Machine learning applied to information extraction in specific domains: an example: gene interaction extraction from bibliography in genomics

Claire Nédellec

Machine Learning applied to Information Extraction in specific domains — an example, gene interaction extraction from bibliography in genomics

Claire Nédellec

Laboratoire Mathématique, Informatique et Génome (MIG), INRA

nedellec@versailles.inra.fr

1. Introduction

As well as the generalization of multimedia communication, the volume of textual information is exponentially increasing. Today mere Information Retrieval technologies are unable to meet specific information needs because they provide information at a document collection level. Developing intelligent tools and methods, which can give access to document content, is therefore more than ever a key issue for knowledge and information management. Text content access is a crucial issue as much in the document engineering system of a small firm as in the document management of a whole scientific domain, whichever the source of information is: an Intranet information system or the "semantic web".

As soon as one wants to automate access to the content of texts in electronic form, one needs semantic knowledge to localize and interpret the relevant information. The acquisition of semantic knowledge is a well-known bottleneck for real-world applications, whichever technology is used (Information Extraction, Question/Answering, and more generally document engineering). There are two main reasons. Firstly, little semantic knowledge specific to application domains has been available because, until now, effort has been mainly devoted to the definition of formal languages for the representation of ontology and to the acquisition of generic knowledge bases, either lexical databases such as WordNet or EuroWordNet or general ontologies; CYC, for instance. In contrast, almost no community effort has been devoted to the acquisition of specific semantic knowledge that is required for particular applications and to the design of the acquisition methods that could be applied. We claim that no generic knowledge can be used as such and that the required semantic knowledge, even if it is derived from a generic source, must be specifically tuned to the application, domain and task that it will be used for. Although the process of acquiring this specific semantic knowledge cannot be fully automatic, methods and tools can be designed to efficiently help its acquisition. Secondly, it is also noticeable that there has been little dialogue between the various disciplines involved in knowledge acquisition and text analysis, although the integration of methods and tools from various disciplines is obviously needed. These disciplines include Information Science, Linguistics, Natural Language Processing, Knowledge Acquisition, Knowledge Representation, Machine Learning, Information Retrieval and Information Extraction. The Caderige project (http://www-caderige.imag.fr) is an example of such a collaboration in the domain of functional genomics. It involves four French laboratories, IRISA (ML and NLP), LIPN (KA, KR and NLP), LRI (ML) and MIG (genomics, ML and IE) and more recently a biotechnology company, Hybrygenics.

After sequencing, the next challenge in genomics is identifying the role of genes in interaction networks. Genome research projects have resulted in new experimental approaches, such as using DNA chips, at the level of whole organisms. Such chips provide comprehensive data about gene activity, so a research team can quickly produce thousands of measurements. More than ever, these new lab technologies are calling for fast and efficient access to previous results to interpret elementary measurements from the laboratory. Unfortunately, most functional genomics knowledge is not described in databanks; it is only available in scientific abstracts and articles written in natural language. For instance, the main generalist bibliographic database, Medline, contains approximately 12 millions entries. Efficiently using previous research results requires automating access to bibliography content. Therefore, exploring bibliographies and extracting knowledge from literature is a major milestone toward developing functional models of gene interactions.

In our opinion this new challenge offers as many benefits and present the same level of technical difficulty as other more popular bioinformatics challenges such as designing predictive algorithmic models. Moreover, AI research in natural language processing (NLP), information extraction (IE), machine learning (ML), and genomics have now reached the stage where automating IE from genomics literature is a realistic and exciting research goal. The specificity of the genomics bibliography, compared to other domains, justifies the expectation for short-term and high-quality results. We will illustrate this claim in the following by a genomic example about information extraction of gene interaction in *Bacillus subtilis*.

2. An example of the IE problem 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 the corresponding papers. To extract the requisite gene interaction knowledge from the retrieved papers, they must identify the relevant fragments (see the bold text in Figure 1). Such manual processing is time consuming and repetitive, because of the bibliography size, the relevant data sparseness, and the database continuous updating.

UI - 99175219
AB - GerE is a transcription factor produced in the mother cell compartment of sporulating Bacillus subtilis. **It is a critical regulator of cot genes encoding proteins that form the spore coat late in development.** Most cot genes, and the gerE gene, are transcribed by sigmaK RNA polymerase. **Previously, it was shown that the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK.** Here, we show that GerE binds near the sigK transcriptional start site, to act as a repressor […]

Figure 1. An extract of a Medline abstract on transcription in Bacillus subtilis.

