semantic category, case information);
- The use of additional lexicons (semantic categories, hyperonym links, thematic roles, case frames);
- The role of the user for annotating the examples and validating the result, (the whole document is classified as relevant or not, the text fragment is labeled with the slot, the sentence is labeled with a central concept, tags are inserted, seed semantic categories or seed patterns are provided, intermediate learned patterns are validated);
- The type of learning algorithm (case-based, naïve Bayes-based, grammatical inference, relational learning, ILP) and the learning steps (building a pool of good rules and then specializing them, refining the boundaries).

4. Information need in Genomics

Biologists can search bibliographic databases via the Internet using keyword queries that retrieve a large superset of relevant papers. Alternatively, they can navigate through hyperlinks between genome databanks and referenced papers. To extract the requisite knowledge from the retrieved papers, they must identify the relevant abstracts or paragraphs. Such manual processing is time consuming and repetitive, because of the bibliography size, the relevant data sparseness, and the database continuous updating. For example, the focused query "Bacillus subtilis and transcription" retrieves 2,209 abstracts. We chose this example because *Bacillus subtilis* is a model bacterium and *transcription* is a central phenomenon in functional genomics involved in genic interaction, a popular IE problem.

"**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.**"

Fig. 4. Example of sentence describing genic interactions.

Once relevant abstracts have been retrieved, there is no operational IE tool available in genomics and forms such as the one of Fig. 5 should be filled by hand.

Interaction	**Type:** positive
	Agent: *GerE*
	Target: `transcription of the gene` **`sigK`**

Fig. 5. Example of form describing a genic interaction.

However, applying IE *à la* MUC to genomics and more generally to biology is not an easy task because IE systems require deep analysis methods for the relevant fragments. As shown in the example Fig. 4, retrieving that GerE is the agent of the inhibition of the transcription of the gene sigK requires at least syntactic dependency analysis and coordination processing. In most of the genomics IE tasks (function, localization, homology) the methods should then combine the semantic-conceptual analysis of text understanding methods with IE through pattern matching.

5. State of the art in genomics

5.1. Document filtering

Information retrieval and more generally the management of document collections in biology are out of the scope of this paper. However, it is a prerequisite step to IE as the lack of robustness of the IE methods and their computational cost make them inapplicable to large corpora and to irrelevant documents. IR can then be viewed as a way to select the appropriate document subset for IE. In most of the applications, the target information is local to the sentence, or to the paragraph. Then, the next step consists of selecting the relevant text fragments within the set of retrieved documents. Classical ML- and statistics-based approach to document and sentence filtering have been applied to genomics. Among SVMs, naïve Bayes (NB) methods, Neural Networks, decision trees (Marcotte *et al*., 2001; Nedellec *et al*., 2001), NB methods coupled with feature selection seems to outperform the other sentence filtering approaches by yielding around 90 % precision and recall. No clear conclusion can be drawn from the linguistic-based representation change such as the use of lemmatization, terminology and named entities, as also observed in other domains.

5.2. Named entity recognition

Most of the work in IE application to genomics is devoted to NER. The main reasons are that this field has been deeply explored in MUC competitions and some of the genomics problems can be solved by a quite direct application of known methods; NER is a prerequisite step for many document processing tasks and not only IE; existing genomics dictionaries can be used as a starting point; the NER task raises difficult research problems because of the high variability in the name spelling and the incredibly large rate of word homology and ambiguity.

The entities to be recognized are mainly gene and protein names (Fukuda *et al.*, 1998; Proux *et al.* 1998; Cohen *et al.*, 2002; Franzen *et al.*, 2002), receptors, promoters, binding-sites, organs, organisms, species, molecular functions, phenotypes, diseases (Rindflesh *et al.*, 2000), syndroms, drugs, chemical compounds and experimental conditions. The limit between named entities and terms is often unclear.

The variations are graphical (*sigma K* / *sigma(K)* / *sigma-K*), morphological (*Down syndrom* / *Down's syndrom*), syntactic including co ordinations (*human cancer* / *cancers in human*, *human B- or T-cell lines* / *human B-cell lines*) and semantic (*rat somatotropin*, *rat growth hormon*). Synonymy may be due to renaming. For instance, genes may be renamed once their function is known (*SpoIIIG* / *sigma G*). Segmentation may be not obvious because of frequent ellipsis (*EPO mimetic peptide* / *EPO*) and syntactic variations. Abbreviations (*Bacillus subtilis* / *B. subtilis*) and acronyms (*chloramphenicol acetyltransferase* / *CAT*) are often used. Imprecise references are frequent, including anaphoric references, references to families and groups (*Rho family, protein kinases, globulins, eukaryotic RhoA-binding kinases*).

