An Automated Semantic Annotation Tool Supported by an Ontology in the Computer Science Domain

Rodrigo Espinoza[1,2] and Andrés Melgar[1,3]

[1] *Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada,*
Pontificia Universidad Católica del Perú, Lima, Peru
[2] *Especialidad de Ingeniería Informática, Facultad de Ciencias e Ingeniería,*
Pontificia Universidad Católica del Perú, Lima, Peru
[3] *Sección de Ingeniería Informática, Departamento de Ingeniería,*
Pontificia Universidad Católica del Perú, Lima, Peru

Keywords: Semantic Annotation, Automated Annotation, Annotation Tool, Ontology, Automated Semantic Tool.

Abstract: The annotation of documents can be performed manually, semi-assisted or automated, also it can use the help of different knowledge resources as a set of rules or ontology. In this paper, we show the design of a semantic annotation tool that works automatically on power in order to efficiently manage academic documents in spanish produced in the university related to computer science. The tool uses an ontology annotations to provide a corpus of documents the necessary attributes to be managed using other tools that use annotations as searchers or indexers. This is done by relating the concepts found in documents with concepts in the ontology performing semantic and syntactic comparisons, it is produced using open source tools for natural language processing and knowledge management.

1 INTRODUCTION

The Web was designed to be understood by humans, so most of the information it contains can be neither understood nor processed by machines. This results in problems in searching, organizing and maintaining pages hosting. The Knowledge Management (KM) problems in it are closely related to the size of the Web; while there are more number of pages, searches and information maintenance become worst (Berners-Lee et al., 2001). This situation causes the upload of redundant information every day, in consequence the efficiency of Knowledge Retrieval (KR) in the Web decreases. Also inefficient searches generate a lot of transactions, which saturates the global network, it cause huge maintenance costs and forcing new solutions on how to improve their infrastructure. On the other hand, not being able to analyze the content of the pages generate many problems in the transference of knowledge (Studer et al., 2000). Knowing the problems with the Web, the solution may be that machines should be able to understand the resources (pages) found on the Web, be able to process and analyze them to perform better searchs and classifications based on the content of the pages. One way to do that is to provide them with properly structured metadata, that it have consistent information on the most important concepts of the documents content by domain. Metadata is information about the content of a document, which facilitate processing by software agents (Wolfe, 2000).

One of the resources capable of providing that enriched information to the pages are the semantic annotation tools that make use of ontologies. The goal is to use metadata to annotate pages and documents according to the information that they contain, that is made using an ontology in the respective domain of information from documents in question. The use of ontologies allows us to unify a single concept in various heterogeneous representations. The analysis can be done in a word or a phrase by linking the main content of the page and the existing elements in the ontology (Corcho, 2006).

This paper propose an automatic semantic annotation tool that uses an ontology whose domain is computer science. The tool will be used to semantically annotate documents produced in university who belong to the domain of ontology. These annotations allow other search tools and information management promote the documents among the university commu-

Espinoza, R. and Melgar, A..
An Automated Semantic Annotation Tool Supported by an Ontology in the Computer Science Domain.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 2: KEOD*, pages 133-138
ISBN: 978-989-758-158-8

nity.

This paper is structured as follows: after this introduction, we present the literature review about semantic web, annotation, annotation with ontologies, metadata and natural language processing. Subsequently, some related works are presented. In the following sections the proposed semantic annotation tools is described. Then, we present the developed prototype, the results and the discussion. Finally, in the last section, we present the conclusions and future works.

2 LITERATURE REVIEW

2.1 Semantic Web

Now what we call Web would become the syntactic Web, which perform searches simply by finding matches of words or phrases that we indicate. The Semantic Web does not have the same nature, but rather it is an extension of syntactic, it provides semantic support to the content of the pages and that allows both people and computers work together with information from the Web (Berners-Lee et al., 2001). This aggregate on pages is compounded for metadata or meta-information that will allow machines to understand and process their content just like a human would (Davies et al., 2003). Actually we can see KR with the behaviour of the Semantic Web in corporate intranets and information systems of large multinationals, the information for these organizations is one of their most important assets (Daconta et al., 2003).

