Discourse Processing

Manfred Stede
Universität Potsdam
stede@uni-potsdam.de

1 Overview

The observation that discourse is more than a mere sequence of utterances or sentences amounts to a truism. But what follows from this? In what way does the "value added" arise when segments of discourse are juxtaposed - how does hierarchical structure originate from a linearized discourse?

While many discourse phenomena apply to dialogue and monologue alike, this tutorial will center its attention on monologue written text. The perspective taken is that of practical language processing: We study methods for automatically deriving discourse information from text, and point to aspects of their implementation. The emphasis is on breadth rather than depth, so that the attendees will get an overview of the central tasks of discourse processing, with pointers to the literature for studying the individual problems in more depth. Much of the tutorial will follow the line of the recent book M. Stede: Discourse Processing. Morgan & Claypool 2011.

Specifically, we will study the most important ways of ascribing structure to discourse. This is, first, a breakdown into functional units that are characteristic for the genre of the text. A news message, for example, is conventionally structured in a different way than a scientific paper is. For grasping this level of structure, the patterns that are characteristic for the specific genre need to be modeled.

Second, an ongoing text, unless it is very short, will cover different topics and address them in a sensible linear order. This is largely independent of genre, and since the notion of topic is relatively vague, it is harder to describe and sometimes difficult to identify. The common approach is to track the distribution of content words across the text, but in addition, overt signals for topic switches can be exploited.

Third, the identification of coreference links is a central aspect of discourse processing, and has received much attention in computational linguistics. We will survey the corpus-based methods that have dominated the field in recent years, and

Tutorials, NAACL-HLT 2013, pages 4–6,
Atlanta, Georgia, June 9 2013.

then look at the ramifications that the set of all coreference links in a text has for its structure.

Fourth, we investigate the structure resulting from establishing coherence relations (e.g., Cause, Contrast) among adjacent text segments. The term ”discourse parsing” is often used for the task of identifying such relations (by exploiting more or less explicit linguistic signals) and building tree structures that reflect the semantic or pragmatic scaffolding of a (portion of) text.

Thus emerges a picture of a text as a series of different, yet related, layers of analysis. The final part of the tutorial addresses the issue of inter-connections between these levels. As a tool for accessing such multi-layered text corpora, we will see how the (open-source) ANNIS2 database allows for querying the data across different layers, and for visualizing different structural layers in appropriate ways.

2 Outline

1. Introduction: Coherence and cohesion. How does a text differ from a ”non-text”?
2. Discourse structure as induced by the genre. Not all texts are created equal: The genre can determine text structure to a large extent. We look at three examples: Court decisions, film reviews, scientific papers.
3. Topics and text structure. Few texts keep talking about just one thing: Methods for finding topic breaks.
4. Coreference and its role for text structure. For understanding a text, we need to know who and what is being referred to: Methods for coreference analysis.
5. Coherence relations and ”rhetorical structure”. Trees resulting from semantic or pragmatic links between text segments: Methods for discourse parsing.
6. Synopsis: Text analysis on multiple levels
7. Accessing multi-layer corpora: The ANNIS2 Database

3 Speaker Bio

Manfred Stede[1], University of Potsdam. After completing his dissertation on the role of lexical semantics in multilingual text generation, Manfred Stede shifted

[1] `http://www.ling.uni-potsdam.de/~stede/`

his research focus towards problems of discourse structure and its role in various applications of text understanding. For discourse structure, his work centered on coherence relations and associated structural descriptions of text, and on the linguistic signals of such relations, especially connectives. From the early 2000s on, he developed the Potsdam Commentary Corpus as an example of (German) texts analyzed simultaneously on multiple levels, including sentential syntax, coreference, and rhetorical structure; in parallel, the technical infrastructure of a database for querying and visualizing multi-layer corpora was developed. In recent years, more analysis levels have been added to the corpus (e.g., content zones, connectives and their arguments). As for applications, Manfred worked on text summarization and various tasks of information extraction; more recently, his focus has been on issues of subjectivity and sentiment analysis.