Discourse Processing for Text Analysis: Recent successes, current challenges

Bonnie Webber[1]

School of Informatics, University of Edinburgh, UK bonnie.webber@ed.ac.uk

Abstract. Computational discourse processing has come a long way in the 10 years since I spoke at ACL'2009 on *Discourse: Early problems, current successes, future challenges*. Much of this progress can be attributed to the vast amounts of textual data that have become available and to a concomitant weakening of theoretical commitments, so as to be able to use the data in information extraction, sentiment analysis, question answering, etc. Along with weakened commitments to the demands of particular theories, has been a greater willingness to consider what can be learned from textual data and from various forms of annotation, in English and in other languages as well.

This paper briefly summarizes (1) changing assumptions about discourse structure; (2) recent work on lexico-syntactic grounding of low-level discourse structure and frameworks for higher-level discourse structure that recognize differences in genre; and (3) suggestions for addressing some of the challenges still facing us. For more detail, the reader is encouraged to go to the references themselves.

Keywords: discourse processing · discourse structure · discourse relations.

1 Introduction

Discourse poses many challenges to text processing systems, beyond those posed by isolated clauses. Firstly, since one can refer to anything mentioned in previous clauses, even things mentioned only implicitly, resolving referring expressions becomes a challenge. Secondly, since there is information embodied in relations that hold between clauses or sentences or larger spans of text — relations that may be signalled explicitly in the discourse or left to inference — they need to be detected, so that the information in the relation can be extracted. Thirdly, since the reason for some piece of text being included in a discourse may reflect communicative goals that are specific to a particular genre, such goals also need to be modelled, recognized and applied to whatever information has been extracted.

My concern here is with discourse structure and discourse relations — what was assumed prior to 2009, how that has changed in the intervening years, and where we are now. A reader who would like a general introduction to discourse processing is referred to Stede's 2012 monograph [30]. A reader interested in discourse structure and its use in language technology prior to 2012 is referred to [35]. Finally, for examples illustrating points made in this brief paper, the reader is referred to the slides of this keynote available on the BIRNDL 2019 website.

2 Early assumptions about discourse and discourse processing

Two early computational assumption about discourse were (1) that it has a simple computational structure, specifiable in the form of a regular expression or context-free grammar (CFG), and (2) that that structure covered the entire text. This could be seen in McKeown's schemas [17], in work on topic segmentation of texts [3, 6, 11, 12, 15], in the tree-structured analyses that followed from Rhetorical Structure Theory (RST) [16] or from seeing a task as comprising a sequence of sub-tasks and a text describing how to carry it out as being similarly composed of a sequence of sub-texts [7]. In the resulting tree structure, the *left-to-right order* of the children of a non-terminal node would correspond to the temporal ordering of sub-tasks and *immediate dominance* between a parent and its children would correspond to sub-task inclusion. The text itself would correspond to simple top-down, L-R tree traversal.

These two assumptions were held so widely that any work on text structure that didn't conform to them was ignored. This was true of work by Sibun [28], which modelled spoken descriptions of complex structures such as house layouts as a *linear traversal of a complex graph*, which required both marking which nodes had already been visited (since they could be reached in multiple ways) and consulting a decision process when more than one node could be visited next. Although Sibun's view of text as a structure that systematically reflected the structure of the world was no different than Dale's, Sibun's work was ignored as not conforming to the view of discourse structure as a tree.

3 Issues at play since then

Subsequent to this early work, computational researchers began to acknowledge (1) that text structure varies with genre such that, for example, persuasive texts differ in structure from instructions, which both differ in structure from descriptive texts; and (2) that simplicity in text structure, whatever the genre, is just a useful simplification that will be violated when needed or else completely discarded. Instead, researchers have accepted different kinds of discourse structure and, adopting a more empirical perspective, have turned to looking at what provides evidence for discourse structure – in particular, lexico-syntactic evidence.

3.1 Multiple kinds of discourse structure

The earliest claim to the existence of multiple kinds of discourse structure was made by Grosz and Sidner [10], who posited a *linguistic structure* (signalled by discourse cues), an *intentional structure* (modelling how the purpose of one segment contributed to that of another), and an *attentional structure* (in the form of a *stack*, reflecting its origin in tree structures for discourse).

