Genic Interaction Extraction by Reasoning on an Ontology

Alain-Pierre Manine, Erick Alphonse
LIPN, Univ. Paris13/CNRS UMR7030
Laboratoire d'Informatique Paris-Nord
Institut Galilée, Université Paris 13
99 ave. Jean-Baptiste Clément
F93430 Villetaneuse
{alainpierre.manine,
erick.alphonse}
@lipn.univ-paris13.fr

Philippe Bessières
MIG, INRA UR1077
Unité Mathématique,
Informatique et Génome
Institut National de la
Recherche Agronomique
F78352 Jouy-en-Josas
philippe.bessieres
@jouy.inra.fr

Abstract

Information Extraction (IE) systems have been proposed in recent years, to extract genic interactions from bibliographical resources. But they are limited to single interaction relations, and have to face a trade-off between recall and precision, by focusing either on specific interactions (for precision), or general and unspecified interactions of biological entities (for recall). Yet, biologists need to process more complex data from literature, in order to study biological pathways, so an ontology is an adequate formal representation to model this sophisticated knowledge. But the tight integration of IE systems and ontologies is still a current research issue, a fortiori with complex ones that go beyond hierarchies. Here, we propose a rich modeling of genic interactions with an ontology, and show how it can be used within an IE system. The ontology is seen as a language specifying a normalized representation of text. IE is performed by first extracting instances from Natural Language Processing (NLP) modules, then deductive inferences on the ontology language are completed. New instances may be infered, bringing together otherwise scattered textual information. We validated our approach on an annotated corpus of gene transcription regulations in *Bacillus subtilis*. We reach a global recall of 89.3% and a precision of 89.6%, with high scores for the ten semantic relations defined in the ontology.

1 Introduction

Interactions between genes and proteins were long studied, while most of their biological knowledge is not described in structured formats of genomic databanks, but scattered in scientific articles. For this reason, numerous works in recent years have been carried out to design Information Extraction (IE) systems, which aim at automatically extracting genic interaction networks from bibliography (Blaschke et al., 1999; Craven and Kumlien, 1999; Friedman et al., 2001; Krallinger et al., 2007). Relations between biological entities are multiple (protein and gene regulations, DNA binding, phosphorylation, homology relations, etc.). Nevertheless, most IE systems are limited to extract unique relations, and face a trade-off between recall and precision. Some focus on precision by extracting specific interactions, for instance between proteins (Craven and Kumlien, 1999; Rindflesch et al., 2000; Blaschke et al., 1999; Ono et al., 2001; Saric et al., 2005), whereas other stress on recall using general relations (Nédellec, 2005; Fundel et al., 2007). However, this does not take into account the complexity of the data processed by biologists, such as biological pathways (Oda et al., 2008). Therefore, ontologies are a well-motivated formal representation able to convey this complex knowledge, but their utilization in IE, beyond mere conceptual hierarchies, is still a research issue. In this paper, we introduce a rich modeling of genic interactions, and a way to fully integrate an ontology within an IE platform.

We refer to an ontology as a thesaurus (concept and relation hierarchies), along with a logical theory given as a set of inference rules (see e.g. (Gómez-Pérez, 1999)). The ontology is seen as a specification of a normalized and decontextualized text representation. A NLP pipeline extracts a first set of ontology instances, then deductive inferences on the ontology language are completed, deriving more instances. IE results

are a set of concept instances linked by semantic relations. Using several well-defined relations gives the opportunity to model more accurately biological domains, and inference rules reasoning on the ontology are able to gather information otherwise scattered throughout bibliographical databases, and to discover knowledge not explicitly stated in texts. Inference rules may be crafted by the domain expert as part of the ontology design, or automatically learnt by Machine Learning (ML) techniques. We focus on this latter case which has been well-motivated in the context of IE systems, as a generic component to easily adapt them to new domains. However, as opposed to previous approaches, learning takes place in the ontology language to produce deductive rules which hold in the domain ontology. From a ML point of view, the learner uses the ontology as hypothesis language, and instantiations of ontology as example language.

