Discourse Relations and Signaling Information: Anchoring Discourse Signals in RST-DT

Yang Liu
Department of Linguistics
Georgetown University
yl879@georgetown.edu

Amir Zeldes
Department of Linguistics
Georgetown University
amir.zeldes@georgetown.edu

1 Introduction

Research on discourse relations between clauses or sentences, such as *cause* or *contrast*, has studied how such relations are established or signaled in discourse. Several corpora exist with discourse relation signaling information such as the Penn Discourse Treebank (PDTB, Prasad et al. 2008) and the Rhetorical Structure Theory Signalling Corpus (RST-SC, Taboada and Das 2013), which added signaling annotations to the RST Discourse Treebank (RST-DT, Carlson et al. 2002). By 'signaling' we refer to identifying the means by which humans recognize that relations hold, e.g. a discourse marker, including a connective such as 'but' or an adverbial 'however' marking *contrast*.

In addition to different inventories of relations, discourse annotation frameworks differ in segmentation and analysis strategies, which Polakova et al. (2017) term "local" and "global" approaches in PDTB and RST-DT respectively. PDTB anchors signaling information by marking explicit connectives such as 'because', 'although' etc. (including for non-adjacent clauses) and implicit relations not actually introduced by connectives in the text in adjacent units within the same paragraph. Discourse annotation is therefore applied to small-scale local structures (pairs of clauses). RST-SC, by contrast, annotates the existence of signaling information between any two units (e.g. clauses) or groups of units (possibly a paragraph or multiple paragraphs), including syntactic, morphological, semantic and other signals, thereby admitting any types of signaling devices, beyond just discourse markers. As a result, RST-SC provides much more signaling information; however, unlike PDTB, it does not anchor signals to tokens: annotations for a relation state only that it is signaled by a certain signal type (e.g. 'lexical/indicative_word') without marking relevant word positions in the text.

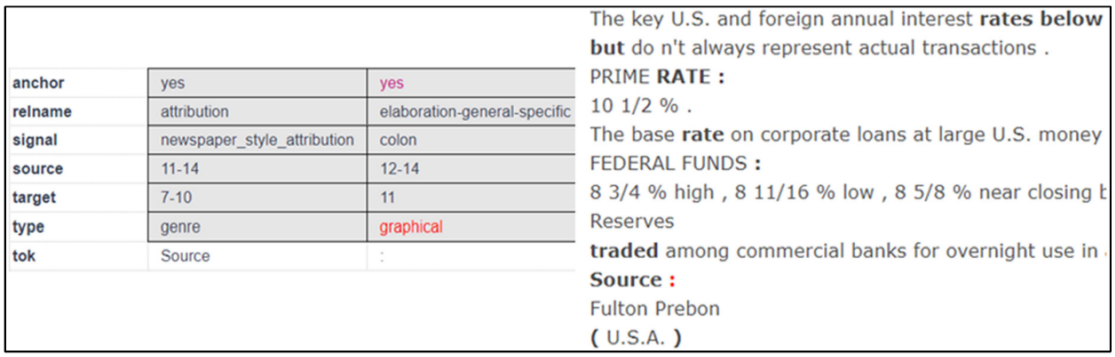
The present project therefore presents an annotation effort to anchor discourse signals at all levels (elementary and complex units), which is open to all types of signals of coherence relations based on RST-SC, with the intention of bridging the gap between the two aforementioned frameworks. Figure 1 illustrates our annotations (see Section 3). Our results show that over 92% of discourse signals can be anchored to specific tokens in the text, with the signal type *semantic* representing the most cases (41.7% of signaling anchors) whereas discourse relations anchored by discourse markers are only about 8.5% of the signal anchoring tokens.

2 Goals

Theoretical frameworks for signaling annotation are of great interest since they provide insights in both Linguistics and Natural Language Processing: from a psycholinguistic point of view, we would like to know how readers recognize relations to obtain almost the same interpretations given the same text. From a computational perspective, understanding signals can help with feature engineering for automatic discourse parsing.

