Learning ontological rules to extract multiple relations of genic interactions from text

Alain-Pierre Manine[a], Erick Alphonse[a], Philippe Bessières[b]

[a]LIPN, Univ. Paris 13/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
[b] MIG, INRA UR1077
Unité Mathématique, Informatique et Génome
Institut National de la Recherche Agronomique, F78352 Jouy-en-Josas
philippe.bessieres @jouy.inra.fr

Abstract

Introduction: Information Extraction (IE) systems have been proposed in recent years to extract genic interactions from bibliographical resources. 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. An ontology is an adequate formal representation to model this sophisticated knowledge. However, the tight integration of IE systems and ontologies is still a current research issue, *a fortiori* with complex ones that go beyond hierarchies.
Method: 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. First, IE is performed by extracting instances from Natural Language Processing (NLP) modules. Then, deductive inferences on the ontology language are completed, and new instances are derived from previously extracted ones. Inference rules are learnt with an Inductive Logic Programming (ILP) algorithm, using the ontology as the hypothesis language, and its instantiation on an annotated corpus as the example language. Learning is set in a multi-class setting to deal with the multiple ontological relations.
Results: We validated our approach on an annotated corpus of gene transcription regulations in the *Bacillus subtilis* bacterium. 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.

Key words: Information Extraction, Ontology, Machine Learning, Genic Interactions, Inductive Logic Programming

1. Introduction

The elucidation of molecular regulations between genes and proteins, as well as the physical interactions associated to it, is essential in the understanding of living organisms, as they underlie the control of biological functions. However, their knowledge is usually not available in structured formats from widely accessed international databanks, which contain generic annotations of genomes. This is basically concerning data collections such as EMBL/GenBank for annotated DNA sequences of complete genomes, SwissProt/UniProt for annotated protein sequences, or KEGG, which is dedicated to the metabolism of the cells and other biological processes. Contrary to this, most of the descriptions of molecular interactions are scattered in the unstructured texts of scientific publications.

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 [1, 2, 3, 4]. 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 [2, 5, 1, 6, 7], and do not handle other biological phenomenons; whereas other stress on recall using general relations [8, 9], but face greater lexical variability which makes extraction more difficult. However, this does not take into account the complexity of the data processed by biologists, such as biological pathways [10].

Ontologies are a suitable 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. [11]). The ontology is seen as a specification of a normalized and decontextualized text representation. A Natural Language Processing (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 more accurately model 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 (e.g. [12, 13]). 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 the ontology as example language.

However, as stated by [14], ontologies are not necessarily useful to IE, in the sense that the granularity of the classes between a conceptual and a sub-language 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 [15, 16], who propose a lexicon model to map expressions in Natural Language to their corresponding ontological structure, although none of them address it in an IE context.

The article will firstly discuss related works using ontologies and ML techniques to support IE systems in section 2. In the following section, we will present our approach where IE is fully specified through the design of a domain ontology along with its lexical layer. We will describe how ML techniques can be applied on the ontology instantiations from a corpus to learn deductive rules that can infer new instances during the extraction process in section 4. Next, we will 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). We will discuss how a complex domain ontology helps extracting information beyond current systems' capabilities. Finally, we will conclude and give perspectives on 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 [17, 18]. A term–concept connection is assumed by several systems, which use mere conceptual hierarchies, without relation [19, 8, 7]. 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, some approaches (e.g. [8]) 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.

[8, 20] use respectively a general "genic interaction" relation or a very specific one, as trade-offs between recall and precision. 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 [19, 21, 22], but using ontology as our representation language, we can reason at a semantic level (see, for instance, the use of inference rules in OWL [23]). 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.