However, as stated by (Friedman et al., 2002), ontologies are not necessarily useful to IE, in the sense that the granularity of the classes between a conceptual and a sublanguage model may differ. We deal with this problem by introducing, along with the ontology, a lexical layer, i.e. relations and classes in an intermediate level of abstraction between raw text and concept. This is in line with (Cimiano et al., 2007; Brickley and Miles, 2005), who propose a lexicon model to map expressions in natural language to their corresponding ontology structure, although none of them address it in an IE context.

We discuss related works using ontologies and ML techniques to support IE systems in section 2. We present our approach where IE is fully specified through the design of a domain ontology along with its lexical layer in the next section. We describe how ML techniques can be applied on the ontology instantiations from a corpus to learn deductive rules which can infer new instances during the extraction process (section 4). And we validate our architecture by defining an ontology of genes transcription in bacteria, and by learning inference rules to extract genic interactions from a corpus of the LLL05 challenge (section 5), to finally give perspectives of our work.

2 Related works

The unifying purpose of the ontology allows us to integrate several aspects not simultaneously handled in related works. Consider the sentence:

> The degR gene is transcribed by RNA polymerase containing sigma D, and the level of its expression is low in a mecA-deficient mutant. (PMID: 10486575.)

Extracting the interaction-related knowledge involves processes occurring in multiple abstraction levels. The biological entities have to be recognized, and properly represented. Simplest lexical variations are captured by Named Entities Recognition (NER), as extensively discussed in (Tanabe and Wilbur, 2002; Park and Kim, 2006). A term–concept connection is assumed by several systems, which use mere conceptual hierarchies, without relation (Miyao et al., 2006; Nédellec, 2005; Saric et al., 2005). Here, we normalize a term as a subgraph of ontology instances, including domain knowledge: in the example, the term “RNA polymerase containing sigma D” may be represented as a *protein complex* relation between an “RNA polymerase” *enzyme* and a “sigma D” *protein*. All the synonyms have to share the same representation (e.g. “EsigmaD” or “RNA polymerase sigma D”). We emphasize the terminology status: while, in the previous expression, (Nédellec, 2005) only tag the “sigma D” protein and inaccurately regard it as the interacting entity, we normalize the full term (“RNA polymerase containing sigma D”). Furthermore, whereas most terminological works focus on nouns, we handle verbal terms: the terms “transcription by EsigmaD” and “transcribed by EsigmaD” will be identically represented.

(Nédellec, 2005; Saric et al., 2004) use respectively a general “genic interaction” relation, or a very specific one. The ontology allows to define various conceptual relations: a transcription event between EsigmaD and degR, and a more general regulation between the mecA mutant and the degR gene.

Furthermore, we do not only provide rules processing on a syntactico-semantic level (Miyao et al., 2006; Alphonse et al., 2004; Daraselia et al., 2004), but using ontology as our representation language, we can reason at a semantic level (see, for instance, the use of inference rules in OWL (Mcguiness et al., 2004)). In the previous sentence, this allows to deduce that, although the second interaction of the example involves an inhibition (“level of its expression is low”), as a mutant

gene is implied, mecA and degR are linked by an activation. Inferences may be achieved on multiple sentences, inducing knowledge not explicitly present in the text as we will show it in section 5.

Ontologies become preeminent in the IE field, while most authors exploit it punctually. Their structure may offer a basis to craft extraction rules (Saric et al., 2005; Friedman et al., 2001), or a useful disambiguation resource. For instance, (Cimiano, 2003; Gaizauskas et al., 2003) use it to solve coreferences, (Daraselia et al., 2004) selects relevant syntactic graphs from a parser using the structure of an ontology, (Saric et al., 2005) stress the benefit of an ontology to solve some syntactical ambiguities relying on concepts arity. In most IE pipelines, ontology (or conceptual hierarchy) is only applied to enrich the text with semantic categories (Alphonse et al., 2004; Saric et al., 2004). On the contrary, we used the ontology structure throughout the extraction process, as a language to make inferences from text.

