Discourse Structures and Language Technologies

Bonnie Webber
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
bonnie.webber@ed.ac.uk

1 Introduction

I want to tell a story about computational approaches to discourse structure. Like all such stories, it takes some liberty with actual events and times, but I think stories put things into perspective, and make it easier to understand where we are and how we might progress.

Part 1 of the story (Section 2) is the past. Here we see early computational work on discourse structure aiming to assign a simple tree structure to a discourse. At issue was what its internal nodes corresponded to. The debate was fierce, and suggestions that other structures might be more appropriate were ignored or subjected to ridicule. The main uses of discourse structure were text generation and summarization, but mostly in small-scale experiments.

Part 2 of the story (Section 3) is the present. We now see different types of discourse structure being recognized, though perhaps not always clearly distinguished. An increasing number of credible efforts are aimed at recognizing these structures automatically, though performance on unrestricted text still resembles that of the early days of robust parsing. Generic applications are also beginning to appear, as researchers recognize the value of these structures to tasks of interest to them.

Part 3 of the story (Section 4) is the future. We now see the need for a mid-line between approaches hostage to theory and empirical approaches free of theory. An empirical approach underpinned by theory will not only motivate sensible back-off strategies in the face of unseen data, but also enable us to understand how the different discourse structures inter-relate and thereby to exploit their mutual recognition. This should allow more challenging applications, such as improving the performance of statistical machine translation (SMT) through the extended locality of discourse structures and the linguistic phenomena they correlate with.

2 Early computatational approaches to discourse structure

Early computational work generally assumed discourse structure had an underlying tree structure, similar to the parse tree of a sentence. At issue was what its internal nodes and other formal properties corresponded to. In Rhetorical Structure Theory (Mann and Thompson, 1988), used in both text generation (Scott and de Souza, 1990; Moore, 1995; O'Donnell et al., 2001) and analysis (Marcu, 1996; Marcu, 2000), an internal node corresponded to a rhetorical relation holding between the text units associated with its daughters, and *precedence* corresponded to their order in the text. In work on generating task instructions (Dale, 1992), each internal node corresponded to the next step to take to accomplish the plan associated with its parent. In (Grosz and Sidner, 1986), which I will return to in Section 4, internal nodes corresponded to speaker intentions, with *dominance* in the tree corresponding to a daughter node's intention supporting that of its parent and *precedence* corresponding to one intention needing to be accomplished before another. The internal nodes in (Moser and Moore, 1996) reflected an attempt to reconcile Grosz and Sidner's approach with that of Mann and Thompson.

Work that attempted to show that a simple linear model might be a better account for types of expository text (Sibun, 1992) was, by and large, ignored.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 12–16

3 Current computatational approaches to discourse

As well as further elaboration of recursive discourse structures (Asher and Lascarides, 2003; Polanyi et al., 2004), current computational approaches have focussed on discourse structures more easily linked to data: structure associated with changes in *topic*, structure associated with the *function* of the parts of a text within a given genre, and structure associated with what one might call *higher-order* predicate-argument relations or *discourse relations*.

3.1 Topic structure

Expository text can be viewed as a linear sequence of *topically coherent* segments (sequences of sentences), where the sequence of topics is either specific to a text or conventionalized (Figure 1).

Interest in topic structure originally came from its perceived potential to improve information retrieval (Hearst, 1994; Hearst, 1997). More recent interest comes from its potential use in segmenting lectures, meetings or other speech events, making them more amenable to search (Galley et al., 2003; Malioutov and Barzilay, 2006).

Computational approaches to topic segmentation all assume that: (1) Relations hold between the topic of discourse segments and the topic of the discourse as a whole (eg, History of Vermont → Vermont). (2) The only relation holding between sister segments, if any, is sequence, though certain sequences may be more common than others (Figure 1). (3) The topic of a segment will differ from those of its adjacent sisters. (Adjacent spans that share a topic will belong to the same segment.) (4) Topic predicts lexical choice, either of all the words of a segment or just of its content words (ie, excluding "stop-words").