Interaction	**Type**: negative **Agent**: GerE protein		
	Target:	**Expression**	**Source**: sigK gene **Product**: sigmaK protein

Figure 2. Information extracted from the second selected fragment

For example, the query "Bacillus subtilis and transcription" retrieves 2,209 abstracts such as the one of Figure 1. We chose this query example because *Bacillus subtilis* is a model bacterium and *transcription* is a central phenomenon in functional genomics. Gene functions are realized through gene transcription and protein production. The example of Figure 1 represents the problems posed by applying IE to a bibliography in genomics. Extraction involves understanding and requires expertise in biology. The information to be extracted is sparse in the document set. For instance, in the set of 2,209 abstracts I mentioned, only 3 percent of the sentences contain relevant information on gene interaction—that is, text that mentions the interaction's agents and type. Hopefully, in biology the bibliography is well structured and the information is local, mainly located in a single sentence or in a part of it as opposed to other domains where it is spread over the document. Many other biological phenomenon, such as translation or gene homology, raise similar IE problems.

3. Limitations of usual IE methods

Up to now, DARPA's MUC (Message Understanding Conference) program has defined automatic IE as the task of extracting specific, well-defined types of information from natural language texts in restricted domains. The objective is to fill predefined template slots and databases, such as shown in Figure 1b. In functional genomics, even such a restrictive view of IE is useful. Until now, no operational

IE tool has been made available in genomics, and extraction has not been automated.

However, applying IE à la MUC to genomics and more generally to biology is not an easy task because deep text analysis methods are needed to handle the relevant fragments. IE systems should combine the semantic-conceptual analysis of text understanding methods with IE through pattern matching, [Thomas *et al.*, 2000], [Blaschke *et al.*, 99], [Sekimizu *et al.*, 98], [Ono *et al.*, 2001]. Indeed, IE approaches to genomics, based either on predefined sets of fixed patterns, or on shallow representations of the text, yield limited results with either a bad recall or a low precision.

Hand-coded sets of patterns based on significant interaction verbs, gene names, or even syntactic tags and dependencies, [Blaschke *et al.*, 99], [Thomas *et al.*, 2000], [Ono *et al.*, 2001], retrieve little high-quality information. Our experiments with such patterns (described in the IE literature in genomics)—for example, [(Protein1/Gene1) *[1] (interact/associate/bind) * (Protein2/Gene2) *]—yield a precision around 98 percent with a recall between 0 and 20 percent. The reason is that, even in technical and scientific domains, there are many ways to express given biological knowledge in natural language. Manually encoding all patterns encountered in a corpus is thus unfeasible due to cost and unreliability. Therefore, automatically learning such IE patterns or rules from corpus seems to be an appropriate solution. Additionally, building IE systems is time consuming if they rely on manually encoded dictionaries and extraction rules or patterns that are specific to the domains and tasks at hand and they are not easily portable.

At the opposite end, some methods are based on statistic measures of keywords and gene name co-occurrences, [Craven, 99] (for example, shallow information-retrieval-based techniques), [Blaschke *et al.*, 99]. They yield high recall and low precision because they assume that any pair of genes encountered in the retrieved sentences interact, which is not always true. Many false positives are thus retrieved because potentially discriminant keywords and gene names occur in sentences where the genes mentioned are *not* semantically related. The following example (Figure 3.a and 3.b) illustrates some of the problems encountered by both hand-coded patterns and statistic –based approaches. Figure 3.a gives an example of a sentence that cannot be handled by these approaches and Figure 3.b represents the correct gene interaction network that should be extracted from this sentence.

"**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**.". The sentence describes five interactions, sigma K with **cotA** and **cotD**, GerE with **cotD**, with **cotA** and with sigK.

Figure 3.a An example of sentence that cannot be handled by hand-coded patterns and pure statistic–based approaches

An intuitive pattern, such as the one mentioned above, (i. e. [(Protein1/Gene1) * (interact/associate/bind) * (Protein2/Gene2) *]), that would match any pair of gene or protein names and interaction verbs or nouns (framed in the figure 3.a), [Craven & Kumlien, 1999], [Blaschke *et al.*, 99], would retrieve many erroneous interactions from this sentence, such as cotD [...] inhibits [...] cotA. Additional criterion such as a maximum number of words between gene names would yield a better precision but would miss some interactions such as the inhibition of sigK gene transcription by GerE (28 words apart). Statistics and keyword-based approaches would select the relevant sentences but would not be able to determine the right interactions between the five different gene and protein names cited in Figure 3.a (in bold-faced text).

