Discourse Structure: Theory, Practice and Use

Bonnie Webber,♡ **Markus Egg,**◇ **Valia Kordoni**♠

♡ University of Edinburgh   ◇ Humboldt University   ♠ Saarland University

bonnie@inf.ed.ac.uk   markus.egg@anglistik.hu-berlin.de   kordoni@dfki.de

1 Introduction

This tutorial aims to provide attendees with a clear notion of how discourse structure is relevant for language technology (LT), what is needed for exploiting discourse structure, what methods and resources are available to support its use, and what more could be done in the future.

2 Content Overview

This tutorial consists of four parts. Part I starts with a brief introduction to different bases for discourse structuring, properties of discourse structure that are relevant to LT, and accessible evidence for discourse structure.

For discourse structure to be useful for language technologies, one must be able to automatically recognize or generate with it. Hence, Part II surveys computational approaches to recognizing and generating discourse structure, both manually-authored approaches and ones developed through Machine Learning.

Part III of the tutorial describes applications of discourse structure recognition and generation in LT, as well as discourse-related resources being made available in English, German, Turkish, Hindi, Czech, Arabic and Chinese. Part IV concludes with a list of future possibilities.

3 Tutorial Outline

1. PART I – General Overview

   (a) Bases for structure in monologic, dialogic and multiparty discourse

   (b) Aspects of discourse structure relevant to Language Technology

   (c) Evidence for discource structure

2. PART II – Computational Recognition and Generation of discourse structure

   (a) Discourse chunking and parsing

   (b) Recognizing arguments and sense of discourse connectives

   (c) Recognizing and generating entity-based discourse structure

   (d) Dialogue parsing

3. PART III – Applications and Resources

   (a) Applications to Language Technology

   (b) Discourse structure resources (monolingual and multilingual)

4. PART IV – Future Developments

4 References

○ Regina Barzilay and Lillian Lee (2004). Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. *Proc. 2nd Human Language Technology Conference and Annual Meeting of the North American Chapter, Association for Computational Linguistics*, pp. 113-120.

○ Regina Barzilay and Mirella Lapata (2008). Modeling Local Coherence: An Entity-based Approach. *Computational Linguistics* 34(1), pp. 1-34.

○ Daniel Marcu (2000). *The theory and practice of discourse parsing and summarization*. Cambridge: MIT Press.

○ Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma and Aravind Joshi (2009). The Hindi Discourse Relation Bank. *Proc. Third Linguistic Annotation Workshop (LAW III)*. Singapore.

○ Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki et al. (2008). The Penn Discourse TreeBank 2.0. *Proc. 6th Int'l Conference on Language Resources and Evaluation*.

○ Manfred Stede (2008). RST revisited: Disentangling nuclearity. In Cathrine Fabricius-Hansen and Wiebke Ramm (eds.), *Subordination versus Coordination in Sentence and Text*. Amsterdam: John Benjamins.

○ Ben Wellner (2008). *Sequence Models and Ranking Methods for Discourse Parsing*. Brandeis University.

○ Deniz Zeyrek, Ümit Deniz Turan, Cem Bozsahin, Ruket Çakici et al. (2009). Annotating Subordinators in the Turkish Discourse Bank. *Proc. Third Linguistic Annotation Workshop (LAW III)*. Singapore.