Correctly typing or categorizing is a much more difficult task than simply recognizing that a given word sequence is a named entity because of the frequent homologies. (Cohen *et al.*, 2002) observed for instance that the names produced by a simple typographic hyphenation variation refer to different entities in 85 % of the cases. This observation is based on LocusLink database and raises the question of the soundness of the source. Typing also includes finding the correct reference to the specie, which is often not trivial, as many gene and protein names are the same in different species.

5.2.1 Hand-coded patterns

Among the methods applied, only very few are ML- and statistics-based. While the pattern learning approach tends to use rather basic information from the text, the hand-coded pattern approach, on contrary, relies on multiple sources of information: on existing dictionaries and lexicon such as SWISSPROT, TREMBL, HUGO, UMLS among

others (Rindflesh *et al.*, 2000; Cohen *et al.*, 2002; Leonard *et al.*, 2002), character and word-based approaches, linguistic processing (Proux *et al.* 1998), contextual disambiguation and domain knowledge (Humphreys *et al.* 2000; Fukuda *et al*. 1998; Hishiki *et al.* 1998; Franzen, 2002 ; Narayanaswamy *et al.*, 2003). The experimental results are difficult to compare because of the lack of standard annotated corpora and share tasks apart the recent GENIA corpus (Ohta *et al*., 2002).

Combination of letters, digits and symbols (including Greek letters for instance) are representative of named entities (Franzen *et al.*, 2002) but also source of ambiguity. Specific patterns must be designed for excluding bibliographic references, chemical or arithmetic formula or sequences. Typographic variations (hyphenation, parenthesis, case) coded in patterns can be productive for named entity recognition from existing dictionaries although main cause of typing ambiguity (Cohen *et al.*, 2002). The only application of a simple edit distance (Cohen *et al.*, 2002) or protein name alignment algorithm such as BLAST (Krauthammer, 2000) for recognizing notational and typographic variations is not realistic without additional knowledge and constraints.

Hand-encoded patterns also include knowledge of the domain. For instance, proteins are often designated by their function (*growth hormon*), their localization or cellular origin (*HIV-1 envelop glycoprotein gp120*), their physical properties (*salivary acidic protein-1*) or homologue proteins (*Rho-like protein*). (Narayanaswamy *et al.*, 2003), among others, uses contextual semantic labeling of terms by domain knowledge to identify and disambiguate NE.

As usually in NER, signal words are very helpful. *Factor*, *receptor*, *enzyme*, *protein*, *particle*, *peptide*, *domain*, *terminal* (Franzen *et al.*, 2002), and *cell*, *clone* and *line for cells* as in the EDGAR system (Rindflesh *et al.*, 2000) can be used for example.

Morphological suffix and prefix can also be discriminant (*e.g. -in*, *-ase* for proteins). The linguistic processing, mainly morphological analysis, POS tagging and chunking must be adapted to the domain as shown by (Majoros *et al.*, 2003) that presents a HMM-based methods for POS-tagging of biomedical texts from an existing general trained HMM and training examples of the biomedical UMLS lexicon phrases. 1 % improvement only has been observed. Manual tuning of general POS taggers appears as more efficient and easier to implement.

The association of acronyms or abbreviations and their definition or expansion can be also done by hand-built regular expression (Pustejovsky *et al*., 2001; Yoshida *et al*., 2000; Schwartz and Hearst, 2003; Nenadic *et al.*, 2003) using external dictionaries, capitalization criteria, edit distance, parenthesis occurrence, distance between the acronym and its candidate expansion or syntactic information. See (Schwartz and Hearst, 2003) for a review of the methods and results. The homonymy problem is not correctly handled by this work.

More generally than entities, terms are extracted, classified and semantically typed by methods that combine dictionaries, distributional semantics and lexico-syntactical patterns in the line of (Hearst, 1992).

(Hishiki *et al.* 1998) gives examples of contextual regular expressions applied to term and entity recognition and categorization that rely, for instance, on:

- Indefinite appositions: the pattern NP(X), a NP(Y) gives X as an instance of Y, if Y is a type. From the sentence "csbB, a putative membrane-bound glucosyl transferase", csbB is interpreted as an instance of transferase if transferase is defined as a type.
- Exemplification of copula constructions: NP(X) be one of NP(Y) or NP(X) e.g. NP(Y). The fact that abrB is an instance of gene is extracted from "to repress certain genes, *e*.g. abrB".