Ontologies become preeminent in the IE field, while most authors exploit it punctually. Their structure may offer a basis to craft extraction rules [7, 3], or a useful disambiguation resource. For instance, [24, 25] use it to solve coreferences, [22] selects relevant syntactic graphs from a parser using the structure of an ontology; [7] stress the benefit of an ontology to solve some syntactical ambiguities relying on concepts arity. In most IE pipelines, an ontology (as a conceptual hierarchy) is only applied to enrich the text with semantic categories [21, 20]. 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 [12, 13, 2, 21, 26], 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 [27, 28]: 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"*). Also, a protein may bind to a site (*"b_to"*) of a promoter (*"s_of"*), meaning that a promoter of a gene (*"p_of"*) may be dependent of a protein (*"p_dep"*), and therefore an interaction exists between a protein and a gene (*"i"*). Furthermore, a protein results from the expression of a gene (*"product_of"*), and a protein complex results from the assembly of several proteins (*"complex_with"*).

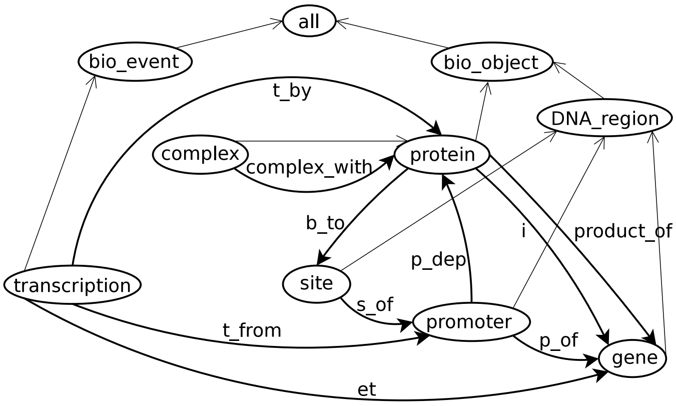
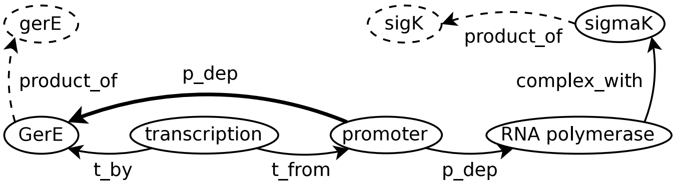


Figure 1: Example of ontology. Labels of "is_a" relations are omitted.

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



The DNA binding protein GerE
stimulates transcription
from several promoters used by E sigmaK

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

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

It means that "if protein A is responsible for a transcription event C from promoter B, then B is dependent on (may be bound 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 gathering 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 [17, 18]. 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 [3, 29, 20, 7], whereas a second class of systems uses full parsers [30, 22, 19, 9]. The latter implies two distinct modules [30]: a linguistic module, that handles domain-independent structural aspects of the sentence (possibly adapted as in [31]); and an IE module, which is a task-dependent parameter. 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

The goal of our IE system is to automatically produce, from Natural Language (NL) inputs, a set of instances compliant with the domain ontology. This requires complex mappings between expressions in NL to ontology structures [15], going beyond mere class/label linking (like the `rdf:label` property of RDF [32], or the more complex properties of SKOS [16]). To bridge the gap, some authors introduce lexicon models [15, 33] to ground the semantic information to the linguistic domain, although not in an IE context.

We follow this approach by providing a *lexical layer* along with the ontology. However, where the previous authors follow a linguistic point of view, by proposing a model to link ontology structures to lexical descriptions, we adopt an application-oriented perspective. Our lexical layer is a task-dependent parameter, it comprises classes and relations required to link the output of NLP modules with the ontology. Its purpose is to provide a representation with sufficient expressiveness for efficient inference. These classes and relations define normalizations of text in intermediate stages of abstraction, between raw text and conceptual level.

For instance, a relation of the lexical layer may associate a syntactic label with an instance, or a syntactic relation between two instances (subject (*"subj"*) and object (*"obj"*) relations in figure 3). The lexical layer is described in the same language as the ontology, so the inference rules can benefit from it. Inference rules do not only need semantic features, but also syntactic ones.

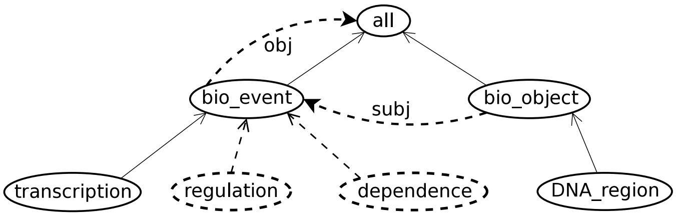


Figure 3: Sample of the lexical layer (elements in dotted line) along with the domain ontology.

