Book Reviews

Discourse Processing

Manfred Stede
University of Potsdam

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 15), 2011, ix+155 pp; paperbound, ISBN 978-1-60845-734-2, $40.00; ebook, ISBN 978-1-60845-735-9, $30.00 or by subscription

Reviewed by
Bonnie Webber
University of Edinburgh

Discourse is coming in from the cold. After years of being ignored by researchers in other areas of computational linguistics and language technology, many of these same researchers are beginning to think that their own work could benefit from treating text as more than just a bag of sentences. That is, they are beginning to think that discourse offers some low-hanging fruit—achievable improvements in system performance that exploit either aspects of text structure or the context that text establishes and uses for efficient referring and/or predicational expressions.

This new monograph on *Discourse Processing* by Manfred Stede both reflects this new *zeitgeist* and provides an introduction to discourse for researchers in computational linguistics or language technology with little or no background in the area. This clear and timely monograph consists of a brief introduction to discourse, a meaty chapter on each of the three aspects of discourse processing that hold most promise for language technology, and a brief conclusion on where discourse research might go in the future. I will go through the three major chapters, and then make some general remarks.

Chapter 2

Chapter 2 addresses two distinct types of large-scale discourse structure: structure that follows from a text belonging to a particular genre, and structure that follows from the topic (or topic mix) of a text. The genre of a text affects features such as style and register. What is relevant here is structure that genre may confer on a text. Stede suggests that some, but not all, texts inherit large-scale structure from their genre, calling some *unstructured*, some *structured*, and some *semi-structured*. As a reader, I did not find this distinction useful, because all text that belongs to a genre seems to get some large-scale structure from it. On the other hand, all or part of this structure might simply not be manifest in the kind of lexico-syntactic features that automated systems regularly rely on for text segmentation. As a case in point, although Stede offers the text *Suffering* (used as a running example throughout the book) as an example of unstructured text, like other instances of *Comments* in the *Talk of the Town* section of the *New Yorker* magazine, its large-scale structure comprises a "hook" aimed at getting the reader's attention, followed by a short essay that concludes with a serious point. Although ways of attracting a reader's attention may not have specific lexico-syntactic features, it might still be possible to recognize the transition between "hook" and essay, and essay structure itself is what ETS's *eRater* system (Burstein and Chodorow 2010) aims to recognize and evaluate.

This first half of Chapter 2 focuses on the genre-based structure of scientific texts and of film reviews. Here researchers have already shown that language technologies such as information extraction and sentiment analysis benefit from taking such structure into account, so this is entirely appropriate for the book's target audience. More on genre-based functional structure and its use in producing structured biomedical abstracts can be found in the recent survey of research on discourse structure and language technology by Webber, Egg, and Kordoni (2012).

The second half of Chapter 2 discusses large-scale discourse structure associated with patterns of topics. Such structure is often found in expository writing such as encyclopedia articles and travel pieces. Here, changing patterns of content words correlate well with changes in topic, rendering them useful for the many approaches to text segmentation that are well-described in this half of the chapter. Because the discussion here of probabilistic models for topic segmentation is rather short, the reader who wants to know more should consult the excellent survey of topic segmentation methods by Purver (2011).

Chapter 3

Chapter 3, entitled *Coreference Resolution*, addresses more than this, dealing with the resolution of other expressions whose reduction is licensed by the discourse context, such as *bridging* reference and *"other"* reference, which Halliday and Hasan (1976) call *comparative reference* because it occurs with comparative forms such as "larger fish" and "a more impressive poodle," as well as with "other," "another," and "such." Stede justifies inclusion of this chapter for two reasons—the close connection between coreference resolution and topic segmentation and the benefits to text analysis provided by having its pronouns resolved. But another reason must be the link mentioned earlier between text and context: Discourse creates the context in which context-reduced expressions make sense, so it falls naturally within the tasks of discourse processing to resolve them, either through modeling context explicitly or through the use of proxies.

