ADROIT: Automatic Discourse Relation Organizer of Internet-based Text

A. S. M. Mahbub Morshed[1] and Mitsuru Ishizuka[2]

Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[1]mmorshed@mi.ci.i.u-tokyo.ac.jp, [2]ishizuka@i.u-tokyo.ac.jp

Abstract

The ADROIT system that we are developing allows automatic discourse analysis of information rich natural language texts extracted directly from the web. We use guidelines and relations of Rhetorical Structure Theory (RST) to decompose texts into elementary segments and to perform the discourse parsing between them. In this paper, we present version 1.0 of ADROIT and focus on the noble technique of cue-phrase disambiguation and machine learning for identification and organization of discourse relations.

Introduction

The study of discourse (Grosz & Sydner 1986) has a lengthy history in various disciplines such as linguistic, psychology and philosophy. The studies show that any coherent text, might be taken from the web or from a personal diary, has internal structures that are characterized by discourse relations, which subsequently describe the information content within the text. Discourse techniques have been used to improve the performance of text processing applications such as text summarization (Marcu 2000; Polanyi et al. 2004), information retrieval (Morato et al. 2003), natural language generation (Moore 1995) and text understanding (Torrance & Bouayad-Agha 2001). Despite the multiple applications of discourse analysis in constructing automatic text processing system, systems using discourse techniques are rare as automatic computation and organization of discourse relation of a plain text is a perplexing task. The main characteristics of our system are as follows:

1. It can analyze texts such as web pages where textual organizations are not always evident.
2. To accommodate semi-coherent texts or large volumes of texts, the system does not enforce on building of only one single RST tree for any given input.
3. It uses a small class of coarse granular relations that are important for all text processing application and permits further specific extensions of relation schema depending upon the application engine.
4. It uses discourse indicators such as anaphora, abstract verbs, verb dependency, etc. to recognize relations when cue-phrases are not available.

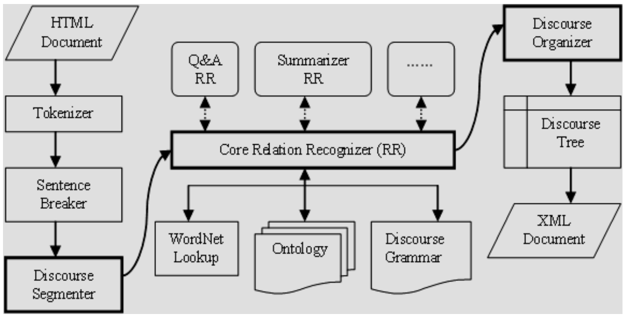

Figure 1: System Architecture of ADROIT 1.0

System Architecture

ADROIT takes preprocessed web pages as the input and derives the discourse tree in the form outlined by the Rhetorical Structure Theory (Mann & Thompson 1988). Tool described by Palmer & Hearst (1997) is used for breaking text into sentences. For brevity, only the 3 major processes of our system's architecture (Figure 1) for automatic discourse relation organization are briefly explained as follows:

1. **Discourse Segmenter (DS)** segments text into minimal non-overlapping units of discourse called elementary discourse units (EDUs) using rules derived automatically by Support Vector Machine (SVM). We use POS tags, syntactic information, discourse cues and punctuation as features. Preliminary results show that our segmenter performs as well as SynDS (Soricut & Marcu 2003), the best reported system that we are aware of.
2. **Core Relation Recognizer (RR)** finds all explicit rhetorical relations between elementary discourse units using both syntactic information such as cue phrases, time relation and semantic information such as word similarity. Core RR is designed to recognize 10 classes of coarse granular discourse relations grouped from 110 relations defined in RST-DT (2002) corpus.
3. **Discourse Organizer (DO)** implements an algorithm to derive discourse structures in the form of RST trees. Our algorithm overcomes many of the problems - i.e. large

search space and combinatorial problems – encountered by systems created by Marcu (2000) and Corston (1998). Search space is reduced by considering the organization of text into sections and paragraphs when building large trees. The algorithm also makes allocation when no such text organization is evident, for instance web pages, by using block comparison to compute correlation coefficient and tf-idf to determine the cohesion between widely separated text segments.

The algorithm design principle for the DS and RR processes was shaped to large extent by RST-DT corpus analysis. Syntactic rules were preferred over shallow processing and pattern recognition as almost 60% of intra-sentence segment boundaries were not marked by any orthographical markers or cue phrases.

Walk-through of Example

Here we describe the operation of ADROIT on an extract text taken from the RST-DT test corpus.

> But Mr. Ortega's threat over the weekend to end a 19-month cease-fire with the rebels seeking to topple him, effectively elevated the Contras as a policy priority just as they were slipping from the agendas of their most ardent supporters. (1)

The DS then splits example (1) into EDUs (a)-(e).

(a) But Mr. Ortega’s threat over the weekend
(b) to end a 19-month cease-fire with the rebels
(c) seeking to topple him,
(d) effectively elevated the Contras as a policy priority
(e) just as they were slipping ···· ardent supporters.

The RR module then is used to figure out which discourse relation holds between text spans as well as their nuclear roles. The discourse relation between a reporting clause (b) and a reported clause (c) is an ELABORATION relation, where EDU (b) is the nucleus and (c) is the satellite [b←c ELAB]. Similarly, there is a TEMPORAL relation between (d) and (e) [d←e TEMP]. Finally, after recognition of all explicit relations, DO identifies, if available, implicit relations – none for example (1) – and performs the following organization: [a←[b←c ELAB] ELAB] [d←e TEMP]. The XML version of the structure produces 2 distinct trees and this output agrees with the corpus as EDUs (d) and (e) actually refers to a sentence appearing much earlier in the text.

Future Activity and Conclusion

We have developed a working prototype of our ADROIT system. After through evaluation, we intend to further develop and enhance each module iteratively. Ultimately, our aim is to plug the system to a dialogue generation engine for multi-modal presentation.

References

Corston, S. O. 1998. *Computing Representations of the Structure of Written Discourse*. Ph.D. Thesis. University of California, Santa Barbara, CA, USA.

Grosz, B. J. and Sydner, C. L. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12:175-204.

Mann, W. C. and Thompson, S. A. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. *Text*, 8:243-281.

Marcu, D. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press, Cambridge, Massachusetts, London, England.

Morato, J., Llorens, J., Genova, G. and Moreiro, J. A. 2003. Experiments in discourse analysis impact on information classification and retrieval algorithms. *Information Processing and Management.* 39(6):825-851.

Moore, J. 1995. *Participating in explanatory dialogues: interpreting and responding to questions in context.* Cambridge, MA: MIT Press.

Palmer, D. D. and Hearst, M. A. 1997. Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics*, 23(2):241-269, June.

Polanyi, L., Culy, C., Thione, G. L., and Ahn, D. 2004. A rule based approach to discourse parsing. In *Proceedings of SigDial2004*, pp.108-117.

RST-DT. 2002. RST Discourse Treebank. Linguistic Data Consortium.

Soricut, R. and Marcu, D. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the HLT/NAACL-2003*, Edmonton, Canada, May-June.

Torrance, M. and Bouayad-Agha, N. 2001. Rhetorical structure analysis as a method for understanding writing processes. In *Proceedings of the International Workshop on Multi-disciplinary Approaches of Discourse (MAD 2001)*, pp.51-59, Amsterdam & Nodus Publications.