Figure 4 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 instance of the "regulation" concept). The lexical layer also allows to introduce classes 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. We exemplify this procedure in figure 5, which shows the definition of a "biological actor" (bio_actor) class. This class is semantically irrelevant, but a "gene", a "protein" and a "genes family" may share common syntactical contexts.

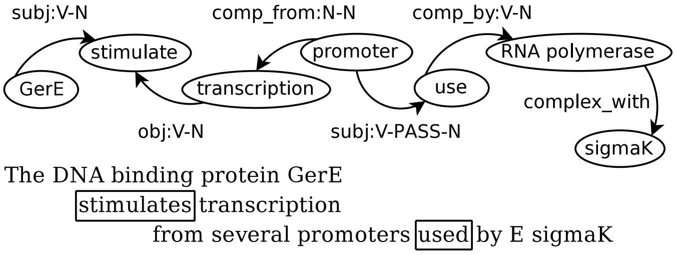


Figure 4: Example of a text representation

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.

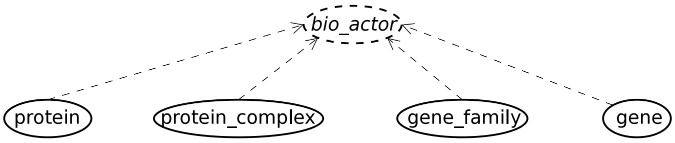


Figure 5: Definition of a syntactico-semantic feature (dotted line) in the ontology.

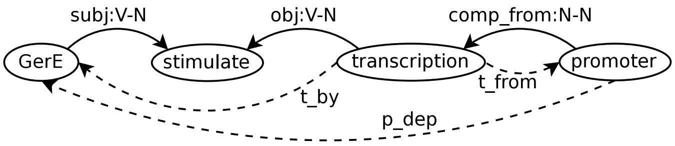


Figure 6: Learning example provided by a semantic annotation.

Learning from such a relational language is known as Inductive Logic Programming (ILP) [34, 35], 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, learning examples are represented as closed Horn clauses, where the head of the clause is the target relation to learn [36]. For instance, the example of the "t_by" relation in figure 6 will be equivalently represented as the following:

$$\begin{aligned} t\_by(id1, id2) \quad \leftarrow \quad & subj\_v\_n(id2, id3), \\ & obj\_v\_n(id1, id3), \\ & transcription(id1), \\ & protein(id2), \\ & regulation(id3). \end{aligned}$$

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:

$$\begin{aligned} & protein(A) \leftarrow protein\_complex(A). \\ & gene\_product(A) \leftarrow protein(A). \\ & gene\_product(A) \leftarrow rna(A). \end{aligned}$$

Processing an example involving a protein complex or a RNA, the learning algorithm now has 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 the architecture of our IE system by designing an ontology of molecular interactions and regulations, used to learn inference rules from a corpus concerning the transcription of genes in the *Bacillus subtilis* model bacterium. Therefore, the ontology is mainly oriented about the description of a structural model of genes, the mechanisms of their transcription, and interactions and regulations associated to it. As a matter of fact, the ontology is catching a model of gene transcription to which authors implicitly refer in their publications.

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, from the most general, such as target event ("et") and interaction ("i"), to the ones specific of gene transcription (especially relations concerning transcription 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 binds near the sigK transcriptional start site", the protein "GerE" (in bold font) binds to (b_to) the site "transcriptional start site" (in italics). Figure 7 shows some normalisations of phrases involving those relations with instantiations of sub-graphs of the ontology.

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 8, 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. From sentence 3, which 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.