Coreference resolution has also been recognized by MUC as necessary part of an IE system. In genomics, (Castano *et al.*, 2002) presents a hand-coded rule-based method for resolving anaphora in the specific cases of bio-entities represented by pronominal anaphors (*The S210A Spo0A mutant exhibited no change from wild-type binding, although* **it** *was defective in* [..]) and sortal anaphors (*Both SigK and gerE were essential for* **ykvP** *expression, and* **this gene** *was transcribed* [..]) but not event anaphora and cataphora, which are also frequent. The features include syntactic information (POS tag, number, person, definite/indefinite) and UMLS type as semantic information although the coverage of UMLS in genomics is quite loose. Resolution includes multiple antecedents (**Both proteins** *could be involved* [..]) and cascades of anaphors. The method weights the candidate antecedents according to classical constraints (same number and person), morphological preference (substring similarity) and semantic similarity according to the UMLS typing. The authors observe that surprisingly, syntactic dependencies such as subject-object badly affect the accuracy, while the type of arguments (subject and object) of some specific biological verbs used as constraints significantly improves it. As opposed to MUC, there is still no tentative in genomics for training such an algorithm.

5.2.2. ML for named entity recognition

Most of the ML- and statistics-based approach developed for the newswires of MUC competitions does not use sophisticated feature sets such as the ones required in genomics. Therefore few works only automates NER for genomics and the methods are more or less the same as presented in section 3 for the general case. The results are not

as good as those of hand-coded patterns and at this stage, these methods should more be seen as a help than as a way to fully automates the NER task. All methods use training corpus and include entity typing.

The work on NER in genomics is mainly by the group of the GENIA project (Collier *et al.*, 2000; Nobata *et al.*, 2000; Takeuchi and Collier, 2002; Kazama *et al.*, 2002). It makes comparison easier because the methods are generally applied to the GENIA corpus or a subset of it (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA). The current version of this corpus of 670 MedLine abstracts on transcription in human blood cells contains 23 793 entities from 24 different semantic classes.

HMM-based methods were the first machine learning methods applied to NE recognition in genomics (Collier *et al.*, 2000). The entities of the corpus of 100 MedLine abstracts from the GENIA corpus has been tagged by a domain expert. The 3500 entities identified belong to 10 categories (proteins, genes, cell line, location, etc.). The HMM is trained by bigrams training examples. The features describing the words are mainly character-based (digit, symbol, punctuation mark, etc.) plus determiner, and conjunction part-of-speech. No domain knowledge and linguistic-based feature is used. HMM post-processing corrects tags by comparing the tags of the different occurrences of the same word through the corpus, increasing the accuracy of 2.3%. The experiment suffers of lack of training data. The best recognized categories (proteins 76%, genes 47%) are also the most frequent and they benefit from the text features as opposed to the other categories. On the same 100 abstract corpus (Nobata *et al.*, 2000) compares a naïve Bayes (NB) based method using term lists and typed head nouns to a decision tree (DT) using chunking (shallow parsing). The NB method performs better on gene names (84%) while the DT method yields better results on protein names (85%) and on the other categories. This could be explained by the lack of data. No conclusion is drawn here on the role of chunking. Later work demonstrates the utility of larger word-window and more word-based and linguistic features such as morphology to the cost of generality.

For example, (Kazama *et al.*, 2002) presents an application of SVMs to the NE task in the GENIA corpus. The class of non-entity words of the corpus is split according to the POS tag information in order to make learning by SVMs tractable and it results in an accuracy improvement. SVM binary classification is extended to multi-class learning by a classical pair-wise with majority voting approach (Weston and Watkins, 1998). The examples are represented by vectors coding the following information for the preceding, current and following words: position of the word in lists of vocabulary, of POS tags, of suffixes, of prefixes, of substrings and of categories. The most informative features seem to be the class of the preceding words and the suffixes. Window of (-3, +3) size yields the best results. Compared with a Maximum Entropy method (ME), the SVM method with the polynomial kernel obtains slightly better results. It is noticed by the authors, that some of the useful character features (hyphens, numbers) used by the ME method (Kazama *et al.*, 2001) have been abandoned in this experiment for comparison reasons.

