Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology*

Massimiliano Ciaramita† **Aldo Gangemi**† **Esther Ratsch**‡ **Jasmin Šarić**‡ **Isabel Rojas**‡
m.ciaramita/a.gangemi@istc.cnr.it eboja@web.de jasmin.saric/isabel.rojas@eml-r.villa-bosch.de
†: Laboratory for Applied Ontology, National Research Council (LOA-CNR), Roma, Italy
‡: EML Research gGmbH, Heidelberg, Germany

Abstract

We present an unsupervised model for learning arbitrary relations between the concepts defined in a molecular biology ontology for the purpose of text data mining and support to manual ontology building. Relations are learned from the GENIA corpus, in which named-entities representing the GENIA ontology concepts have been tagged, by means of several natural language processing techniques. We carry out an in-depth analysis of the results of the system, from both the perspective of biologists and ontologists, which shows that the model is both accurate and consistent.

1 Introduction

There is an increasing interest in the application of text mining techniques to the biomedical domain, motivated by the massive and fast-growing body of scientific work accumulating in collections, such as MEDLINE and SwissProt, of millions of documents. In this kind of data mining the goal is not so much that of indexing documents so that those relevant to a query can be accessed. The ultimate purpose is to access and operate on the information contained "inside" the documents to discover new information.

To achieve this objective a conceptualization of the domain is necessary: relevant *concepts* as well as *semantic relations* such as is-a, part-of, and more complex relations which encode important interactions between concepts need to be specified. In other words, it is necessary to build a *domain ontology*. It is also necessary to apply natural language processing (NLP) techniques to recognize the linguistic structures where target concepts are instantiated by specific *entities*, and important interactions expressed by semantic patterns. Several ontologies which define concepts and basic semantic relations such as is-a are available. Instead there is a need for ontologies that define relevant arbitrary semantic relations between concepts; e.g., that "cells express-the-receptor-for proteins" or that "viruses replicate-in cells".

In this paper we investigate the problem of enriching an existing ontology with directed relations representing arbitrary semantic dependencies which are strongly associated with ordered pairs of ontology concepts. We design a completely unsupervised system that combines an array of off-the-shelf NLP techniques such as syntactic parsing, collocations extraction and selectional restrictions learning. Then we apply our system to a corpus of molecular biology literature, the GENIA corpus [Ohta *et al.*, 2002], and generate a list of labeled binary relations between pairs of GENIA ontology concepts. We then provide an in-depth analysis of the learned templates. The analysis of the biological content expressed by the relations shows that a good fraction are correct and relevant, above 83%. Another analysis from the perspective of ontology engineering shows that the learned ontology possesses a high level of consistency when aligned to another more general ontology for consistency checking. Therefore the model we present, which is characterized by a very simple architecture, offers good potential for applications to text mining and ontology development.

The paper is organized as follows. In the next Section we describe the problem of learning relations from text and related work. Then in Section 3 we describe our system and the data used in our experiment in details. In Section 4 we discuss our evaluation of the system's generated ontological relations, and finally we present our conclusions.

2 Problem and related work

The GENIA ontology contains concepts related to gene expression and its regulation, binding interactions involving proteins, DNA, and RNA, signaling pathways and the like. The goal of our system is to support biological text data mining and automatic ontology building by exhaustively discovering relevant relations between the defined concepts.

Text data mining in bioinformatics aims at knowledge discovery by means of natural language processing and machine learning (cf. for example [Swanson and Smalheiser, 1997]). Much work has focused on the problem of *named-entity recognition* (NER), or Information Extraction (IE), where the goal is the identification of sequences of words that represent instances of a set of target concepts. As an example one would like to recognize that "NS-Meg cells", "mRNA" and "EPO receptor" are, respectively, instances of the GENIA classes "Cell line", "RNA family or group" and "Protein molecule" in the following text (Example 1):

*Accepted for publication in the proceedings of IJCAI 2005. This is not the final version of the paper please do not cite.

