The Theory and Practice of Discourse Parsing and Summarization

Daniel Marcu
(Information Sciences Institute, University of Southern California)

Cambridge, MA: The MIT Press, 2000,
xix+248 pp; hardbound, ISBN
0-262-13372-5, $39.95

Reviewed by
Udo Hahn
Albert-Ludwigs-Universität Freiburg

Marcu's monograph is based on his Ph.D. thesis—research carried out at the Department of Computer Science, University of Toronto—and subsequent work conducted at the Information Sciences Institute, University of Southern California. It argues for the idea that discourse/rhetorical relations that connect text spans of various length can be computed *without* a complete semantic analysis of sentences that make up these text segments. As an alternative, a formal specification of admissible text structures is provided, which constrains the range of possible semantic and functional connections between text spans and imposes strict well-formedness conditions on valid discourse structures. For effectively computing these text structures, mainly surface-oriented lexical cues and shallow text-parsing techniques are used. Complementary to these formal and computational considerations, Marcu reports on various evaluations, both intrinsic and extrinsic, in order to assess the strengths and weaknesses of his approach and the generality of the principles it is based on. These experiments were mostly carried out on *Scientific American*, TREC, MUC, *Wall Street Journal*, and Brown corpora.

The book consists of three main parts. In the first part, linguistic and formal properties of coherent texts are discussed, with a focus on high-level discourse structures. This theoretical framework serves, in the second part, as the background for developing discourse structure parsing algorithms that compute rhetorical relations in real-world free texts. The benefits of such algorithms for building a high-performance text summarization system are dealt with in the third part.

In the first part, the author factors out a set of assumptions that are common to prominent approaches to discourse structure. So, consensus has been reached that texts can be segmented into nonoverlapping, elementary textual units, that discourse relations of different types link (elementary and complex) textual units of various sizes, that some textual units are more important to the writer's communicative intentions and goals than others, and that trees are a good approximation of the abstract structure of most texts. These considerations lead to a compositionality criterion that requires that discourse relations that link two large text spans can be explained by discourse relations that hold between at least two of the most salient text units of the constituent spans. This notion then forms the basis for a first-order logic axiomatization that captures formal properties of valid text structures. Although this formalization is independent of the set of rhetorical relations actually considered, it yields, by proper relation instantiation, a formal characterization of the structural properties that are specific to Rhetorical Structure Theory (RST) (Mann and Thompson 1988). Building on these formal considerations, the author discusses three (nonincremental) algorithmic paradigms that compute some or all valid discourse structures of a text. Two of them employ model-theoretic techniques and encode the problem of text-structure

derivation as a classical constraint satisfaction problem and as a propositional satisfiability problem. The other one is grammar-based and builds on a proof theory for solving the text-structure derivation problem (demonstrated to be sound and complete with respect to the given logical formalization). The performance of these algorithms is compared empirically on a benchmark of eight manually encoded text-structure derivation problems.

Marcu uses logic to distinguish between discourse structures that are valid and those that are not, so that all valid discourse structures of a text can be determined. In the second part of the monograph, attention then shifts to alternative approaches to deriving valid discourse structures. The first approach relies primarily on discourse markers for shallow rhetorical parsing and employs, as a result of an in-depth corpus analysis, manually designed rules covering more than 450 English cue phrases such as *because*, *however*, and *in addition*, as well as punctuation marks. The second approach adds to plain discourse markers knowledge of surface-oriented lexical co-occurrence data, syntactic criteria (such as part-of-speech categories), and lexical similarity measures based on semantic relation information in order to identify text segments and their rhetorical organization. Given this knowledge-richer setting, discourse parsing rules were automatically derived by applying machine learning techniques (the C4.5 decision-tree algorithm) to data obtained from three corpora of manually annotated discourse trees. All these approaches are meticulously and lucidly described by providing various algorithm schemata for relevant computation steps. Empirical studies are then concerned with the role that discourse markers play in properly segmenting texts into elementary text units and in signaling rhetorical relations that hold between the text segments they connect. The correctness of the discourse trees built by the parser is judged intrinsically, by comparing automatically derived trees with ones that have been built manually, as well as extrinsically, by evaluating the impact automatically derived discourse trees have on properly solving natural language processing problems such as the summarization of texts.