(Takeuchi and Collier, 2002) conclusions on SVMs are similar. They also noticed that SVMs seem to be sensitive to the problem of segmentation (dealing with complex expressions and hyphens). Compared to HMM, SVM obtain slightly better results that should be even improved by an adapted POS tagger and a better segmentation.

(Hanish *et al.*, 2003) proposes a hybrid approach including the use of dictionaries and hand-coded rules in combination with the optimization of the parameters of the score measure through a machine learning method, Robust linear programming (RLP). The results obtained for human are encouraging, but the problem of unspecific synonyms stays partially unsolved because of the lack of contextual linguistic analysis.

Tagging training corpora is time-consuming and is an obstacle to the popularity of ML-based methods. (Hatzivassiloglou *et al.*, 2001) proposes to use as positive examples the entity name recognized with the help of GeneBank database and that are directly followed by their types as *cwlH gene*. 2,65% of the entity occurrences are expressed in this way. Three ML methods have been applied, naïve Bayes-based, decision tree (C4.5) and rule-based (Ripper, by Cohen, 1996). They yield comparable results on the task of disambiguating protein, gene and RNA references. Text preprocessing consists of tokenization, stemming, stop-word removal, feature selection and POS-tagging. The word features include nearby words, distance from the nearby words, case and POS. Position information and feature selection decrease the accuracy for data sparseness reasons, while capitalization, stop word removal and stemming has a little positive effect but notably reduce the feature space, POS slightly improve the accuracy (1%). The performance is difficult to evaluate since human experts are less reliable than for other tasks. This is probably due to the fact that in many cases, the distinction between genes, proteins and messengers is irrelevant, as for instance in *The S210A Spo0A mutant exhibited no change from wild-type binding* [..]. In genic interactions, it is not necessary to explicitly distinguish the gene that expresses the protein from the protein itself.

Few works on IE only present hybrid approaches involving both hand-coded patterns and machine learning methods. However, (Tanabe and Wilbur, 2002) presents an interesting combination of the application of successive hand-coded heuristics and training phases for identifying gene and protein names. The first step trains Brill POS tagger augmented by the UMLS SPECIALIST lexicon, then false positive names are filtered through an anti-list and false negative names are filtered through LocusLink and GeneOntology. Compound word names are recovered with the

help of classical character and word-based criterion. Relevant trigger words and suffixes are identified by occurrence counting in UMLS. Bayesian learning is applied at a document level for discarding documents and then false positive names. Incorrect tagging of verbs as adjectives that yield to wrongly include verbs into terms is corrected by training a SVM as for instance in, *inhibiting NF-kappaB*.

Some automatic methods have been designed for retrieving acronyms or abbreviations and their definition or expansion (Chang *et al.*, 2002; Adar, 2002). According to (Schwartz and Hearst, 2003), they require time-consuming training data and the results are similar to those obtained by hand-designed algorithm and patterns. However, results are difficult to compare in genomics as the methods are applied on very different sets of data. No comparison has been done on a standard set.

For identifying synonyms (Nenadic *et al.*, 2002; Nenadic *et al.*, 2003) method does not use patterns in isolation but in combination with a distributional semantics based approach because synonymy extraction patterns are not as reliable as for hyponymy. Extraction patterns capture syntagmatic information whereas synonymy is a paradigmatic relation[1]. Similarities between terms are computed on the base of contextual (POS tag), lexical (same head / modifier) and syntactical cooccurrence counting (with the help of lexico-syntactic patterns such as enumeration and coordination patterns). These similarities are then combined in a hybrid *CLS measure* that computes the semantic similarity between pairs of terms.

5.3 Extraction rules

In a very similar way to what has been presented in section 5.2.2 for NE recognition task, the methods currently applied to the event extraction task in genomics are mainly based either on manual patterns including more or less linguistic processing, lexicon and domain knowledge, or on statistics-based techniques applied to very shallow representations of the text. Some notable effort is done in research projects such as Caderige (http://caderige.imag.fr), or BioMint (http://cui.unige.ch/AI-group/biomint) to apply ML methods such as ILP to more complex representations of the text after a deep morpho-syntactic and semantic analysis based on lexical and semantic resources specific to the domain.

