Iria da Cunha

Towards Discourse Parsing in Spanish

Universidad Nacional de Educación a Distancia (UNED), Spain
iria.dacunha@upf.edu

Introduction

Texts can be analysed from different perspectives. One of the most difficult phenomena to process is discourse structure (Hovy 2010). In recent years, one of the main challenges in the field of Natural Language Processing (NLP) has been discourse parsing. Research on this topic has been done for several languages, such as Japanese (Sumita et al. 1992), English (Marcu 2000) and Portuguese (Pardo 2008), among others. Also, for English, the CoNLL-2015 Shared Task focused on Shallow Discourse Parsing.[9] Discourse annotated corpora have been created too, for example for English (Carlson et al. 2002), German (Stede 2004), Portuguese (Pardo 2008) and French (Afantenos 2012). Discourse parsing tools and resources are used to develop NLP applications; for example, automatic summarization, information extraction, text generation, machine translation and sentiment analysis (Taboada & Mann 2004).

The aim of this paper is to present the advances in discourse parsing for Spanish. Specifically, after explaining our theoretical framework, we will detail the tools we have developed for the automatic annotation of discourse information in texts in Spanish and the discourse annotated resources we have created.

Theoretical Framework

Most discourse NLP tools are based on Rhetorical Structure Theory (RST, Mann & Thompson 1988). This is a language independent theory based on the idea that a text can be segmented into Elementary Discourse Units (EDUs)

[9] http://www.aclweb.org/anthology/K/K15/

linked by means of nucleus-satellite or multinuclear discourse relations. In the first case, the satellite gives additional information about the other one, the nucleus, on which it depends (e.g. Result 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 text author (e.g. Contrast or Sequence). RST discourse parsing includes three stages: *a)* segmentation, *b)* relations detection and *c)* building of hierarchical rhetorical trees.

Discourse Tools and Resources for Spanish

In this section we explain the discourse tools and resources we have developed for Spanish, in the framework of RST. First, we have developed the discourse segmenter DiSeg (da Cunha et al. 2012), which can be used online.[10] It is based on shallow parsing and a set of linguistic rules that insert segment boundaries into sentences, following specific criteria.[11] DiSeg performance was evaluated using a corpus of manually annotated texts (a gold standard).[12] The system obtained an F-score between 80% and 96% in experiments with a corpus containing medical texts, and an F-Score of 91% with a corpus of texts about terminology.

Second, we have developed a discourse corpus containing texts manually annotated, the RST Spanish Treebank (da Cunha et al. 2011), which can be consulted and downloaded online.[13] The texts have been annotated with the RST-Tool (O'Donnell 2000). The corpus includes 267 specialised texts (from several domains and genres), 52,746 words, 2,256 sentences and 3,349 discourse segments. It is divided into a learning corpus (183 texts) and a test corpus (84 texts).

Third, we have developed a sentence-level discourse parser, DiSeg2 (da Cunha et al. 2012a), which can also be consulted online.[14] To do this, we have analysed the learning corpus of the RST Spanish Treebank in order to manually detect all the markers that show discourse relations. We divided the markers into 3 categories: *1)* traditional discourse markers, *2)* markers including lexical units

[10] http://dev.termwatch.es/esj/DiSeg/WebDiSeg/
[11] Similar to the ones used in: da Cunha & Iruskieta 2010.
[12] http://dev.termwatch.es/esj/DiSeg/index.html
[13] http://corpus.iingen.unam.mx/rst/
[14] http://diseg2.termwatch.es/

(nouns and verbs), and *3)* markers including verbal structures. We obtained 778 markers. Taking these markers into account, we have designed an algorithm to automatically detect intra-sentence RST relations and nuclearity. It is based on linguistic rules including discourse patterns and the aforementioned discourse segmenter. We have evaluated the system with the test corpus, obtaining an accuracy of 81.75 regarding EDUs, SPANs (that is, sets of EDUs) and nuclearity, and 81.75 with regard to relations.