But the Semantic Web goes beyond that enterprise's benefits, his goal is that knowledge can reach all and use in the best possible way the computing resources; economizing the Web and allowing it to find useful information without being redundant which is a goal that will require hard work and commitment of the entire community, although the benefits that would bring are incalculable (Daconta et al., 2003).

2.2 Annotation

They are a source of information that can be captured in comments, notes, explanations referring to a document or part of a document. They can be considered external type if the do not modify the document or internal if they do. Conceptually, annotations are considered as metadata which we provide with information about a piece of data (Meena et al., 2004). People in the academic segment have been using the annotations in books, papers, magazines. with various purposes such as marking information that requires our attention for future reviews, mark sections where additional references are needed to understand its content, highlighting the most important text, annotate any idea regarding what they read (Wolfe, 2000). Annotations can be used to manage the content that is in the Web pages, but not all of them are useful, for this we need to have a level of formality. Following this approach we can classify the annotations in formal and informal annotations. Formal annotations have a level of formality that ensures interoperability among different agents. Theoretically these annotations are more apt to be interpreted in the same way by different consultation mechanisms, an example of this type of annotation is a metadata would following specific standards in structure and assigned their values using conventional authorized names. On the other hand, informal annotations would become notes or annotations that you write in a book or article while you read; these notes may have different utilities such as reminders, quotes, reviews (Marshall, 1998).

2.3 Ontologies in Annotation

According to Gruber, to define what an ontology is, we must first understand the meaning of conceptualization. Conceptualization can be defined as an abstract representation of a world we want to represent, namely representing existing objects or concepts in certain areas and the relationships between them. Therefore ontologies would become an explicit specification of a conceptualization, namely in a formal way (Gruber, 1995). For artificial intelligence ontologies refers to a specialized vocabulary for a certain domain of knowledge. Language could be changed without affecting the ontology conceptualization. Identify vocabulary and conceptualizations requires a thorough analysis of the types of objects and relations of their domain (Studer et al., 2000). Being able to manage a clear definition whatever the vocabulary or who is using it is one of the reasons why ontologies are becoming very popular in KR (Davies et al., 2003). To the Semantic Web, ontologies are important to support the information seeking in the delimit the domain searches and reach sources that are actually useful for the query being performed. They also help in the reuse and classification of information and to be able to handle concepts in a clearer manner regardless of the source or where agents come.

2.4 Metadata

Metadata is information about information, a part of a secondary information refers to primary resource. Examples of metadata include schema, integrity constraints, comments on the data, ontologies, quality parameters, comments, notes, sources and security policies (Srivastava and Velegrakis, 2007). In information management, metadata is very useful to clarify the information meanings, to prevent misunderstandings and facilitate their handling and extraction. Another aspect that favors their use is that they can be added to a variety of documents on the Web, on our computers, on physical books. Also it can be expressed in many languages and vocabularies also be available in both hard and electronic (Corcho, 2006). It offers great advantages in KR in the Web as providing formalization to the contents of the annotated documents for facilitate their searching and sorting, also emphasize that the metadata are very flexible tools that can be easily understood by humans and by machines (Agosti and Ferro, 2007). This flexibility and simplicity in their performance favors the use of annotation metadata using ontologies. They used together are especially useful in the semantic annotation because they are easy to understand, simple to build and maintain, and is easy to reach a consensus on the information provided (Uren et al., 2006).

3 RELATED WORKS

Publications related to semantic annotation tools include the use of NLP for the treatment of various information sources on the web (Joksimovic et al., 2013). This used APIs for processing plain text and then proceed to their respective semantic annotation using a specific knowledge base. In (Chechev et al., 2012) the API used was Gate an open-source framework for NLP, but for reasons of language's corpus, a library that works best with the Spanish language will be used. Respect to use of ontologies in (Pipitone and Pirrone, 2010), using upper-level ontologies for realization of semantic annotations is recommended. The ontology will serve us to infer the semantic meaning in previously processed texts. And as for the identification of the correct meaning for texts analyzed in (Hotho et al., 2003), different strategies disambiguation of terms which make use of ontologies and may be useful in conjunction with other knowledge basis.