A key limitation of PDTB is that signaling information only covers discourse connectives in three categories: subordinating conjunctions (e.g. 'because'), coordinating conjunctions (e.g. 'and'), and adverbials (e.g. 'instead') (Prasad et al., 2008). However, Taboada and Das (2013) found that such markers signal only 22% of relations in RST-DT, with the remainder being more complex. Thus, it is necessary and important to develop a scheme to include all other types of signals. By anchoring RST-SC annotations to specific tokens, we aim to study the types of signal anchors in the corpus and their distribution, as well as to establish

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 314-317.
New York City, New York, January 3-6, 2019



anchor	yes	yes
relname	attribution	elaboration-general-specific
signal	newspaper_style_attribution	colon
source	11-14	12-14
target	7-10	11
type	genre	graphical
tok	Source	:

The key U.S. and foreign annual interest **rates below** **but** do n't always represent actual transactions .
PRIME **RATE :**
10 1/2 % .
The base **rate** on corporate loans at large U.S. money
FEDERAL FUNDS **:**
8 3/4 % high , 8 11/16 % low , 8 5/8 % near closing t
Reserves
traded among commercial banks for overnight use in
Source :
Fulton Prebon
(U.S.A.)

Figure 1: Left: Anchored signal annotations for a genre specific newspaper-style attribution (the word *Source*), and a colon used as a graphical signal of elaboration; Right: Excerpt highlighting anchored signals.

what proportion of relations is signaled by means that are attributable to a specific span of tokens in the text.

3 Annotation Scheme

This pilot study annotated 11 Wall Street Journal documents with 4,732 tokens in the RST-SC corpus. Two annotators developed the scheme as they annotated the documents and adjudicated questionable cases. The annotation was done using the GitDOX interface (Zhang and Zeldes, 2017). For each instance of a signal, the following categories are annotated (cf. Figure 1):

- **Type**: RST-SC types, e.g. *syntactic, genre*
- **Signal**: sub-categories of RST-SC types such as *relative clause, tense* etc.
- **Anchor**: do tokens correspond to the signal?
- **Source/Target**: the related discourse units
- **Relname**: the relation name being signaled
- **Discontinuous**: a co-index for signals anchored to discontinuous token spans

According to the RST Signalling Corpus (Taboada and Das, 2013), signals can be *single*, *combined*, *multiple* or *unsure*. The category *single* is self-explanatory: the discourse relation is signaled by one and only one type of signal; the category *combined* means that two or more *single* signals are combined with each other in order to jointly signal the relation; the category *multiple* means that a discourse relation can be signaled by different kinds of signals independently; and the category *unsure* is used to indicate that no signals seem to signal the relation clearly. Moreover, the *single* signal types include *discourse markers (DMs), reference, lexical, semantics, morphological, syntactic, graphical, genre*, and *numerical*. The attested *combined* signals in the corpus are *reference+syntactic*, *semantic+syntactic*, *lexical+syntactic*, *syntactic+semantic*, and *graphical+syntactic*.

It is worth noting that since we adopted the RST Signalling Corpus, we assumed that the gold standard signalling information is 'correct' without questioning it further; however, our annotation left out instances of the category *unsure*, as we cannot be certain whether potentially anchorable signals can be found in these cases.

4 Results

With 11 documents and 4,732 tokens, 923 instances of signals were anchored in this pilot study: approx. 92.2% of the total number of signals. As Table 1 shows, the type *semantic* covers the most cases, most often corresponding to *lexical chains* in which related items indicate a relation (e.g. a phrase such as "rates below...", foreshadowing the appearance of "prime rate" further on) or cases of co-referring expressions, including pronominal anaphora. It is interesting to see that discourse markers only cover about 8.5% of anchor tokens in this study. Table 1 shows how often each signal type is anchored. The top three types of anchored signals are *semantic*, *syntactic*, and the *combined* signal *semantic+syntactic*[1]. Figure 2 provides an example of how the signal type *syntactic* is marked.

