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

Massimiliano Ciaramita† **Aldo Gangemi**† **Esther Ratsch**‡ **Jasmin Šarić**§ **Isabel Rojas**§
† Institute for Cognitive Science and Technology (CNR), Roma, Italy, *firstname.lastname*@istc.cnr.it
‡ University of Würzburg, Würzburg, Germany, esther.ratsch@biozentrum.uni-wuerzburg.de
§ EML-Research gGmbH, Heidelberg, Germany, *lastname*@eml-r.de

Abstract

In this paper we present an unsupervised model for learning arbitrary relations between concepts of a molecular biology ontology for the purpose of supporting text mining and manual ontology building. Relations between named-entities are learned from the GENIA corpus by means of several standard natural language processing techniques. An in-depth analysis of the output of the system shows that the model is accurate and has good potentials for text mining and ontology building applications.

1 Introduction

Bioinformatics is one of the most active fields for text mining applications because of the fast rate of growth of digital documents collections such as Medline and SwissProt. The ultimate goal of text mining in bioinformatics is knowledge discovery by means of natural language processing (NLP) and machine learning. To achieve this objective it is necessary to access the information contained "inside" the documents, i.e. their content. One possible strategy to get to the content is through *information extraction*. One starts with a conceptualization of the domain; i.e., a *domain ontology*, which specifies relevant *concepts* as well as *semantic relations*, such as is-a, part-of, and more complex relations encoding important interactions between concepts. Then it is necessary to apply extraction techniques to recognize where, in the documents, concepts are instantiated by specific entities, and where important interactions are expressed by linguistic structures.

Several ontologies which define concepts and structural semantic relations, e.g., is-a, are available. Instead there is a need for ontologies that specify relevant arbitrary semantic relations between concepts; for example, that "Cell express-the-receptor-for Protein" or that "Virus replicate-in Cell". In this paper we investigate the problem of enriching an existing ontology with arbitrary semantic relations which are strongly associated with ordered pairs of concepts. We design an unsupervised system that combines an array of off-the-shelf NLP techniques such as syntactic parsing, collocation extraction and selectional restriction learning. 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. An in-depth analysis of the learned templates shows that the model, characterized by a very simple architecture, has good potentials for text mining and ontology building applications.

In the next section we describe the problem of learning relations from text and related work. In Section 3 we describe our system and the data used in our study in detail. In Section 4 we discuss the evaluation of the system's output.

2 Problem statement and related work

The GENIA ontology contains concepts related to gene expression and its regulation, including cell signaling reactions, proteins, DNA, and RNA. Much work in bioinformatics has focused on *named-entity recognition* (NER), or *information extraction* (IE), where the task is the identification of sequences of words that are instances of a set of 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):

(1) "Untreated [$_{\text{Cell\_line}}$ NS-Meg cells] expressed [$_{\text{RNA\_family\_or\_group}}$ mRNA] for the [$_{\text{Protein\_molecule}}$ EPO receptor]"

A natural extension of NER is the extraction of relations between entities. NER and relation extraction could provide a better support for mining systems; e.g., patterns of entities and relations could be compared across document collections to discover new informative pieces of information (as in [Swanson and Smalheiser, 1997] for example). Currently most of the work on relation extraction applies hand-built rule-based extraction patterns; e.g., Friedman et al. [2001] on identifying molecular pathways and Šarić et al. [2004b] on finding information on protein interactions which use a manually-built ontology similar to that described in [Ratsch *et al.*, 2003]. One problem with rule-based IE is that systems tend to have good precision but 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

*We would like to thank our colleagues in the Laboratory for Applied Ontology (LOA-CNR), and the Klaus Tschira Foundation for their financial support.

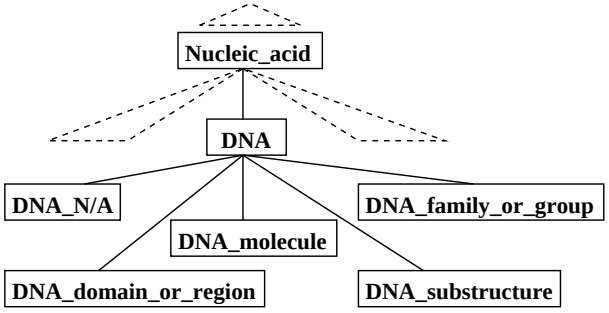

Figure 1: A small fraction of the GENIA ontology. Continuous lines represent unspecified taxonomic relations, dashed lines represent other regions.

