POPULATING BIOMEDICAL ONTOLOGIES FROM NATURAL LANGUAGE TEXTS

Juana María Ruiz-Martínez, Rafael Valencia-García, Rodrigo Martínez-Béjar
Computer Science Faculty, University of Murcia, Campus de Espinardo, 30100 Murcia, Spain

Achim Hoffmann
School of Computer Science and Engineering, University of New South Wales, Sydney 2052, Australia

Keywords: Ontology Population, Semantic Role, Knowledge Acquisition.

Abstract: Ontology population is a knowledge acquisition activity that relies on (semi-) automatic methods to transform unstructured, semi-structured and structured data sources into instance data. In this work, a semantic-role based process for ontology population is presented that provides a suitable framework for textual knowledge acquisition in the biological domain. In particular, with our approach, a given ontology can be enriched by adding instances gathered from biological natural language texts. Our system’s modular architecture provides a greater versatility than current approaches in the mentioned domain, as the process of ontology population is not directly dependent on the linguistic rules developed from the corpus.

1 INTRODUCTION

The exponential growth of scientific literature in the biomedical domain makes it difficult for researchers to access the massive amounts of online information and to keep abreast of biomedical knowledge that spreads at an increasing rate.

In recent years, several techniques for discovering, accessing, and sharing knowledge from medical literature have been developed, including a remarkable number of studies on discovering various kinds of knowledge by mining the medical literature (Friedman et al., 2006).

The Semantic Web is viewed as an extension of the current Web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The knowledge representation technology used in the Semantic Web is the ontology, which provides the meaning and facilitates the search for contents and information (Jiang et al., 2009). So, in complex domains such as biomedicine, ontologies are used for organizing and sharing biological knowledge as well as integrating different sources of knowledge in order to provide interoperability among different research communities.

Ontologies have captured the interest of the entire biomedical community (Rubin et al., 2007), in part due to impact and success of the Genome Ontology (GO) (Lewis, 2005). But the manual building of ontologies is a tedious, time consuming task which results in a knowledge acquisition bottleneck. So, automated ontology learning methods have been proposed to allow a reduction in the time and effort needed in the ontology development process (Valencia-García et al., 2008).

Ontology Learning (also named ontology generation or ontology extraction) is a knowledge acquisition activity that relies on (semi-) automatic methods to transform unstructured (e.g. corpora), semi-structured (e.g. folksonomies, html pages, etc.) and structured data sources (e.g. databases) into conceptual structures. Some ontology learning approaches, such as TERMINAE (Aussenac-Gilles et al., 2008), provide guidance to conceptualization from natural language text integrating functions for linguistic analysis and conceptual modelling. Ontology Population, on the other hand, is a knowledge acquisition activity that relies on (semi-) automatic methods to transform unstructured, semi-structured and structured data sources into instance data. Thus, while Ontology Learning deals with the acquisition of new concepts and relations with the...
consequence of changing the definition of the ontology itself, the goal of Ontology Population is the extraction and classification of instances of the concepts and relationships defined in the ontology. The instantiation of the ontology with new knowledge is a relevant step towards the provision of valuable ontology-based knowledge services.

In this work, a scalable methodology for ontology population from textual resources based on Natural Language Processing (NLP) and ontological engineering techniques is proposed. This methodology attempts to support expert communities in building ontologies from natural language texts. It allows several semantic relations to be used and reduces the degree of expert participation during the ontology construction process. Our methodology has been implemented in the form of a software prototype and tested in the biomedical domain.

The structure of the paper can be described as follows. In Section 2 related works are shown, whereas Section 3 presents the Technical Background. In Section 4, the Ontology population process is described. Finally, some conclusions are put forward in Section 5.

2 RELATED WORK

During the last decade several approaches that (semi-)automatically build domain-specific ontologies have been proposed, some of them just generating hierarchies (taxonomies) of concepts, (Cimiano et al., 2005) or relating concepts with a reduced set of semantic relationships (Maedche et al., 2001) (He, 2006). In biomedicine, taxonomy and partonomy relationships are an important starting point, although they are not enough for modelling such a complex domain.

Some ontology learning methods enrich pre-existing ontologies. For example, in (Agirre et al., 2000) a methodology for enriching the concepts of WordNet is presented. In (Sánchez and Moreno, 2008), a domain ontology is enriched by discovering non-taxonomic relationships from the web using patterns based on verb phrases. Other works such as (Bada et al., 2007) focus on the enrichment of biological ontologies through integration processes.