One important result is that a large number of

[1]Though RST-SC treats *semantic+syntactic* and *syntactic+semantic* as distinct types, our annotation collapsed the two into one, namely *semantic+syntactic*.

Type of Signals	Count	%Anchored
semantic	385	100
syntactic	172	85.57
semantic+syntactic	126	100
dm	78	100
graphical	57	79.17
lexical	36	100
morphological	25	100
reference+syntactic	16	100
genre	3	8.11
reference	14	100
graphical+syntactic	5	100
lexical+syntactic	4	100
numerical	2	100

Table 1: Distribution of Signals and Anchoring.

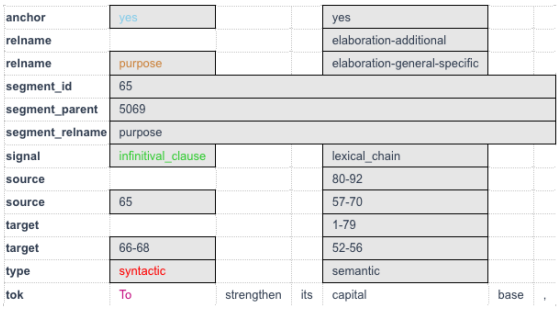


Figure 2: The PURPOSE relation is signaled by a syntactic feature, the infinitival clause (i.e. ***To*** *strengthen its capital base*), which is one of the sub-categories of the signal type *syntactic* in RST-SC.

signals rely on relations between open classes of tokens, such as repetition or relatedness, which cannot be modeled using word embeddings (e.g. a system fed only word embeddings cannot learn that repetition of an arbitrary noun is significant). Example (1) illustrates this point.

(1) Congress gave **Senator Byrd**'s state 21.5 million. **Senator Byrd** is chairman of the Appropriations Committee.

In this example, the Signalling Corpus suggests that the repetition of Senator Byrd's name signals an elaboration. While a system trained on word embeddings could learn that words such as 'Senator' or 'Bird' signal elaborations in certain environments, word embeddings alone cannot capture the importance of repeating an arbitrary name verbatim, including for novel names not seen in training data.

5 Evaluation & Error Analysis

Inter-Annotator Agreement. In order to evaluate the reliability of the scheme and the complexity of the task, we conducted an agreement study on two documents, which include 958 tokens. Agreement is calculated based on token spans, i.e., we would like to see whether the two annotators agreed on the annotated categories (see Section 3) for each token, and we assume that the number of decisions is fixed at the number of tokens, allowing us to calculate kappa.[2]

As this is the first study attempting to anchor RST-SC-style annotations, we discovered that annotator practices differ substantially in the absence of very clear guidelines. For Annotator A, there were 36 unique span annotations, while Annotator B identified 108 unique annotations. This discrepancy is due in part to whether or not the first member of a repetition or chain is annotated as part of the signal, or only the repeated mention, a guideline which must be clarified for future analyses. Moreover, there were 104 exact token matches with positive anchoring information, and 717 tokens for which annotators agreed that no signaling information was present. The raw agreement on all the tokens is 86%, and Cohen's Kappa is 52%, due to the high probability of the negative class (i.e. chance agreement may be high).

Error Analysis. As can be seen from the data presented above, the agreement level in this pilot study is modest. There are several reasons for this: First of all, the guidelines are not clear on whether or not referential entities are annotated for all their occurrences. For instance, in one of the documents, the word *Congress* was annotated as the signal of one relation by one annotator but as the signal of two relations by the other annotator, based on a shared reference signal annotation. Secondly, the similar nature of the sub-categories *lexical chain* and *repetition* makes it difficult to draw a clear distinction between them. As a result, they were not consistently annotated in the original corpus, which resulted in confusions about our annotations that involved them.