ML techniques have often been used to acquire resources for IE systems, like extraction patterns or rules (Huffman, 1996; Riloff, 1996; Craven and Kumlien, 1999; Alphonse et al., 2004), which are related to our approach. However, they are limited to learn from enriched text representation, as opposed to our approach, where learning takes place in the ontology language.

3 Knowledge representation language of an IE system based on an ontology

Historically, following the "General Theory of Terminology" created by Eugene Wüster from the late 1930s, a term is defined as a word or a group of words which correspond to a concept in a pre-existing conceptual model. More recently, some have criticized this doctrine (Rastier, 1995; Bourigault and Jacquemin, 2000): the conceptual model and the terms are not seen anymore as absolute notions, but as the result of an artificial and application-oriented construction process based on a domain-related corpus. In other words, the terminology is not *discovered*, but *constructed*. We follow this latter conception: our conceptual model, the ontology, is seen as a specification of a normalized representation of a text, neglecting some aspects of the discourse, and keeping some other ones. By designing it, we specify an IE system. Hence, the IE process is equivalent to an automatic semantic annotation of text, into which sentence fragments (terms) are normalized as ontology instances.

3.1 Ontology as a representation language

Figure 1 exemplifies a simplified ontology of transcription in bacteria. In this model, the "transcription" of a gene (*"et"*) from a promoter (*"t_from"*) may happen due to the action of a protein (*"t_by"*). Furthermore, a protein results from

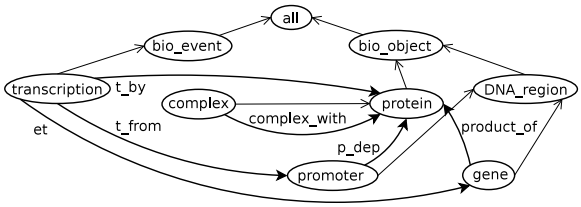

Figure 1: Example of ontology. Labels of "is_a" relations are omitted.

the expression of a gene (*"product_of"*), and a protein complex results from the assembly of several proteins (*"complex_with"*). Figure 2 shows, on an example sentence, the result of the IE system provided as instances of the ontology. Note that, as a normalized representation of the text, not all the meaning is kept: for instance, we do not stress anymore about the "DNA binding" nature of the "GerE" protein; the fact that the transcription happens from "several" promoters is lost. The semantic relations at the bottom of the figure, in plain line, were extracted from text. From the term "transcription from several promoters", a terminological module has extracted instances of "transcription" and "promoter". Then, inferences rules have extracted from text a "t_from" ("transcription from") semantic relation between them. The "p_dep" relation, in bold line in the middle of the figure, is inferred from instances previously extracted from the text, by applying deductive rules on the normalized text representation. This representation fits the specifications

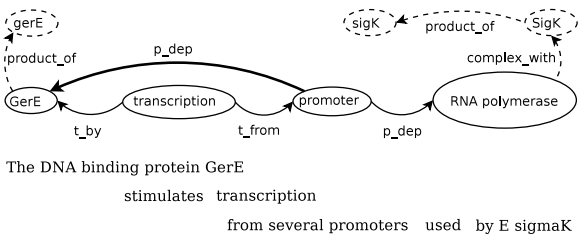

Figure 2: Example of a semantic representation resulting from the IE system.

of the ontology shown in figure 1. Such a rule is the following:

$$\begin{aligned} p\_dep(B, A) \quad \leftarrow \quad & t\_by(C, A), \\ & t\_from(C, B), \\ & protein(A), \\ & promoter(B), \\ & transcription(C). \end{aligned}$$

It means that "if protein A is responsible for a transcription event C from promoter B, then B is dependent on (may be binded by) protein A". Additionally, instances in dotted lines result from domain knowledge: the "GerE" protein is encoded by the "gerE" gene, and the "E sigmaK" protein is a RNA polymerase complexed with the "SigK" protein, itself encoded by the "sigK" gene.

