Representing discourse coherence: A corpus-based analysis

Florian WOLF
MIT NE20-448
Cambridge, MA 02139, USA
fwolf@mit.edu

Edward GIBSON
MIT NE20-459
Cambridge, MA, 02139, USA
egibson@mit.edu

Abstract

We present a set of discourse structure relations that are easy to code, and develop criteria for an appropriate data structure for representing these relations. Discourse structure here refers to informational relations that hold between sentences in a discourse (cf. Hobbs, 1985). We evaluated whether trees are a descriptively adequate data structure for representing coherence. Trees are widely assumed as a data structure for representing coherence but we found that more powerful data structures are needed: In coherence structures of naturally occurring texts, we found many different kinds of crossed dependencies, as well as many nodes with multiple parents. The claims are supported by statistical results from a database of 135 texts from the Wall Street Journal and the AP Newswire that were hand-annotated with coherence relations, based on the annotation schema presented in this paper.

1 Introduction

An important component of natural language discourse understanding and production is having a representation of discourse structure. A coherently structured discourse here is assumed to be a collection of sentences that are in some relation to each other. This paper aims to present a set of discourse structure relations that are easy to code, and to develop criteria for an appropriate data structure for representing these relations.

Discourse structure relations here refer to informational relations that hold between sentences or other non-overlapping segments in a discourse monologue. That is, discourse structure relations reflect how the meaning conveyed by one discourse segment relates to the meaning conveyed by another discourse segment (cf. Hobbs, 1985; Marcu, 2000; Webber et al., 1999).

Accounts of discourse structure vary greatly with respect to how many discourse relations they assume, ranging from two (Grosz & Sidner, 1986) to over 400 different coherence relations, reported in Hovy and Maier (1995). However, Hovy and Maier (1995) argue that taxonomies with more relations represent subtypes of taxonomies with fewer relations. This means that different taxonomies can be compatible with each other.

We describe an account with a small number of relations in order to achieve more generalizable representations of discourse structures; however, the number is not so small that informational structures that we are interested in are obscured. The next section will describe in detail the set of coherence relations we use, which are mostly based on Hobbs (1985). Additionally, we try to make as few a priori theoretical assumptions about representational data structures as possible. These assumptions will be outlined in the next section. Importantly, however, we do not assume a tree data structure to represent discourse coherence structures. In fact, a major goal of this paper is to show that trees do not seem adequate to represent discourse structures.

2 Collecting a database of texts annotated with coherence relations

This section describes (1) how we define discourse segments, (2) which coherence relations we used to connect the discourse segments, and (3) how the annotation procedure worked.

2.1 Discourse segments

Discourse segments can be defined as non-overlapping spans of prosodic units (Hirschberg & Nakatani, 1996), intentional units (Grosz & Sidner, 1986), phrasal units (Lascarides & Asher, 1993), or sentences (Hobbs, 1985). We adopted a sentence unit-based definition of discourse segments. However, we also assume that contentful coordinating and subordinating conjunctions (cf. Table 1) can delimit discourse segments.

2.2 Coherence relations

We assume a set of coherence relations that is similar to that of Hobbs (1985) and Kehler (2002). Table 1 shows the coherence relations we assume, along with contentful conjunctions that can signal the coherence relation.

cause-effect	because
violated expectation	although; but
condition	if…then; as long as
similarity	(and) similarly
contrast	but; however
elaboration	also, furthermore
attribution	…said, according to…
temporal sequence	before; afterwards

Table 1. Coherence relations with contentful conjunctions for determining coherence relations.

Below are examples of each coherence relation.