Making topic structure explicit (ie, topic segmentation) is based on either **semantic-relatedness**, where each segment is taken to consist of words more related to each other than to words outside the segment (Hearst, 1994; Hearst, 1997; Choi et al., 2001; Bestgen, 2006; Galley et al., 2003; Malioutov and Barzilay, 2006) or **topic models**, where each segment is taken to be produced by a distinct, compact lexical distribution (Purver et al., 2006; Eisenstein and Barzilay, 2008; Chen et al., 2009).

3.2 Function-based structure

Texts within a given genre (eg, news reports, errata, scientific papers, letters to the editor, etc.) generally share a similar structure that is independent of topic and reflects the **function** played by each of its parts. Best known is the *inverted pyramid* of news reports, consisting of a headline; a lead paragraph, conveying *who* is involved, *what* happened, *when* it happened, *where* it happened, *why* it happened, and (optionally) *how* it happened; a body that provides more detail; and a tail, containing less important information. This is why the first (ie, lead) paragraph can provide the best *extractive summary* of a news report.

In the genre of scientific papers (and, more recently, their abstracts), high-level structure comprises the following ordered sections: *Objective* (also called *Introduction*, *Background*, *Aim*, or *Hypothesis*); *Methods* (also called *Method*, *Study Design*, or *Methodology*); *Results* (also called *Outcomes*); *Discussion* and optionally, *Conclusions*. This does not mean that every sentence within a section realises the same function: Fine-grained functional characterizations of scientific papers (Liakata et al., 2010; Teufel, 2010) show a range of functions served by the sentences in a section.

Interest in automatic annotation of functional structure comes from its value for summarization (noted above), sentiment analysis, where words may have an objective sense in one section and a subjective sense in another (Taboada et al., 2009), and citation analysis, where a citation may mean different things in different sections (Teufel, 2010).

As with computational models of topic-based structure, computational models of function-based structure make assumptions that may or may not actually hold: (1) Relations hold between the function of a segment and that of the discourse as a whole: While relations may hold between sisters (eg, *Methods* constrain *Results*), only sequence has been used in modelling. (2) Function predicts more than lexical choice: it can predict indicative phrases such as "results show" (→ *Results*) or indicative stop-words such as "then" (→

	Wisconsin	Louisiana	Vermont
1	Etymology	Etymology	Geography
2	History	Geography	History
3	Geography	History	Demographics
4	Demographics	Demographics	Economy
5	Law and government	Economy	Transportation
6	Economy	Law and government	Media
7	Municipalities	Education	Utilities
8	Education	Sports	Law and government
9	Culture	Culture	Public Health

Figure 1: Structure of Wikipedia articles about US states, as shown in sub-headings

Method). (3) Functional segments usually appear in a specific order, so either sentence position is a feature used in modelling or sequential models are used..

While the internal structure of a functional segment has usually been ignored in high-level modeling (Chung, 2009; Lin et al., 2006; McKnight and Srinivasan, 2003; Ruch et al., 2007), (Hirohata et al., 2008) found that assuming that properties of the first sentence of a segment differ from those of the rest (as in 'BIO' approaches to Named Entity Recognition) leads to improved performance in segmentation (ie, 94.3% per sentence accuracy vs. 93.3%).

While most functional modelling has been on biomedical text, where texts with explicitly labelled sections serve as "free" training data for segmenting unlabelled texts, there has also been some work on functional segmentation of legal texts and student essays.

3.3 "Higher-order" pred-arg structure

The third type of discourse structure receiving significant attention from the computational world is what can be called *higher-order* predicate-argument structure, or structure associated with *discourse relations*. Whereas at the sentence level, pred-arg structures are usually headed by a verb (Kingsbury and Palmer, 2002) or a noun (Gerber et al., 2009), predicate-argument structures in discourse are usually headed by a discourse connective — eg, a conjunction like *because* or *but*, or a discourse adverbial like *nevertheless* or *instead*.

And just as pred-arg relations within a sentence can conveyed through adjacency (eg, English noun-noun modifiers such as *container ship crane operator courses* – courses to train operators of cranes that load/unload ships whose cargo is packed in containers), pred-arg relations in discourse can be conveyed through adjacency between clauses or sentences.