Regarding the biological domain, a number of biological NLP and text mining systems have been developed to extracting biological information and knowledge. Most of them include a module that recognizes biological entities or concepts in text, usually called Named Entity Recognition (NER). Biological NER is the task of identifying the boundary of a substring and then mapping it onto a predefined category (e.g., Protein, Gene or Disease). Once the biological terms have been extracted, the semantic relations can then be detected. In (Bundschus et al., 2008), semantic relations between diseases and treatments are classified using Conditional Random Fields. In (Chun et al., 2006), relationships between genes and diseases from MedLine abstracts are obtained by studying the co-occurrence of terms.

Other approaches (He, 2006) use machine learning techniques and discourse analysis methods in order to extract protein to protein interactions, however they just considered taxonomy and partonomy relationships.

In (Rosario and Hearst, 2004), the authors identify semantic relations between “treatment” and “disease” in bioscience texts by means of graphical models and a neural network. Semantic role labelling approaches have been developed to extract the semantic relations. For example, PASbio (Wattarjeekrit et al., 2004) is an extended model of PropBank applied to Molecular Biology. The work introduces the notion of semantic analysis of argument roles in biological texts and proposes the construction of Predicate Argument Structures (PAS) for Molecular Biology. PAS are knowledge structures that represent the relations between a verb and its arguments. These predicates describe the roles of genes and gene products in mediating their biological functions. BioFrameNet (Dolbey et al., 2006) is an extension of FrameNet that has added semantic frames relevant to the Molecular Biology domain. The semantics have been implemented in OWL DL to facilitate links to domain ontologies like GO or EntrezGene. Finally, BIOSMILE (Tsai et al., 2007) is another semantic role labelling system that was trained on BioPop, a biomedical proposition bank semi-automatically annotated consisting of 30 biomedical verbs that were annotated into 500 abstracts. This system incorporates lemmatized forms together with Part-of-Speech tags and NE types.

3 TECHNICAL BACKGROUND

3.1 Ontologies

In this work, an ontology is seen as “a formal and explicit specification of a shared conceptualisation” (Studer et al., 1998). Ontologies provide a formal and structured knowledge representation that has the
advantage of being reusable and shareable. In our methodology, ontologies are obtained as a result of knowledge extraction processes.

The Web Ontology Language (OWL) has been used to represent the biological ontologies that are to be populated from texts. In OWL, the main ontological entities are classes, “subclass of” relationships, datatype properties, object properties, and individuals. OWL provides a formal theory for taxonomy, whereas any other semantic relationship, like partonomy or topology, must be manually defined and implemented using object properties for that.

In (Smith et al., 2005), the most common relations used in biomedical domain ontologies were presented and formalized. As a result of this effort, the OBO ontology of biomedical relations was produced. The OBO Relation Ontology (http://www.obofoundry.org/ro) comprises ten different types of relations including taxonomic and partonomic relations. In this work, an ontological model based on the different types of relations defined in OBO Relations Ontology, has been defined. This ontological model has been implemented using the new version of the OWL language, namely OWL 2, that adds several new features to OWL, including increased expressive power for properties and extended support for datatypes.

Table 1: OWL 2 Property axioms of the semantic relations in OBO relation ontology.

<table>
<thead>
<tr>
<th>Relation</th>
<th>T</th>
<th>S</th>
<th>R</th>
<th>I</th>
<th>A</th>
<th>F</th>
<th>IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>is a</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>part_of</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>located_in</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>contained_in</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>adjacent_to</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>transformation_of</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>derives_from</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>preceded_by</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>has_participant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>has_agent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>instance_of</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

These relations are binary, and they have been implemented using object properties and the property axioms that can be defined in OWL 2 as shown in Table 1. The main OWL 2 property axioms are described in the following:

- Reflexive (R). (X Relation X)
- Irreflexive (I). not(X Relation X).
- Symmetric (S). (X Relation Y) ↔ (Y Relation X)
- Asymmetric (A). (X Relation Y) → not(Y Relation X).
- Transitive (T). (X Relation Y) and (Y Relation Z) → (X Relation Z).
- Functional (F). (X Relation Y) and (X Relation Z) → (Y = Z)
- Inverse Functional (IF). (X Relation Y) and (X Relation Y) → (X = Z)