(1) Cause-Effect

[There was bad weather at the airport]$_a$ [and so our flight got delayed.]$_b$

(2) Violated Expectation

[The weather was nice]$_a$ [but our flight got delayed.]$_b$

(3) Condition

[If the new software works,]$_a$ [everyone will be happy.]$_b$

(4) Similarity

[There is a train on Platform A.]$_a$ [There is another train on Platform B.]$_b$

(5) Contrast

[John supported Bush]$_a$ [but Susan opposed him.]$_b$

(6) Elaboration

[A probe to Mars was launched this week.]$_a$ [The European-built 'Mars Express' is scheduled to reach Mars by late December.]$_b$

(7) Attribution

[John said that]$_a$ [the weather would be nice tomorrow.]$_b$

(8) Temporal Sequence

[Before he went to bed,]$_a$ [John took a shower.]$_b$

The *same* relation, illustrated by (9), is an epiphenomenon of assuming contiguous distinct elements of text. (a) is the first segment and (c) is the second segment of what is actually one single discourse segment, separated by the intervening discourse segment (b), which is in an *attribution* relation with (a) (and therefore also with (c), since (a) and (c) are actually one single discourse segment).

(9) Same

[The economy,]$_a$ [according to some analysts,]$_b$ [is expected to improve by early next year.]$_c$

Cause-effect, *violated expectation*, *condition*, *elaboration*, *temporal sequence*, and *attribution* are asymmetrical or directed relations, whereas *similarity*, *contrast*, *temporal sequence*, and *same* are symmetrical or undirected relations (Mann & Thompson, 1988; Marcu, 2000). The directions of asymmetrical or directed relations are as follows: cause → effect for *cause-effect*; cause → absent effect for *violated expectation*; condition → consequence for *condition*; elaborating → elaborated for *elaboration*, and source → attributed for *attribution*.

2.3 Coding procedure

In order to code the coherence relations of a text, annotators used a procedure consisting of three steps. In Step One, a text is segmented into discourse segments as described above. In Step Two, adjacent discourse segments that are topically related are grouped together. For example, if a text discusses inventions in information technology, there could be groups of a few discourse segments each talking about inventions by specific companies. There might also be subgroups of several discourse segments each talking about specific inventions at specific companies. Thus, marking groups determines a partially hierarchical structure for the text. In Step Three, coherence relations are determined between discourse segments and groups of discourse segments. Each previously unconnected (group of) discourse segment(s) is tested to see if it connects to any of the (groups of) discourse segments in the already existing representation of discourse structure.

In order to help determine the coherence relation between (groups of) discourse segments, the (groups of) discourse segments under consideration are connected with a contentful conjunction like the ones shown in Table 1. If using a contentful conjunction to connect (groups of) discourse segments results in an acceptable passage, this is used as evidence that the coherence relation corresponding to the contentful conjunction holds between the (groups of) discourse segments under consideration.

2.4 Statistics on annotated database

In order to evaluate hypotheses about appropriate data structures for representing coherence structures, we annotated 135 texts, from the Wall Street Journal 1987-1989 and the AP Newswire 1989 (Harman & Liberman, 1993), with the coherence relations described above. For the 135 texts, the mean number of words was 545 (min.: 161; max.: 1409; median: 529), the mean number of discourse segments was 61 (min.: 6; max.: 143; median: 60).

Each text was independently annotated by two annotators. In order to determine inter-annotator agreement for the database of annotated texts, we computed kappa statistics (Carletta, 1996). For all annotations of the 135 texts, the agreement was 88.45%, per chance agreement was 24.86%, and kappa was 84.63%. Annotator agreement did not differ by text length ($\chi^2 = 1.27$; $p < 0.75$), arc length ($\chi^2 < 1$), or kind of coherence relation ($\chi^2 < 1$).

3 Data structures for representing coherence relations

Most accounts of discourse coherence assume tree structures to represent coherence relations between discourse segments in a text (Carlson et al., 2002; Corston-Oliver, 1998; Lascarides & Asher, 1993; Longacre, 1983; Grosz & Sidner, 1986; Mann & Thompson, 1988; Marcu, 2000; Polanyi, 1988; van Dijk & Kintsch, 1983; Walker, 1998; Webber et al., 1999). Other accounts assume less constrained graphs (Hobbs, 1985). The proponents of tree structures argue that trees are easier to formalize and derive than less constrained graphs (Marcu, 2000). We tested whether coherence structures of naturally occurring texts can be represented by trees, i.e. if these structures are free of crossed dependencies or nodes with multiple parents. However, we found a large number of both crossed dependencies as well as nodes with multiple parents in the coherence structures of naturally occurring texts. Therefore we argue for less constrained graphs as an appropriate data structure for representing coherence, where an ordered array of nodes represents discourse segments and labeled directed arcs represent the coherence relations that hold between these discourse segments.[1] The following two sections will give examples of coherence structures with crossed dependencies and nodes with multiple parents. The section after that will present statistical results from our database of 135 coherence-annotated texts.