4 SEMANTIC ANNOTATION TOOL

The proposed tool, seeks to facilitate the management of academic papers produced at the university through semantic annotations and ontologies. These documents are in Spanish and mostly in PDF format. The structure of the tool consists of 6 components (see figure 1). Below is a brief description of each component.

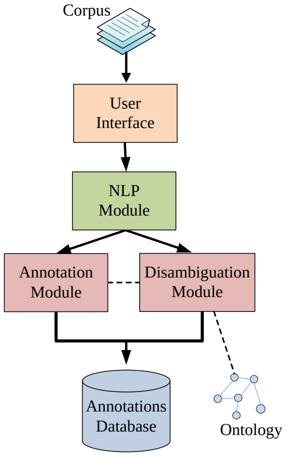

Figure 1: Architecture of the Tool.

4.1 User Interface

It is basically a simple user interface which allows load the ontology to be used, the documents that make up the corpus will process and the connection information for the database where the annotations will be saved. The interaction between the user and the tool will be minimal, because the annotations will be made automatically.

4.2 NLP Module

This module is in charge of the first processing performed of the corpus to be annotated. His first task is to transform the contents of the corpus in plain text to facilitate their treatment. Then, it will produce a list of terms for each of the documents in the corpus. To make this list, the plaintext obtained tokenization process, separation of prayers and part-of-speech tagging is submitted. The library that perform these processes must have support for the Spanish language since the NLP mechanisms vary depending on the language being analyzed. The terms obtained will be related to the document from which they were extracted and will work both with them and with their lemmas for easy identification with the concepts in the ontology.

4.3 Ontology

The structure of the ontology to be used is related to the curricula of courses in college, this will allow them to be used for different fields of study but in this paper we will limit the field of computer science. The ontology consists of 6 classes (see figure 2) according to the organizational structure of the subjects taught in college. These classes are:

- **Concept.** The most basic kind of ontology represents all concepts pertaining to courses.
- **Learning Unit.** Represents a specific topic containing a set of concepts.
- **Program of the Course.** It consists of learning units and represents the program of a course in a given period of time.
- **Course.** Represents the courses at the university, is composed of units of learning, for example: `Programming Languages I`,`Fundamentals of Programming`.
- **Department.** Represents the department that dictates the respective course.
- **Faculty.** Is composed of a set of departments.

Each of these classes contains the property `Has` which indicates that it contains another class of lower rank. Also for the `Concept` class have another property called terms which contains explicit representations of it.

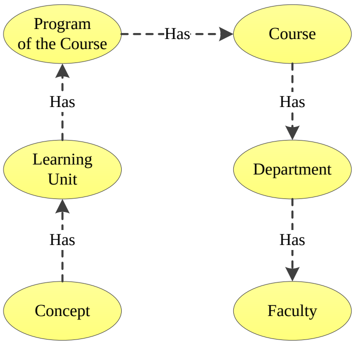

Figure 2: Hierarchy of the Ontology.

4.4 Disambiguation Module

This module would become the core of the tool, it will link the terms obtained in the previous module with the concepts of ontology. To make this task, it use libraries to navigate between classes in the ontology. And to choose the right concepts, one of the disambiguation strategy described in (Hotho et al., 2003) were applied, it called disambiguation by context. This strategy helps to define the correct concept of a term according to a vicinity semantic concepts (see figure 3). The process begins by finding the possible concepts of the term under review (would each term extracted from the documents of the initial corpus) in the ontology, then take as its vicinity concepts belonging to the learning unit, analytical program, course and faculty according to the unit Learning to which it belongs. Finally using the property in terms of concepts, be checked if the context of the document where the term is obtained coincides with the context belonging to the learning unit and course concept. The concept chosen is determined by a merit function based on the terms of matching. This process is repeated for every term from the corpus and the output will be a of concepts belong to the ontology.

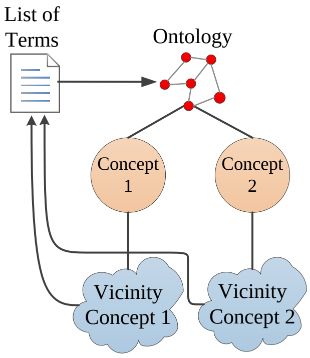

Figure 3: Disambiguation Process.