Name	Example
et	**expression** of *yvyD*
i	**KinC** was responsible for Spo0A˜P *production*
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**
p_of	the *araE* **promoter**
p_dep	*sigmaA* recognizes **promoter elements**
t_from	**transcription** from the Spo0A-dependent *promoter*
t_by	**transcription** by final *sigma(A)-RNA polymerase*

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

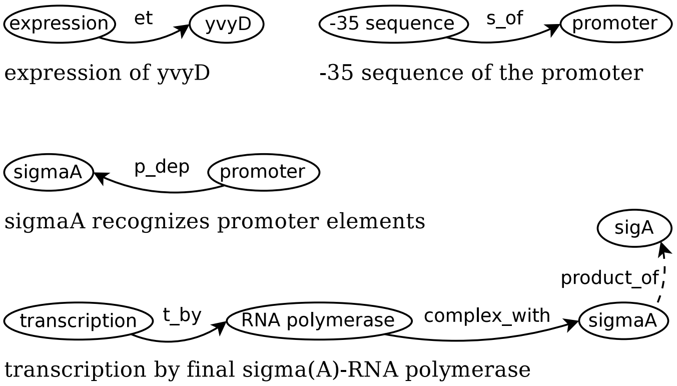


Figure 7: Normalisation of phrases as ontology instances (dashed lines represent domain knowledge relations).

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. Learning the 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

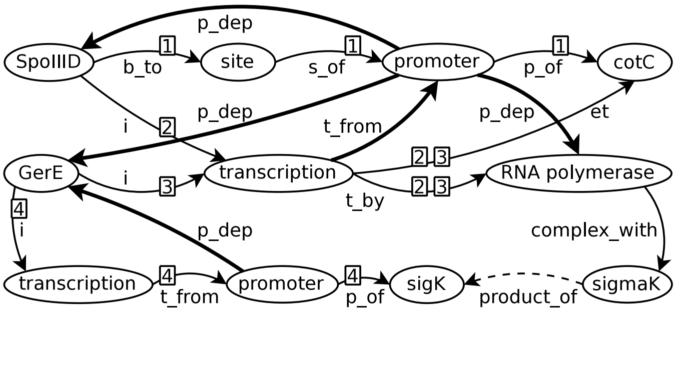


Figure 8: Extracted network from: (1) SpoIIID binds strongly to two sites in the cotC promoter region; (2) SpoIIID represses cotC transcription by sigma(K) RNA polymerase; (3) Transcription of cotC by sigmaK RNA polymerase is activated by GerE; (4) GerE represses transcription from the sigK promoter. Dashed lines represent domain knowledge relations, and bold lines inferred ones.

LLL05 challenge [8]. It contains 160 sentences, in which we annotated terms, concepts and relations; 587 relations were labeled. This corpus provides dependency like parsing of the sentences, following a normalised set of syntactic relations [37], with resolved coreferences. 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. Syntactic relations were limited to the members of the syntactic path between the two entities implied in each semantic relations. This was proved to be a useful bias in several previous studies (see, e.g., [38]). 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*, meaning that everything not tagged as true in the corpus is false. This assumption allows to use any untagged tuples of arguments as negative examples of the target relations. For example, according to figure 6, *t_by(transcription,GerE)* will be a positive example of the *t_by* relation, whereas *t_by(GerE,transcription)* will be a negative one[1]. According to section 4, this negative example is represented as the following Horn clause (only the order of arguments differs):

$$\begin{aligned} \mathit{t\_by}(\mathit{id}2,\mathit{id}1) \quad \leftarrow \quad & \mathit{subj\_v\_n}(\mathit{id}2,\mathit{id}3),\\ & \mathit{obj\_v\_n}(\mathit{id}1,\mathit{id}3),\\ & \mathit{transcription}(\mathit{id}1),\\ & \mathit{protein}(\mathit{id}2),\\ & \mathit{regulation}(\mathit{id}3). \end{aligned}$$

We applied the multi-class ILP learner Propal [39] to acquire a set of rules for each relation; the non-interacting class was used as negative examples each time but was not learnt. From the 587 relation examples, we excluded 46 of them, as they were matched by expert rules which exhibited recursion patterns like the transitivity of the general interaction relation "i". Learning such recursive dependencies is a very interesting follow-up but it is out of the scope of the paper. We provided Propal with 541 examples from ten classes, and 10155 from the non-interacting class.

5.3. Information extraction

We used ten-fold cross-validation (stratified), averaged ten times, to evaluate recall and precision of the extraction process. The results are shown in table 2.

