The Good, the Bad, and the Disagreement: Complex ground truth in rhetorical structure analysis

Debopam Das
Dept. of Linguistics
University of Potsdam
Potsdam, Germany
debdas@uni-potsdam.de

Maite Taboada
Dept. of Linguistics
Simon Fraser University
Burnaby, BC, Canada
mtaboada@sfu.ca

Manfred Stede
Dept. of Linguistics
University of Potsdam
Potsdam, Germany
stede@uni-potsdam.de

Abstract

We present a proposal to analyze disagreement in Rhetorical Structure Theory annotation which takes into account what we consider "legitimate" disagreements. In rhetorical analysis, as in many other pragmatic annotation tasks, a certain amount of disagreement is to be expected, and it is important to distinguish true mistakes from legitimate disagreements due to different possible interpretations of the structure and intention of a text. Using different sets of annotations in German and English, we present an analysis of such possible disagreements, and propose an underspecified representation that captures the disagreements.

1 Introduction

The past ten years have seen continuous interest in RST-oriented discourse parsing, which aims at automatically deriving a complete and well-formed tree representation over coherence relations assigned to adjacent spans of text. For various downstream applications (e.g., summarization, essay scoring), such a complete structure is more useful than the purely localized assignment of individual relations, as it is done in PDTB-style analysis (Prasad et al., 2008).

At the same time, it is well known that RST parsing is difficult, and furthermore, it is more difficult to achieve good human agreement on RST trees, as compared to PDTB annotation. This latter problem has not been in the spotlight of attention, though, while the computational linguistics community developed a series of parsing approaches over the years (Hernault et al., 2010; Ji and Eisenstein, 2013; Feng and Hirst, 2014; Braud et al., 2016). Part of the reason for the focus on data-oriented automatic parsing is the availability of the RST Discourse Treebank (Carlson et al., 2003), a corpus large enough to supply training/test data in supervised machine learning (ML).

The central thesis of our paper is that the fundamental questions of RST annotation and agreement deserve to be re-opened. With powerful ML and parsing technology in place, it is timely to give more attention to the nature of the underlying data, and to its descriptive and theoretical adequacy. Our claim is that the "single ground truth asssumption" is essentially invalid for an annotation task such as rhetorical structure, which inevitably includes a fair amount of subjective decisions on the part of the annotator. As we will emphasize later, we regard this not as a fault of Rhetorical Structure Theory (Mann and Thompson, 1988; Taboada and Mann, 2006), but as a reality to accept, shared with labelling of other pragmatic phenomena, such as speech acts or presuppositions.

Specifically, we will argue that a certain amount of ambiguity is to be regarded as part of the "gold standard" or "ground truth". At the same time, it is clear that RST annotation is not a matter of "anything goes". So, the central challenge in our view is to differentiate between good and bad disagreement: Two annotators may legitimately disagree on some part of the analysis, when both alternatives are in line with the annotation guidelines, and they arise from, for instance, different background knowledge. This needs to be kept separate from disagreement with a not-so-well-educated annotator who misread the guidelines and thus sometimes makes analysis decisions that should not be regarded as legitimate.

Our overall project has two parts: Teasing apart the two types of disagreement, and adequately repprovide a brief sketch of the second.

In the next section, we discuss relevant related work, and then present two agreement studies we undertook on German and English texts (Section 3). We draw conclusions from both in Section 4 and then sketch our framework for technically representing alternative analyses in Section 5. A brief summary (Section 6) concludes the paper.

2 Related work

In Computational Linguistics, a discussion on ambiguity in RST started shortly after Mann and Thompson (1988) was published, mostly in the Natural Language Generation community. The well-known proposal by Moore and Pollack (1992) argued that certain text passages can systematically have two different analyses, one drawing on the intentional, the other on the subject-matter (informational) subset of coherence relations. In a pair of two sentences, for example, when the first states a subjective claim, the second might be interpreted as EVIDENCE for the first, or as merely providing ELABORATION. Moore and Pollack also gave examples where the alternative analyses coincide with conflicting nuclearity assignments.

These questions were never really resolved; instead, with the availability of the RST Discourse Treebank (RST-DT), attention shifted to automatic parsing with ML techniques, starting with Marcu (2000), who also suggested a way of measuring agreement between competing analyses, splitting the overall task into four subtasks (units, spans, nuclearity, relations); we will also use this approach below in our experiments. As to the results achieved, Carlson et al. (2003) reported these kappa results for an experiment with pre-segmented text (i.e., where there is no point in computing unit agreement): spans .93, nuclearity .88, and relations .79. Note that these results were obtained after annotators had already worked for several months on many texts.