(1) “Untreated NS-Meg_cells$_{\text{Cell\_line}}$ expressed mRNA$_{\text{RNA\_family\_or\_group}}$ for the EPO_receptor$_{\text{Protein\_molecule}}$”

A natural extension of this line of research is the extraction of relations associated with important interactions between entities. NER and relation extraction could support together higher-level inferences: patterns of entities and relations could be compared across document collections to discover and relate new informative pieces of knowledge. Currently most of the work on relation extraction applies hand-built rule-based extraction patterns; e.g., Friedman et al. [2001] on finding molecular pathways and Šarić et al. [2004] on the problem of extracting information about protein interactions which use a manually-built ontology [Ratsch *et al.*, 2003]. One problem with rule-based Information Extraction is that systems tend to have good precision but very low recall. Machine learning oriented work has focused on extracting manually-compiled lists of target relations; e.g., Rosario and Hearst [2004] address the relation extraction problem as an extension of NER and use sequence learning methods to recognize instances of a set of 6 predefined relations about “Diseases” and “Treatments”. These systems yield good precision and recall but still need that sets of relations between classes be defined first. Yet another problem which deals with semantic relations is that addressed by Craven and Kumlien [1999] who present a model for finding extractor patterns for 5 binary relations involving proteins. A similar work is that of Pustejovsky et al. [Pustejovsky *et al.*, 2002] on automatically extracting “inhibit” relations. Semantic relations have also been used as templates, or guiding principles, for the generation of database schemata [Rojas *et al.*, 2002]. Another application of ontological relations is that of consistency checking of data in molecular biology databases to individuate errors in the knowledge base (for example by checking the consistency of the arguments) or to align different databases.

Current biological text mining systems that work with relations require predefined sets of relations that have to be manually encoded, a work which is complex and tedious and that as such can only have a narrow coverage, typically a handful of relations and one pair of classes. Our aim is to automatically generate all relevant relations found in a corpus between all ontological concepts defined in an ontology. This work is also valuable to ontologists since ontology building and evaluation are becoming more and more automatized activities and most of the corpus-based work has focused only on structural relations such as is-a and part-of [Pantel and Ravichandran, 2004; Berland and Charniak, 1999].

3 Learning relations from text

Our model takes as input a corpus of documents in which named-entities corresponding to ontology concepts have been tagged, the GENIA corpus. In this case the tagging has been carried out manually but in the future we plan to train a NER system to generate additional data. The model outputs a set of templates that involve pairs of GENIA ontology classes and a semantic relation. A generic template of this kind is defined as a triple (r, c_1, c_2), where r is a lexico-syntactic pattern, and c_1 and c_2 are respectively the first and second arguments of the relation. For example, a reasonable template might be $(express, Cell\_type, Protein\_molecule)$.

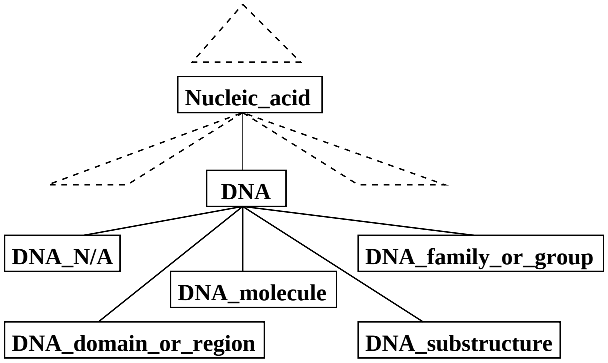

Figure 1: The GENIA ontology around the concept “DNA”, the edges represent generic taxonomic relations.