An OWL formal model allows for performing automatically a set of Description Logic inference services, which can be supported by DL reasoners (e.g., HermiT, Pellet2, Fact++, Racer) (Sirin and Parsia, 2004):

- Consistency checking, which ensures that an ontology does not contain any contradictory facts.
- Concept satisfiability, which checks whether it is possible for a class to have any instances. If a class is unsatisfiable, then defining an instance of the class will cause the whole ontology to be inconsistent.
- Classification, which computes the subclass relations between every named class to create the complete class hierarchy. The class hierarchy can be used to answer queries such as getting all or only the direct sub-classes of a class.
- Realization, which finds the most specific classes that an individual belongs to; or in other words, it computes the direct types for each of the individuals.

An OWL ontology can be viewed from a logical point of view as a collection of axioms that must be satisfied. This does not only include classes and properties, but also restrictions such as disjoint classes. Consistency is a critical issue in Ontology Engineering. We say that an ontology is internally inconsistent when some parts of it are inconsistent with other parts of itself. For instance, an ontology is internally inconsistent if one of the properties concerning relationships between concepts is not satisfied. The property axioms defined in Table 1 help to detect any inconsistency in the populated ontology. For example, the part-of relation holds both transitive and asymmetric properties, so it is not possible to have a cycle inside a conceptual partonomy. This ensures the correct results of the knowledge that can be inferred from the ontology by applying the corresponding axioms. Moreover, the
existence of such restrictions is useful to grant the consistency of the individuals built, which must satisfy the restrictions defined for their corresponding class. Moreover, the collection of conditions defined for the classes can be used by the reasoner for the automatic classification of individuals.

3.2 Named Entity Recognition (NER)

The gap between linguistic biomedical texts and the extraction and organization of their knowledge in ontologies has been addressed primarily from the extraction of terms and relations between them. Terms extraction is a prerequisite for all aspects of ontology learning from text. Terms are linguistic realizations of domain-specific concepts and are therefore central to further, more complex tasks. In relation to current systems and frameworks related to this work, UMLS stands out as it merges information from more than 100 biomedical vocabularies, which makes existing terminologies both easier to use and more useful. However, not all these terms will be names of biomedical entities which are essential for populating the ontology.

In biomedical literature, NER refers to the task of recognizing entity-denoting expressions such as genes, proteins, cells and diseases. In this work, NER is performed by the GENIA Named Entities module, based on the GENIA corpus, combined with grammatical patterns.

3.3 Semantic Role Labelling

A semantic role is the relationship between a syntactic constituent and a predicate. It defines the role of a verbal argument in the event expressed by the verb. The semantic roles set developed in the PropBank (Palmer et al., 2005) and in the FrameNet project (Filmore, 2002) are the most widely used in the literature.

In the biological domain, the most important sets of semantic roles are PASbio (Wattarujeekrit et al., 2004), BioFrameNet (Dolbey et al., 2006) and BIOSMILE (Tsai et al., 2007). In this work, the semantic roles provided by PASbio are used for extracting and detecting ontological relationships (amongst those defined in Table 1) between the named entities extracted in order to obtain and insert the individuals of the domain ontology. In Figure 1, an example of a part of the transform frame in PASbio is shown. This pattern models the relation between an entity that is going to be transformed into a new state by an agent through the verb transform.

For expression of the recombinant protein, \[ \text{[pET28a-5] was transformed into Escherichia coli strain BL21(DE3)}. \]

Figure 1: An example of the transform frame in PASbio.
relationship transformation_of in the following manner:

\[ \text{transformation.01}(\text{role}_0, \text{role}_1, \text{role}_2) \Rightarrow \text{role}_2 \]

That is, in this frame there exists an ontological relation transformation_of between the role_2 and role_1. For example, in the sentence shown in Figure 1 the relation \[\text{Escherichia coli strain BL21(DE3)}\] transformation_of \[\text{pET28a-5}\], would be obtained.

A mapping between verbal expressions and PASbio frames is also necessary. In Table 2 an excerpt of these mappings are showed.

Table 2: An excerpt of the mapping between verb expressions, PASbio frames and the ontological relationships.