4.5 Annotation Module

This module is responsible for assembling the annotations based on the terms and concepts linked in the phase of disambiguation. Those annotations will be in RDF format and contain information on the concept of the term, the learning unit concept,the ontology on which was built and the document to which it belongs.

4.6 Annotations Database

The last proposed module is responsible for the persistence of annotations made on processed corpus. They will be in a relational database which can be used for queries on semantic annotation of documents produced in the University. This technology is chosen because it is easy to use and it would be difficult to transform the metadata stored in it whether it is in RDF format or another markup language.

5 TOOL IMPLEMENTATION

In the implementation of the tool we use open-source resources in general. Java was used to create the interface and for the interaction between the libraries used. The library Apache Tika[1] was used to trans-

[1] Apache Tika

form the content of the corpus to plain text. When already has the plain text, is subjected to corresponding NLP tasks, to do that we use the Freeling(Padro and Stanilovsky, 2012)[2] processing library because it has excellent results in the analysis contained in the Spanish language. With the help of Freeling were able to extract the terms of the corpus, which were limited to nouns and adjectives to facilitate their relationship with the concepts in the ontology, to have that rule the extraction of terms is not limited to the domain of knowledge ontology used.

Regarding the creation of ontology, Protégé(Jain and Singh, 2013)[3] tool was used because it was easy to use and it have a lot of documentation. The interaction of ontology with the other components of the tool was performed using the Jena[4] library and its engine SPARQL(Pérez et al., 2006) for query language with which navigate in the ontology to find concepts to assign to the terms. Finally the annotations will be stored in RDF format in a relational database, in this case a mysql engine was used.

Figure 4-1 shows the main classes including: `Concepto` (concept), `Curso` (course), `Especialidad` (academic units), `Facultad` (faculty), `Programa Analitico` (syllabus) and `Unidad Aprendizaje` (learning unit). Because the ontology aims query expansion, we added the properties lemma, preferred name and synonyms for all classes (see figure 4-3). For example the `Archivos` (files in english) class, has *archivos* (plural form of file in Spanish) as a preferred name, *archivo* (singular form of file in Spanish) as lemma and *fichero* (synonym of file in Spanish) as a synonym.

The object properties can be seen in the figure 4-2. The main property is `tieneConcepto` (haveConcept). This property, associates learning units with certain concepts in the computer science domain. Through this relationship is possible to perform QE. The other properties allow linking other concepts. Learning units are part of syllabus which in turn are made for a specific course. The courses belong to one academic unit that make up a specific faculty.

6 RESULTS

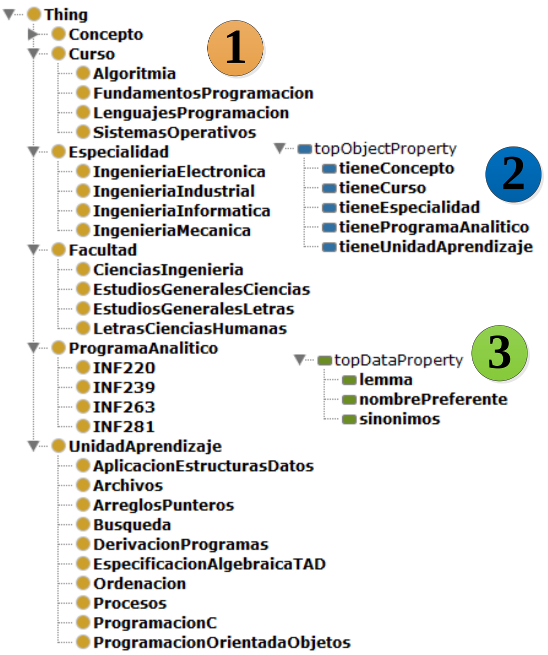

Figure 4: Ontology for CC Curricula.