3.1 Data

The GENIA ontology was built to model cell-signaling reactions in humans with the goal of supporting NLP systems for information extraction. It consists of a taxonomy of 46 nominal concepts with underspecified taxonomic relations, see Figure 1. The ontology was used to semantically annotate biological entities in the GENIA corpus. We used the G3.02 version of the corpus consisting of 2,000 articles, 18,546 sentences, 49,0941 types, and 36 types of labels. The GENIA corpus has complex annotations for disjunctive/conjunctive entities, for cases such as “erythroid, myeloid and lymphoid cell types”. We excluded sentences that contained only such instances and also excessively long sentences (more than 100 words). The final number of sentences was 18,333, 484,005 tokens and 91,387 tags. Many tags have nested structures; e.g. “[$_{\text{Other\_name}}$ [$_{\text{DNA}}$ IL-2 gene] expression]”. In this paper we only considered the innermost labels and ignore the external ones, although they contain useful information and should eventually be used.

One potential drawback of the GENIA ontology is the relatively small number of biological concepts and their coarse granularity which causes groups of similar but distinct entities to be assigned to the same class. Some relations fit very well to part of the entities of the concepts related by this relation, whereas they don’t fit well for other entities of the same concept. For example, the concept “DNA domain or region” contains sequences with given start and end positions, as well as promoters, genes, enhancers, and the like. Even if promoters, genes, and enhancers are pieces of sequences too (with start and end positions), they also are functional descriptions of sequences. Therefore, different statements can be made about such kinds of “DNA domain or regions” and (pure) sequences. The relation “DNA domain or region encodes Protein molecule” makes sense for genes, but not for enhancers, and may make sense or not for (pure) sequences, depending on their (unknown) function. On the other hand any NLP oriented resource cannot have many fine-grained concepts defined, otherwise IE wouldn’t be accurate. In this respect the GENIA corpus is unique in that provides exten-

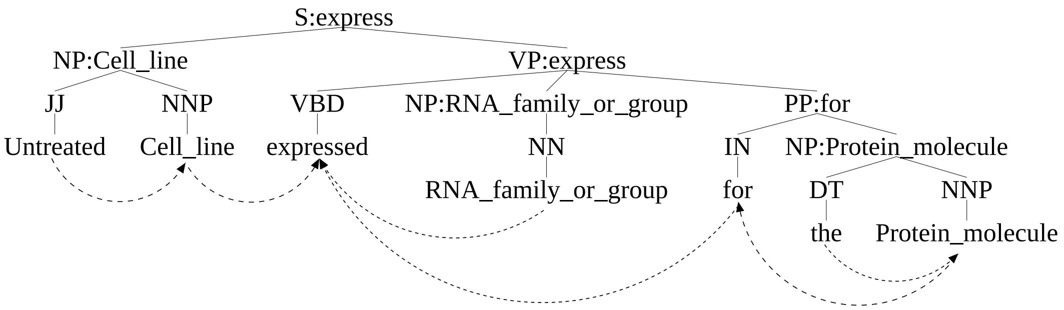

Figure 2: Parse tree for the sentence of Example 1. Entities are substituted with their tags. Phrases are labeled with their syntactic heads. The dependency graph is depicted with dashed directed edges pointing to the governing elements.

sive name-entity annotations which can be used to train state of the art NER systems (cf. [Kazama *et al.*, 2002].

3.2 Relations as dependency paths

We parsed the 18,333 sentences with a statistical parser [Charniak, 2000][1]. Since we are interested in relations that connect entities as chunks we want to avoid that the parser analyzes an entity that is split among different phrases. This can happen because GENIA entity instances can be fairly long, complex and contain words that are unknown to the parser. To avoid this problem we substituted the entity tags for the actual named-entities; the result can be seen in Figure 2 which shows the transformation and relative parse tree for the sentence of Example 1. The trees obtained in this way don't split entities across different phrases and therefore look simpler and cleaner. For each tree we generated a *dependency graph* where each word[2] is associated with exactly one *governor*, defined as the *syntactic head*[3] of the phrase closest to the word that differs from the word itself. For example, in Figure 2 "Cell type" is governed by "transcribe", while "lipid" is governed by the preposition "with". In this way we can formalize a linguistically motivated representation of the semantic dependencies between entities. A similar application of dependency paths is extraction of paraphrases [Lin and Pantel, 2001].

