On the Development of the RST Spanish Treebank

Iria da Cunha
Institute for Applied Linguistics (UPF), Spain
Instituto de Ingeniería (UNAM), Mexico
Laboratoire Informatique d'Avignon (UAPV), France
iria.dacunha@upf.edu

Juan-Manuel Torres-Moreno
Laboratoire Informatique d'Avignon (UAPV), France
Instituto de Ingeniería (UNAM), Mexico
École Polytechnique de Montréal, Canada
juan-manuel.torres@univ-avignon.fr

Gerardo Sierra
Instituto de Ingeniería (UNAM), Mexico
gsierram@iingen.unam.mx

Abstract

In this article we present the RST Spanish Treebank, the first corpus annotated with rhetorical relations for this language. We describe the characteristics of the corpus, the annotation criteria, the annotation procedure, the inter-annotator agreement, and other related aspects. Moreover, we show the interface that we have developed to carry out searches over the corpus' annotated texts.

1 Introduction

The Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a language independent theory based on the idea that a text can be segmented into Elementary Discourse Units (EDUs) linked by means of nucleus-satellite or multinuclear rhetorical relations. In the first case, the satellite gives additional information about the other one, the nucleus, on which it depends (ex. Result, Condition, Elaboration or Concession). In the second case, several elements, all nuclei, are connected at the same level, that is, there are no elements dependent on others and they all have the same importance with regard to the intentions of the author of the text (ex. Contrast, List, Joint or Sequence). The rhetorical analysis of a text by means of RST includes 3 phases: segmentation, detection of relations and building of hierarchical rhetorical trees. For more information about RST we recommend the original article of Mann and Thompson (1988), the web site of RST[1] and the RST review by Taboada and Mann (2006a).

RST has been used to develop several applications, like automatic summarization, information extraction (IE), text generation, question-answering, automatic translation, etc. (Taboada and Mann, 2006b). Nevertheless, most of these works have been developed for English, German or Portuguese. This is due to the fact that at present corpora annotated with RST relations are available only for these languages (for English: Carlson et al., 2002, Taboada and Renkema, 2008; for German: Stede, 2004; for Portuguese: Pardo et al., 2008) and there are automatic RST parsers for two of them (for English: Marcu, 2000; for Portuguese: Pardo et al., 2008) or automatic RST segmenters (for English: Tofiloski et al., 2009). Scientific community working on RST applied to Spanish is very small. For example, Bouayad-Agha et al. (2006) apply RST to text generation in several languages, Spanish among them. Da Cunha et al. (2007) develop a summarization system for medical texts in Spanish based on RST. Da Cunha and Iruskieta (2010) perform a contrastive analysis of Spanish and Basque texts. Romera (2004) analyzes coherence relations by means of RST in spoken Spanish. Taboada (2004) applies RST to analyze the resources used by speakers to elaborate conversations in English and Spanish.

We consider that it is necessary to build a Spanish corpus annotated by means of RST. This corpus should be useful for the development of a rhetorical parser for this language and several other applications related to computational linguistics, like those developed for other languages

[1] http://www.sfu.ca/rst/index.html

Proceedings of the Fifth Law Workshop (LAW V), pages 1–10,
Portland, Oregon, 23-24 June 2011.