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 extraction 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 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 (e.g., by checking the consistency of the arguments) or to align different databases.

Biological text mining systems that involve relations require predefined sets of relations that have to be manually encoded, a job which is complex, expensive and tedious, and that as such can only guarantee 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 texts in which named-entities, corresponding to ontology concepts, have been identified. Here we use the GENIA corpus, for which the tagging has been carried out manually, but the corpus data can be also generated automatically using an appropriate NER system. The model outputs a set of templates that involve pairs of GENIA ontology classes and a semantic relation. For example, a template might be "Virus infect Cell".

3.1 Data

The GENIA ontology was built to model cell-signaling reactions in humans with the goal of supporting information extraction systems. 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 consisting of 2,000 articles, 18,546 sentences, roughly half a million word tokens, and 36 types of labels. This corpus has complex annotations for disjunctive/conjunctive entities, for cases such as "erythroid, myeloid and lymphoid cell types". We excluded sentences that contained only instances of complex embedded conjunctions/disjunctions and also excessively long sentences (more than 100 words). The final number of sentences was 18,333 (484,005 word tokens, 91,387 tags). Many tags have nested structures; e.g. "[$_{\mathrm{Other\_name}}$ [$_{\mathrm{DNA}}$ IL-2 gene] expression]". Here we only considered the innermost labels, although the external labels 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 subsets of the entities of the related concepts, 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 domains 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 it provides extensive named-entity annotations which can be used to train appropriate NER systems (cf. [Kazama *et al.*, 2002].

3.2 Relations as dependency paths

The 18,333 sentences were parsed 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 entity names 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 substitution and the relative parse tree for the sentence of Example 1. Trees obtained in this way are simpler and don't split entities across phrases. For each tree we generated a *dependency graph*: each word[2] is associated with one *governor*, defined as the *syntactic head*[3] of the phrase closest

[1] Which takes roughly three hours on a Pentium 4 machine.

[2] Morphologically simplified with the "morph" function from the WordNet library, plus 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.

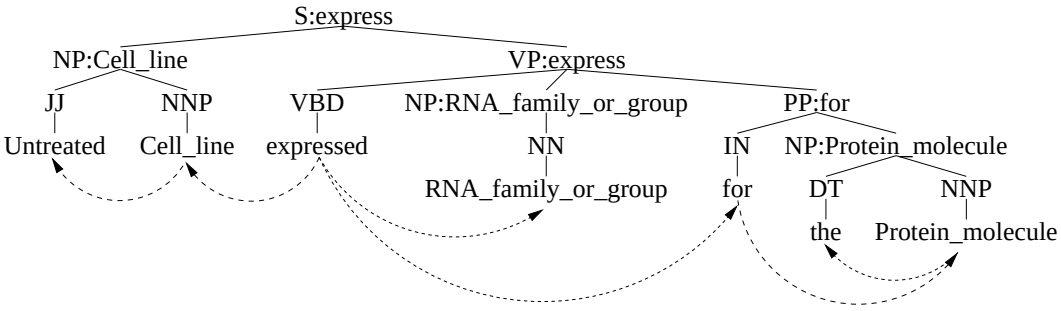

Figure 2: Parse tree for the sentence of Example 1. Entities are substituted with their tags. Phrases are labeled with their syntactic heads. The dependency graph is depicted with dashed directed edges pointing to the governed element.

to the word that differs from the word itself. For example, in Figure 2 "Cell_line" is governed by "express", while "Protein_molecule" is governed by the preposition "for".

In this way it is possible to formalize the notion of semantic relation between two entities. A related application of dependency relations concerns the recognition of paraphrase sentences [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. For example, in Figure 2 the path between "Cell_line" and "Protein_molecule" is "←express→for→". There is a path for every pair of entities in the tree. Paths can be considered from both directions since the reverse of a path from A to B is a path from B to A. A large number of different patterns can be extracted, overall we found 172,446 paths in the data. For the sake of interpretability, by inspection, of the outcome of the model in this paper we focused on a subset of these patterns. We selected 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 imposed the following constraints: c_i is governed by v under an S phrase (i.e., is v's surface subject, SUBJ), e.g., "Cell_line" in Figure 2; and one of the following six constraints holds:

1. 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;
2. 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;
3. 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";
4. 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";
5. 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*";
6. 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

The data contained 485 types of entity pairs, 3,573 types of patterns and 5,606 entity pair-pattern types.