The main attempts to information extraction in genomics aim at identifying the protein localization in the cell and at building enzymes and metabolic pathways, or regulation networks. Such networks are described by complex graphs of interactions between genes, proteins and environmental factors such as drugs or stress and can include phenotypic effects. The complete scenario should represent at least the entities, their reactions, their properties, their relations and, at a higher level, feedback cycles. In fact, single elementary and binary relations between entities are independently extracted by current IE methods. The integration of these elementary relations into a conceptual model highly depends on the other extracted facts and on wider knowledge of the domain. Few works address this interpretation and integration question. IE mainly adds new instances of the interaction relation in most of the cases. For instance, from the sentence "SpoIIID represses spoVD transcription" the new event Agent(Repress, SpoIIID) and Target(Repress, spoVD) is extracted (Roux *et al.* 2000).

We will first briefly sketch what has been done with hand-encoded patterns in order to give examples of the type of text feature that could be useful for automating the extraction.

5.3.1 Hand-coded patterns

Basically, hand-coded sets of patterns for genic interaction extraction are based on significant interaction verbs, entity names (protein and genes), POS-tagging, and possibly syntactic dependencies (Sekimizu *et al.*, 1998; Blaschke *et al.*, 1999; Rindflesh *et al.*, 2000; Thomas *et al.*, 2000; Ono *et al.*, 2001). Such patterns retrieve high-quality information but with a very poor recall. Our own experiments with such patterns —for example, [(Protein1/Gene1) *[2] (interaction verb) * (Protein2/Gene2) *]—yield a precision around 98 percent with a recall between 0 and 20% if the distance between the verb and the entities is constraints, otherwise, both precision and recall are low. The reason is that, even in technical and scientific domains, there are many ways to express given biological knowledge in natural language. In our corpus, only very few of the genic interactions are expressed by verbs but rather by names or more complex forms. Even in the case where the interaction is expressed by a verb, all the correct information may not so easy to extract because it requires to correctly identifying syntactic dependencies in complex expressions including coordination and embedded clauses as illustrated by the example Fig. 6.

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**..

Fig. 6. An example of a complex genic interaction sentence.

[1] Along the paradigmatic axis, the terms can substitute to each other; along the syntagmatic axis, terms rather tend to combine.

[2] * matches any string of any length (including zero).

The sentence describes five interactions, `sigma K` with `cotA` and `cotD` and `GerE` with `cotD`, `cotA` and `sigK`. GerE is the subject of the three interaction verbs although it occurs only once at the beginning of the sentence. Patterns able to handle such cases must include conditions on syntactic dependencies that are difficult to parse correctly. Some of the recent of works are based on predicate-argument structures (P-A structures), also referred as subcategorization frames that describe the number, the type and the syntactic construction of the predicate arguments (Yakushiji *et al.* 2001; Pustejovsky *et al.*, 2002). The P-A structures are used for extracting gene and protein interactions as shown in Fig. 7. The mapping between P-A structures and IE event frames is explicit and different P-A structures can be associated to a same event frame. For instance, the extraction of gene/protein interactions is viewed as the search for the subject and the object of an interaction verb that are interpreted as the agent and the target of the interaction. In these works, parsing is done by shallow, robust or full parsers, which handle or not coordinates, anaphora, passive mood and nominalization (Sekimizu *et al.* 1998; Thomas *et al.* 2000; Roux *et al.* 2000; Park *et al.* 2001; Leroy and Chen 2002). Additional semantic constraints may be added as selectional restrictions[3] for disambiguation purposes.

activate is an interaction verb
P-A structure of activate:

Predicate activate	**Frame: activate**
args: subject (1)	slot: agent (1)
object (2)	slot: target (2)

Fig 7. Example of a predicate-argument driven rule in functional genomics.

These approaches rely on the assumption that semantic relations (*e.g.* agent, target) are fully determined by the verb/noun predicate, its syntactic dependencies and optionally the semantic categories of its arguments, (Pustejovsky *et al*. 1993; Gildea and Jurafsky, 2002).

5.3.2 Statistics-based approach and shallow representation

In many cases, the genomic information is very redundant because papers will mention explicitely previous results that they complement or extend. Hopefully, the expression form changes from one occurrence to another, and one may expect that some of the forms are simple to handle. Thus, an attractive alternative to hand-coded patterns and deep syntactic analysis consists on applying robust statistics-based methods searching for relevant word co-occurrences in texts represented as bags of words (Blaschke *et al.*, 1999). For instance, if pairs of gene/protein names are encountered enough frequently in different sentences, one may conclude that they interact at a molecular level. Unfortunately such cooccurrence may reflect other relations than genic interaction, such as sequence or structure homology or co-localization. Moreover newly discovered interactions may not be retrieved because of the lack of citations, although they are the most interesting for the biologist. The nature of the genic interaction, positive or negative, direct or indirect is not easily identified once a significant level of cooccurrence is pointed out. Such an approach usually yields a rather high recall but a poor precision.