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

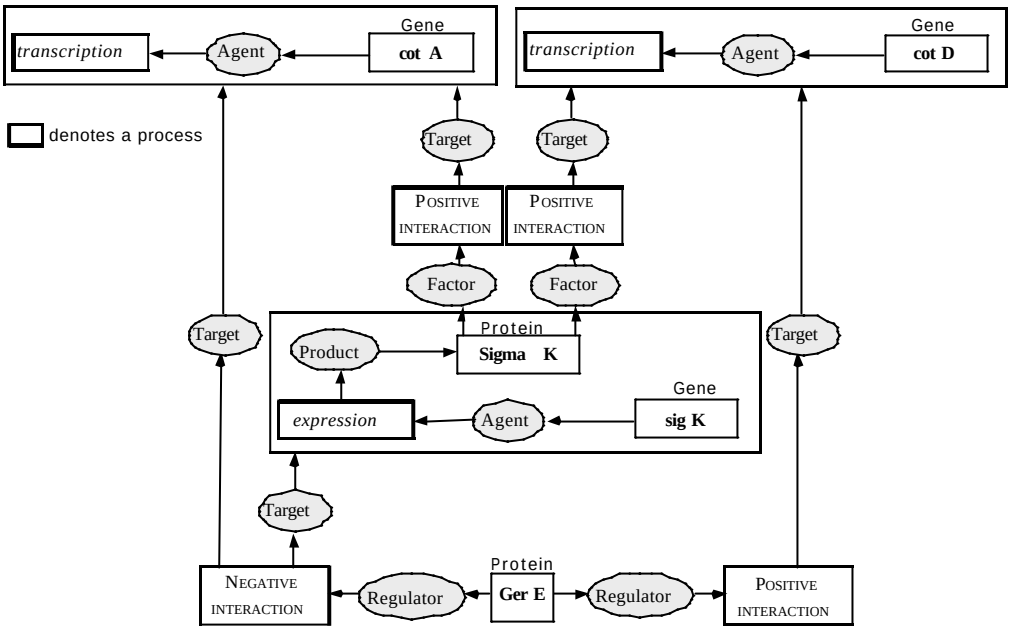

Figure 3.b The gene interaction network to be extracted.

Extracting relevant knowledge in the selected documents thus requires deeper syntactic and semantic analysis based on lexical and semantic resources specific to the domain. For instance, in Figure 2.a, identifying that GerE is the subject of the verb "inhibit" and that sigK is its direct object, and given that these relations are compatible with the conceptual agent and target roles, would improve the extraction's quality. To summarize, extraction patterns should be learned, because their manual development is unfeasible, and the learning should be based on syntactic–semantic regular expressions.

4. ML and IE today

Since the beginning of the nineties, automatically learning extraction rules from examples of pairs of filled patterns and annotated documents seemed like an attractive approach.[6] However, by the end of the decade, people were questioning the relative merits of the trainable and the knowledge engineering approaches—Doug E Appelt and David J. Israel, for example, discussed this issue at an IJCAI-99 (Int'l Join Conference on Artificial Intelligence) tutorial on IE (http://www.ai.sri.com/~appelt/ie-tutorial/). According to them, trainable (that is, statistics and ML-based) approaches should be preferred when the training data is cheap and plentiful, the extraction specifications are stable, and obtaining the highest possible performance is not a critical issue. They consider that the best recall the ML-based systems obtained is quite low compared to hand-coded IE systems. Appelt and Israel's analysis is based on the current state of the art in IE, in which existing ML-based systems exploit little, if any, background knowledge for guiding learning. The systems are often applied to a rather shallow representation of the training texts, and most of them are based on general-purpose ML algorithms —mainly *K* nearest-neighbor, grammatical inference, naïve Bayes methods, and top-down or bottom-up relational learning based on an exhaustive search or a local information gain measure.