<table>
<thead>
<tr>
<th>Verb expression</th>
<th>PASbio Frame</th>
<th>Ontological relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>be transform into</td>
<td>Transform</td>
<td>Transformation_of</td>
</tr>
<tr>
<td>is altered by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>was mutated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>changes in</td>
<td>Modify</td>
<td>Transformation_of</td>
</tr>
<tr>
<td>be susceptible to modify</td>
<td></td>
<td></td>
</tr>
<tr>
<td>may develop</td>
<td>Develop</td>
<td>Derives_from/</td>
</tr>
<tr>
<td>be altered</td>
<td>Alter</td>
<td>transformation_of</td>
</tr>
<tr>
<td>be generated by/from</td>
<td>Generate</td>
<td>Derives_from</td>
</tr>
<tr>
<td>is the result of</td>
<td>Result</td>
<td>Derives_from</td>
</tr>
<tr>
<td>resulting in</td>
<td></td>
<td></td>
</tr>
<tr>
<td>are transcribed from</td>
<td>Transcribe</td>
<td>Derives_from</td>
</tr>
</tbody>
</table>

4 ONTOLOGY POPULATION PROCESS

The main aim of the process proposed here is to populate biological domain ontologies from natural language text using NLP and semantic technologies. The architecture of the process is shown in Fig. 2. It is composed of three main sequential phases: NLP Phase, NER and relation extraction Phase and the Ontology Population Phase. In a nutshell, the process works as follows: in the NLP phase a lexical and syntactical analysis of the corpus is done. Here the grammar category of the words and the constituents of each sentence are obtained. In the second phase, the possible NEs and the semantic relationships between them are extracted using NER and semantic roles technologies. In the final phase, the instances of the domain ontology are obtained from the semantic annotations of the previous phase. In this phase, the consistency of the instances and its classification in the ontology are also addressed. An enriched consistent domain ontology is obtained at the end of this phase. In Figure 2, the referred phases are shown.

4.1 NLP Phase

The main objective of this phase is to obtain the morphologic and syntactic structure of each sentence. For this, a set of NLP tools including a sentence detection component, tokenizer, POS taggers, lemmatizers and syntactic parsers has been developed using the GATE framework (http://gate.ac.uk/). GATE is an infrastructure for developing and deploying software components that process human language. GATE helps scientists and developers in three ways: (i) by specifying an architecture, or organizational structure, for language processing software; (ii) by providing a framework, or class library, that implements the architecture and can be used to embed language processing capabilities in diverse applications; (iii) by providing a development environment built on top of the framework made up of convenient graphical tools for developing components.

Some aspects of biomedical texts may affect the morphologic and syntactic analysis, such as the ambiguity caused by names and abbreviations that begin with capital letters; chemical and numeric expressions including non-alphanumeric characters such as commas, parentheses, and hyphens; participles of unfamiliar verbs that describe domain-specific events; and fragments of words (Tateisi and Tsujii, 2004). The GENIA tagger (Tsuruoka et al.,...
is able to manage these problems more efficiently than a general POS Tagger. Nevertheless, not all the ambiguities are solved and, in some cases, these unsolved issues can affect to later stages in the process. Next, an example of a sentence analyzed with GENIA is shown:

After/IN treatment/NN with/IN 5-azacytidine/NN , the/DT adult/JJ mesenchymal/JJ stem/NN cells/NNS were/VBD transformed/VBN into/IN cardiomyocytes/NNS .

Once the sentences have been analysed, a lemmatization and the shallow parsing syntactic analysis is done using Freeling (Atserias et al., 2006) to obtain the main chunks of the sentence:

[After] [treatment] [with] [5-azacytidine] , [the adult mesenchymal stem cell] be transformed into [cardiomyocytes].

4.2 NER and Relation Extraction Phase

During this phase firstly the NE candidates are identified by making use of the GATE Framework. The output produced by each component of GATE is a set of annotations, namely metadata associated with a particular section of the document content. Each annotation in the text is then merged into a unified representation for each entity. All the occurrences of identified NEs (NE mentions) in the text will be candidate instances in the ontology. A combination of JAPE rules and lists of Gazetters are also used to perform the processes associated with this phase.