3.2 Features choice for text extraction

Inferences from text require more features. Basically, normalizing a text to a conceptual representation is equivalent to gather multiple lexical forms into a single semantic representation. Hence, the difficulty of the task is related to the complexity of the encountered types of variations. Methods aiming at capturing orthographical and morphological variations are related to Named Entities Recognition (NER), described in (Tanabe and Wilbur, 2002; Park and Kim, 2006). The more complex types of variations are related to relational IE, and processing them involves using NLP tools to enrich the text with syntactic and semantic features. A first set of works builds syntactico-semantic parsers (Friedman et al., 2001; McDonald et al., 2004; Saric et al., 2004; Saric et al., 2005), whereas a second class of systems uses full parsers (Yakushiji et al., 2001; Daraselia et al., 2004; Miyao et al., 2006; Fundel et al., 2007). The latter implies two distinct modules (Yakushiji et al., 2001): a linguistic module, that handles domain-independent structural aspects of the sentence; and an IE module, which is a task-dependent parameter (possibly adapted to the task (Pyysalo et al., 2004)). We follow this general approach which does not involve designing a new syntactico-semantic parser for each new application. This impacts the design of the lexical layer we describe in the next section.

3.3 Lexical layer

We introduce a lexical layer along with the ontology, in which we define relevant semantic features. In figure 3, the concept of "regulation" (and in the example, its instance "stimulate"), and the concept of "dependence" (and its instance "use"), are obviously required. Inference rules do not only need semantic features, but also syntactic ones. To specify them, we introduce syntactico-semantic classes and relations in the lexical layer. Following our conception about ontologies, these classes and relations will define normalizations of text in intermediate states of abstraction, between raw text and conceptual level. They are specified in the ontology shown in figure 4, and will be instantiated by a parser and a terminological module. The layer also allows to introduce classes

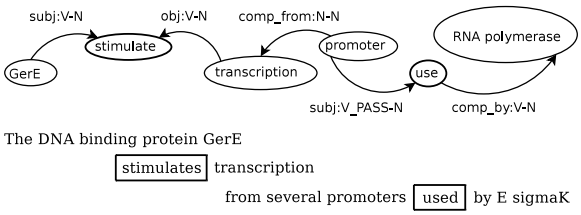

Figure 3: Exemple of a text representation

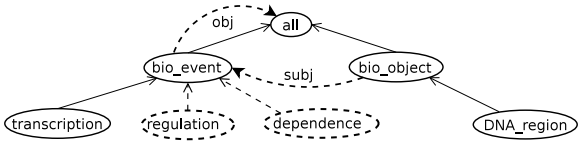

Figure 4: Sample of the lexical layer (elements in dotted line) along with the domain ontology.

which may be semantically irrelevant from a domain ontology point of view but factorize concepts that share common properties, and thus, factorize together otherwise multiple inference rules. This is exemplified in figure 5, which shows the definition of a "biological actor" (bio_actor) class, where a "gene", a "protein" and a "gene family" share common syntactical contexts in biological articles. Figure 3 illustrates a final representation combining semantic features (a protein instance "GerE"), and syntactic ones (a subject "subj:V-N" relation between "GerE" and "stimulate", an

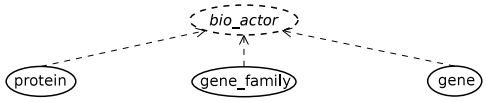

Figure 5: Definition of a syntactico-semantic feature (dotted line) in the ontology.

instance of the "regulation" concept).