Two related facts explain the limited range of these approaches, despite the rich spectrum of the modern state of the art in ML. First, according to the limited experiments performed, [Freitag, 98], on the common and quite simple IE tasks (MUC tasks, IE on the job, and seminar announcements), approaches based on linguistic analysis, lexical semantics, and informative representation of the training data do not perform much better than more shallow approaches. This does not encourage the design and application of novel symbolic and relational ML methods, which would be suitable for richer text analysis. Second, until recently, the main stream in text processing was mainly linguistic and statistic but not ML-based, besides

some notable exceptions such as S. Soderland's work and T. Mitchell's group research, [Soderland, 99] [Freitag, 98]. A large part of the effort in learning for IE, including genomic applications, has also been devoted to lower-level tasks such as named entity recognition, [Fukuda, 98]. This situation is evolving with the growing interest of the ML community in text processing and in IE in particular. Moreover, the growing demand for applications brings many new IE tasks, such as IE in functional genomics, that require a deeper understanding and consequently call for more sophisticated linguistics- and ML-based approaches [Craven & Kumlien, 99]. Additionally, in real-world applications, training complements rather than opposes knowledge engineering, as ontology-based and interactive approaches illustrate.

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

In Caderige, we view the genomics IE of the future as a three-step method. In the first step, we select the relevant textual fragments from all sentences in the papers, based on shallow criteria (for example, discriminant keywords or gene and protein names) to deal with the relevant data's sparseness. In the second step, we build a representation of the content of the fragments using successive interpretation operations based on syntactic–semantic lexicon, [Sekimizu *et al.*, 99], [Rindflesh *et al.*, 2000], following a classical approach in text understanding. Figure 4 shows an example of this phase's output. This step should involve terminology, ontologies, and predicate argument structures to label the relevant terms and syntactic dependencies with the appropriate concepts. In doing so, we rely on the fact that in the language of a given specific domain there exists strong syntactic regularities, which make it possible to build a semantic structure.

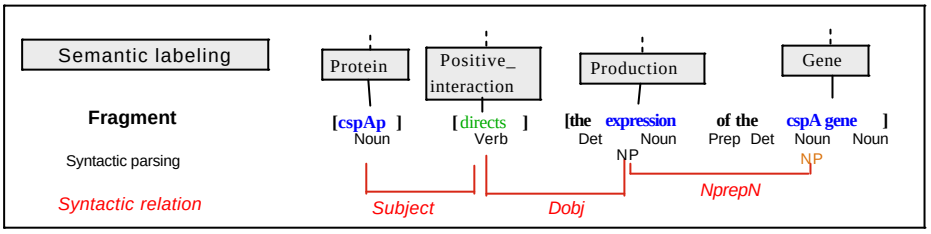

Figure 4. Example of syntactic-semantic interpretation (NP denotes noun phrases and Dobj denotes the Direct Object).

Finally, we apply extraction rules (see Figure 5) to the resulting text interpretation to identify the relevant information and store it in a database in the suitable format, or to fill forms as in MUC case. In this example, the IE is realized by transducers designed by Intex software, that insert XML labels in the text fragments when the syntactic and semantic conditions are verified. For example, the transducer in Figure 5 says states conditions, among others, there must be a noun phrase, subject of the verb and representing a protein (denoted by variable $1), and a noun phrase, direct object of the interaction verb (denoted by variable $2), representing a gene expression (denoted by variable $3). The gray boxes represent subtransducers. If all conditions are true, then XML protein, interaction and gene expression tags should be inserted (for example, see <protein>, <interaction> and <gene_expression> tags in the figure).

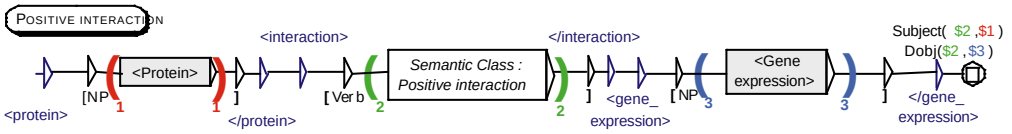

Figure 5. Example of extraction rule in the form of transducers for extracting gene interactions in functional genomics.

ML methods can help develop the knowledge bases needed for each step. For sentence filtering,

discriminant keywords are learnable by classification methods such as naive Bayes or Support Vector Machines, [Marcotte *et al.*, 2001], [Nedellec *et al.*, 2001]. For building terminologies and ontologies from parsed corpora or assisting their design, unsupervised methods such as conceptual clustering are appropriate [Nedellec & Faure, 98]. The many methods designed for semantic class learning, query expansion, word sense disambiguation, or for building restrictions of selection are easily applicable to ontology and subcategorization frame learning. Then, predicate argument structures are learnable from subcategorization frame clustering or from semantically labeled corpora. Finally, extraction rules or automata (see Figure 5) are learnable from annotated corpora (see Figure 6) at the suitable level of linguistic interpretation (see Figure 4), [Sasaki & Matsuo, 2000]. The feasibility of such learning tasks from parsed corpora has been shown many times in the framework of specific domains such as scientific ones.

```
<SENTENCE                name      = "2" >
        <INTERACTION
            id               = "1"
            type   = "Y"
            Previous studies showed that <Agent1 type="Protein" func="Factor"> spoIIID </Agent1> <Interaction> is
            needed to produce </Interaction> <Target1 type="SigmaFactor">sigma K</Target1> [...]
        </INTERACTION>
```

Figure 5. Example of annotated sentence for IE rule learning. The highlights indicate the graphic attributes of the XML tags. For example, the regions tagged as "Interaction" are underlined, and the regions tagged as Agent are in bold.