While Grosz and Sidner were primarily focussed on accounting for and modelling *intentional structure*, Moore and Pollack [19] wanted to break out of the requirement in RST [16] of only one sense relation holding between any two (adjacent) discourse segments. This often forced annotators to choose either a relation between the information conveyed in consecutive elements of a coherent discourse (*informational relations*) or a relation reflecting the aim of discourse to effect changes in the mental state of the discourse participants, through a textual plan whose consecutive elements relate in terms of their roles in the plan (*intentional relations*). Instead, Moore and Pollack proposed one discourse structure that reflected *informational relations* between the elements, and a separate one that reflected *intentional relations* between those same elements. Importantly they pointed out that these structures may not be *isomorphic*, even though they cover the same text. This could mean that a text segment that was prominent in one structure could be less so in the other.

Genre is clearly at play in the structures proposed for discourse. While the texts considered by Moore and Pollack [19] were persuasive texts, Knott and his colleagues [14] aimed to generate descriptions of objects in museum displays that were appropriate in the context of other objects that had already been described to the visitor and other objects in the same display. Although subscribing to RST [16], Knott and his colleagues had to face the problem that the texts they were modelling violated RST's assumption that the spans linked by a discourse relation had to be *adjacent*, or if interrupted by another span, had to be linked to the initial span by a relation of the same type.

However, after noticing that all violations of this assumption involved RST's OBJECT-ATTRIBUTE ELABORATION relation (where one segment presents an object, and the next presents one of its attributes), Knott et al [14] proposed a hybrid structure for discourse, taking it to be structured as a *sequence* of RST trees (minus ELABORATION), supplemented by an *entity-based model of focus structure*. That is, these descriptive texts were structured as a sequence of RST trees. each of whose top nodes focussed on some entity that had been mentioned previously and was then further described in the rest of its tree-structured segment.

More recently, researchers concerned with argumentation such as Stede and his colleagues [31], Stab and Gurevych [29] and others have been exploring how argumentation structure can be grounded in an RST-based coherence structure. This is currently a very active area of research, so links between other forms of discourse and dialogue structures are being explored as well.

3.2 Empirical bases for discourse structure

In the early 90s, researchers were exploring the idea that sentence-level syntactic structure was projected from structures associated with lexical items. This was true of both lexicalized Tree-Adjoining Grammar [37] and Combinatory Categorial Grammar [33]. This encouraged researchers to ask whether the same could hold of discourse and to build corpora based on the notion that low-level discourse structure was signalled either by explicit lexico-syntactic phrases or constructions or by adjacency that would leading readers to infer a relation between the adjacent units [1, 2, 20–22, 36, 39–43]. This in turn led to researchers developing lexicons of discourse connectives such as [8, 18, 26, 32] and even a re-annotation of the RST Corpus to identify the likely linguistic signals for the annotated relations [9].

4 Current challenges

To my mind, there are at least two areas in which initial efforts require more investment, in order to see a pay-off: (1) More general acceptance of segments contributing their semantics and pragmatics to multiple discourse relations; and (2) exploring the stance/sentiment associated with discourse connectives and discourse relations, so as to more accurately describe speakers' and writers' attitudes towards their subjects.

With respect to segments simultaneously linked by multiple sense relations, despite there being only one (or possibly even no) explicit discourse connective between them, evidence comes both from experiments using crowdsourcing [23–25] and from corpus annotation [22, 36]. Other evidence comes from cross-lingual parallel texts, which often differ in their signalling of discourse relations [27].

With respect to exploring the stance/sentiment associated with discourse connectives and discourse relations, there are some obvious examples, such as the preposition *thanks to*. While it clearly indicates that one clause is seen as the REASON for the other clause holding, as in

> *Operations are running smoothly* <u>thanks to</u> *decentralizing the company's computer system before the quake*

thanks to also indicates the speaker's positive attitude to what is expressed in the latter clause, which would not be evident if the phrase *as result of* had been used instead.

Another example is the connective *but then* (also phrased *but then again*). While it signals a CONCESSION relation, with one clause denying an expectation raised by the conceded clause, as in

> *To many, it was a ceremony more befitting a king than a rural judge seated in the isolated foothills of the southern Allegheny Mountains.* <u>But then</u> *Judge O'Kicki often behaved like a man who would be king – and, some say, an arrogant and abusive one.*

<u>but then (again)</u> also indicates that the speaker's attitude that the listener shouldn't <u>be surprised</u>.