Jape is a rich and flexible regular expression-based rule mechanism offered by the GATE framework (Sabou et al., 2005). On the other hand, the gazetteer consists of a list of entities that are relevant in the domain under question. Several lists containing biological terms extracted from GeneOntology (http://www.geneontology.org/) and UMLS (http://www.nlm.nih.gov/research/umls/) have been created. Examples of the general lists that have been created are: Lipid, DNA, Aminoacid monomer, Peptide, Organic compound, Multicellular organisms, Cell type, etc. In Figure 3 an example of the named entities recognized using GATE is shown.

Once the NEs have been extracted, the PASbio frames are detected in the text in order to extract the possible relations between these named entities. For example, in the previous example verb expression be transform into is found, so the PASbio frame transform has to be applied and that indicates that there exists an ontological relationship transformation_of between the named entities obtained. That is cardiomyocytes transformation_of mesenchymal stem cells.

Figure 3: Obtaining named entities with GATE.

4.3 Ontology Population Phase

Given a set of NEs and the ontological relationships between them, during this phase the system firstly determines if they are individuals of the ontology. If so, the system must assign each individual to a particular class of the domain ontology and inserted it into the ontology. If some of the individual already exist in the ontology then they are not inserted.

Each type of NE is associated with a Class of the Ontology, so that, those NE mentions that in the previous phase have been classified into a type of NE become candidates for Ontology Instances. These candidates are provisionally inserted in the corresponding class or classes. Once the reasoner checks the consistency of the ontology, those instances that are inconsistent are deleted.

For example, Myosin heavy chain and troponin I are NE mentions or occurrences of the type Protein. The NE Protein is associated in the ontology with the class Protein. So these NE mentions become individuals of the Protein class. At the end of the population process, the reasoner checks if the new instances are consistent and if new knowledge can be inferred.
Regarding the relationships between the individuals, they are represented by means of object properties between classes in the domain ontology. Due to the fact that the domain ontology has been developed using the ontological model described in section 3.1, the object properties defined between the classes have to be associated with any of the relationships described in Table 1.

Once the individuals have been inserted into an ontology, they have to be related using the relationships identified in the previous phase. For each of these relationships, the individuals that participate in it are obtained from the ontology and the system checks if they are already related by an object property of the same type (i.e. part_of, located_in, derives_from, etc). The relationship is only inserted if the relationship does not exist in the ontology, yet.

After that, a reasoner such as Pellet2 is executed in order to (1) check for the consistency of the ontology and (2) compute inferred types. If the ontology is inconsistent, the last relationship inserted into the ontology will be removed. In case that the ontology is consistent, and the reasoner has inferred that one individual belonging to the relationship can be classified into a new class, this new classification is done.

5 EVALUATION

A software prototype that implements this methodology by means of a platform has been developed for validating our approach. The platform has been developed in Java, and the OWLAPI has been used for processing the content of the ontologies. The OWLAPI does not only provide a rich API for dealing with OWL ontologies, but also facilitates the use of OWL reasoners, so making it possible to employ query languages such as SPARQL. The NLP part is done using the GATE framework along with some of the resources it provides. A GENIA POSTagger and Freeling plugin has been developed to integrate with GATE.

An ontology extracted from the xGENIA ontology (Rak et al., 2007) has been used in the experiment. The xGENIA ontology is an OWL-DL ontology based on the GENIA taxonomy that was developed as a result of manual annotation of the GENIA corpus, which is a subset of the MEDLINE one. Both the ontology and the corpus have been used as a benchmark to test and develop biological information extraction tools.

In Figure 4, the ontology that has been used for validation is shown. The object properties between classes in the xGENIA ontology are not based on the most common relationships modelled by the OBO relation ontology (see Table 1), so some changes in these properties have been done in order to represent them as a subset of the relations proposed in Table 1. The ontology represents two hierarchies:

- **Protein**: Proteins are organic compounds made of amino acids arranged in a linear chain and folded into a globular form. Proteins include protein groups, families, molecules, complexes, and substructures. In the ontology the following protein types have been modelled: protein family or group, that is, a family or a group of proteins; protein complex, which includes conjugated proteins such as lipoproteins and glycoproteins; individual protein molecules, which are individual members of a group of non-complex proteins, and subunits of a protein complex.
- **Natural source**: Natural sources are biological locations where substances are found and their reactions take place. In this hierarchy only the types, body part, cell type and organism have been taken into account.