A relation r between two entities c_i and c_j in a tree is the shortest path between c_i and c_j following the dependency relations; the arrows are directed from the governor to the governed element. For example, in Figure 2 the path between "Cell line" and "Protein molecule" is "$\leftarrow \text{express} \rightarrow for \rightarrow$". There is a path for every pair of entities in the tree. Also, paths can be considered from both directions since the reverse of a path from A to B is a path from B to A. Hence a large number of different types of patterns can be extracted. Overall we found 172,446 paths between entities in the GENIA corpus. For the sake of interpretability of the outcome of the model in this paper we focus on a subset of these patterns. We limit ourselves to paths from c_i to c_j where $j > i$ and the pivotal element, the word with no incoming arrows, is a verb v, in addition we impose the following constraints:

1. c_i is governed by v under an S phrase (i.e., is v's surface subject, SUBJ), e.g., "Cell type" in Figure 2; AND
2. c_j is governed by v under a VP (i.e., is v's direct object, DIR_OBJ), e.g., "RNA family or group" in Figure 2; OR
3. c_j is governed by v under a PP (i.e., is v's indirect object, IND_OBJ), e.g. "Protein molecule" in Figure 2; OR
4. c_j is governed by v's direct object noun (i.e., is a modifier of the direct object, DIR_OBJ_MOD), e.g. "Virus" in "... influenced virus replication"; OR
5. c_j is governed by v's indirect object noun (i.e., is the indirect object's modifier, IND_OBJ_MOD), e.g., "Protein molecule" in "..was induced by protein molecule stimulation"; OR
6. c_j is governed by a PP which modifies the direct object (DIR_OBJ_MOD_PP); e.g., "Protein molecule" in ".. induce overproduction of protein_molecule"; OR
7. c_j is governed by a PP which modifies the indirect object (IND_OBJ_MOD_PP); e.g., "Lipid" in "..transcribed upon activation with Lipid".

For the sentence in Figure 2 we identify two good patterns: "SUBJ←express→DIR_OBJ" between "Cell line" and "RNA family or group", and "SUBJ←express→for→IND_OBJ", between "Cell line" and "Protein molecule". Overall we found 7,189 instances of such relations distributed as follows:

Type	Counts	RelFreq
SUBJ-DIR_OBJ	1,746	0.243
SUBJ-IND_OBJ	1,572	0.219
SUBJ-DIR_OBJ_MOD_PP	1,156	0.161
SUBJ-DIR_OBJ_MOD	943	0.131
SUBJ-IND_OBJ_MOD_PP	911	0.127
SUBJ-IND_OBJ_MOD	861	0.120

we found also 485 types of entity pairs, 3,573 patterns pairs and 5,606 entities-pattern types.

3.3 Stage 1: learning relations

Let us take A to be a generic ordered pair of GENIA classes; e.g. A = (Cell_type,RNA_family_or_group), and B to be a

[1] It took roughly three hours on a Pentium 4 machine.

[2] Morphologically simplified with Wordnet's library's "morph" function and lists of morphological simplifications from UMLS.

[3] The word whose syntactic category determines the syntactic category of the phrase; e.g., a verb for a verb phrase (VP), a noun for a noun phrase (NP), etc.

generic pattern; e.g., B = SUBJ←express→DIR_OBJ. Our goal is to find relations that are strongly associated with ordered pairs of classes, that is bi-grams AB. This problem is equivalent to that of finding *collocations*; e.g., multi-word expressions such as ”real estate”. Accordingly the simplest method would be to select the most frequent AB bi-grams. Unfortunately there are many bi-grams that are very frequent because either A or B, or both, are frequent. The pattern SUBJ←induce→DIR_OBJ for example is among the most frequent patterns for 37 different A pairs. Since high frequency can be accidental this method is prone to produce noisy results, additionally it doesn't provide a well-motivated way for distinguishing relevant from irrelevant bi-grams.

