

OntoTag's Linguistic Ontologies as a Reference for Semantic Web Annotations

Guadalupe Aguado de Cea
DLACT, UPM
lupe@fi.upm.es

Inmaculada Álvarez de Mon
DLACT, UPM
ialvarez@euitt.upm.es

Antonio Pareja-Lora
SIP, UCM / LIA, UPM
apareja@sip.ucm.es

Campus de Montegancedo, s/n. 28660 – Boadilla del Monte (Spain)

Tel: (+34) 91 336 74 13 Fax: (+34) 91 336 54 72

Abstract

Following the road in-between purely linguistic annotation and solely ontology-based annotations for the Semantic Web, a hybrid (ontological and linguistic) model and platform, called OntoTag, has been created, aiming at better machine communication, interoperability and language understanding; these capabilities are derived from the incorporation into the platform of a set of linguistic ontologies, the main topic of this demonstration, which are the main referent for the generation of multi-leveled and standardized annotations of Semantic Web documents within OntoTag.

Introduction

Many are the schemas developed so far for the different kinds of annotation required in the field of **Corpus Annotation**. Besides, with the appearance of the **Semantic Web** (Berners-Lee *et al.*, 1999) many other schemas have been devised (most of them based on *ontologies* (Gruber 1993; Borst 1997)) for *web page annotation*. Thus far, on the one hand, Corpus Linguistics researchers are trying to cover as many levels and aspects of annotation –from a linguistic point of view– as possible to describe language phenomena (Wilson & Thomas, 1997; Schmidt, 1988); on the other hand, researchers in the Semantic Web area are focusing on achieving a sound model of semantic annotation for web pages, that is able to capture as much knowledge from these pages as possible, so that computers can process them in a much smarter way (Benjamins *et al.*, 1999, Motta *et al.*, 1999, Luke *et al.*, 2000, Staab *et al.*, 2000). However, there is an emerging road in-between, nowadays, that seeks to merge and sum up both kinds of annotations, combining them in order to bear a new, unified, multilingual, flexible, extensible and fully semantic model of annotation, useful for both communities (Aguado *et al.*, 2003a). Moreover, as shown by the ISO - TC37SC4 (2003) “there is an increasing need for new standardization as well as urgent recognition of existing *de facto* standards and their transformation into International Standards”. In fact, one of the main aims of this committee is “to develop standards and related documents to maximize the applicability of language resources”. The OntoTag model for Semantic Web Annotation (Aguado *et al.*, 2003b), whose Linguistic Ontologies we present here, is being developed following this in-between road aforementioned, as well as a number of guidelines hitherto published (EAGLES 1996a, 1996b; CES 1999; MILE 2003; GDA 2002), in order to achieve the goal of standardisation sought within the ISO - TC37SC4 committee.

OntoTag's Linguistic Ontologies

One of the main components of the OntoTag model is its set of linguistic ontologies, devised to represent the

structure and relationships between the elements of language at different linguistic levels. The kind of elements and relationships considered in them are the ones usually included in existing annotation schemas and also those already discussed in the literature but not implemented yet (Wilson & Thomas, 1997; Schmidt, 1988) as well as some others, determined by our research team.

OntoTag's Core Linguistic Ontologies

First of all, a Linguistic Level Ontology (LLO) has been implemented both to capture the stratification of natural language analysis and generation and to simplify the study of the other elements. Then, following the EAGLES guidelines for morpho-syntactic annotation of corpora (EAGLES 1996a), but obviously broadening its scope, three different ontologies have been implemented to represent the category-attribute-value formalism at all levels of annotation (morpho-syntactic, syntactic, semantic, discourse and pragmatic): a Linguistic Unit Ontology (LUO), a Linguistic Attribute Ontology (LAO), and a Linguistic Value Ontology (LVO).

The Linguistic Unit Ontology (LUO) includes all the units (categories) identified at the different levels of annotation considered in the LLO, and incorporates an adaptation of the SIMPLE (2000) ontologies at the semantic level; the Linguistic Attribute Ontology (LAO) includes the various attributes associated to the units in the LUO; and the Linguistic Value Ontology (LVO) accounts for the possible values of the attributes in the LAO.

OntoTag's Supplementary Linguistic Ontologies

Complementing these four ontologies, a fifth one (the Linguistic Pattern Ontology, LPO) has been designed for the representation of the patterns that these units follow when combined in an utterance. Finally, the OntoTag Integration Ontology (OIO) establishes the main relationships between documents (annotated and non-annotated), units, attributes and values both in the linguistic and in the ontological areas of annotation.

OntoTag's Linguistic Ontologies: Application

The application of these six ontologies in the OntoTag annotation model is twofold: first, as discussed above, they identify the different elements (mostly linguistic, but also ontological) that are annotable in the Semantic Web field; second, once the ontology has been populated (instantiated) by the annotations obtained with OntoTag, they will also act as a repository or database of these annotations.

Further information about OntoTag's linguistic ontologies, their respective roles and interaction, as well as their properties and application (to pragmatic purposes or with automatic means, for instance) can be found in Aguado *et al.* (2004a; 2004b).

Conclusions

To conclude, we could say that, derived from the extensibility and flexibility capabilities of the Linguistic Ontologies presented here, the OntoTag model of annotation inherits these properties as well. It can also be considered as domain independent in the sense that these source ontologies can be replaced and, still, meaningful annotations would be obtained. Due to the multilingual nature of the EAGLES guidelines, followed (and broadened) in the design of the different Linguistic Ontologies, OntoTag becomes also applicable to the annotation of the languages studied in these guidelines. The consensual nature of ontologies and the sources used in their construction (EAGLES 1996a, 1996b; CES 1999; MILE 2003; GDA 2002; Dubuc & Lauriston 1997; Faber & Tercedor 2000; Mel'cuk 1996, 1988; Pustejovsky 1998) enables them (and the annotations obtained with them) so as to be considered standardised.