3.1 Crossed dependencies

Crossed dependencies are rampant and occur in many different forms in the coherence structures of naturally occurring texts. Here we will give some examples. Consider the text passage in (10). Figure 1 represents the coherence relations in (10). The arrowheads of the arcs represent directionality for asymmetrical relations (*elaboration*) and bidirectionality for symmetrical relations (*contrast*).

(10) Example text (from SAT practicing materials)
0. Schools tried to teach students history of science.
1. At the same time they tried to teach them how to think logically and inductively.
2. Some success has been reached in the first of these aims.
3. However, none at all has been reached in the second.

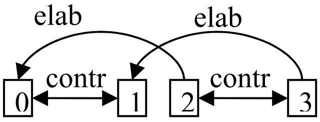

Figure 1. Coherence graph for (10).

The coherence structure for (10) can be derived as follows: there is a *contrast* relation between 0 and 1; 0 and 1 describe teaching different things to students. There is another *contrast* relation between 2 and 3; 2 and 3 describe varying degrees of success (some vs. none). 2 provides more details (the degree of success) about the teaching described in 0, so there is an *elaboration* relation between 2 and 0. Furthermore, in another *elaboration* relation, 3 provides more details (the degree of success) about the teaching described in 1. In the resultant coherence structure for (10), there is a crossed dependency between {2, 0} and {3, 1}.

In order to be able to represent the crossed dependency in the coherence structure of (10) in a tree without violating validity assumptions about tree structures, one might consider augmenting a tree with feature propagation (Shieber, 1986) or with a coindexation mechanism (Chomsky, 1973). But the problem is that both the tree structure itself as well as the features and coindexations represent the same kind of information (coherence relations). It is unclear how one could decide which part of a text coherence structure should be represented by the tree structure and which by the augmentation.

As pointed out above, coherence structures of naturally occurring texts contain many different kinds of crossed dependencies. This is important because it means that one cannot simply make special provisions to account for list-like structures like the structure of (10) and otherwise assume tree structures. As an example of a non-list-like structure with a crossed dependency (between {3, 1} and {2, 0-1}), consider (11).

[1] Other accounts also acknowledge examples that cannot be represented in tree structures (Webber et al., 1999). In order to maintain trees, these accounts distinguish non-anaphoric coherence structures, represented in a tree, and anaphoric coherence structures, which are not subject to tree constraints. However, e.g., Haliday & Hasan (1976) stress the importance of anaphoric links as a cue for coherence structures. Therefore, by Occam's Razor, we assume a single level of representation for coherence rather than multiple levels.

(11) Example text
0. Susan wanted to buy some tomatoes
1. and she also tried to find some basil
2. because her recipe asked for these ingredients.
3. The basil would probably be quite expensive at this time of the year.

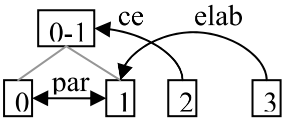

Figure 2. Coherence graph for (11).

The coherence structure for (11) can be derived as follows: there is a *parallel* relation between 0 and 1; 0 and 1 both describe shopping for grocery items. There is a *cause-effect* relation between 2 and 0-1; 2 describes the cause for the shopping described by 0 and 1. Furthermore, there is an *elaboration* relation between 3 and 1; 3 provides details about the basil in 1.

(12) from the AP Newswire1989 corpus is an example with a similar structure:

(12) Example text (from text ap890109-0012)
0. The flight Sunday took off from Heathrow Airport at 7:52pm
1. and its engine caught fire 10 minutes later,
2. the Department of Transport said.
3. The pilot told the control tower he had the engine fire under control.