The chapter starts with an overview of coreference and anaphora that covers both their forms and their functions. This is followed by an important section on corpus annotation (Section 3.2), included because (as Stede notes) what has been annotated and why it has been annotated strongly determines what expressions are resolved and how. This section identifies many of the problems in coreference annotation that have been raised in the literature, but recognizes that research has to make use of the resources that exist and not just the resources it wants. Several of these are indicated at the end of the section, reminding one that it would have been useful to have some pointers in Chapter 2 to corpora available for genre-based segmentation (such as Liakata's ART corpus)[1] or for topic-based segmentation.

Stede then links the current chapter to the previous one through a discussion of entity-based coherence (Section 3.3) and then discusses how to identify when a pronoun or definite noun phrase should be treated as anaphoric (Section 3.4) as groundwork for discussion of anaphora resolution (Sections 3.5–3.7). Missing from the discussion of detecting non-anaphoric (pleonastic) pronouns is mention of Bergsma's recent system *NADA* for doing this (Bergsma and Yarowsky 2011).[2]

[1] Downloadable from `http://www.aber.ac.uk/en/ns/research/cb/projects/art/art-corpus/`

[2] Downloadable from `http://code.google.com/p/nada-nonref-pronoun-detector/`

The discussion of anaphora resolution covers rule-based approaches to resolving nominal anaphora (Section 3.5) and then supervised machine learning methods for anaphora resolution (Section 3.6). The latter follows the structure (albeit not the content) of Ng's survey (2010), in discussing *mention-pair models*, and then *entity-mention models*. Whereas Ng then discusses *ranking models*, including his *cluster ranker* (Rahman and Ng 2009), which is conceptually similar to the Lappin and Leass (1994) approach described in Section 3.5, Stede discusses a range of more recent models, most of which are subsequent to Ng's survey.

Section 3.8 surveys methods evaluating coreference resolution and some of the known problems in doing so. A good complement to this is Byron's too-little-known discussion of problems in the consistent reporting of such results (Byron 2001). Chapter 3 concludes with a section on *Recent Trends*, which would also have been useful in Chapter 2.

Chapter 4

The fourth and longest chapter deals with semantic or pragmatically oriented *coherence relations* that hold between adjacent text spans or *discourse units*. Whereas the previous two chapters were essentially theory-neutral, the presentation in Chapter 4 largely reflects the perspective of *Rhetorical Structure Theory* (Mann and Thompson 1988). RST takes a text to be a sequence of elementary discourse units that comprise the leaves of a tree structure of coherence relations between recursively defined discourse units. RST also assumes that one of the arguments to a coherence relation may be more important to the speaker's purpose than the other, calling the former the *nucleus* and the latter, the *satellite*.

This RST framework dictates the structure of the chapter: Following an introductory section that explains and motivates coherence relations, each subsequent section considers the next task in an RST analysis—segmenting a text into elementary discourse units (Section 4.2), recognizing which (adjacent) units stand in a coherence relation and what (single) relation holds between them (Section 4.3), and finally, inducing the overall tree structure of coherence relations that hold between recursively defined discourse units (Section 4.4). All these tasks are well described, both from a theoretical perspective and in terms of automated procedures for carrying them out. Coverage of relevant work is very high.

Where the reader may get confused, however, is that a good proportion of the more recent work on identifying coherence relations does not fall within the framework of RST, and thus doesn't adhere to several of its assumptions—in particular, that a text is divisible into a covering sequence of elementary discourse units, that only one relation can hold between discourse units, that the arguments to a coherence relation must be adjacent, that one argument to a coherence relation may intrinsically convey information that is more important to the speaker's purpose than the other, and that coherence relations impose an overall tree structure on a text in terms of recursively defined discourse units.