4 Acquisition of inference rules

As opposed to previous approaches (see section 2), learning takes place in the ontology language to produce deductive rules which hold in the domain ontology and in the lexical layer. A domain expert has to provide learning examples defined as instantiations of the ontology. He creates instances of concepts and relations of the ontology from a corpus, some instances being output by NLP modules. Target relations are specified to be logically implied by the inference rules. Figure 6 exemplifies such annotation, the dashed lines corresponding to relations to learn.

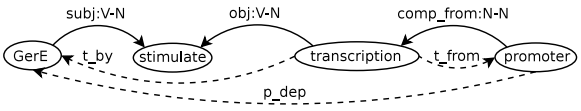

Figure 6: Learning example provided by a semantic annotation.

Learning from such a relational language is known as Inductive Logic Programming (ILP) (Muggleton and Raedt, 1994), where the hypothesis and the example languages are subsets of first-order logic. Most learners handle learning in Datalog which is expressive enough for the task. In Datalog, examples are represented as closed Horn clauses, where the head of the clause is the target relation to learn. For instance, the example of the "t_by" relation in figure 6 will be equivalently represented as the following (relation names have been shorten for presentation):

$$\begin{aligned} t\_by(id1, id2) \quad \leftarrow \quad & subj(id2, id3), obj(id1, id3), \\ & tra(id1, transcription), \\ & pro(id2, "GerE"), \\ & reg(id3, stimulate). \end{aligned}$$

As several relations have to be learnt, learning is set into the multi-class setting where each target relation is learnt in turn, using the other ones as negative examples. Note that all the ontological knowledge is given as background knowledge to the ILP algorithm, like the generalisation relation between concepts. For instance, specifying that a protein complex is a protein, and a protein or a RNA are a gene product, will be represented by a clausal theory:

$$protein(A) \leftarrow protein\_complex(A).$$
$$gene\_product(A) \leftarrow protein(A).$$
$$gene\_product(A) \leftarrow rna(A).$$

Processing an example involving a protein complex or a RNA, the learning algorithm now have the opportunity to choose the most relevant generality level (e.g. "protein complex", "protein" or "gene product") to learn the rules.

5 Results

We validate our architecture by designing an ontology of transcription in bacteria, used to learn inference rules from a *Bacillus subtilis* corpus.

5.1 Ontology encoding biological knowledge

The ontology includes some forty concepts, mainly about biological objects (gene, promoter, binding site, RNA, operon, protein, protein complex, gene and protein families, etc.), and biological events (transcription, expression, regulation, binding, etc.). In the following, we will focus on the ten relations of the ontology.

We defined ten relations: a general interaction relation ("i"), and nine relations specific to some aspects of the transcription (binding, regulons and promoters). Table 1 lists the set of relation names with an example of term. For instance, the third line in the table states that, in the sentence "GerE

Name	Example of related term
p_dep	*sigmaA* recognizes **promoter elements**
p_of	the *araE* **promoter**
b_to	**GerE** binds near the sigK *transcriptional start site*
s_of	*-35 sequence* of the **promoter**
rm	*yvyD* is a member of sigmaB **regulon**
r_dep	*sigmaB* **regulon**
t_from	**transcription** from the Spo0A-dependent *promoter*
t_by	**transcription** by final *sigma(A)-RNA polymerase*
et	**expression** of *yvyD*
i	**KinC** was responsible for Spo0A~P *production*

Table 1: List of relations defined in the ontology, and the corresponding examples of term. Arguments of the relation are shown in italic and bold fonts. The relations are: promoter dependence (p_dep), promoter of (p_of), bind to (b_to), site of (s_of), regulon member (rm), regulon dependence (r_dep), transcription from (t_from), transcription by (t_by), event target (et). "i" is a general interaction relation.