The Penn Discourse TreeBank is currently the largest resource manually annotated for discourse connectives, their arguments, and the senses they convey (Prasad et al., 2008). Related resources are also being created for Modern Standard Arabic (Al-Saif and Markert, 2010), Chinese (Xue, 2005), Czech (Mladová et al., 2008), Danish and Italian parallel treebanks (Buch-Kromann and Korzen, 2010), Dutch (van der Vliet et al., 2011), German (Stede, 2004; Stede, 2008), Hindi (Oza et al., 2009), and Turkish (Zeyrek et al., 2010).

The potential value of being able to automatically recognize these discourse relations, their arguments and their senses comes from their help in question generation (Mannem et al., 2010), extractive summarization (Louis et al., 2010) and sentiment detection (Taboada et al., 2009). So efforts are increasing to automatically recognize them (Elwell and Baldridge, 2008; Lin et al., 2010; Pitler et al., 2008; Pitler et al., 2009; Pitler and Nenkova, 2009; Prasad et al., 2010; Wellner and Pustejovsky, 2007; Wellner, 2008).

4 Future computatational approaches to discourse

This story closes with some speculations about the future. I have sketched a past in which computational approaches to discourse structure were hostage to theory and a present in which they are essentially free of theory. What we really want is an empirical approach underpinned by theory, that allows us to understand (at the very least) the ways in which the various types of discourse structures fit together. Early on, (Grosz and Sidner, 1986) attempted to meld a theory of intention-

based discourse structure with a theory of attentional structure (ie, what the conversational participants were attending to), but the link between theory and data was not sufficiently robust. Later attempts to link multiple discourse structures were motivated by purely practical concerns. (Marcu, 2000) used semantic-relatedness methods from topic segmentation to decide what RST-relation to assign to adjacent non-elementary text spans because he could find no other way to do so reliably. (Schilder, 2002) just assumed that RST-relations could only be computed reliably for elementary spans (ie, single clauses or sentences), and used semantic-relatedness methods for other decisions. More recently, (Louis et al., 2010) have shown that features based on RST text structures complement those from *discourse relations* when it comes to choosing sentences for extractive summaries that are similar to those chosen manually.

While these purely practical links between discourse structures clearly lead to better performance in applications, extensive improvements can, I think, only come with a more theoretically-grounded understanding of how the different types of discourse structure fit together.

References

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic Discourse Treebank. In *Proc. of the 7^{th} Int'l Conference on Language Resources and Evaluation*, Valletta, Malta.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Yves Bestgen. 2006. Improving text segmentation using Latent Semantic Analysis: A reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, 32(1):5–12.

Matthias Buch-Kromann and Iørn Korzen. 2010. The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. In *Proc. 4^{th} Linguistic Annotation Workshop*, pages 127–131, Uppsala, Sweden.

Harr Chen, S. R. K. Branavan, Regina Barzilay, and David Karger. 2009. Global models of document structure using latent permutations. In *Proc. HLT/NAACL*, pages 371–379.

Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent Semantic Analysis for text segmentation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 109–117.

Grace Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC Medical Informatics and Decision Making*, 10(9), February.

Robert Dale. 1992. *Generating Referring Expressions*. MIT Press, Cambridge MA.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343.

Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Proc. IEEE Conference on Semantic Computing (ICSC-08)*, Santa Clara CA.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proc. 41^{st} Annual Meeting of the ACL*, pages 562–569.

Matt Gerber, Joyce Chai, and Adam Meyers. 2009. The role of implicit argumentation in nominal srl. In *Proc. HLT/ACL*, pages 146–154, Boulder CO.

Barbara Grosz and Candy Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Marti Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proc. 32^{nd} Annual Meeting of the ACL*, pages 9–16.

Marti Hearst. 1997. TextTiling. *Computational Linguistics*, 23(1):33–64.

Kenji Hirohata, Naoki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientic abstracts using conditional random fields. In *Proc 3^{rd} Int'l Joint Conference on Natural Language Processing*, pages 381–388.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proc. 3^{rd} Int'l Conference on Language Resources and Evalution*, Las Palmas.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proc.7^{th} Conference on Language Resources and Evaluation*, Valletta, Malta.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proc. HLT-NAACL Workshop on BioNLP*, pages 65–72.