3.3 Learning relations (Stage 1)

Let us take A to be an ordered pair of GENIA classes; e.g. A = (Protein_domain,DNA_domain_or_region), and B to be a pattern; e.g., B = SUBJ←bind→DIR_OBJ. Our goal is to find relations strongly associated with ordered pairs of classes, i.e., bi-grams AB. This problem is similar to finding *collocations*; e.g., multi-word expressions such as "real estate", which form idiomatic phrases. Accordingly the simplest method would be to select the most frequent bi-grams. However many bi-grams are frequent because either A or B, or both, are frequent; e.g., SUBJ←induce→DIR_OBJ is among the most frequent pattern for 37 different pairs. Since high frequency can be accidental and, additionally, the method doesn't provide a natural way for distinguishing relevant from irrelevant bi-grams, we use instead a simple statistical method.

As with collocations a better approach is to estimate if A and B occur together more often than at chance. One formulates a *null hypothesis* H_0 that A and B do not occur together more frequently than 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 bi-gram we create a contingency table of the frequencies AB, ¬AB, A¬B, and ¬A¬B; e.g., for A = Protein_molecule-DNA_domain_or_region and B = 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

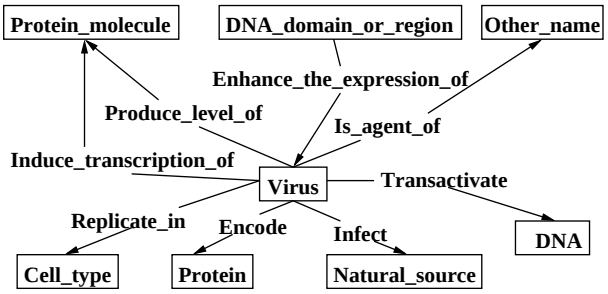

Figure 3: The "Virus" concept with the selected and generalized relations, and related concepts, in the enriched ontology.

the log-likelihood chi-squared statistic: [4]

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

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 table. 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 is accepted as a relevant pattern for A. The following table shows the three highest ranked class pairs for pattern B. 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 this study we used $\alpha = 0.005$. In general, α is an adjustable parameter which might be set on held-out data in order to maximize an objective function. We also ignored bi-grams occurring less than 2 times and pairs A, patterns B, occurring less than 4 times. Overall there are 487 suitable AB pairs, 287 (58.6%) have a value for G^2 higher than α.

3.4 Generalization of relations (Stage 2)

Relations can share similar arguments as in "bind" above where, in both significant cases, the direct object 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 holding between superordinates of the arguments found in the first stage. It is desirable, when possible, to learn more general relations such as "Protein SUBJ←bind→DIR_OBJ DNA", because the learned ontology is more compact and has greater generalization power, i.e., relations apply to more entities. Finding such generalizations is similar to learning *selectional restrictions* of predicates, that is, the preferences that predicates place on the semantic category of their arguments; e.g., that "eat" prefers objects that are "foods". Several methods have been proposed for learning such restrictions. We use here the method proposed in [Clark and Weir, 2002] which is accurate and simple. We use the taxonomy defined in the GENIA ontology, see Figure 1, to generalize arguments of the learned patterns.[5]

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 is 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 supposed to provide a good approximation for $p(r|c, s)$, 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, from the GENIA corpus, of all descendants of that class for that pattern.

For each relation r, slot s and class c, learned in stage 1, we used Clark and Weir's method to map c to $top(c, r, s)$. We again used the G^2 statistic and the same α value of 0.005. Using these maps we generalized, when possible, the original 287 patterns learned. The outcome of this process is a set of 240 templates 153 of which have generalized arguments. As an example, the templates above "Protein_domain binds DNA_domain_or_region" and "Protein_family_or_group binds DNA_domain_or_region" are mapped to the generalized template "Protein binds DNA". Figure 3 depicts the set of labeled relations the concept Virus is involved in, and the respective paired concepts, after stage 1 and 2.