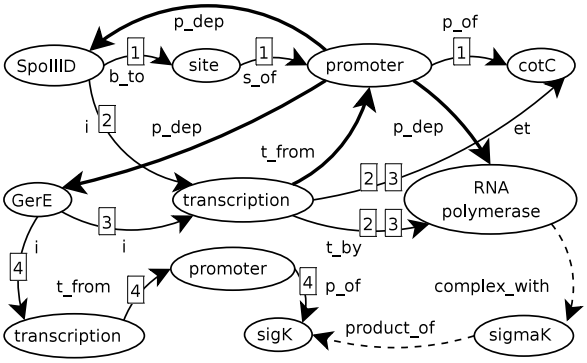

Figure 7: Extracted network from: (1) SpoIIID binds strongly to two sites in the cotC promoter region; (2) SpoIIID represses cotC transcription by sigma(K) RNA polymerase; (3) Transcription of cotC by sigmaK RNA polymerase is activated by GerE; (4) GerE represses transcription from the sigK promoter. Dashed lines represent domain knowledge relations, and bold lines infered ones.

binds near the sigK transcriptional start site", the protein "GerE" (in bold font) binds to (b_to) the site "transcriptional start site" (in italics).

Using an ontology including inference rules, to describe some aspects of the transcription, allows to model biological knowledge more accurately. This is exemplified in figure 7, which shows the instances extracted from four sentences. From the first sentence, inference rules provide the following normalization: SpoIIID binds to (b_to) a site of (s_of) the promoter of (p_of) cotC. The well-defined nature of the involved relations allows to deduce that the cotC promoter is dependent (p_dep) of SpoIIID, as the latter binds to one of its sites. Inferences are not restricted to a sentence: for instance, as the sentence 3 asserts that cotC transcription is activated by GerE, it is possible to deduce that it happens from the cotC promoter (t_from). This latter deduction permits to conclude that the cotC promoter is dependent (p_dep) of GerE. Implicit knowledge distributed into two sentences is therefore made explicit. If less descriptive knowledge is needed, it is easy, by defining a general transitive relation, to provide a database with the genic interacting couples (spoIIID,cotC), (gerE,cotC), (gerE,sigK) and (sigK,cotC). Relations between interacting entities and genes are provided by domain knowledge, as illustrated in the figure with "sigmaK RNA polymerase". The protein complex is known to include protein sigmaK, which is the product of the sigK gene.

5.2 Ontology to learn inference rules

We want to validate the interest of using multiple relations, defined with an ontology, to learn inference rules by ML. In order to test the ontology relevance, we reused the corpus of the LLL05 challenge (Nédellec, 2005), containing 160 sentences, in which we annotated terms, concepts and relations. 541 relations were labeled. Output of NLP tools is complex and heavily noisy, making errors difficult to trace. Thus, to focus exclusively on the rules acquisition task, we only chose to allow as parameters the representation choice and the learning algorithm, the remaining having to be constants and as noiseless as possible. Hence, we enriched and manually curated the linguistic annotations of the LLL05 corpus (parse trees, syntactic categories, lemmas). The representation of the examples was defined following the procedure described in 3.3. We introduced syntactic relations between classes, and syntactico-semantic classes, meant for factorizing entities which may share the same syntactical context: namely, gene and protein, gene family and protein family, transcription and expression events. Eventually, the annotated corpus was used to produce the learning set. To help learning, we added a class of non-interacting biological entities which was generated using the closed-world assumption. We applied the multi-class ILP learner PROPAL (Alphonse and Rouveirol, 2006) to acquire a set of rules for each relation; the non-interacting class was used as negative examples each time but was not learnt. Currently, we only automatically acquire rules involving syntactico-semantic attributes. We will remove this limitation by stratification learning. We provided PROPAL with 541 examples from ten classes, and 10155 from the non-interacting class, and used ten-fold cross-validation, averaged ten times, to evaluate recall and precision of the extraction process. The results are shown in table 2.