Ziheng Lin, Hwee Tou Ng, , and Min-Yen Kan. 2010. A PDTB-styled end-to-end discourse parser. Technical Report 1011.0835, TRB8/10, arXiv.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proc. SIGDIAL*, pages 147–156, Tokyo.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proc. 21^{st} International COLING Conference and 44^{th} Annual Meeting of the ACL*.

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

Prashanth Mannem, Rashmi Prasad, and Aravind Joshi. 2010. Question generation from paragraphs at upenn. In *Proc. 3rd Workshop on Question Generation (QG2010)*, Pittsburgh PA.

Daniel Marcu. 1996. Building up Rhetorical Structure Trees. In *Proc. AAAI*, pages 1069–1074, Portland OR.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts. *Computational Linguistics*, 26(3):395–448.

Larry McKnight and Padmini Srinivasan. 2003. Categorization of sentence types in medical abstracts. In *Proceedings of the AMIA Annual Symposium*, pages 440–444.

Lucie Mladová, Šárka Zikánová, and Eva Hajičová. 2008. From sentence to discourse: Building an annotation scheme for discourse based on the Prague Dependency Treebank. In *Proc. 6th Int'l Conference on Language Resources and Evaluation*.

Johanna Moore. 1995. *Participating in Explanatory Dialogues*. MIT Press, Cambridge MA.

Megan Moser and Johanna Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

M. O'Donnell, C. Mellish, J. Oberlander, and A. Knott. 2001. Ilex: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7:225–250.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The Hindi Discourse Relation Bank. In *Proc. 3rd ACL Language Annotation Workshop*, Singapore.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proc. 47th Meeting of the ACL and 4th Int'l Joint Conference on Natural Language Processing*.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of COLING*, Manchester.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proc. 47th Meeting of the ACL and 4th Int'l Joint Conference on Natural Language Processing*.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. Sentential structure and discourse parsing. In *In Proceedings of the ACL 2004 Workshop on Discourse Annotation*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and et al. 2008. The Penn Discourse Treebank 2.0. In *Proc. of the 6th Int'l Conference on Language Resources and Evaluation*, Marrakech.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Exploiting scope for shallow discourse parsing. In *Proc. 7th Int'l Conference on Language Resources and Evaluation*, Valletta, Malta.

Matthew Purver, Konrad Körding, Thomas Griffiths, and Joshua Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proc. 21st COLING and 44th Annual Meeting of the ACL*, pages 17–24, Sydney.

Patrick Ruch, Celia Boyer, Christine Chichester, Imad Tbahriti, Antoine Geissbühler, Paul Fabry, and et al. 2007. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2–3):195–200.

Frank Schilder. 2002. Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering*, 8(3):235–255.

Donia Scott and Clarisse Sieckenius de Souza. 1990. Getting the message across in RST-based text generation. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*, pages 47–73. Academic Press, London, England.

Penni Sibun. 1992. Generating text without trees. *Computational Intelligence*, 8(1):102–122.

Manfred Stede. 2004. The Potsdam Commentary Corpus. In *ACL Workshop on Discourse Annotation*, Barcelona, Spain.

Manfred Stede. 2008. Disambiguating rhetorical structure. *Research on Language and Computation*, 6:311–332.

Maite Taboada, J. Brooke, and Manfred Stede. 2009. Genre-based paragraph classification for sentiment analysis. In *Proc. 10th SIGDIAL Conference on Discourse and Dialogue*, pages 62–70, London.

Simone Teufel. 2010. *The Structure of Scientific Articles*. CSLI Publications, Stanford CA.

Nynke van der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker. 2011. Building a discourse-annotated Dutch text corpus. In *Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*, Gottingen.

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *Proc. Conference on Empirical Methods in Natural Language Processing*.

Ben Wellner. 2008. *Sequence Models and Ranking Methods for Discourse Parsing*. Ph.D. thesis, Brandeis University.

Nianwen Xue. 2005. Annotating discourse connectives in the Chinese Treebank. In *Proc. ACL Workshop in Frontiers in Annotation II*, Ann Arbor MI.

Deniz Zeyrek, Işın Demirşahin, Ayışığı Sevdik-Çallı, Hale Ögel Balaban, İhsan Yalçınkaya, and Ümut Deniz Turan. 2010. The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In *Proc. 4th Linguistic Annotation Workshop*, Uppsala, Sweden.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955