Acknowledgements

This research has partly been supported by the ministry of Science and Technology grant (Reference TIC2001-2745 CONTENTWEB project) and by the UPM grant (Reference 14286 PLAN-H-SEMWEB project)

References

- Aguado de Cea, G., Álvarez de Mon, I., Pareja-Lora, A. 2003a. "Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de Web Semántica: OntoTag" in *Revista Iberoamericana de Inteligencia Artificial*, Vol 1, pp 37-49.
- Aguado de Cea, G., Álvarez de Mon, I., Gómez-Pérez, A., Pareja-Lora, A. 2003b. "OntoTag: XML/RDF(S)/OWL Semantic Web Page Annotation in ContentWeb" in *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2003) - Language Technology and the Semantic Web*, pp. 25-32. 10th Conference of the European Chapter of the Association for Computational Linguistics. EACL'03. Budapest, Hungary.
- Aguado de Cea, G., Álvarez de Mon, I., Gómez-Pérez, A., Pareja-Lora, A. 2004a. "OntoTag's Linguistic Ontologies: Improving Semantic Web Annotations for a Better Language Understanding in Machines" in *Proceedings of the International Conference on Information Technology (ITCC 2004)*, pp. 124-128. IEEE Computer Society: Washington, Brussels, Tokyo.
- Aguado de Cea, G., Álvarez de Mon, I., Pareja-Lora, A. 2003b. "OntoTag's Linguistic Ontologies: Enhancing Higher Level and Semantic Web Annotations" in *Proceedings of the 4th Language Resources and Evaluation Conference (LREC 2004)*. Lisbon, Portugal.
- Álvarez de Mon-Rego, I. 2003. La cohesión del texto científico-técnico: un estudio contrastivo inglés-español. Universidad Complutense de Madrid (forthcoming).
- Benjamins, V.R., Fensel, D., Decker, S., Gómez-Pérez, A. 1999. (KA)²: Building Ontologies for the Internet: a Mid Term Report. *IJHCS, International Journal of Human Computer Studies*, 51, pp. 687-712.
- Berners-Lee, T., Fischetti, M. 1999. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper. San Francisco.
- Borst, W. N. 1997. Construction of *Engineering Ontologies*. PhD thesis, University of Twente, Enschede.
- CES. 1999. *Corpus Encoding Standard*. <http://www.cs.vassar.edu/CES/>
- Dubuc, R. and Lauriston, A. 1997. "Terms and Contexts" in Wright, S.E. and G. Budin, *Handbook of Terminology management* Vol 1, John Benjamins: Amsterdam, Philadelphia, pp. 80-87.
- EAGLES. 1996a. *EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG--TCWG--MAC/R.
- EAGLES. 1996b. *EAGLES: Recommendations for the Syntactic Annotation of Corpora*. EAGLES Document EAG--TCWG--SASG/1.8.
- Faber, P. and Tercedor, M. 2000. "Codifying conceptual information in descriptive terminology management" in *Meta*, XLVI, 1, pp. 192-204.
- GDA. 2002. Global Document Annotation Initiative: The GDA Tag Set. <http://www.i-content.org/GDA/tagset.html>
- Gruber, T. R. 1993. "A Translation Approach to Portable Ontologies" in *Journal on Knowledge Acquisition*, Vol. 5(2), 199-220
- ISLE. 2003. http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm
- ISO - TC37SC4. 2003. <http://www.tc37sc4.org>
- Luke S., Heflin J. 2000. *SHOE 1.01. Proposed Specification*. SHOE Project. <http://www.cs.umd.edu/projects/plus/SHOE/spec1.01.htm>
- Motta, E., Buckingham Shum, S. Domingue, J. 1999. "Case Studies in Ontology-Driven Document Enrichment" in *Proceedings of the 12th Banff Knowledge Acquisition Workshop*, Banff, Alberta, Canada.
- Mel'cuk, I. A. 1988. *Dependency Syntax*, New York: State University of New York Press.

- Mel'čuk, I. A. 1996. "Lexical functions: a tool for the description of lexical relations in a lexicon", in Wanner, L. *Lexical functions in lexicography and natural language processing*, John Benjamins: Amsterdam, Philadelphia.
- Pustejovsky, J. 1998. *The generative lexicon*, MIT Press: Cambridge, Massachusetts.
- Schmidt, K. M. 1988. Der Beitrag der begriffsorientierten Lexicographie zur systematischen Erfassung von Sprachwandel und das Begriffswörterbuch zur mhd. Epik. *Mittelhochdeutsches Wörterbuch in der Diskussion*, ed. by Bachofer, W. Max Niemeyer: Tübingen, 35–49.
- SIMPLE Project. 2000. <http://www.ub.es/gilcub/SIMPLE/simple.html>
- Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Mädche, A., Schnurr, H.-P., Studer, R. 2000. Semantic Community Web Portals. *WWW'9*. Amsterdam.
- Wilson, A., Thomas, J. 1997. Semantic Annotation. *Corpus Annotation: Linguistic Information from Computer Text Corpora*, R. Garside, G. Leech & A. M. McEnery (ed.). Longman: London.