As expected, the more specific relations (et, r_dep, rm), assumed to have little lexical variability, are rather trivial to learn, and reach especially high scores. On the contrary, more general ones (i, t_by), exhibiting greater variability, are noticeably harder to learn. We also experiment the two-class case, merging the ten conceptual relations into a positive label, and as shown in table 3, we obtain good recall and precision. Scores are much better than in prelimi-

Relation	Recall	Prec.	Numb.
i	76.4	73.5	161
rm	90.0	90.0	17
r_dep	95.0	100.0	12
b_to	75.0	90.0	14
p_dep	91.5	94.3	47
p_of	87.5	85.2	39
s_of	61.7	80.7	21
et	95.8	99.4	168
t_from	85.0	96.7	18
t_by	65.5	82.6	44

Table 2: Multi-class learning results, for ten fold cross validation averaged ten times, with Recall and Precision in %, and the Number of examples by relation.

nary experiments implying the unique and general "genic interaction" relation from the LLL05 challenge. This corroborates the benefit of using multiple specific relations to model biological knowledge, which involves less complex rules. For instance, in the unique "genic interaction" relation case, the sentences "sigma(H)-dependent expression of spo0A" and "sigma(K)-dependent cwlH gene" would need two rules to be matched (typically, patterns like "A-dependent expression of B" and "A-dependent B"); however, in the multiple relation case, the first sentence would be matched by the patterns "A-dependent B" ("i" relation) and "B of C" ("et" relation), and the second sentence by "A-dependent B" ("i" relation). Thus, in the second case, the "i" rule matches two sentences, where two "genic interaction" rules were needed. By allowing more general rules, the ontology-based approach decreases the required number of examples to be used by the ML algorithm, improving its results.

6 Conclusion and Perspectives

Ontology is a well-motivated formalism to model biological knowledge, and we showed how a domain ontology allows access to knowledge, beyond the capability of current IE systems. However, complex ontologies are not yet fully exploitable in IE systems, which often limit their use to enrich textual data. In this paper, we proposed an original integration of ontology into IE systems. We use the ontology as a language to make inferences on the semantic level, as well as the syntactico-semantic level, thanks to the addition of a lexical layer. IE is performed by first extracting a set of instances from NLP modules,

Recall (%)	Prec. (%)
89.3	89.6

Table 3: Results for two classes learning, using ten fold cross validation averaged ten times.

then deductive inferences on the ontology language are performed, to complete the extraction process. We validated the approach by designing an ontology of genic interactions, and used Machine Learning techniques to learn inference rules from a *Bacillus subtilis* corpus. From a ML point of view, we use the ontology as hypothesis language, and instances of this ontology as example language.

We are currently extending the ontology to handle more phenomenons, especially inhibition/activation distinction, and non-genic actors (e.g. environmental factors). Also, from an operational perspective, we aim at fully automatizing our system by linking the lexical layer to an available NLP pipeline. Notably, as the representation choice is a crucial step in ML, its declarative definition through the ontology is a significant contribution. We then plan to work on text representation, through a comparative study of several lexical layers.

Acknowledgements

We thanks Thierry Poibeau for his useful comments and suggestions on the manuscript. We are grateful to INRA for awarding a Doctoral and Postdoctoral Fellowship to Alain-Pierre Manine.

References

E. Alphonse and C. Rouveirol. 2006. Extension of the top-down data-driven strategy to ILP. In *Proc. Conf. Inductive Logic Programming*. pages 49–63.

E. Alphonse, S. Aubin, Ph. Bessières, G. Bisson, T. Hamon, S. Laguarigue, A. Nazarenko, A.-P. Manine, C. Nédellec, M. Ould Abdel Vetah, T. Poibeau, and D. Weissenbacher. 2004. Event-based information extraction for the biomedical domain: the Caderige project. In *Proc. Intl. Joint Workshop NLP in Biomedicine and its Applications*, pages 43–49.

C. Blaschke, M.A. Andrade, C. Ouzounis, and A. Valencia. 1999. Automatic extraction of biological information from scientific text: Protein-Protein interactions. In *Proc. Seventh Intl. Conf. Intelligent Systems for Molecular Biology*, pages 60–67.