![Figure 4: An excerpt of the ontology.](image)

Apart from that, four non-taxonomic relations have been defined:

- Protein is located in Natural source
- Protein derives from Protein
- Protein_molecule part of Protein_complex
- Protein_subunit part of Protein_complex

A small corpus extracted from the GENIA corpus containing 3,798 words has been used for populating
and extracting the relations between the individuals in the ontology.

Due to the fact that the prototype uses the GENIA NER, the precision and recall of the NE extraction is close to 96%. The aim of this experiment was to evaluate the precision and recall of the relation detection between instances in the population process.

The resulting populated ontology has been compared to the part of the xGENIA ontology which is already instantiated.

Some frames representing part_of and is located_in relations have been manually developed since PASbio frames are only based on event detection relationships in molecular biology.

The precision score has been defined as the number of ontological relationships extracted that exist in the xGENIA ontology divided by the total number of relationships extracted:

$$\text{precision} = \frac{\text{correct relationships extracted}}{\text{total relationships extracted}}$$

Another evaluation parameter used has been the recall score, which has been defined as the number of ontological relationships extracted divided by the total number of ontological relationships that exist in the corpus:

$$\text{recall} = \frac{\text{correct relationships extracted}}{\text{total relationships in corpus}}$$

The prototype achieved a precision of 79.02% and a recall of 65.4%. These values are significant because (1) the domain is quite specific, and (2) the semantic roles used have been designed for the biomolecular domain.

6 CONCLUSIONS AND FUTURE WORK

The semantic-role based process for ontology population presented here provides a suitable framework for textual knowledge acquisition in the biological domain. In particular, with our approach, a given ontology can be enriched by adding instances gathered from biological natural language texts.

In this work, the NER process is performed using the GENIA Named Entities module, which is based on machine learning techniques. Currently, there exist many knowledge bases in the biomedical domain such as UMLS and GeneOntology and the use of such controlled vocabularies would be very helpful for identifying NE and terms in biomedical text. It is planned to develop a new NER module that can also use these ontologies.

The performance of our system depends heavily on the performance during the NER phase. Poor performance during this phase limits significantly the system’s recall and precision.

On the other hand, the semantic relation extraction is done using the reduced set of semantic roles defined in PASbio. In order to improve the semantic role detection, we are planning to include more frames from BioFrameNet, BIOSMILE, FrameNet and VerbNet to improve this issue.

Therefore, a set of predefined semantic relations based on the OBO relation ontology have been defined. However, some of the relationships between the instances of the GENIA corpus cannot be modelled with this set. BIOTOP (Beißwange et al., 2008) is an upper domain ontology for biology that adds some semantic relationships to the OBO relation ontology such as, for example, has-inherence, realization-of or has-grain relationships.

The system’s modular architecture gives our system a greater flexibility, as the process of ontology population is not directly dependent on the linguistic rules developed from the corpus, such as the approaches presented in (Tanev and Magnini, 2006) and (Amardeilh et al., 2005). The results of the validation seem promising although they should be compared with those of other ontology construction methods. A more in-depth evaluation of the system is planned, comprising the application of the whole GENIA corpus and xGENIA ontology, the use of statistical methods for analyzing the results, and the comparison with some other ontology population methods is also planned. The validation of the proposed methodology within the scope of other related domains such as the biomedical domain is also left for future work.

ACKNOWLEDGEMENTS

This work has been possible thanks to the Spanish Ministry for Education through grant JC2009-00194 under the program “José Castillejo”, and the Regional Government of Murcia under project BIO-TEC 06/01-0005. Juana María Ruiz-Martínez is supported by the Fundación Séneca through grant 06857/FPI/07.
REFERENCES


Aguirre, E., Ansa, O., Hovy, E. and Martinez, D. 2000, Enriching very large ontologies using the WWW, Proceedings of the ECAI Ontology Learning Workshop in conjunction with the 14th European Conference on Artificial Intelligence (ECAI 2000).


Buitelaar, P., Cimiano, P. and Magnini, B. 2005, Ontology learning from text: An overview, Ontology learning from text: Methods, evaluation and applications, , pp. 3-12.


He, X. 2006, A protocol for constructing a domain-specific ontology for use in biomedical information extraction using lexical-chaining analysis. Thesis presented at University of Waterloo.


Sánchez, D. and Moreno, A. 2008, Learning non-taxonomic relationships from web documents for
domain ontology construction, *Data & Knowledge Engineering*, vol. 64, no. 3, pp. 600-623.


