CSTParser – a multi-document discourse parser

Erick Galani Maziero, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística Computacional – NILC
Instituto de Ciências Matemáticas e 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 CSTParser, a multi-document discourse parser. Based on machine learning techniques and hand-crafted rules, the system identifies a set of relations predicted by CST (Cross-document Structure Theory) among sentences of different texts on the same topic.

Keywords: multi-document parsing, cross-document relationships

1 Introduction

In the electronic media, there are many sources reporting the same topic from the same or different perspectives. Online newspapers are an example: the same event is reported on different news portals. In general, the documents are produced soon after some event and, subsequently, other documents are generated to update the news. For instance, the two sentences below, S1 and S2, from different documents, are contradictory regarding the number of bombs in an attack, but both also present overlapping information (that there was a bomb):

S1: The downtown Public Finance Department building was hit by three home-made bombs.

S2: The Public Finance Department was also hit by a bomb.

It is believed that, when readers know how the parts of multiple documents are related, they can, for example, ignore redundancy, find contradictions, and understand the temporal evolution of a fact or event, which would allow them to approach the information in which they are interested in a more organized way. In another vein, this type of knowledge might also be useful for several computer applications, such as web browsers and automatic summarizers, which would have more information available to produce their results and meet the users' needs more efficiently. Some theories or models on multi-document relationships have been proposed for this purpose. One of the most used ones is the Cross-document Structure Theory (CST) [3].

This paper presents the CSTParser[1] [2], an on-line multi-document discourse parser based on CST. Using machine learning techniques and hand-crafted rules, the system

[1] Available at www.nilc.icmc.usp.br/~erick/CSTParser

indicates which CST relations occur among sentences of different texts on the same topic. The system was developed/trained and evaluated using the CSTNews corpus [1], composed of news texts in Brazilian Portuguese.

2 The Parser

The input to the parser is a group of texts on the same topic (Figure 1), coming from a web search or indicated by the user.

Fig 1. First step in the parser – selecting texts to analyze

To start the analysis, the parser segments each text in sentences. In order to select the most probable sentence pairs to be related (this is a necessary step, since there may be too many sentence combinations), the parser selects sentence pairs that have some lexical similarity according to the traditional word overlap measure, since it has been empirically shown in the area that CST-related sentences show some lexical similarity.

The selected pairs are then analyzed by several tools and resources in order to extract relevant features. Such features are the input date for machine learning classifiers and rules, which predict possible CST relations for the corresponding sentence pairs. The relations *Overlap*, *Subsumption*, *Elaboration*, *Equivalence*, *Historical-background* and *Follow-up* may be found by the classifiers, while the relations *Contradiction*, *Attribution*, *Indirect-speech* and *Translation* may be identified by rules. The decision on which relations would be detected by the classifiers and the rules was based on their frequency in the corpus (since the more data there is, the better the machine learning may be) and on recurrent lexical patterns the related sentences presented (which may be codified in rules).

The output of the parsing process is a graph, whose nodes are sentences from the several documents under analysis and the edges are the identified relations. Figure 2 shows a sentence of a document and its relations with other sentences from other documents. The general accuracy of the parser is 68.57%, which is better than the current state of the art for other works in the area.

Although the system was trained with Portuguese data, we believe that its method is general enough to be applied to other languages.

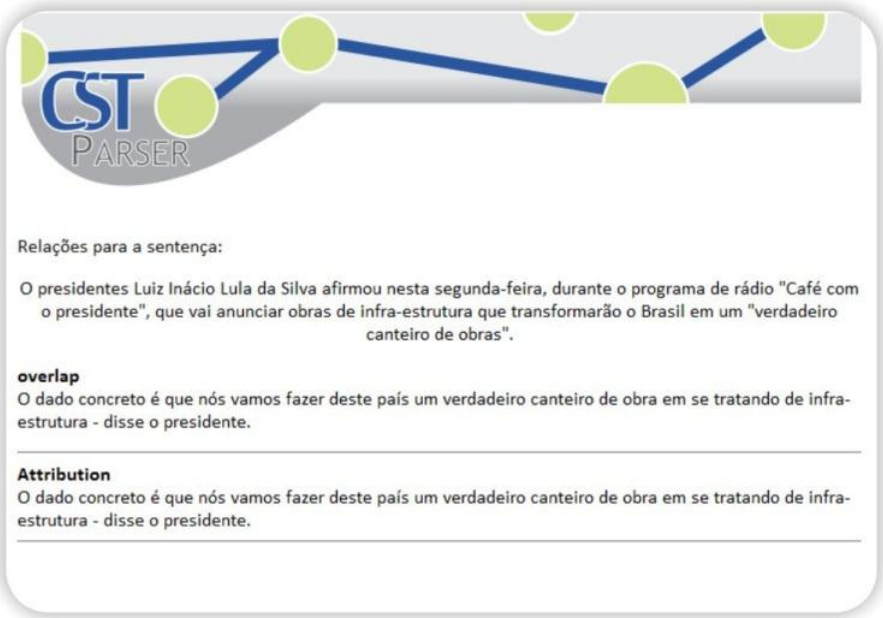

Fig 2. Example of the results of the CSTParser

This system will be demonstrated on-line in a notebook. Some groups of texts will be used to demonstrate the CST parsing.

References

1. Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105.
2. Maziero, E.G. and Pardo, T.A.S. (2011). Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning. In the *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, pp. 1-10.
3. 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*.

Acknowledgments

The authors are grateful to FAPESP for supporting this work.