Cast3LB-CoNNL-SemRol: A Seed Corpus for Machine Learning Experiments

Roser Morante[1]

Abstract

In this paper we will present the annotation scheme used to annotate the Cast3LB–CoNLL–SemRol Corpus of Spanish with semantic roles, and the application of the corpus in machinelearning experiments. The annotation has been carried out as part of the project *T´ecnicas semiautom´aticas para el etiquetado de roles sem´anticos en corpus del espa˜nol*, which focuses on researching semiautomatic techniques for semantic role labelling. The corpus is used in the project as seed corpus to train a semantic role labeller in order to annotate a bigger corpus of seventy million words using active learning techniques.

The Cast3LB–CoNLL Corpus is a revised version of the converted version of the Cast3LB treebank (Civit, 2003, Navarro et al., 2003) used in the CoNLL Shared Task 2006 (Buchholz and Marsi, 2006) as train corpus. The corpus contains 89199 words in 3303 sentences. As for verbs it contains 11023 forms, and 1443 lemmas. Like in the CoNLL Shared Task 2006 a sentence consists of tokens, each one starting on a new line. A token consists of 9 fields which contain information about morphosyntactic features, dependencies and semantic roles. The set of semantic roles that has been used is the following one: ARG0, ARG1, ARGM, Atributive, Benefactive, Cause, Company, Concessive, Condition, Consequence, Destination, Extent, Instrument, Location, Manner, Means, Opposition, Origin, Predicative, Purpose, Quantity, Result, Source, State, Temporal, Topic.

The main reference projects for semantic role labeling are FrameNet (Fillmore et al., 2003, Johnson et al., 2002) and PropBank (Palmer and Gildea, 2005), both for the English language. In the paper we will elaborate on the characteristics of our annotation as compared to the annotation in these projects. The goal of the annotation of the Cast3LB–CoNLL corpus is to generalise as much as possible the semantic relation that holds between a predicate (only verbs for now) and its complements, in order to have as few roles as possible, since increasing the number of roles increases also the complexity of the classification task for the automatic semantic role labeller.

The Cast3LB–CoNLL–SemRol seed corpus has been used to train two dependency parsers (Canisius et al., 2006, Nivre et al., 2006) and a semantic role labeller based on TiMBL (Daelemans and van den Bosch, 2005). The paper will put forward the initial results of the experiments performed to annotate a big corpus starting from a small seed corpus, and will extract conclusions about the relation between the annotation scheme used in the seed corpus and the results of the experiments.

[1] *e-mail*: R.Morante@uvt.nl

References

S. Buchholz and E. Marsi. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the X CoNLL Shared Task*. SIGNLL, 2006.

S. Canisius, T. Bogers, A. van den Bosch, J. Geertzen and E. Tjong Kim Sang. Dependency parsing by inference over high-recall dependency predictions. In *Proceedings of the Tenth Conference on Computational Natural Language Learning, CoNLL-X*, pages 3–8, New York City, NY, 2006.

M. Civit. Guía para la anotación sintáctica de Cast3LB: un corpus del español con anotación sintáctica, semántica y pragmática. X-TRACT-II WP-03-06 y 3LB-WP-02-01, CliC–UB, 2003.

W. Daelemans and A. van den Bosch. *Memory-based language processing*. Cambridge University Press, Cambridge, UK, 2005.

Ch. Fillmore, Ch. Johnson and M. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250, 2003.

Ch. Johnson, Ch. Fillmore, M. Petruck C. Baker, M. Ellsworth, J. Ruppenhofer and E. Wood.

FrameNet: Theory and practice. Techinical report, ICSI, Berkeley, 2002.

B. Navarro, M. Civit, M.A. Martí, R. Marcos and B. Fernández. Syntactic, semantic and pragmatic annotation in cast3lb. In *Proceddigns of the Shallow Processing of Large Corpora (SPro-LaC) Workshop of Corpus Linguistics 2003*, Lancaster,UK, 2003.

J. Nivre, J. Hall, J. Nilsson, G. Eryigit and S. Marinov. Labeled pseudo–projective dependency parsing with support vector machines. In *Proceedings of the Tenth Conference on Computational Natural Language Learning, CoNLL-X*, New York City, NY, June 2006.

M. Palmer and D. Gildea. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105, 2005.