4 Evaluation

We discuss now an evaluation of the model carried out by a biologist and an ontologist, both familiar with GENIA. The biological evaluation focuses mainly on the *precision* of the system; namely, the percentage of all relations selected by the model that, according to the biologist, express correct biological interactions between the arguments of the relation. From the ontological perspective we analyze semantic aspects of the relations, mainly the consistency with the GENIA classes.

4.1 Biological evaluation

The output of stage 1 is a set of 287 patterns, composed of an ordered pair of classes and a semantic relation. 91 of these patterns, involving in one or both arguments the class "Other_name", were impossible to evaluate and excluded altogether. This GENIA class is a placeholder, for very different sorts of things, which needs partitioning and structuring. Relations involving "Other_name" (e.g., "treat") might prove correct for a subset of the entities tagged with this

[4] Dunning [1993] argues that G^2 is more appropriate than Pearson's X^2 with sparse data; here they produce similar rankings.

[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".

label (e.g., "inflammation") but false for a different subset (e.g., "gene expression"). Of the remaining 196 patterns 76.5% (150) are correct, i.e., express valid biological facts such as "Protein_molecule induce-phosphorylation-of Protein_molecule", while 23.5% (43) are incorrect, e.g. "Protein inhibit-expression-of Lipid". Evaluation involved the exhaustive inspection of the original sentences to verify the intended meaning of the pattern and spot recurring types of errors. Half of the mistakes (22) depend on how we handle coordination, which causes part of the coordinated structure to be included in the relation. For example, the first two DNA entities in the noun phrase "DNA, DNA, and DNA" are governed by the head DNA rather than by, say, the main verb. Thus wrong relations such as "Protein bind-DNA DNA" are generated in addition to good ones such as "Protein bind DNA". Fixing this problem involves simply the specification of a better dependency structure for coordinations. Finally, 5 errors involved the class "Other_name" embedded somewhere within the relation, suggesting again generalizations that cannot be judged with enough confidence. The remaining errors are probably due to sparse data problems. In this respect we plan to implement an NER system to produce more data and consequently more reliable distributional information. Finally, we notice that, 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 errors as in the following case: "An intact TCR signaling pathway is required for p95vav to function.". In this case we derive the relation: "Protein_molecule is-required-for Protein_molecule" since only "TCR" is annotated as "Protein_molecule" neglecting *signaling pathway*.

To the best of our knowledge we can compare these results with one other study. Reinberger et al. [2004] evaluate (also by means of experts) 165 subject-verb-object relations, extracted from data similar to ours[6] but with a different approach. They report an accuracy of 42% correct relations. Their method differs from ours in three respects: relations are extracted between nouns rather than entities (i.e., NER is not considered), a shallow parser is used instead of a full parser, and relations are selected by frequency rather than by hypothesis testing. A direct comparison of the methods is not feasible. However, if the difference in accuracy reflects the better quality of our method this is likely to depend on any, or on a combination, of those three factors.

As far as stage 2 is concerned we first removed all relations involving "Other_name" (40 out of 153), which does not have superordinates nor subordinates, and evaluated if the remaining 113 generalized patterns were correct. Of these, 60 (53.1%) provided valid generalizations; e.g., "Protein_molecule induce-phosphorylation-of Amino_acid_monomer" is mapped to "Protein induce-phosphorylation-of Amino_acid_monomer". Excluding mistakes caused by the fact that the original relation is incorrect, over-generalization is mainly due to the fact that the taxonomy of the GENIA ontology is not a is-a hierarchy; e.g., "DNA_substructure" is not a kind of "DNA", and "Protein" is not a kind of "Amino_acid". Generalizations such as selectional restrictions instead seem to hold mainly between classes that share a relation of inclusion. In order to support this kind of inference the structural relations between GENIA classes would need to be clarified.

4.2 Ontological assessment

The 150 patterns validated by the expert are potential new components of the ontology. We compiled GENIA, including the newly learned relations, in OWL (Ontology Web Language [McGuinness and van Harmelen, 2004]) to assess its properties with ontology engineering tools. Ignoring "Other_name", the GENIA taxonomy branches from two root classes: "Source" and "Substance". GENIA classes, by design, tend to be *mutually exclusive*, meaning that they should be logically disjoint. Our main objective is to verify the degree to which the new relations adhere to this principle.