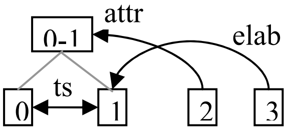

Figure 3. Coherence graph for (12).

The coherence structure for (12) can be derived as follows: 1 and 0 are in a *temporal sequence* relation; 0 describes the takeoff that happens before the engine fire described by 1 occurs. 2 and 0-1 are in an *attribution* relation; 2 mentions the source of what is said in 0-1. 3 and 1 are in an elaboration relation; 3 provides more detail about the engine fire in 1. The resulting coherence structure, shown in Figure 3, contains a crossed dependency between {3, 1} and {2, 0-1}.

3.2 Nodes with multiple parents

In addition to crossed dependencies, many coherence structures of natural texts include nodes with multiple parents. Such nodes cannot be represented in tree structures. For instance, in the coherence structure of (10), nodes 0 and 2 have two parents. Similarly, in the coherence structure of (13) from the AP Newswire 1989, node 1 has one *attribution* and one *condition* ingoing arc (cf. Figure 4).

(13) Example text (from text ap890103-0014)
0. “Sure I’ll be polite,”
1. promised one BMW driver
2. who gave his name only as Rudolf.
3. “As long as the trucks and the timid stay out of the left lane.”

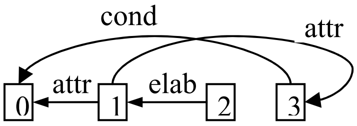

Figure 4. Coherence graph for (13).

The coherence structure for (13) can be derived as follows: 1 states the source of what is stated in 0 and in 3, so there are *attribution* relations between 1 and 0 and 1 and 3 respectively. 2 and 1 are in an *elaboration* relation; 2 provides additional detail about the BMW driver in 1. 3 and 0 are in a *condition* relation; 3 states the BMW driver’s condition for being polite, stated in 0; the *condition* relation is also indicated by the phrase “as long as”.

4 Statistics

4.1 Crossed dependencies

An important question is how frequent the phenomena discussed in the previous sections are. The more frequent they are, the more urgent the need for a data structure that can adequately represent them.

This section reports counts on crossed dependencies in the annotated database of 135 texts. In order to track the frequency of crossed dependencies for the coherence structure graph of each text, we counted the minimum number of arcs that would have to be deleted in order to make the coherence structure graph free of crossed dependencies (i.e. the minimum number of arcs that participate in crossed dependencies). The example graph in Figure 10 illustrates this process. This graph contains the following crossed dependencies: {1, 3} crosses with {0, 2} and {2, 4}. By deleting {1, 3}, both crossed dependencies can be eliminated. The crossed dependency count for the graph in Figure 5 is thus “one”.

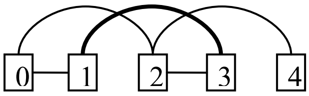

Figure 5. Example graph with crossed dependencies.

On average for the 135 annotated texts, 12.5% of arcs in a coherence graph have to be deleted in order to make the graph free of crossed dependencies (min.: 0%; max.: 44.4%; median: 10.9%). Seven texts out of 135 had no crossed dependencies. The mean number of arcs for the coherence graphs of these texts was 36.9 (min.: 8; max.: 69; median: 35). The mean number of arcs for the other 128 coherence graphs (those with crossed dependencies) was 125.7 (min.: 20; max.: 293; median: 115.5). Thus, the graphs with no crossed dependencies have significantly fewer arcs than those graphs that have crossed dependencies (χ^2=15330.35; $p < 10^{-4}$). Text length is hence a likely explanation for why these seven texts had no crossed dependencies.

Linear regressions show that the more arcs a graph has, the higher the number of crossed dependencies ($R^2 = 0.39$; $p < 10^{-4}$). Also, the longer a text, the more crossed dependencies are in its coherence structure graph (for text length in discourse segments: $R^2 = .29$, $p < 10^{-4}$; for text length in words: $R^2 = .24$, $p < 10^{-4}$).