According to the RST Siganlling Corpus, *lexical chain* is defined as words or phrases in the

[2]We recognize that this is not an entirely natural interpretation for signals spanning multiple tokens, where we might want to give full or partial credit for agreement on non-identical but overlapping spans. In the present paper no partial credit is given, but we are considering different metrics as well.

respective spans being identical or semantically related, and according to our guidelines, lexical chains are annotated for words with the same lemma or for synonyms or other non-identical but semantically terms. Even though RST-SC distinguishes *lexical chain* from other sub-categories such as *synonymy*, *antonym*, *meronymy* and *indicative word pair*, our guidelines do not explicitly specify whether antonyms are treated as a separate category or fall into *lexical chain* (though we did decide to add a notation of *non-ident* for chains with distinct lemmas). Based on our experience with RST-SC annotations, some instances of *lexical chain* cannot be resolved unless we select antonyms, meronyms, etc. (i.e. no other plausible lexical chain members can be found), and annotation guidelines must clarify whether or not we allow resolution of lexical chain to such instances.

As a result, annotation in this pilot study was not consistent with such instances, which make up a large proportion as shown in Section 4. The subcategory *repetition*, on the other hand, is defined as entities being introduced in one span and repeated in the other span. It is a relatively frequent pattern that *lexical chain* and *repetition* co-occurred. Therefore, it is likely that the same token was doubly annotated in the original corpus because sometimes it is hard to find lexical items that are not the same ones already instantiating *repetition*.

In addition to the unclear guidelines mentioned above, the low agreement is also likely due to some non-core reasons. As mentioned in Section 3, the guidelines were developed as the annotation proceeded. These two documents were annotated within different time frames. One annotator annotated at an earlier stage of the project whereas the other annotator annotated them after annotating other documents and gathering more experience with these issues. Moreover, since we do not give partial credit, typos, missed items, or mismatch in the categories like *Source/Target* would all lead to low agreement. This suggests that we require a better tool to annotate discourse signals with, a task which we are currently pursuing.

The low agreement also indicates the fact that signal anchoring based on a third party's unanchored signal annotations is a very difficult task, due to its complex nature and dynamics of language in general. In particular, a lot of these tokens belong to open classes, and it is nearly impossible for everyone to agree on cases that have many varieties and variations. However, it is these categories that help us establish discourse relations and that represent part of the content of texts. We therefore feel strongly that signaling annotation should be pursued, but that signal identification and anchoring should be performed simultaneously by one person in subsequent work on new datasets.

6 Conclusion

This research presents an annotation effort to anchor all types of signals that establish discourse relations. Even though this preliminary pilot study only consists of 11 documents with 4,732 tokens, the results reveal a wide variety of signal anchors, and the vast range of relevant signaling information that is not represented by classic discourse markers, but is still anchorable to tokens in text. It is clear that the nature of discourse relation signaling is highly complex, meaning that focusing only on discourse markers cannot achieve a full picture. In future work, we will expand to cover more documents and present a revised inter-annotator agreement study using lessons from the adjudication in this pilot study. Moreover, we are exploring how the (anchored) signaling annotation scheme can work with other genres outside RST-DT, which would provide more richly annotated data for discourse parsing as well as empirical evidence for the range of signaling strategies for theoretical research.

References

Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Lucie Polakova, Jiri Mirovsky, and Pavlina Synkova. 2017. Signalling implicit relations: A pdtb-rst comparison. *Dialogue & Discourse*, 8(2):225–248.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Maite Taboada and Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *D&D*, 4(2):249–281.

Shuo Zhang and Amir Zeldes. 2017. Gitdox: A linked version controlled online xml editor for manuscript transcription. In *Proceedings of FLAIRS 2017, Special Track on Natural Language Processing of Ancient and other Low-resource Languages*, pages 619–623.