To analyze the relations we *align*, i.e., map, "Source" and "Substance" to equivalent classes of another more general ontology. Ideally, the alignment should involve an ontology of the same domain such as TAMBIS [Stevens *et al.*, 2000]. Unfortunately TAMBIS scatters the subordinates of "Source" (organisms, cells, etc.) across different branches, while "Substance" in TAMBIS does not cover protein and nucleic acid-related subordinates of "Substance" in GENIA.[7] In GENIA substances are classified according to their chemical features rather than 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 chemical level. This feature of GENIA makes it suitable for alignment with DOLCE-Lite-Plus (DLP, http://dolce.semanticweb.org), a simplified translation of the DOLCE foundational ontology [Gangemi *et al.*, 2003]. DLP specify a suitable distinction between "chemical" and "biological" objects. It features about 200 classes, 150 relations and 500 axioms and has been used in various domains including biomedicine [Saric *et al.*, 2004a].

We aligned "Source" and "Substance" to the biological and chemical classes in DLP. There are 78 types of relations out of 150, 58% of them (45) occur only with one pair of classes, i.e., are monosemous, while 33 have multiple domains or ranges, i.e., are polysemous. Since the root classes of GENIA are disjoint we checked if there are polysemous relations whose domain or range mix up subclasses of "Source" with subclasses of "Substance". Such relations might not imply logical inconsistency but raise doubts because they suggest the possibility that a class of entities emerged from the data, which is the union of two classes that by definition should be disjoint. Interestingly, there are only 4 such relations out of 78 (5.1%); e.g., "encode", whose subject can be either "Virus" or "DNA". In biology, DNA encodes a protein, but biologists sometimes use the verb "metonymically". By saying that a virus encodes a protein, they actually mean that a virus' genome contains DNA that encodes a protein. The small number of such cases suggests that relations emerging

[6] The SwissProt corpus, 13 million words of Medline abstracts related to genes and proteins.

[7] Notice that we are not questioning the quality of TAMBIS, but only its fitness for aligning GENIA.

from corpus data are consistent with the most general classes defined in GENIA.

At a finer semantic level relations are composed as follows: 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, supporting the claim made in GENIA that biologically relevant chemical reactions involve both a biological and chemical object. Non-eventive relations have either a structural (e.g. "Consist-of"), locative ("Located-in"), or epistemological meaning ("identified-as").

5 Conclusion and future work

We presented a study on learning semantic relations from text in the domain of molecular biology. We investigated an unsupervised approach using the GENIA ontology and its corpus. Our model is based on a representation of relations as syntactic dependency paths between an ordered pair of named-entities. To select "good" relations, class pairs and dependency paths can be interpreted as bi-grams, and scored with statistical measures of correlation. We showed that it is also possible to generalize over the arguments of the relation using a taxonomy and algorithms for selectional restrictions learning. The results of a biological and ontological analysis of the acquired relations are positive and promising.

Other aspects need to be addressed beyond precision, in particular we are interested in evaluating the recall, i.e., the coverage, of the system,[8] the precision of alternative selection criteria, and the usefulness of automatically learned relations in information extraction tasks. The latter will imply the identification of synonymic relations; e.g., in the context of Protein-Protein interaction "positively-regulate" is equivalent to "activate", "up-regulate", "derepress", "stimulate" etc. Representing relations as dependency paths one can frame this problem straightforwardly as that of finding paraphrases (e.g. as in [Lin and Pantel, 2001]).

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.

[8] Which is more problematic because it requires, in principle, considering a very large number of discarded relations.

[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.

[Reinberger *et al.*, 2004] M-L Reinberger, P. Spyns, and A.J. Pretorius. Automatic Initiation of an Ontology. In *Proceedings of ODBase 2004*, 2004.

[Rojas *et al.*, 2002] I. Rojas, L. Bernardi, E. Ratsch, R. Kania, U. Wittig, and J. Saric. 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.

[Saric *et al.*, 2004a] J. Saric, A. Gangemi, E. Ratsch, and I. Rojas. Modelling gene expression. In *Proceedings of the Workshop on Models and Metaphors from Biology to Bioinformatics Tools (NETTAB 2004)*, 2004.

[Saric *et al.*, 2004b] J. Saric, 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.

[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 Scientifi c Discovery. *Artificial Intelligence*, 91(2), 1997.