Another important question is whether certain types of coherence relations participate more or less frequently in crossed dependencies than other types of coherence relations. In other words, the question is whether the frequency distribution over types of coherence relations is different for arcs participating in crossed dependencies compared to the overall frequency distribution over types of coherence relations in the whole database.

Results from our database indicate that the overall distribution over types of coherence relations participating in crossed dependencies is not different from the distribution over types of coherence relations overall. This is confirmed by a linear regression, which shows a significant correlation between the two distributions of percentages ($R^2 = 0.84$; $p < .0001$). Notice that the overall distribution includes only arcs with length greater than one, since arcs of length one could not participate in crossed dependencies.

However, some types of coherence relations occur considerably less frequently in crossed dependencies than overall in the database. The proportion of *same* relations is 15.21 times greater, and the percentage of *condition* relations is 5.93 times greater overall than in crossed dependencies. We do not yet understand the reason for these differences, and plan to address this question in future research.

Another question is how great the distance or arc length typically is between sentences that participate in crossed dependencies. It is possible, for instance, that crossed dependencies primarily involve long-distance arcs and that more local crossed dependencies are disfavored. However, the distribution over arc lengths is practically identical for the overall database and for coherence relations participating in crossed dependencies ($R^2 = 0.937$; $p < 10^{-4}$), with short-distance relations being more frequent than long-distance relations for coherence relations overall as well as for those participating in crossed dependencies. The arc lengths are normalized in order to take into account the length of a text; the absolute length of an arc is divided by the maximum length that that arc could have, given its position in a text. Furthermore, we exclude arcs of (absolute) length 1 from the overall distribution, since such arcs could not participate in crossed dependencies.

Taken together, statistical results on crossed dependencies suggest that crossed dependencies are too frequent to be ignored by accounts of coherence. Furthermore, the results suggest that any type of coherence relation can participate in a crossed dependency. However, there are some cases where knowing the type of coherence relation that an arc represents can be informative as to how likely that arc is to participate in a crossed dependency. The statistical results reported here also suggest that crossed dependencies occur primarily locally, as evidenced by the distribution over lengths of arcs participating in crossed dependencies.

4.2 Nodes with multiple parents

Above we provided examples of coherence structure graphs that contain nodes with multiple parents. Nodes with multiple parents are another reason why trees are inadequate for representing natural language coherence structures. The mean in-degree (=mean number of parents) of all nodes in the investigated database of 135 texts is 1.6 (min.: 1; max.: 12; median: 1). 41% of all nodes in the database have an in-degree greater than 1. This suggests that even if a mechanism could be derived for representing crossed dependencies in (augmented) tree graphs, nodes with multiple parents present another significant problem for trees representing coherence structures. Results from our database indicate that the overall distribution over types of coherence relations ingoing to nodes with multiple parents is significantly correlated with the distribution over types of coherence relations overall ($R^2 = 0.967$; $p < 10^{-4}$).

As for crossed dependencies, we also compared arc lengths. Here, we compared the length of arcs that are ingoing to nodes with multiple parents to the overall distribution of arc length. Again, we compared normalized arc lengths. By contrast to the comparison for crossed dependencies, we included arcs of (absolute) length 1 because such arcs can be ingoing to nodes with either single or multiple parents. The distribution over arc lengths is practically identical for the overall database and for arcs ingoing to nodes with multiple parents (R^2

= 0.993; $p < 10^{-4}$), suggesting a strong locality bias for coherence relations overall as well as for those participating in crossed dependencies.

In sum, statistical results on nodes with multiple parents suggest that they are a frequent phenomenon, and that they are not limited to certain kinds of coherence relations. Additionally, the statistical results reported here suggest that ingoing arcs to nodes with multiple parents are primarily local.

5 Conclusion

The goals of this paper have been to present a set of coherence relations that are easy to code, and to illustrate the inadequacy of trees as a data structure for representing discourse coherence structures. We have developed a coding scheme with high inter-annotator reliability and used that scheme to annotate 135 texts with coherence relations. An investigation of these annotations has shown that discourse structures of naturally occurring texts contain various kinds of crossed dependencies as well as nodes with multiple parents. Both phenomena cannot be represented using trees, which implies that existing databases of coherence structures that use trees are not descriptively adequate.