In the third part of the book, the utility of computing discourse structures is empirically assessed in the context of such a text summarization (i.e., extraction) task. The approach advocated by Marcu is readily applicable to this problem, since the representation structures it yields offer implicit content salience orderings in terms of the hierarchical tree structure and the distinction of important information contained in the nucleus and less-important information contained in the satellite portion of text spans, all of which are of immediate relevance for summarization purposes. The main hypothesis to be confirmed is whether or not discourse structures can be successfully exploited in a practical summarization setting. In a methodological experiment, evidence is gathered that text structures such as those mentioned above indeed effectively contribute to identifying the most important units of a text. A discourse-structure-based summarization algorithm that builds on these principles implements a simple salience metric that interprets the tree structure generated by the simple cue-phrase-based text-structure parser. A comparative evaluation reveals that this approach significantly outperforms two baseline algorithms (lead sentence and random sentence selection) and Microsoft's Office 97 summarizer. Considering the structure of discourse to be the paramount factor in determining salience, and incorporating a variety of additional position-, title-, text-tree-, and lexically-based summarization heuristics, a simple GSAT-style learning mechanism is presented that optimizes a linear combination of seven single salience metrics (in terms of combined recall and precision). This way, a significant increase in the performance of the discourse-based summarizer is achieved (yet parameter tuning is clearly dependent on the given text genre and compression rate!).

Marcu's monograph presents a cornerstone in the computational treatment of texts. It has formal merits, as it provides a model-theoretic framework for the study of text coherence structures, in general, and the study of RST, in particular. It has computational merits, as it provides alternative ways of deriving text-structure descriptions automatically and inexpensively (i.e., avoiding full, in-depth text understanding) and distinguishes, given the *a priori* axiomatization, valid text structures from invalid ones. It has methodological merits, as it incorporates machine learning techniques for automatically acquiring the rules needed for discourse parsing and discourse-structure-driven summarization. Finally, it has empirical merits, as algorithms are tested and validated under different experimental conditions.

Marcu also frankly admits that his work ignores the wealth of linguistic constructs that have been shown to be important in text understanding. Such phenomena include focus, topicality, cohesion and reference, pragmatics, and so on. Hence, the notion of validity being proposed is a constrained one, and it has to be weighed carefully against the notion of adequacy and expressiveness of the representation structures derivable therefrom. Still, the author claims that these phenomena can be couched in his formal framework as well. Additionally, one might mention the crucial role of domain-knowledge-dependent inferences and their interaction with building text structures in the absence of explicit cue phrases. Further open issues are the granularity of the text units that span rhetorical relations (e.g., the phrasal as opposed to the clausal or "clause-like" level) and the impact of the text genres under scrutiny. Finally, the dependence on basic assumptions and constructs underlying RST, despite the author's attempt to abstract away from it as much as possible, might be more prevalent than is acknowledged.

The book spans a wide variety of issues in a well-structured, reader-friendly way, and it is easy to understand even in its technical passages. Hence, it can be highly recommended for graduate courses on text analysis. Students are given an outstanding example of the current research paradigm of computational linguistics, which includes formal, algorithmic, methodological, and empirical contributions. And they also may learn how scientific results can be communicated in a rigorous though comprehensible manner.

Reference

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

Udo Hahn is a professor of computational linguistics at Albert-Ludwigs-Universität Freiburg, Germany. His methodological interests include text parsing, knowledge and discourse representation, and learning from texts. He has worked mainly on text analysis applications such as text summarization, knowledge extraction and text mining, and document retrieval. Hahn can be contacted via www.coling.uni-freiburg.de/∼hahn.