Relation	Recall (%)	Prec. (%)	Numb.
et	95.8	99.4	168
i	76.4	73.5	161
b_to	75.0	90.0	14
s_of	61.7	80.7	21
rm	90.0	90.0	17
r_dep	95.0	100.0	12
p_of	87.5	85.2	39
p_dep	91.5	94.3	47
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.

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, in line with the best results reported on the original LLL05 challenge (see e.g. [9, 40]).

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.

6. Conclusion and Perspectives

We introduced an ontology-based IE platform, which is not limited to extract a single interaction, but allows to handle multiple biological relations. Specific relations are defined in an ontology, which is an appropriate formalism to model biological knowledge. We showed how a domain ontology allows access to knowledge beyond the capability of current IE systems, by allowing 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, 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.

In the future, we plan to extend the ontology to handle more

[1] IDs are substituted by terms for clarification.

Recall (%)	Prec. (%)
89.3	89.6

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

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, before evaluating its performances. 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 thank Thierry Poibeau for his useful comments and suggestions on earlier drafts. We are grateful to INRA for awarding a Doctoral and Postdoctoral Fellowship to Alain-Pierre Manine. This work is supported by the BaSysBio project, European Commission research grant (LSHG-CT2006-037469).

References

[1] C. Blaschke, M. Andrade, C. Ouzounis, A. Valencia, Automatic extraction of biological information from scientific text: Protein-protein interactions, in: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, AAAI Press, 1999, pp. 60–67.

[2] M. Craven, J. Kumlien, Constructing biological knowledge bases by extracting information from text sources, in: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, AAAI Press, 1999, pp. 77–86.

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

[4] M. Krallinger, F. Leitner, A. Valencia, Assessment of the second BioCreAtIvE PPI task: Automatic extraction of protein-protein interactions, in: Proceedings of the Second BioCreAtIvE Challenge Evaluation Workshop, 2007, pp. 41–54.

[5] T. Rindflesch, L. Tanabe, J. Weinstein, L. Hunter, EDGAR: extraction of drugs, genes and relations from the biomedical literature, in: Proceedings of the Fifth Pacific Symposium on Biocomputing (PSB'03), 2000, pp. 517–528.

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

[7] J. Saric, L. Jensen, R. Ouzounova, I. Rojas, P. Bork, Large-scale extraction of protein/gene relations for model organisms, in: First International Symposium on Semantic Mining in Biomedicine, 2005, p. 50.

[8] C. Nédellec, Learning Language in Logic — Genic interaction extraction challenge, in: J. Cussens, C. Nédellec (Eds.), Proceedings of the Fourth Learning Language in Logic Workshop (LLL05), 2005, pp. 31–37.

[9] K. Fundel, R. Küffner, R. Zimmer, RelEx — Relation extraction using dependency parse trees, Bioinformatics 23 (3) (2007) 365–371.

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

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

[12] S. Huffman, Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing, Springer Verlag, 1996, Ch. Learning Information Extraction Patterns from Examples.

[13] E. Riloff, Automatically generating extraction patterns from untagged text, in: Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), AAAI Press / The MIT Press, 1996, pp. 1044–1049.

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

[15] P. Cimiano, P. Haase, M. Herold, M. Mantel, P. Buitelaar, LexOnto: A model for ontology lexicons for ontology-based NLP, in: Proceedings of the OntoLex07 Workshop held in conjunction with ISWC'07, 2007.

[16] D. Brickley, A. Miles, SKOS core vocabulary specification, W3C working draft, W3C (Nov. 2005).

[17] L. Tanabe, W. Wilbur, Tagging gene and protein names in biomedical text, Bioinformatics 18 (8) (2002) 1124–1132.

[18] J. Park, J.-J. Kim, Text Mining for Biology, Artech House Books, Norwood, MA, 2006, Ch. Named Entity Recognition.

[19] Y. Miyao, T. Ohta, K. Masuda, Y. Tsuruoka, K. Yoshida, T. Ninomiya, J. Tsujii, Semantic retrieval for the accurate identification of relational concepts in massive textbases, in: Proceedings of COLING-ACL, 2006, pp. 1017–1024.