Our statistical results suggest that crossed dependencies and nodes with multiple parents are not restricted phenomena that could be ignored or accommodated with a few exception rules. Furthermore, even if one could find a way of augmenting tree structures to account for crossed dependencies and nodes with multiple parents, there would have to be a mechanism for unifying the tree structure with the augmentation features. Thus, in terms of derivational complexity, trees would just shift the burden from having to derive a less constrained data structure to having to derive a unification of trees and features or coindexation.

Because trees are neither a descriptively adequate data structure for representing coherence structures nor easier to derive, we argue for less constrained graphs as a data structure for representing coherence structures. Such less constrained graphs would have the advantage of being able to adequately represent coherence structures in one single data structure (cf. Skut et al., 1997). Furthermore, they are at least not harder to derive than (augmented) tree structures. The greater descriptive adequacy might in fact make them easier to derive. However, this is still an open issue and will have to be addressed in future research.

References

Jean Carletta. 1996. Assessing agreement on classifi-cation tasks: the kappa statistic. *Computational Linguistics, 22*(2): 249-254.

Lynn Carlson, Daniel Marcu, and Mary E. Okurowski. 2002. *RST Discourse Treebank*. Philadelphia, PA: LDC.

Noam Chomsky. 1973. Conditions on transformations. In: Anderson, S. & Kiparsky, P., eds., *A Festschrift for Morris Halle*, 232-286. New York: Holt, Rinehart and Winston.

Simon Corston-Oliver. 1998. *Computing representations of the structure of written discourse*. Microsoft Research Technical Report MSR-TR-98-15. Redmont, WA, USA.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics, 12*(3): 175-204.

Michael A.K. Haliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Donna Harman and Mark Liberman. 1993. *TIPSTER complete*. Philadelphia, PA: LDC.

Marti Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics, 23*(1): 33-64.

Julia Hirschberg and Christine H. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In: *Proceedings of the 34th Annual Meeting of the ACL*, 286-293. Santa Cruz, CA, USA.

Jerry R. Hobbs. 1985. *On the coherence and structure of discourse*. CSLI Technical Report 85-37. Stanford, CA, USA.

Eduard Hovy and Elisabeth Maier. 1995. *Parsimonious or profligate: How many and which discourse relations?* Unpublished manuscript.

Andrew Kehler. 2002. *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI Publications.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations, and common sense entailment. *Linguistics and Philosophy, 16*(5): 437-493.

Robert E. Longacre. 1983. *The grammar of discourse*. New York: Plenum Press.

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

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. Cambridge, MA: MIT Press.

Mitchell Marcus, Grace Kim, Mary A. Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In: *Proceedings of the ARPA Human Language Technology Workshop*. San Francisco, CA: Morgan Kaufman.

Livia Polanyi. 1988. A formal model of the structure of discourse. *Journal of Pragmatics, 12*: 601-638.

Stuart M. Shieber. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy, 8*: 333-343.

Stuart M. Shieber. 1986. *An introduction to unification-based approaches to grammar*. Stanford University: CSLI Lecture Notes 4.

Wojciech Skut, Brigitte Krenn, Thorsten Brants and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In: *Proceedings of the 5th ANLP Conference*. Washington, DC, USA.

Teun A. van Dijk and Walter Kintsch. 1983. *Strategies of discourse comprehension*. New York: Academic.

Marilyn A. Walker. 1998. Centering, anaphora resolution, and discourse structure. In: Prince, E., Joshi, A.K. & Walker, M.A., eds., *Centering Theory in discourse*. Oxford: Oxford University Press.

Bonnie L. Webber, Alastair Knott, Stone, M. & Joshi, A.K. 1999. Discourse relations: A structural and presuppositional account using lexicalized TAG. In: *Proceedings of the 37th Annual Meeting of the ACL*, 41-48. College Park, MD, USA.