As with collocations a better approach is to estimate if A and B occur together more often than at chance. One formulates the *null hypothesis* H_0 that A and B do not occur together more frequently than would be expected at chance. Using corpus statistics the probability of $P(AB)$, under H_0, is computed and H_0 is rejected if $P(AB)$ is beneath the significance level. We use a chi-square test for this purpose. For each observed AB pair we create a contingency table of the frequencies of AB, ¬AB, A¬B, and ¬A¬B; e.g., for A$_i$ = Protein_molecule-DNA_domain_or_region and B$_j$ = SUBJ←bind→DIR_OBJ the table computed from the corpus would contain respectively the values 6, 161, 24 and 6,998. The chi-square test compares the observed frequencies vs. the frequencies expected under H_0. Together with the test we use the log-likelihood chi-squared statistic: [4]

$$G^2 = \sum_{i,j} o_{ij} \log \frac{o_{ij}}{e_{ij}} \tag{2}$$

where i and j range over the rows and columns of the contingency table, and the expected frequencies are computed off the marginal frequencies. In the previous example G^2 is equal to 16.43, which is above the critical value 7.88 for $\alpha = 0.005$, hence B$_j$ is accepted as a relevant pattern for A$_i$. The following table shows the three highest ranked class pairs for pattern B$_j$. There is strong evidence that entities of the protein type tend to bind DNA locations, which is a reasonable conclusion.

B = SUBJ←bind→DIR_OBJ		
A	G^2	Sig
Protein_domain-DNA_domain_or_region	16.43	YES
Protein_family_or_group-DNA_d._or_r.	13.67	YES
Virus-Protein_molecule	7.84	NO

In all the experiments reported we used a fixed α value of 0.005, in addition we ignored bi-grams which occurred less than two times. Overall there are 490 such AB pairs.

3.4 Stage 2: generalization of relations

Relations can share very similar arguments as in “SUBJ←bind→DIR_OBJ” above: in both significant cases the direct object of “bind” is “DNA domain or region” while the subject is some kind of protein. This can be evidence that in fact we are facing a more general template which holds between superordinate concepts of the ones found in this first stage. It would be desirable to learn more general relations such as “Protein SUBJ←bind→DIR_OBJ DNA”, because in this way the learned ontology is more compact and would have a higher generalization power, i.e., the relations apply to more entities. This problem is similar to that of finding *selectional restrictions* of predicates, that is, the preferences that predicates such as verbs place on the semantic category of their arguments; e.g., that “eat” prefers direct objects that are “foods”. Several statistical methods have been proposed for selectional restrictions learning from text. We use here the method proposed by [Clark and Weir, 2002] which is both accurate and simple. We use the taxonomy defined in the GENIA ontology, see Figure 1, to generalize arguments of the learned patterns.[5]

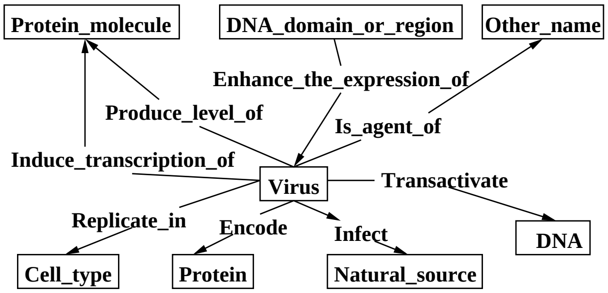

Figure 3: The “Virus” concept with the selected and generalized relations, and related concepts, in the enriched ontology.

Clark and Weir define an algorithm $top(c, r, s)$ which (adjusting the terminology to our case) takes as input a relation r, a class c and a syntactic slot s, and returns a class c' which can be c itself or one of its ancestors, whichever provides the best generalization for $p(r|c, s)$. The method uses the chi-squared test to check if the probability $p(r|c, s)$ is significantly different from $p(r|c', s)$, where c' is the parent of c. If this is false then $p(r|c', s)$ is a good approximation of $p(r|c, s)$, which is interpreted as evidence that (r, s) holds for c' as well. The procedure is iteratively applied until a significant difference is found. The last class considered is the output of the procedure, the concept that best summarizes the class that r “selects” in syntactic slot s. We computed the frequencies of patterns involving superordinate classes summing over the frequencies of all descendants of that class for that pattern found in the GENIA corpus.