5.3.3 ML-based approach

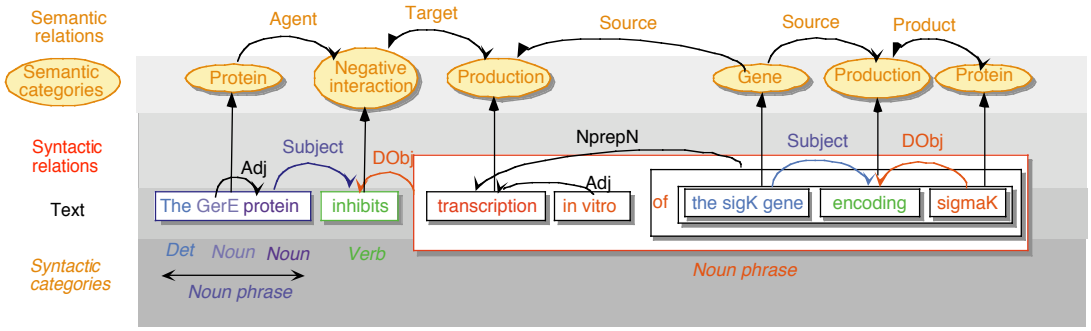
The ML-based approach appears as an attractive alternative to hand-coded patterns and statistics-based learning because it should be able to be more exhaustive than hand-coded patterns under the assumption of useful training example availability and it should be able to handle the complex text features that are needed for high precision. However, the cost of precisely annotating training examples is very high in the general case. There are very few publications on such attempts although some running projects explicitly include ML-IE based approach in the objectives. The training example annotation problem can be usefully overcome in the case where a subset of the target information is already available in a structured database. (Craven and Kumlien, 1999) illustrates this strategy on protein localization. Training examples are tagged with the help of the YPD database that describes protein localization and refers to the relevant bibliography. The sentences that include both a pair of protein name and a subcellular localization are tagged as positive. Examples are represented as bags of words. The classification algorithm is based on a NB method. Inter-corpus validation yields disappointing results because of YPD bias that focuses on yeast specie. Other experiments with an ILP-based method on parsed (POS, dependencies) and hand-annotated training examples result in more understandable IE rules with a better precision but a lower recall. The best compromise is obtained with the NB method.

6. Linguistics- and ML-based approach of IE in future genomics

Recent developments in IE involve more and more morpho-syntactic and semantic linguistic preprocessing and interpretation of text understanding methods (Yakushiji *et al.* 2001; Pustejovsky *et al.*, 2002, Tanabe and Wilbur

[3] A selectional restriction is a semantic type constraint that a given predicate enforces on its arguments.

2002, Franzen *et al.*, 2002; Nenadic *et al.*, 2003). In parallel, in NE recognition, as well as in IE rule, the applied ML-methods such as ILP, ME and SVMs tend to take into account more and more text features compared to the early works (Collier *et al.*, 2000; Craven and Kumilien, 1999). One of the main reasons is the lack of annotated training examples. The normalization of training examples using successive interpretation operations based on morpho-syntactic and semantic lexicon and processing, augments the regularities, reduces the need for training examples and makes learning easier. Fig. 8 shows the result of such a normalization on an example. This step can involve terminology, ontologies, and predicate argument structures to label the relevant terms and syntactic dependencies with the appropriate concepts. It relies on the fact that, in given specific domain languages, strong syntactic regularities make it possible to build a useful semantic structure.



Semantic relation	agent(Ger_protein, inhibit), target(transcription, inhibit), …
Semantic category	concept(Ger_protein,protein),concept(inibit,negative_interaction), …
Syntactic relation	subject(Ger_protein, inhibit), DObj(transcription, inhibit), …
Text	token(the), token (Ger_protein), token (inhibit), …
Syntactic category	cat(the, det), cat(Ger_protein, term), cat(inhibit, verb), …

Fig. 8. Example of sentence morpho-syntactic and semantic normalization.

High-level IE rules, with conditions that include abstract text features such as concepts, instead of a disjunction of specific words, can be learned from such representations. This eases learning, but also the readability, the revision and the maintenance of the rules.

Such normalization requires fine-tuned parsing tools, specific lexicons and dictionaries. More and more promising results, as shown above, demonstrate that these resources can be acquired with semi-automatic methods at a low cost. In the near future, these attempts should extend.