Fourth, we have created DiZer 2.0 (Maziero et al. 2011), an adaptable online platform designed to develop discourse parsers in any language, which integrates a language-independent algorithm to build discourse trees (Marcu 2000). In order to automatically obtain hierarchical rhetorical trees from full texts in Spanish, we have included our discourse segmenter and patterns in this platform. Currently, we are evaluating the performance of this discourse parser for Spanish.

Conclusions and Future Work

The aim of this paper has been to show the main automatic tools and resources related to discourse parsing for Spanish: the discourse segmenter, the RST Spanish Treebank, the sentence-level discourse parser, and the platform to build rhetorical trees. As future work, we plan to evaluate the complete discourse parser and to develop several NLP applications. Also, we plan to research about the cross-linguistic applicability of these tools.[15]

References

- Afantenos, S., Asher, N. & Benamara, F. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In: *Proceedings of the 8th Conference of LREC*, 2012.
- Carlson, L., Marcu, D. & Okurowski, M. E. 2002. *RST Discourse Treebank.* Pennsylvania: Linguistic Data Consortium.

[15] This work has been partially supported by a Ramón y Cajal research contract (RYC-2014-16935) and the research project APLE 2 (FFI2009-12188-C05-01) of the Institute for Applied Linguistics (IULA).

- da Cunha, I. & Iruskieta, M. 2010. Comparing rhetorical structures of different languages: The influence of translation strategies. *Discourse Studies* 12 (5): 563–598.
- da Cunha, I., Torres-Moreno, J-M., Sierra, G. 2011. On the Development of the RST Spanish Treebank. In: *Proceedings of the 5th Linguistic Annotation Workshop (ACL 2011)*. 1–10.
- da Cunha, I., SanJuan, E., Torres-Moreno, J-M., Cabré, M. T. & Sierra, G. 2012a. A Symbolic Approach for Automatic Detection of Nuclearity and Rhetorical Relations among Intra-sentence Discourse Segments in Spanish. *LNCS* 7181: 462–474.
- da Cunha, I., SanJuan, E., Torres-Moreno, J-M., Lloberes, M. & Castellón, I. 2012b. DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications (ESWA)* 39 (2): 1671–1678.
- Feng, V. W. & hirst, G. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In: *Proceedings of the 52nd Annual Meeting of the ACL (2014)*. 511–521.
- Hovy, E. 2010. Annotation. A Tutorial. *Presented at the 48th Annual Meeting of the ACL*.
- Joty, Sh., Carenini, G. NG, R. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics* 41 (3): 385–435.
- Mann, W. C. & Thompson, S. A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8 (3): 243–281.
- Marcu, D. 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach, *Computational Linguistics* 26 (3): 395–448.
- Maziero, E. G. & Pardo, Th. A. S., da Cunha, I., Torres-Moreno, J-M. & San-Juan, E. 2011. DiZer 2.0 – An Adaptable On-line Discourse Parser. In: *Proceedings of the III RST Meeting (STIL 2011)*.
- O'Donnell, M. 2000. RSTTOOL 2.4 – A markup tool for rhetorical structure theory. In: *Proceedings of the International Natural Language Generation Conference (2000)*. 253–256.
- Pardo, Th. A.S. & Nunes, M. G. V. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing* 15(2): 43–64.
- Soricut, R. & Marcu, D. 2003. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In: *Proceedings of 2003 HLT and North American ACL Conference (2003)* 149–156.
- Stede, M. 2004. The Potsdam commentary corpus. In: *Proceedings of the Workshop on Discourse Annotation (ACL 2004)*.

- Subba, R. & Di Eugenio, B. 2009. An effective discourse parser that uses rich linguistic information. In: *Proceedings of 2009 HLT and North American ACL Conference (2009).* 566–574.
- Sumita, K., Ono, K., Chino, T., Ukita, T. & Amano, Sh. 1992. A discourse structure analyzer for Japanese text. In: *Proceedings of the International Conference on Fifth Generation Computer Systems 2.* 1133–1140.
- Taboada, M. & Mann, W. C. 2006. Applications of rhetorical structure theory. *Discourse Studies* 8 (4): 567–588.