DiZer 2.0 – a Web Interface for Discourse Parsing

Erick Galani Maziero, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400. P.O.Box. 668. 13560-970 - São Carlos/SP, Brazil
{erickgm,taspardo}@icmc.usp.br

Abstract. This paper presents the DiZer 2.0, a freely available web interface for discourse parsing. Different from its first version, it is easier to use and allows easy customization for other text genres, domains and languages.

Keywords: discourse parsing, RST

1 Introduction

Discourse parsing aims at automatically identifying the discourse structure of a text. RST (Rhetorical Structure Theory) [1] has been the most followed discourse model. According to RST, a text may be structured as a sequence of segments/Elementary Discourse Units (EDUs) (usually clauses or sentences) connected by rhetorical relations, forming a tree-like structure. See, e.g., Fig. 1 below, where each segment is numbered. N indicates the segments that are nuclear and, therefore, are more important in the text; S, otherwise, indicates satellite segments, which are considered complementary information.

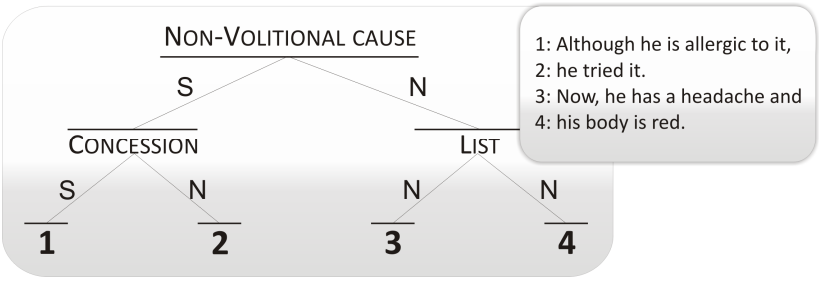

Fig. 1. Example of rhetorical structure

Some discourse parsers exist for English, Portuguese, Spanish, and Japanese, to the best of our knowledge. DiZer [2] is the only known parser for Portuguese. DiZer was originally designed to academic texts and uses a set of discourse patterns (that correlate text markers and discourse relations) to identify text structure. Its main drawbacks are that it is a heavy system (requiring the installation of several tools and resources) and is hard to adapt and port to any other text genre, domain, and language. In order to deal with this, we proposed DiZer 2.0, a web interface that (i) do not require any tool or resource to be locally installed, (ii) is easier to use, and (iii) may be easily adapted to other texts and languages. DiZer 2.0 is introduced in the next section.

2 DiZer 2.0

Fig. 2 shows the steps and resources used in DiZer 2.0. The text may be manually or automatically segmented in order to identify the EDUs. A syntactical parser is used to assign part-of-speech and lemma information to each word. After this, a module carries out the identification of all possible relations between the EDUs using a discourse pattern repository and complementary word lists, and these information are used to generate a DCG (Definite Clause Grammar) grammar that, when executed, produces the possible rhetorical structures of the text. Optionally, statistics (automatically derived from corpus analysis) are used to rank the produced structures.

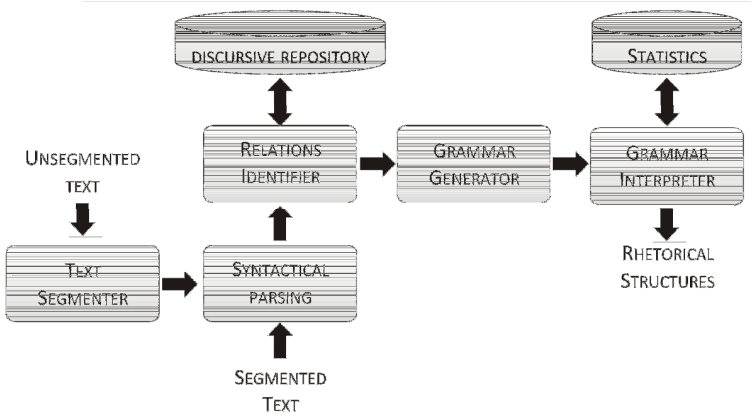

Fig. 2. **DiZer 2.0 architecture**

DiZer 2.0 allows the user to create and edit its own discourse pattern repository and word lists, and to use the system with its native language (nowadays DiZer 2.0 naturally supports Portuguese and Spanish), as well as to use DiZer traditional repository. It is also possible to easily extend DiZer 2.0 to include new parsing and segmentation tools. Also, the system can be adapted to batch processing.

Although DiZer 2.0 is much more flexible and user friendly than the original DiZer, it has a new drawback: processing speed. The system was already time consuming, given that some of its modules had high complexity. Now, due to its web facilities, even more time is required for performing the parsing. We are currently working on this issue.

DiZer 2.0 is freely available for use at www.nilc.icmc.usp.br/dizer2.

References

1. Mann, W.C. and Thompson, S.A. Rhetorical structure theory. Toward a functional theory of text organization. Text, Vol. 8, N. 3, pp. 243-281. (1998)
2. Pardo, T.A.S. and Nunes, M.G.V. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. Journal of Theoretical and Applied Computing, Vol. 15, N. 2, pp. 43-64. (2008)

Acknowledgments. The authors are grateful to FAPESP and CAPES for supporting this work.