Multiple Signals of Coherence Relations

Debopam Das and Maite Taboada

Introduction

Coherence relations (relations between text segments, also known as discourse or rhetorical relations) are often signalled by discourse markers[1] (henceforth DMs). Besides DMs, coherence relations can also be indicated by other textual signals, such as syntactic, semantic, graphical or genre-related features (Das & Taboada, 2017). For example, a syntactic signal such as *parallel syntactic construction* can indicate a *List* relation, or a graphical feature such as *semi-colon* can act as a signal for an *Elaboration* relation. Furthermore, in a great many cases, relations can simultaneously be indicated by multiple signals; for example, a *Contrast* relation can be signalled by a DM like *but*, and also by a semantic feature like *antonyms* in the respective text segments at the same time.

In this paper, we investigate the signalling of coherence relations by multiple signals, especially when relations are accompanied by both a DM and some other signal(s). We examine what DMs frequently occur with other signals, and also what relations involve signalling by other signals in addition to DMs. Based on the corpus evidence of multiple signals in a discourse-annotated corpus, we analyze the semantics of co-occurring DMs and other signals, and also examine the semantics and functions of coherence relations that are frequently realized by multiple signals.

DMs are generally considered to be the most prototypical signals of coherence relations. Yet the fact that relations are indicated by other signals in addition to DMs raises a few important questions: (1) Do DMs that co-occur with other signals constitute weak or underspecified signals, so that the relations which already include those DMs also require some sort of extra signalling by other devices? (2) Could the use of multiple signals (specifically with DMs) be triggered not by the signal types, but by the relation types themselves? This could be the case because relations often differ in terms of intention, subjectivity, veridicality, polarity or other features and some relation types may necessitate additional signalling.

In order to address these questions, we begin with a two-fold hypothesis: (A) Because some DMs are inherently ambiguous (or underspecified) and they can indicate more than one relation type (e.g., the DM *and* can signal both *List* and *Elaboration* relations), relations which include these DMs also need the presence of some other signal(s) to convey their specific meanings and intended effect; (B) Some relations types require multiple signals in cases where it is important to distinguish, e.g., between a semantic and pragmatic reading of a relation, and the other signals help express that difference.

Methodology

We conduct a study based on a corpus analysis of multiple signals in the RST Signalling Corpus (Das et al., 2015). The RST Signalling Corpus (henceforth RST-SC) is a corpus developed over the RST Discourse Treebank (henceforth RST-DT) (Carlson et al., 2002), and is annotated for signals of coherence relations already present in the RST-DT. The RST-SC provides annotation for DMs as well as for a wide range of other signals such as reference, lexical, semantic, syntactic, graphical and genre-related features which function as potential indicators of coherence relations.

The RST-SC is annotated with UAM CorpusTool (O'Donnell, 2008), which is also used to view the annotations in the corpus, and to search for instances of particular signals or relations present therein. We extract instances of those DMs that co-occur with other signal(s) in the corpus. Furthermore, we separately extract instances of those relations that contain multiple signals including DMs. We closely examine the semantics of co-occurring DMs and other signals on the one hand, and

[1] DMs are also known by (or comparable to) terms like discourse connectives, discourse relational devices or cue phrases.

study the semantics and discourse functions of coherence relations containing DMs and other signals on the other hand.

Results and Discussion

The RST-SC includes annotations of 21,400 relations, out of which 1,616 relation instances (7.55% of all relations) contain both a DM and some other signal(s). Our analysis shows that over 20 DMs (with a minimum of 10 instances) co-occur with other signal(s). The most frequently occurring DMs in this category are *and (631), but (309), while (73), however (56)* and *because (52)*[2]. We observe that most of these DMs are highly ambiguous in their signalling nature. For example, the distribution of the DM *while* illustrates that the DM, when accompanied by some other signal(s), is used to indicate *seven* different relation types, including *Comparison, List, Contrast and Antithesis*. The distribution is even wider for the more frequent DMs: *and* and *but*[3]. On the other hand, we find that over 15 relation types (with a minimum of 10 instances) contain multiple signals including a DM. The most common relation types are *List (533), Elaboration-addition (221), Contrast (167), Antithesis (128)* and *Circumstance (90)*.

