A Web Interface for Browsing the CSTNews Corpus

Pedro Paulo Balage Filho, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-Carlense, 400. P.O.Box 668. 13566-590 - São Carlos/SP, Brazil
pedrobalage@gmail.com, taspardo@icmc.usp.br

Abstract. This paper presents a web interface for browsing the CSTNews Corpus, a collection of news texts annotated according to the CST (Cross-document Structure Theory) model.

Keywords: corpus, CST, web interface

1 Introduction

We present in this paper a web interface for browsing the CSTNews corpus, which consists in a collection of texts annotated according the CST (Cross-document Structure Theory) [1] model. This model allows establishing relations between parts of distinct documents that are about the same topic. These relations are useful for several tasks in natural language processing, e.g., multidocument summarization, question answering, and information extraction, among others.

The CSTNews corpus [2] contains 50 Brazilian Portuguese text collections. Each collection has approximately 3 documents on the same subject but from different sources. Each collection is also accompanied by its human summary. The corpus was annotated by 4 computational linguists and produced satisfactory annotation agreement.

2 The Web Interface

The purpose of this work is to develop a web interface for browsing CST relations present in the corpus. For this, we chose a web platform that could be easily shared with all community and could allow the corpus navigation. The operations that can be performed in the interface are: to browse by relation; to visualize the complete corpus; and to download the corpus.

For the first operation (to browse by relation), the user may browse through all CST relations. For each relation, the interface shows the sentences that were annotated with it. Fig. 1 shows a screen dump for the equivalence CST relation.

Fig. 1. Screen dump

The user may also open the texts from where the related sentences were retrieved. For the second operation (corpus visualization), the user may see the entire corpus grouped by collection and by source. By selecting a sentence, the user may view all relations assigned to that sentence. The third operation (download the corpus) allows the user to fully download the corpus. In the second and third operations, it is also possible to visualize the corpus in its original version (without CST annotation), the segmented texts, and their annotated versions (in XML).

The web interface is available at www.nilc.icmc.usp.br/nilc/tools/CSTNews.

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

References

1. Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue. Hong Kong.
2. Aleixo, P. e Pardo, T.A.S. (2008). CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory). Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos-SP, Maio, 12p.