Thus for each relation r, syntactic slot s and class c learned in the first stage we used Clark and Weir's method to learn a mapping from c to $top(c, r, s)$. We again used the G^2statistic and a fixed α level of 0.005. Using these maps we generalized, when possible, the original 287 patterns learned. The outcome of this process is a new set of 240 templates 153 of which have generalized arguments. As an example, the

[4] Dunning [1993] argues that G^2 is more appropriate than Pearson's X^2 with sparse data. Here they produce similar rankings while G^2 yields more conservative estimates; i.e., lower values than X^2.

[5] Four of the 36 GENIA corpus class labels, namely, “DNA substructure”, “DNA N/A”, “RNA substructure” and “RNA N/A”, have no entries in the GENIA ontology, we used them as subordinates of ” DNA” and “RNA”, consistently with “Protein N/A” and “Protein substructure” which in the ontology are subordinates of “Protein”.

two significant templates above “Protein domain binds DNA domain or region” and “Protein family or group binds DNA domain or region” are collapsed together in the generalized template “Protein binds DNA”. Figure 3 plots the labeled relations Virus is involved in and the respective classes after stage 1 and generalization.

4 Evaluation

Evaluating the enriched GENIA ontology is not a straightforward task. One could try to carry out an indirect evaluation by using the relations or their selectional restriction as features in an NLP task such as NER. The problem with indirect evaluations is that success or failure do not necessarily mean that the acquired knowledge is meaningful and relevant or not, but rather if it is useful or not. While it is important that the learned relations will also be useful we are first of all interested in understanding their degree of meaningfulness. For this reason a biologist, a linguist and an ontologist, who are all familiar with the GENIA ontology, carried out a direct assessment of the model’s outcome.

4.1 Biological evaluation

The templates learned in stage 1 were evaluated for their biological correctness with the following results: out of 287 tuples, 150 were meaningful, 30 made no sense, and 107 couldn’t be evaluated. Almost all of those that couldn’t be evaluated (96 cases), involved in one or both arguments the class “Other name”, which is an underdeveloped dyshomogeneous class which contains all sorts of things, from diseases to events such as gene expression. Relations involving this class are too generic to committ about the biological correctness of the tuple (e.g. “Protein molecule SUBJ← $induce \rightarrow DIR\_OBJ$ Other name”). Disregarding these cases 83.3% of the relations learned are correct while only 16.7% are wrong. In about half the nonsensical cases (14 cases), the original sentence contained a list of entities belonging to one class, which was learned as a tuple, where part of the list was included in the pattern and another part was included in the class pair; e.g. “Protein subunit SUBJ← $interact \rightarrow with \rightarrow Proteinmolecule$ →IND_OBJ_MOD Protein molecule”. These tuples are not strictly speaking wrong, but they do not seem very meaningful. In 17 cases the patterns didn’t make any sense probably due to wrong dependency paths.

Half of the errors can be reconduced to a relatively minor problem which has to do with better handling of coordination. The remaining errors might be due to wrong parses of the sentences and consequential wrong dependencies. This problem is actually not so bad given that the parser is trained on the Wall Street Journal. Parsing “pre-chunked” entities is likely to help here. While this problem does not have a simple fix in terms of the parser it might be argued that this it is made worse by the rather sparse data, therefore more data should help getting more signal. We plan to implement a NER system to tag more MEDLINE documents. We also identified a recurrent problem. Although the GENIA ontology was intended to be a model of cell signaling reactions it lacks important concepts such as *signaling pathway*. This leads to some wrong pattern mappings, as in the following example: ”An intact TCR signaling pathway is required for p95vav to function.”. In this case we derive the following relation: “Protein molecule is-required-for Protein molecule” since only “TCR” is annotated as a “Protein molecule” neglecting *signaling pathway*. Thus it is important to introduce a concept for “signaling pathway” and possibly others.