The finding that multiple signals frequently include ambiguous DMs seems to indicate that DMs which signal different relations might be underspecified in their meaning, and as a result, relations might require extra signalling by other signals in addition to those DMs. In other words, the ambiguity of DMs is compensated by using other signals. We also examine if the set of DMs co-occurring with other signals in the RST-SC corresponds to any categories or taxonomies mentioned in the literature (Halliday & Hasan, 1976; Sanders et al., 1992), and discuss relevant observations.

We also study the relations that include a DM and some other signal(s), with respect to features such as subjectivity (semantic-pragmatic) and additive-causal distinction. In Rhetorical Structure Theory or RST (Mann & Thompson, 1988), relations, based on their intended effect on the reader, are grouped into *subject matter relations* (the intended effect is the reader recognizes the relations) and *presentational relations* (the intended effect is to increase some inclination in the reader, such as positive regard, belief, or acceptance of the nucleus). We observe that our extracted relations mostly belong to the class of subject matter relations (*List, Elaboration-additional, Contrast* or *Circumstance*). This may imply that if the goal of the writer is to make the reader understand the relation, this intended effect might necessitate the use of extra signals besides DMs. On the other hand, according to the *basic operation* feature as proposed in the Cognitive approach of Coherence Relations or CCR (Sanders et al., 1992), these relations are mostly *additive* relations rather than *causal* (or implicational) ones. Furthermore, as CCR suggests that additive relations are weakly connected relations (in contrast to causal relations, which are strongly connected), our findings seem to indicate that relations that are weakly connected might require additional signalling beyond DMs.

One of our future goals includes exploring the categories of possible combination types of DMs and other signals. In a recent work, Hoek et al. (2018) suggest a three-way classification of the combinations of DMs and segment-internal elements[4]: division of labor, agreement and general collocation. We are interested to see if this (or a modified version) also holds for DM plus other signal combinations in the RST-SC. Another challenging area in our future research is the investigation of multiple signals vs. no signals[5] for the otherwise same relation types. It would be worthwhile to study what might necessitate same relation types to employ signalling by multiple signalling devices as opposed to no signals at all.

[2] The number within parentheses denotes the number of instances in the RST-SC.
[3] Knott (1996) also finds *and* and *but* to constitute more general or superordinate classes of cue phrases which, in their meanings, include many specific relational indicators (e.g., *but* → *despite this, whereas*).
[4] Segment-internal elements roughly constitute a subset of other signals in the RST-SC.
[5] In the RST-SC, there are 1,553 relations (7.26% of all the relations) without any identifiable signals.

References

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). RST Discourse Treebank, LDC2002T07. from https://catalog.ldc.upenn.edu/LDC2002T07

Das, D., & Taboada, M. (2017). Signalling of Coherence Relations in Discourse, Beyond Discourse Markers. *Discourse Processes*, 1-29. doi: 10.1080/0163853X.2017.1379327

Das, D., Taboada, M., & McFetridge, P. (2015). RST Signalling Corpus, LDC2015T10. from https://catalog.ldc.upenn.edu/LDC2015T10

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London, UK: Longman.

Hoek, J., Zufferey, S., Evers-Vermeul, J., & Sanders, T. (2018). *The linguistic marking of coherence relations: The interaction between segment-internal elements and connectives.* Paper presented at the TextLink2018 - Final Action Conference, Toulouse, France.

Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations.* (Ph.D. dissertation), University of Edinburgh, Edinburgh, UK.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.

O'Donnell, M. (2008). *The UAM CorpusTool: Software for corpus annotation and exploration.* Paper presented at the the XXVI Congreso de AESLA, Almeria, Spain.

Sanders, T., Spooren, W., & Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes, 15*, 1-35.