Tool tests were conducted with a corpus composed of 20 documents produced at the university in the faculty of computer engineering. The processing of these present some complications because some of these were contained in Spanish and English, even some words like `shell` and `void` is often used along with the rest of the content in Spanish because they have no exact translations in academic context. The issue of language influences the efficiency of the phase of NLP, because when English words are analyzed, inferring that are in Spanish, you can take these as nouns which would render the remaining tasks of the tool those terms. However, queries using ontology facilitate the debugging process much of terms thanks to the efficiency of the engine used in SPARQL queries. To cite a few examples, in Spanish there are words with different meanings like *DERIVACIÓN* (in English derivation), which is related to grammar, mathematics and algorithms. The efficiency of the tool will measure based on the values of precision and recall on concepts that could be identified in the corpus. Their values are calculated based on the number of retrieved concepts that are semantically to the term, the recovered concepts that do not correspond and concepts that could not be retrieved from the corpus. The concepts recovered in the corpus of 20 papers were 133, with a value of 81 of precision and 86 of recall . The texts used were taken to test students in college.

7 CONCLUSION

About development of the tool we can conclude that his accuracy depends on the efficiency of the libraries

[2] http://nlp.lsi.upc.edu/freeling/
[3] http://protege.stanford.edu/
[4] https://jena.apache.org/

used in the phase of NLP as well as the strategy disambiguation of words used when choosing the best concept to translate into an annotation. Also the language adds a bit of difficulty, which when we are working in languages like English have more resources than when we work in Spanish. It is also worth noting the structure of the ontology which allows its extension to other subjects or areas of study within the university. Finally, the proposed architecture is designed so that we can use other resources both to analyze the corpus as the creation and interaction of other sources of knowledge than an ontology.

8 FUTURE WORKS

In the development of this tool, we focus on an ontology of the domain of computer science but domain knowledge used can vary, as the language of the corpus we process. The architecture of this tool is designed to work with any type of ontologies and other strategies disambiguation of words to help us perform automatic annotations. Could improve tool performance enhancing NLP phase and testing new strategies disambiguation of words.

REFERENCES

Agosti, M. and Ferro, N. (2007). A formal model of annotations of digital content. *ACM Transactions on Information Systems (TOIS)*, 26(1):3.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

Chechev, M., Gonzàlez, M., Màrquez, L., and España-Bonet, C. (2012). The patents retrieval prototype in the molto project. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 231–234. ACM.

Corcho, O. (2006). Ontology based document annotation: trends and open research problems. *International Journal of Metadata, Semantics and Ontologies*, 1(1):47–57.

Daconta, M. C., Obrst, L. J., and Smith, K. T. (2003). *The semantic web: a guide to the future of XML, web services, and knowledge management*. John Wiley & Sons.

Davies, J., Fensel, D., and Van Harmelen, F. (2003). Towards the semantic web. *Ontology-Driven Knowledge Management. Chichester*.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5):907–928.

Hotho, A., Staab, S., and Stumme, G. (2003). Ontologies improve text document clustering. In *Third IEEE International Conference on Data Mining*, pages 541–544. IEEE.

Jain, V. and Singh, M. (2013). Ontology development and query retrieval using protégé tool. *International Journal of Intelligent Systems and Applications (IJISA)*, 5(9):67.

Joksimovic, S., Jovanovic, J., Gasevic, D., Zouaq, A., and Jeremic, Z. (2013). An empirical evaluation of ontology-based semantic annotators. In *Proceedings of the seventh international conference on Knowledge capture*, pages 109–112. ACM.

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, pages 40–49. ACM.

Meena, E., Kumar, A., and Romary, L. (2004). An extensible framework for efficient document management using rdf and owl. In *Proceeedings of the Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology*, pages 51–58. Association for Computational Linguistics.

Padro, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Pérez, J., Arenas, M., and Gutierrez, C. (2006). Semantics and complexity of sparql. In *The Semantic Web-ISWC 2006*, pages 30–43. Springer.

Pipitone, A. and Pirrone, R. (2010). A framework for automatic semantic annotation of wikipedia articles. In *6th Workshop on Semantic Web Applications and Perspectives*.

Srivastava, D. and Velegrakis, Y. (2007). Intensional associations between data and metadata. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 401–412. ACM.

Studer, R., Decker, S., Fensel, D., and Staab, S. (2000). Situation and perspective of knowledge engineering. *Knowledge Engineering and Agent Technology*, pages 237–252IOS.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: science, services and agents on the World Wide Web*, 4(1):14–28.

Wolfe, J. L. (2000). Effects of annotations on student readers and writers. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 19–26. ACM.