As far as stage 2 is concerned we checked all the generalized patterns excluding those involving “Other name”, 113 out of 153. Of these, 60 (53.1%) were correctly generalized; e.g., “Protein family or group activates DNA domain or region” is mapped to “Protein activates DNA”. At least 24 of the mistakes (21.2%) are due to over-generalizations; e.g., “Protein is-activated-in Source”. This often depends on the fact that the GENIA ontology doesn’t contain the desired level of abstraction.

An important aspect that needs to be addressed is the identification of synonymic relations. We would like to generalize over patterns and map them to synonym clusters thus defining relations as higher-order concepts. Thus the negated regulation in ”A negatively-regulates B” would be recognized as “A repress B”. Similarly, in the context of Protein-Protein interaction “positively-regulate” should be mapped to “activate”, same as “up-regulate”, “derepress”, “stimulate” etc. Representing relations as dependency paths allows to frame the problem of finding synonymic relations straightforwardly as that of finding paraphrases (cf. [Lin and Pantel, 2001]).

4.2 Ontological evaluation

In this section we provide an evaluation of the outcome of the system from the perspective of ontology engineering. We take the 153 learned patterns (78 relation types) that have passed the expert’s evaluation which now enter the ontology formalization and refinement life-cycle. For this purpose we compiled the GENIA ontology, including the newly learned relations, in OWL (Ontology Web Language [McGuinness and van Harmelen, 2004]). The GENIA taxonomy branches from three root classes: “Source”, “Substance”, and “Other name”. The latter is a placeholder which we ignore because too underdeveloped.

In order to evaluate the relations we start by *aligning*, i.e., mapping, “Source” and “Substance” to equivalent classes from another more general ontology. Ideally, the alignment of GENIA should be carried out against a more general ontology of the same domain such as TAMBIS [Stevens *et al.*, 2000]. Unfortunately TAMBIS scatters the subordinates of “Source” (including organisms, cells, etc.) across different branches, while “Substance” in TAMBIS does not cover the protein-related, and nucleic-acid-related, subordinates of “Substance” in GENIA.[6] In GENIA substances are classified according to their chemical characteristics rather than their biological role, while sources are biological locations where substances are found and their reactions take place. This distinction assumes a stacking of ontology layers within the physical domain where the biological is superimposed to the

[6] Notice that we are not questioning the quality of TAMBIS, but only its fitness for aligning GENIA, which shows nicely the kind of issues involved in ontology refinement.

chemical level. This feature of GENIA makes it suitable for alignment with another ontology called DOLCE [Gangemi *et al.*, 2003] which makes a similar distinction between two kinds of physical objects, “chemical” and “biological”, based on their different properties. DOLCE contains about 200 classes, 150 relations and more than 500 axioms, and has been used in domains including biomedicine and law. Thus we aligned “Source” and “Substance”to the biological and chemical classes in DOLCE and identified the 78 relations.

Since the root classes of GENIA are disjoint we check if there are relations whose domains or ranges mix up subclasses of “Source” with subclasses of “Substance”. Such relations do not imply logical inconsistency but rather reflect ontological inaccuracy. A small number of such relations is evidence for both effectiveness of the learning method with respect to GENIA, and for the fitness of GENIA classes with respect to textual evidence. The results of this generalization are quite positive: only 5 relations out of 78 (6.4%) have either a domain or range that is the union of disjoint classes.

Finally, we examine the relations at a finer semantic level: 54 (68%) are *eventive*, they encode a conceptualization of chemical reactions as events taking place in biological sources; 81% of the relations between biological and chemical classes are eventive, thus supporting the claim made in GENIA that biologically relevant chemical reactions involve both a biological and chemical object. The non-eventive relations have either a structural (e.g. “Consists-of”), locative (e.g. “Located-in”), logical (e.g. “is-one-of”), or epistemological meaning (e.g. “identified-as”).

