Book Reviews

The Structure of Scientific Articles: Applications to Citation Indexing and Summarization

Simone Teufel
(University of Cambridge)

Stanford, CA: CSLI Publications (CSLI Studies in Computational Linguistics), 2010, xii+518 pp; hardbound, ISBN 978-1-57586-555-3, $70.00; paperbound, ISBN 978-1-57586-556-0, $32.50

Reviewed by
Robert E. Mercer
University of Western Ontario

Discourse models have received significant attention in the computational linguistics community with some important connections to the non-computational discourse community. More recently, the importance of discourse annotation has increased as models generated with supervised machine learning techniques are being used to annotate text automatically. A primary area for annotation is science. The theme of Teufel's book is an important contribution in these areas: discourse models, annotation schemes, and applications.

The book is a substantial work, approximately 450 pages of text and appendices. It extends Teufel's Ph.D. thesis (Teufel 2000) with a decade of new work and updated references. The book is content-rich and meticulously written. In addition to presenting Teufel's discourse model, it also works as a good entry point into discourse models and annotation. Because each chapter is structured with background, new material, and a summary, each chapter can be read somewhat independently. Cross-references to other parts of the book are carefully included where warranted. This structure lends itself to using the book as a reference for each of the subtopics or as an introduction to the subject area as a whole, suitable as a textbook.

Chapter 1 sets the stage for the rest of the book. The author sets out her fundamental assumptions and hypotheses. The fundamental assumptions arise from three observations that she has made regarding the literature. Scientific discourse contains descriptions of positive and negative states, contains references to others' contributions, and is the result of a rhetorical game intended to promote one's contribution. Chapter 2, on information retrieval and citation indexes, and Chapter 3, on summarization, provide the motivation for the main theme of the book: These two information-based endeavors can be enhanced with automated tools that incorporate an understanding of the rhetorical aspects of science writing.

Whereas Chapters 2 and 3 give an overview of current methodologies, Chapter 4, "New Types of Information Access," introduces two new techniques, rhetorical extracts and citation maps, that are suggested as information navigation methods enhanced by knowledge of the discourse that contains the information being accessed. Rhetorical extracts are snippets that can be tailored to user expertise and navigation task. Citation maps are interactive citation indexes that have their citation links augmented with rhetorical or sentiment information.

Chapter 5 gives a detailed description of the five scientific text corpora that are used in the research described throughout the book: computational linguistics,

chemistry, genetics, cardiology, and agriculture. The chapter focuses primarily on the computational linguistics corpus, on which most of the results in the book are based. SciXML, Teufel's markup language for science articles, is described.

Chapter 6 contains an in-depth description of the Knowledge Claim Discourse Model (KCDM). Teufel gives reasons why the traditional discourse models are abandoned in favor of her new model. In addition to it being a shallow method, she points out the important aspects of KCDM (compared to Rhetorical Structure Theory): It is text-type-specific (scientific articles); no world knowledge is required; it has global (top-down) not local (bottom-up) relations; it is non-hierarchical (citation and summarization applications do not require a rich hierarchical structure).

Chapter 7 presents three annotation schemes based on the KCDM: Knowledge Claim Attribution (KCA), Citation Function Classification (CFC), and Argumentative Zoning (AZ). The background and purpose of the schemes are carefully laid out. The annotation guidelines (coding manuals) are given in Appendix C.

Chapter 8 reports on the reliability studies that use human annotators and gauge the quality of the annotation scheme using agreement among the annotators as a proxy for this measure. A good discussion of the measures of annotator agreement opens the chapter, followed by a detailed analysis of the four studies. Three of the four studies used three annotators, the other used 18 annotators. All studies used the computational linguistics corpus.

Chapters 9 and 10 discuss the features that will be used by the machine implementations of AZ, KCA, and CFC that are described in Chapter 11. Chapter 9 provides a comprehensive discussion of the various embodiments of meta-discourse, the text that concerns itself with the dialogue between the author and the reader rather than content-bearing text. Chapter 10 discusses the computable surface features that capture the important aspects of meta-discourse that are used by the automatic annotation methods. Chapter 11 then introduces the reader to the standard supervised machine-learning methodology used to generate the statistical models that implement the automatic AZ, KCA, and CFC annotators. Chapter 12 presents gold-standard, and extrinsic and subjective evaluations of these automatic methods. The gold standard is the human-annotated computational linguistics articles and the extrinsic task is rhetorical extracts.

Chapter 13 investigates the universality of the KCDM. The earlier chapters' results were based on the computational linguistics corpus. This chapter considers the disciplines of chemistry, computer science, biology, astrophysics, and legal texts. Two issues surface: the need to modify the original KCDM slightly, and the move from an absolutely domain-knowledge-free annotation to one which includes some high-level facts about research practices in the discipline.

Chapter 14 pushes the frontiers of potential uses of the KCDM methodology: support tools for scientific writing, automatic review generation, scientific summary generation that moves beyond simple sentence extraction methods and summaries of multiple scientific documents, as well as integration of automatic AZ into a large-scale digital library. Chapter 15 provides the conclusion. In the first section it recapitulates the main themes of the book. This section also nicely serves as an introduction to the book, if so desired. Section 2 lists a number of areas that could lead to an improved automatic system.

The four appendices contain a list of the CmpLG-D articles, the DTD for SciXML, the annotation guidelines, and a catalog of lexical items and patterns useful in the discourse setting.

The book makes an important and powerful statement in the field of discourse modeling and annotation, and provides an important body of work to which other

researchers can add or compare their work. I think it is important to keep in mind the following few points while reading the book: First, Teufel comments that she is interested in a discourse model for the experimental sciences, yet her focus for much of the book is a corpus of computational linguistics papers. Also, the discourse model proposed is based on knowledge claims and rhetorical moves. This catholic view of what is science and the narrow view of structure may surprise some readers given the title of the book. Next, some of the fundamental decisions regarding the discourse model are heavily influenced by the requirements of the two motivational topics, leading one to question the full generality of the discourse model. As well, the range of rhetoric in science writing may be broader than anticipated by Teufel's model—for example, the style found in the geology discipline is more cumulative than critical (Heather Graves, personal communication). And finally, some researchers (White 2010) argue that the domain-knowledge-free annotation dictum, although loosened slightly by Teufel, may need to be further relaxed in order to produce a more accurate gold standard, regardless of the automatic system's access to the same domain knowledge.

References

Teufel, Simone. 2000. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

White, Barbara. 2010. Identifying sources of inter-annotator variation: Evaluating two models of argument analysis. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*, Uppsala, pages 132–136.

Robert E. Mercer is a Professor of Computer Science at the University of Western Ontario. His research interests include argumentation in science writing and annotation. Mercer's e-mail address is mercer@csd.uwo.ca.