Although Chapter 4 discusses the Penn Discourse TreeBank (Prasad et al. 2008) and its "somewhat modest annotations" (page 126), the discussion is framed in terms of RST tasks, whereas the assumptions underlying the Penn Discourse TreeBank reflect its concerns with a quite different set of tasks involved in recognizing coherence relations. The first task requires finding evidence for a coherence relation (in the form of a discourse connective such as a coordinating or subordinating conjunction or a discourse adverbial, or in the form of sentence adjacency) and then determining (1) if the evidence does indeed signal a coherence relation, given that evidence is often ambiguous; (2) if it does, what constitutes its arguments; and (3) what is its sense. Although Chapter 4 covers some of this work (Dinesh et al. 2005; Wellner and Pustejovsky 2007; Elwell and Baldridge 2008; Pitler and Nenkova 2009; Prasad, Joshi, and Webber 2010), its appearance within the context of a discussion of RST-tasks may lead to some confusion.

Chapter 4 concludes with a brief discussion of some important open issues regarding coherence relations, including problems with associating a large text span with a single recursive structure of coherence relations and problems with inter-annotator agreement.

Summary

For its intended audience, this monograph will serve as a compact, readable introduction to the subject of discourse processing. The relevant phenomena are presented clearly, as are many of the computational methods for dealing with them. What readers won't get is criteria for choosing among the methods or an understanding of what each method is good for. This problem may reflect the absence of comparable performance results and useful error analyses in the original publications, however.

Also missing from the monograph is discussion of applications of discourse processing, and pointers to more of the resources available to researchers interested in discourse structure. This is where the additional resources I have mentioned may prove complementary.

Finally, a plea to the series editor: Monographs such as this one really need an *index*. Some monographs in the series have one, whereas others (like this one) don't. Because the series appears in both electronic and physical format, one could excuse the former not having an explicit index, since in most cases, one can get away with the basic search facility in the Adobe Reader. Nothing similar is available for the nicely sized physical monographs. Their authors should be strongly encouraged to provide them.

References

Bergsma, Shane and David Yarowsky. 2011. Nada: A robust system for non-referential pronoun detection. In *Proceedings of DAARC*, 12 pages, Faro.

Burstein, Jill and Martin Chodorow. 2010. Progress and new directions in technology for automated essay evaluation. In R. Kaplan, editor, *The Oxford Handbook of Applied Linguistics*. Oxford University Press, 2nd edition, pages 487–497.

Byron, Donna. 2001. The uncommon denominator: A proposal for consistent reporting of pronoun resolution results. *Computational Linguistics*, 27(4):569–577.

Dinesh, Nikhil, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Attribution and the (non-)alignment of syntactic and discourse arguments of connectives. In *ACL Workshop on Frontiers in Corpus Annotation*, pages 29–36, Ann Arbor, MI.

Elwell, Robert and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Proceedings of the IEEE Conference on Semantic Computing*, 8 pages, Santa Clara, CA.

Halliday, Michael and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Lappin, Shalom and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Mann, William and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Ng, Vincent. 2010. Supervised noun phrase coreference research: The first 15 years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala.

Pitler, Emily and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *ACL-IJCNLP '09: Proceedings of the 47th Meeting of the*

Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, pages 13–16, Singapore.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*, pages 2961–2968, Marrakech.

Prasad, Rashmi, Aravind Joshi, and Bonnie Webber. 2010. Exploiting scope for shallow discourse parsing. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2076–2083, Malta.

Purver, Matthew. 2011. Topic segmentation. In Gokhan Tur and Renato de Mori, editors, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, Chapter 11. Wiley, Hoboken, NJ.

Rahman, Altaf and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977, Singapore.

Webber, Bonnie, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, doi:10.1017/S1351324911000337.

Wellner, Ben and James Pustejovsky. 2007. Automatically identifying the arguments to discourse connectives. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 92–101, Prague.

This book review was edited by Pierre Isabelle.

Bonnie Webber is a Professor of Informatics at Edinburgh University. She received both her MSc and PhD from Harvard University. She is a Fellow of the Royal Society of Edinburgh and of the American Association for Artificial Intelligence. Both her early and her recent research have focused on computational approaches to discourse and question answering. In between, she has carried out research on animation from instructions, medical decision support systems, and biomedical text processing. Webber's e-mail address is bonnie.webber@ed.ac.uk.