With respect to genomics, most of the work in IE has been done on human genic interaction. Human is a favorite specie because of the high expectation on short-term results for human therapies. However, there are more biological results in functional genomics available in databanks, today, about bacteria than about eucharyotes and these results could be usefully exploited for research in ML application to IE. Complementary to the bibliography, databanks are obviously useful sources of information at least for tagging the training examples.

Genic interactions might seem easier to extract because one could believe that most of them are described by a limited number of interaction verbs. Unfortunately, it is not the case and limiting information extraction to verbal forms would greatly affect the coverage of the results. Other very useful pieces of information, such as sequence homologies, functions, localizations are not expressed in a more complex way than genic interactions, and could therefore be extracted using the same technology.

MedLine is considered the main source of textual information for IE, although biologists view textual comments such as the ones of SwissProt database as important as well. Unfortunately, they seem to be more complex to process because they are in the form of short comments rather than well-formed sentences. Such sources should become more popular in IE in the future because of their high relevancy in genomics information discovery.

To summarize, the trend observed in recent publications is for the technology to meet the needs of the biologists for more precise and broad coverage information extraction. The availability of standard corpora and the organization of scientific events in text and bioinformatics such as workshop and conference text sessions should popularized this research domain in the near future.

References

Adar E. (2002). *S-RAD: A Simple and Robust Abbreviation Dictionary*. HP Laboratories Technical Report, Sept.

Bikel D. M., Miller S., Schwartz R., Weischedel R. (1997). Nymble: a High-Performance Learning Name-finder. *Conference on Applied Natural Language Processing*.

Blaschke C., Andrade M. A., Ouzounis C. and Valencia A. (1999). Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions," *Proc. Int'l Symp. Molecular Biology* (ISMB'99), AAAI Press, USA pp. 60-67.

Borthwick A. (1999). A Maximum Entropy Approach to Named Entity Recognition. *Ph.D. thesis*, Computer Science Department, New York University.

Collier N., Nobata C., Tsujii J. (2000). Extracting the Names of Genes and Gene Products with a Hidden Markov Model. *Proceedings of COLING-2000*, Sarrebrück.

Castaño J., Zhang J., Pustejovsky J. (2002). Anaphora Resolution in Biomedical Literature. *International Symposium on Reference Resolution*. Alicante, Spain.

Chang J. T., Schutze H. and RB Altman (2002). "Creating an online dictionary of abbreviations from MEDLINE". *J. Am. Med. Inform. Assoc*. **9**(6): 612-620.

Chieu H. L., and Ng H. T. (2002). Named Entity Recognition: A Maximum Entropy Approach Using Global Information. Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). (pp. 190-196). Taiwan.

Cohen K. B., Dolbey A. E., Acquaah-Mensah G. K. and Hunter L. (2002). Contrast and variability in gene names. *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*. pp. 14-20.

Cowie J., Wilks Y. (2000). Information Extraction. In R. Dale, H. Moisl and H. Somers (eds.) *Handbook of Natural Language Processing*. New York: Marcel Dekker.

Craven M. and Kumlien J. (1999). Constructing Biological Knowledge Bases by Extracting Information from Text Sources," *Proc. 7th Int'l Conf. Intelligent Systems for Molecular Biology* (ISMB-99), AAAI Press, USA, pp. 77-86, Heidelberg, Germany.

Franzen K., Eriksson G., Olsson F., Asker L., Liden P. and Coster J. (2002). Protein names and how to find them. *Int J Med Inf*. **67**(1-3): pp 49-61.

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

Fukuda K., Tamura A., Tsunoda T., Takagi T. (1998). Toward information extraction: identifying protein names from biological papers. *PSB'98*. pp 707-18.

Gildea D., Jurafsky D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245-288.

Hanisch D., Fluck J., Mevissen H. T., Zimmer R. (2003). Playing Biology's Name Game: Identifying Protein Names in Scientific Text *Pacific Symposium on Biocomputing* 8:403-414.

Hatzivassiloglou V. and Duboue P. A .and Rzhetsky V. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*. 17 Suppl 1: S97-S106.

Harris Z., Gottfried M., Ryckman T., Mattick P., Daladier A., Harris T. N., Harris S. (1989). *The Form of Information in Science: Analysis of an Immunology Sublanguage*, Kluwer Academic Publishers, Dordrecht.

Hearst M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of COLING'92*, pp. 539-545.