While papers have been written about the contribution of discourse relations to the expression of sentiment (e.g, [4, 5, 13, 34], this should be complemented by aggregating descriptions of the range of sentiments (stances) conveyed by discourse connectives and in discourse relations across multiple languages.

With the appearance of new discourse annotated corpora such as the TED-MDB [41] and the expanded Penn Discourse TreeBank [36] and with new discourse-related Shared Tasks [38], discourse-based information should become more integral to Natural Language Processing and hence more available for use by any technologies such as Information Retrieval and Question Answering that use and thereby add value to text.

References

1. Al-Saif, A., Markert, K.: The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In: Proceedings, 7th International Conference on Language Resources and Evaluation (LREC 2010) (2010)
2. Al-Saif, A., Markert, K.: Modelling discourse relations for Arabic. In: Proceedings, Empirical Methods in Natural Language Processing. pp. 736–747 (2011)
3. Barzilay, R., Lee, L.: Catching the Drift: Probabilistic content models, with applications to generation and summarization. In: Proceedings of the 2nd Human Language Technology Conference and Annual Meeting of the North American Chapter, Association for Computational Linguistics. pp. 113–120 (2004)
4. Bhatia, P., Ji, Y., Eisenstein, J.: Better document-level sentiment analysis from rst discourse parsing. In: Proceedings, Empirical Methods in Natural Language Processing (EMNLP). pp. 2212–2218 (2015)
5. Chardon, B., Benamara, F., Mathieu, Y., Popescu, V., Asher, N.: Measuring the effect of discourse structure on sentiment analysis. In: Proceedings, 14^{th} International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013) (2013)
6. Chung, G.: Sentence retrieval for abstracts of randomized controlled trials. BMC Medical Informatics and Decision Making **10(9)** (February 2009)
7. Dale, R.: Generating Referring Expressions. MIT Press, Cambridge MA (1992)
8. Das, D., Scheffler, T., Bourgonje, P., Stede, M.: Constructing a lexicon of english discourse connectives. In: Proceedings of the 56^{th} Annual Meeting of the ACL (August 2018)
9. Das, D., Taboada, M.: Rst signalling corpus: a corpus of signals of coherence relations. Language Resources and Evaluation **52**, 149–184 (2018)
10. Grosz, B., Sidner, C.: Attention, intention and the structure of discourse. Computational Linguistics **12(3)**, 175–204 (1986)
11. Hearst, M.: TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics **23**(1), 33–64 (1997)
12. Hirohata, K., Okazaki, N., Ananiadou, S., Ishizuka, M.: Identifying sections in scientific abstracts using conditional random fields. In: Proceedings of the 3^{rd} International Joint Conference on Natural Language Processing. pp. 381–388 (2008)
13. Ji, Y., Smith, N.: Neural discourse structure for text categorization. In: Proceedings, Association for Computational Linguistics (ACL). pp. 996–1005 (2017)
14. Knott, A., Oberlander, J., O'Donnell, M., Mellish, C.: Beyond elaboration: The interaction of relations and focus in coherent text. In: Sanders, T., Schilperoord, J., Spooren, W. (eds.) Text Representation:Linguistic and psycholinguistic aspects, pp. 181–196. John Benjamins Publishing (2001)