D. Bourigault and C. Jacquemin. 2000. Construction de ressources terminologiques. In J.-M. Pierrel, editor, *Ingénierie des langues*, pages 215–233.

D. Brickley and A. Miles. 2005. SKOS Core Vocabulary Specification. *Technical report, W3C Working Draft*.

P. Cimiano, P. Haase, M. Herold, M. Mantel, and P. Buitelaar. 2007. LexOnto: A model for ontology lexicons for ontology-based NLP. In *Proc. OntoLex07 Workshop*.

P. Cimiano. 2003. Ontology-driven discourse analysis in GenIE. In *Proc. Intl. Conf. Applications of Natural Language to Information Systems, LNI-29*, pages 77–90.

M. Craven and J. Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proc. Intl. Conf. Intelligent Systems for Molecular Biology*, pages 77–86.

N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo. 2004. Extracting human protein interactions from MedLine using a full-sentence parser. *Bioinformatics*, 20(5):604–611.

C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl. 1):S74–S82.

C. Friedman, P. Kra, and A. Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J. Biomedical Informatics*, 35(4):222–235.

K. Fundel, R. Küffner, and R. Zimmer. 2007. RelEx — relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

R.J. Gaizauskas, G. Demetriou, P.J. Artymiuk, and P. Willett. 2003. Protein structures and information extraction from biological texts: The PASTA system. *Bioinformatics*, 19(1):135–143.

A. Gómez-Pérez. Ontological engineering: A state of the art. *Expert Update*, 2(3):33–43, 1999.

S.B. Huffman. 1996. Learning Information Extraction patterns from examples. *LNCS*, 1040:246–260.

M. Krallinger, F. Leitner, and A. Valencia. 2007. Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proc. Second BioCreative Challenge Evaluation Workshop*, pages 41–54.

D.M. McDonald, H. Chen, H. Su, and B.B. Marshall. 2004. Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics*, 20(18):3370–3378.

D. L. McGuinness and F. van Harmelen. 2004. OWL web ontology language overview. *W3C Recommendation*.

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

S. Muggleton and L. De Raedt. 1994. Inductive logic programming: Theory and methods. *J. Logic Programming*, 19,20:629–679.

C. Nédellec. 2005. Learning language in logic — genic interaction extraction challenge. In *Proc. Fourth Learning Language in Logic Workshop*, pages 31–37.

K. Oda, J.-D. Kim, T. Ohta, D. Okanohara, T. Matsuzaki, Y. Tateisi, and J. Tsujii. 2008. New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics*, 9(Suppl. 3):S5.

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

J.C. Park and J.-J. Kim, 2006. *Text Mining for Biology*, chapter Named Entity Recognition. Artech House Books.

S. Pyysalo, F. Ginter, T. Pahikkala, J. Koivula, J. Boberg, J. Järvinen, and T. Salakoski. 2004. Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In *Proc. Intl. Joint Workshop NLP in Biomedicine and its Applications*, pages 15–21.

F. Rastier. 1995. Le terme: entre ontologie et linguistique. In *La banque des mots*, pages 35–65. CILF.

E. Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proc. Natl. Conf. Artificial Intelligence (AAAI)*, pages 1044–1049.

T.C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter. 2000. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Proc. Fifth Pacific Symp. Biocomputing*, pages 517–528.

J. Saric, L.J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. 2004. Large-scale extraction of gene regulation for model organisms in an ontological context. In *Ontology Special of the Third Workshop on Ontology and Genome — Development and Applications of Ontologies on OMICS Research*.

J. Saric, L.J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. 2005. Large-scale extraction of protein/gene relations for model organisms. In *Intl. Symp. Semantic Mining in Biomedicine 2005*.

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

A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. 2001. Event extraction from biomedical papers using a full parser. In *Proc. Sixth Pacific Symp. Biocomputing*, pages 408–419.