Isozaki H., Kazawa H. (2002). Efficient Support Vector Classifiers for Named Entity Recognition. *Proceedings of COLING-2002*, pp. 390-396.

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.

Hobbs J. R., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., Tyson M. (1997). FASTUS: A Cascaded Finite-State Transducer for Extraction Information from Natural Language Text. In E. Roche and Y. Schabes (eds.), *Finite-State Language Processing*, chapter 13, pp. 383-406. MIT Press.

Humphreys K., Demetriou G., Gaizauskas R. (2000). Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. *PSB'2000*, 5:502-513.

Kazama J., Makino T., Ohta Y. and Tsujii Y. (2002). Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the Workshop of the Natural Language Processing in the Biomedical Domain in ACL '02*, Philadelphia, PA, USA, July.

Krauthammer M., Rzhetsky A., Morozov P. and Friedman C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene*. 259(1-2):245-252.

Leroy G., Chen H. (2002). Filling preposition-based templates to capture information for medical abstracts. *PSB'2001*, Kaua'i, January.

Majoros W. H. and Subramanian G. M. and Yandell M. D. (2003). Identification of key concepts in biomedical literature using a modified Markov heuristic. *Bioinformatics*. 19(3): 402-407.

Marcotte E. M., Xenarios I., and Eisenberg, D. (2001). Mining litterature for protein-protein interactions. In *Bioinformatics*, vo. 17 n° 4, pp. 359-363.

Mikheev A. (1998). Feature Lattices for Maximum Entropy Modelling. In proceedings of COLING-ACL, pp. 848-854.

MUC Proceedings (1987-) Message Understanding conference.

Narayanaswamy M., Ravikumar K. E., Vi jay-Shanker K. (2003). A Biological Named Entity Recognizer. *Pacific Symposium on Biocomputing 8*.

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, Sept.

Nenadic G., Mima H., Spasic I., Ananiadou S. and Tsujii J. (2002). Terminology-driven literature mining and knowledge acquisition in biomedicine. *Int J Med Inf*. 67(1-3): 33-48.

Nenadic G., Spasic I. and Ananiadou S. (2003). Terminology-driven mining of biomedical literature. *Bioinformatics*. 19(8): 938-943.

Nobata C., Collier N. and Tsujii J. (1999). Automatic Term Identification and Classification in Biology Texts. In the *Proceedings of the fifth Natural Language Processing Pacific Rim Symposium* (NLPRS). Beijin, China. pp. 369-374.

Ohta T., Tateisi Y., Mima H. and Tsujii J. (2002). GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. *Proceedings of the Human Language Technology Conference*.

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*.

Proux D., Rechenmann F., Julliard L., Pillet V. and Jacq B. (1998). Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Informatics*. 9:72-80.

Pustejovsky J., Bergler S. and Anick P. (1993). Lexical Semantic Techniques for Corpus Analysis, in Computational Linguistics. *Special Issue on Using Large Corpora*: II, 19(2) pp. 331-358.

Pustejovsky J., Castano J., Cochran B., Kotecki M., Morrell M. and Rumshisky A. (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. Medinfo. 10(Pt 1):371-5.

Pustejovsky J., Castaño J., Zhang J., Kotecki M. and Cochran B. (2002). Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations. *PSB'2002*, 7:362-373.

Riloff E. (1993). Automatically constructing a Dictionary for Information Extraction Tasks. *Proceedings of AAAI'93*, Washington DC, pp 811-816.

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.

Schwartz A.S., Hearst M.A. (2003). A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. *Pacific Symposium on Biocomputing* 8:451-462.

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.).

Sekimizu T., Park H. S., Tsujii J. (1998). Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in MedLine Abstracts. In *Genome Informatics*. Universal Academy Press Inc., Tokyo, Japan.

Takeuchi K. and Collier N. (2002). Use of Support Vector Machines in Extended Named Entity Recognition. *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002),* Taipei, Taiwan, August.

Tanabe L. and Wilbur W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics*. 18(8): 1124-1132.

Thomas J. *et al.*, (2000). Automatic Extraction of Protein Interactions from Scientific Abstracts. *Proc. Pacific Symp. Biocomputing* (PSB'2000), vol. 5, pp. 502–513.

Weston J. and Watkins C. (1998). *Multi-class support vector machines*. Technical Report CSD-TR-98-04, Dept. of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England.

Wilks Y. (1997): Information Extraction as a core language technology. In *Information Extraction*, M. T. Pazienza (ed), Springer, Berlin.

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