15. Malioutov, I., Barzilay, R.: Minimum cut model for spoken lecture segmentation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (2006)
16. Mann, W., Thompson, S.: Rhetorical Structure Theory: Toward a functional theory of text organization. Text **8(3)**, 243–281 (1988)
17. McKeown, K.: Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Texts. Cambridge University Press, Cambridge, England (1985)
18. Mírovský, J., Synková, P., Rysová, M., Polílová, L.: Czedlex: A lexicon of czech discourse connectives. In: Prague Bulletin of Mathematical Linguistics. vol. 109, pp. 61–91 (2017)
19. Moore, J., Pollack, M.: A problem for RST: The need for multi-level discouse analysis. Computational Linguistics **18(4)**, 537–544 (1992)
20. Oza, U., Prasad, R., Kolachina, S., Sharma, D.M., Joshi, A.: The hindi discourse relation bank. In: Proc. 3^{rd} ACL Language Annotation Workshop (LAW III). Singapore (August 2009)
21. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse TreeBank 2.0. In: Proceedings, 6th International Conference on Language Resources and Evaluation. pp. 2961–2968. Marrakech, Morocco (2008)
22. Prasad, R., Webber, B., Joshi, A.: Reflections on the Penn Discourse Treebank, comparable corpora and complementary annotation. Computational Linguistics **40(4)**, 921–950 (2014)
23. Rohde, H., Dickinson, A., Schneider, N., Clark, C., Louis, A., Webber, B.: Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In: Proceedings of the Tenth Linguistic Annotation Workshop (LAW-X. pp. 49–58. Berlin (2016), http://www.aclweb.org/anthology/W16-1707
24. Rohde, H., Dickinson, A., Schneider, N., Clark, C., Louis, A., Webber, B.: Exploring substitutability through discourse adverbials and multiple judgments. In: Proceedings, 12th International Conference on Computational Semantics (IWCS 2017). Montpellier, France (2017)
25. Rohde, H., Johnson, A., Schneider, N., Webber, B.: Discourse coherence: Concurrent explicit and implicit relations. In: Proceedings of the 56^{th} Annual Meeting of the ACL (August 2018)
26. Roze, C., Danlos, L., Muller, P.: Lexconn: A french lexicon of discourse connectives. Discours **10** (2012)
27. Shi, W., Yung, F., Demberg, V.: Acquiring annotated data with cross-lingual explicitation for implicit discourse relation classification. In: Proceedings, Workshop on Discourse Relation Parsing and Treebanking (DISRPT) (2019)
28. Sibun, P.: Generating text without trees. Computational Intelligence **8(1)**, 102–122 (1992)
29. Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Computational Linguistics **43**, 619–659 (2017)
30. Stede, M.: Discourse Processing. Morgan & Claypool Publishers (2012)
31. Stede, M., Afantenos, S., Peldszus, A., Asher, N., Perret, J.: Parallel discourse annotations on a corpus of short texts. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC) (2016)
32. Stede, M., Umbach, C.: Dimlex: A lexicon of discourse markers for text generation and understanding. In: Proceedings, 36^{th} Annual Meeting of the ACL (1998)

33. Steedman, M.: Surface Structure and Interpretation. Linguistic Inquiry Monograph 30, MIT Press, Cambridge MA (1996)
34. Taboada, M., Voll, K., Brooke, J.: Extracting sentiment as a function of discourse structure and topicality. Tech. Rep. 2008-20, School of Computing Science, Simon Fraser University (2008)
35. Webber, B., Egg, M., Kordoni, V.: Discourse structure and language technology. Natural Language Engineering **18**(4), 437–490 (2012)
36. Webber, B., Prasad, R., Lee, A., Joshi, A.: The Penn Discourse Treebank 3.0 Annotation Manual. Tech. rep., University of Pennsylvania (2019), available at https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf
37. XTAG-Group, T.: A Lexicalized Tree Adjoining Grammar for English. Tech. Rep. IRCS 01-03, University of Pennsylvania (2001), see ftp://ftp.cis.upenn.edu/pub/ircs/technical-reports/01-03
38. Zeldes, A., Das, D., Maziero, E.G., Antonio, J., Iruskieta, M.: The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In: Proceedings, Workshop on Discourse Relation Parsing and Treebanking 2019. pp. 97–104. Minneapolis, MN (2019)
39. Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Ögel Balaban, H., İhsan Yalçınkaya, Turan, U.D.: The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In: Proceedings of the 4th Linguistic Annotation Workshop (LAW III) (2010)
40. Zeyrek, D., Kurfalı, M.: An assessment of explicit inter- and intra-sentential discourse connectives in turkish discourse bank. In: Proceedings of the 11th Language Resources and Evaluation Conference. European Language Resource Association, Miyazaki, Japan (May 2018), https://www.aclweb.org/anthology/L18-1634
41. Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., Ogrodniczuk, M.: Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. Language Resources and Evaluation (april 2019). https://doi.org/10.1007/s10579-019-09445-9
42. Zhou, Y., Xue, N.: Pdtb-style discourse annotation of chinese text. In: Proc. 50^{th} Annual Meeting of the ACL. Jeju Island, Korea (2012)
43. Zhou, Y., Xue, N.: The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. Journal of Language Resources and Evaluation **49**, 397–431 (2015)