Superficial approaches will not sufficiently resolve the problem of building IE systems for genomics. However, given the specificity of the language used in genomics texts, we can solve the task by combining ML from a corpus of annotated and un-annotated texts with syntactic–semantic analysis. Genomics provides demanding problems that will stimulate the development of more sophisticated approaches in IE. Although these aspects of extraction are not yet in the mainstream of IE research, this seems a promising direction not only for genomics but more generally for biology and for other perhaps less technical domains. Preliminary work in this area has produced encouraging results that we should now extend and deepen.

Acknowledgments

The author thanks her collaborators in the Caderige project, in particular Philippe Bessières, Gilles Bisson, Adeline Nazarenko and Mohamed Ould Abdel Vetah for their contribution to the scientific program presented in the paper. Caderige project is partially funded by the French inter-EPST program in bioinformatics.
This paper is a completed and revised version of the paper published by the IEEE Intelligent System Journal (May-June 2002).

References

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

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

Faure D. and Nédellec C.,"Knowledge Acquisition of Predicate-Argument Structures from technichal Texts using Machine Learning" in Proceedings of *Current Developments in Knowledge Acquisition: EKAW-99*, p. 329-334, Fensel D. & Studer R. (Ed.), Springer Verlag, Karlsruhe, Germany, April 1999.

Freitag D., "Multistrategy Learning for Information Extraction," *Proceedings of the 15th International Machine*

Learning Conference (ML'98), A. Danyluk (ed.) Morgan Kaufmann, pp. 100-107, Madison, Wisconsin, 1998.

Fukuda K., Tsunoda T., Tamura A. and Takagi T., "Toward Information Extraction: Identifying Protein Names from Biological Papers," *Proceedings of the 3d Pacific Symposium on Biocomputing* (PSB'1998), 3:705-716 (http://www-smi.stanford.edu/projects/helix/psb98/), 1998.

Marcotte E. M., Xenarios I., and Eisenberg D., "Mining Literature for Protein-Protein Interactions," *Bioinformatics Journal*, vol. 17, no. 4, pp. 359–363, Oxford University Press Applications, 2001.

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

Ono T., Hishigaki H., Tanigami A. and Takagi T., "Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature," *Bioinformatics Journal*, vol. 17, no. 2, pp. 155–161, Oxford University Press Applications, 2001.

Riloff E., "Automatically Constructing a Dictionary for Information Extraction Tasks," *Proceedings of the 11th National Conference on Artificial Intelligence* (AAAI-93), pp. 811–816, AAAI Press/The MIT Press, Cambridge, Mass., 1993.

Rindflesh T., Tanabe L., Weinstein J.N. and Hunter L., "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature," *Proceedings of the 5th Pacific Symposium on. Biocomputing* (PSB'2000), (http://www-smi.stanford.edu/projects/helix/psb00/), pp. 514–525, 2000.

Sasaki Y. and Matsuo Y., "Learning Semantic-Level Information Extraction Rules by Type-Oriented ILP," *Proceedings of the 18th International Conference on Computational Linguistics* (COLING-2000), Morgan Kaufmann, Saarbrüken, Germany, 2000

Sekimizu T., Park H. S. and Tsujii J., "Identifying the Interaction Between Genes and Gene Products Based on Frequently Seen Verbs in MedLine Abstracts," *Genome Informatics*, pp. 62–71, Universal Academy Press, Tokyo, Japan, 1998.

Soderland S., "Learning Information Extraction Rules for Semi-Structured and Free Text," *Machine Learning Journal*, vol. 34, pp. 233-272, 1999.

Thomas, J., Milward, D., Ouzounis C., Pulman S. and Caroll M., "Automatic Extraction of Protein Interactions from Scientific Abstracts," *Proceedings of the 5th Pacific Symposium on Biocomputing* (PSB'2000), vol. 5, pp. 502–513, http://wwwsmi.stanford.edu/projects/helix/psb00/, 2000.