[20] J. Saric, L. J. Jensen, I. Rojas, Large-scale extraction of gene regulation for model organisms in an ontological context., In Silico Biology 5.

[21] E. Alphonse, S. Aubin, P. Bessières, G. Bisson, T. Hamon, S. Lagarrigue, A. Nazarenko, A.-P. Manine, C. Nédellec, M. O. A. Vetah, T. Poibeau, D. Weissenbacher, Event-based information extraction for the biomedical domain: the Caderige project, in: Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), 2004, pp. 43–49.

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

[23] D. McGuinness, F. van Harmelen, OWL web ontology language overview: W3C recommendation 10, Tech. rep., W3C (February 2004).

[24] P. Cimiano, Ontology-driven discourse analysis in GenIE., in: A. Düsterhöft, B. Thalheim (Eds.), Proceedings of the Eighth International Conference on Applications of Natural Language to Information Systems, Vol. 29 of LNI, GI, 2003, pp. 77–90.

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

[26] M. Miwa, R. Sætre, Y. Miyao, T. Ohta, J. Tsujii, Combining multiple layers of syntactic information for protein-protein interaction extraction, in: Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM), Turku Centre for Computer Science (TUCS), 2008, pp. 101–108.

[27] F. Rastier, Le terme: entre ontologie et linguistique, in: La banque des mots, CILF, 1995, pp. 35–65.

[28] D. Bourigault, C. Jacquemin, Construction de ressources terminologiques, in: J.-M. Pierrel (Ed.), Ingénierie des langues, Hermès Science, 2000, pp. 215–233.

[29] D. McDonald, H. Chen, H. Su, B. Marshall, Extracting gene pathway relations using a hybrid grammar: the arizona relation parser, Bioinformatics 20 (18) (2004) 3370–3378.

[30] A. Yakushiji, Y. Tateisi, Y. Miyao, J. Tsujii, Event extraction from biomedical papers using a full parser., in: Proceeding of the Sixth Pacific Symposium on Biocomputing (PSB'01), 2001, pp. 408–419.

[31] S. Pyysalo, F. Ginter, T. Pahikkala, J. Koivula, J. Boberg, J. Järvinen, T. Salakoski, Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions, in: Proceedings of Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA'04), 2004, pp. 15–21.

[32] D. Brickley, R. Guha, RDF vocabulary description language 1.0: RDF schema: W3C recommendation 10, Technical report, W3C, Cambridge, MA (February 2004).
URL `http://www.w3.org/TR/rdf-schema/`

[33] P. Buitelaar, M. Sintek, M. Kiesel, A multilingual/multimedia lexicon model for ontologies., in: Y. Sure, J. Domingue (Eds.), ESWC, Vol. 4011 of Lecture Notes in Computer Science, Springer-Verlag, 2006, pp. 502–513.

[34] S. Muggleton, L. D. Raedt, Inductive Logic Programming: Theory and methods, Journal of Logic Programming 19,20 (1994) 629–679.

[35] T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.

[36] S.-H. Nienhuys-Cheng, R. de Wolf, Foundations of Inductive Logic Programming, Vol. 1228 of Lecture Notes in Artificial Intelligence, Springer-Verlag, 1997.

[37] S. Aubin, LLL challenge — Syntactic analysis guidelines, Tech. rep., LIPN, Université Paris 13, Villetaneuse (2005).

[38] J. Ding, D. Berleant, J. Xu, A. W. Fulmer, Extracting biochemical interactions from MedLine using a link grammar parser., in: Proceeding of the Fifteenth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03), 2003, pp. 467–473.

[39] E. Alphonse, C. Rouveirol, Extension of the top-down data-driven strategy to ILP, in: Proceedings of the Conference on Inductive Logic Programming, Springer Verlag, Santiago de Compostela, Spain, 2006, pp. 49–63.

[40] S. Van Landeghem, Y. Saeys, B. De Baets, Y. Van de Peer, Extracting protein-protein interactions from text using rich feature vectors and feature selection, in: Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM), Turku Centre for Computer Science (TUCS), 2008, pp. 77–84.