5 Conclusion

We presented here a research in learning ontological relations from text applied to the domain of molecular biology. We designed a method and investigated its properties using the GENIA ontology and the relative manually annotated corpus. Our model is based on a representation of relations as a syntactic dependency paths between two named-entities. The method for finding “good” relations is based on the idea that ordered pairs of classes and relations can be seen as a bigrams, which allowed us to frame the problem as a collocations finding for which we applied simple statistical hypothesis testing. We also showed that it is possible to generalize over the arguments of the relation using a taxonomy and an algorithm for selectional restrictions learning. The results of both a biological and ontological analysis of the output of the system, which is completely unsupervised besides the decision of the critical value for α, are positive and promising. As future research we plan to investigate further the model, and the full range of potential relations, by producing more data by NER and testing its usefulness directly in information extraction tasks including relations extraction.

References

[Berland and Charniak, 1999] M. Berland and E. Charniak. Finding Parts in Very Large Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, 1999.

[Charniak, 2000] E. Charniak. A Maximum-Entropy-Inspired Parser. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, 2000.

[Clark and Weir, 2002] S. Clark and D. Weir. Class-Based Probability Estimation Using a Semantic Hierarchy. *Computational Linguistics*, 28, 2002.

[Craven and Kumlien, 1999] M. Craven and J. Kumlien. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB 1999)*, 1999.

[Dunning, 1993] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 1993.

[Friedman *et al.*, 2001] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. GENIES: A Natural-Language Processing System for the Extraction of Molecular Pathways from Journal Articles. *Bioinformatics*, 17(1), 2001.

[Gangemi *et al.*, 2003] A. Gangemi, Guarino N., C. Masolo, and A. Oltramari. Sweeting WordNet with DOLCE. *AI Magazine*, 24(3), 2003.

[Kazama *et al.*, 2002] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. Tuning Support Vector Machines for Biomedical Named Entity Recognition. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, 2002.

[Lin and Pantel, 2001] D. Lin and P. Pantel. DIRT - Discovery of Inference Rules from Text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2001)*, 2001.

[McGuinness and van Harmelen, 2004] D. McGuinness and F. van Harmelen. Owl Web Ontology Language Overview. In *W3C Recommendations: http://www.w3c.org/TR/owl-features/*, 2004.

[Ohta *et al.*, 2002] Y. Ohta, Y. Tateisi, J. Kim, H. Mima, and J. Tsujii. The GENIA Corpus: An Annotated Research Abstract Corpus in the Molecular Biology Domain. In *Proceedings of Human Language Technology (HLT 2002)*, 2002.

[Pantel and Ravichandran, 2004] P. Pantel and D. Ravichandran. Automatically Labeling Semantic Classes. In *Proceedings of HLT-NAACL 2004*, 2004.

[Pustejovsky *et al.*, 2002] J. Pustejovsky, J. Castaño, J Zhang, B. Cochran, and M. Kotechi. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In *Proceedings of the Pacific Symposium on Biocomputing, 2002*, 2002.

[Ratsch *et al.*, 2003] E. Ratsch, J. Schultz, J. Saric, P. Cimiano, U. Wittig, U. Reyle, and I. Rojas. Developing a Protein Interactions Ontology. *Comparative and Functional Genomics*, 4(1):85–89, 2003.

[Rojas *et al.*, 2002] I. Rojas, L. Bernardi, E. Ratsch, R. Kania, U. Wittig, and J. Šarić. A Database System for the Analysis of Biochemical Pathways. *In Silico Biology 2, 0007*, 2002.

[Rosario and Hearst, 2004] B. Rosario and M. Hearst. Classifying Semantic Relations in Bioscience Text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, 2004.

[Stevens *et al.*, 2000] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N.W. Paton, C.A. Goble, and A. Brass. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, 16(2), 2000.

[Swanson and Smalheiser, 1997] D.R. Swanson and N.R. Smalheiser. An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. *Artificial Intelligence*, 91(2), 1997.

[Šarić *et al.*, 2004] J. Šarić, L.J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Extraction of Regulatory Gene Expression Networks from PubMed. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, 2004.