More recently, van der Vliet et al. (2011) annotated a Dutch corpus, and computed agreement following Marcu's method, also using pre-segmented text. They report an average kappa agreement of .88 on spans, .82 on nuclearity, and .57 for relations. These figures should not be directly compared to those of Carlson et al., because there are differences in the relation set, the guidelines, and the amount of annotator training.

The problem of ambiguity was again studied by Schilder (2002), who worked in the framework of Segmented Discourse Representation Theory or SDRT (Asher and Lascarides, 2003) and approached the problem from a semantic viewpoint. He proposed that certain aspects of the analysis could be left unannotated. For instance, nuclearity may be assigned, but the specific relation between nucleus and satellite may be left blank, if a decision cannot be reached.

Around the same time, Reitter and Stede (2003) proposed the Underspecified Rhetorical Markup Language (URML), an XML language for encoding competing analyses in a single representation. We will describe this in more detail in Section 5.

More recently, Iruskieta et al. (2015) proposed a qualitative method for analysis comparison, teasing apart constituency, relation, and attachment. The most important aspect of their comparison method is that nuclearity and relation label are separated, unlike in Marcu's quantitative agreement metric.

3 Empirical studies

Both of our studies are attached to existing RST-annotated corpora, so that our results can be related to the earlier work. Also, we used nearly-identical annotation guidelines, which we describe first, before we turn to the actual experiments.

3.1 Annotation guidelines

In contrast to the RST-DT project of Carlson et al. (2003), our annotation guidelines follow the original RST paper (Mann and Thompson, 1988) relatively closely. This means that our relation set is much smaller than that of the RST-DT (31 relations instead of 78). We do not use the many nucleus-satellite variants, and we deliberately left out suggestions like TOPIC-COMMENT or ATTRIBUTION, which we do not regard as coherence relations in the same way as those of "classic" RST.[1] We group the relations in a slightly different way from Mann & Thompson into subject-matter and presentational ones, and we have an extra category for textual relations (LIST, SUMMARY).

For technical reasons, at the moment we avoid the SAME-UNIT relation of the RST-DT by not

[1] We are of course not claiming that phenomena of Topic/Comment and Attribution do not exist. Instead, notions of information structure in our view belong to a separate level of analysis—not to that of coherence relations.

separating center-embedded segments. This decision may be revised later, and it is not critical for the purposes of this paper.

For the German experiment, we used the annotation guidelines developed for the Potsdam Commentary Corpus (Stede, 2016) and which are publicly available. Then, for annotating the English texts, we produced an English version of those guidelines and made minimal changes to the descriptions of relations (clarifications on how to distinguish between certain contrastive and argumentative relations). Further, we used language-specific segmentation guidelines that we borrowed from the implementation of SLSeg (syntactic and lexically based discourse segmenter) (Tofiloski et al., 2009).[2] In addition to many individual examples for the relations, the guidelines finish with a sample analysis of a complete text with 14 elementary discourse units (EDUs).

The guidelines merely guide the annotators in their task. They could in principle be written in such a way as to "strongly encourage" agreement when cases of ambiguity arise (e.g., by specifying preference hierarchies), but they make only minimal use of that move. The interesting issue from a theoretical viewpoint is that the same general guidelines can give rise to what we consider as legitimate disagreements.

3.2 Study I: German

For the German study (see Fodor (2015)), we selected ten texts from the publicly available Potsdam Commentary Corpus[3], which has been annotated at various levels of linguistic description, including RST. They are editorials or "pro and con" commentaries from local newspapers, with a typical length of 8 to 10 sentences (with an average length of 16 words, sentences often consist of more than one EDU). We picked texts of general-interest topics and which do not make too many references to local events or people, which might confuse annotators.

The idea of the annotation experiment was to assess the influence of the amount of training that annotators receive. Thus we worked with four annotators, all with university education. Two sions, since each group consisted of just two annotators, but the result indicates that the difference in training time and content—in particular, we surmise, the difference in the number of jointly-discussed sample analyses—leads to a marked difference in annotator agreement.

In order to measure the agreement between expert and non-expert annotators, we computed the precision and recall values for GE1 and GL1, following the method documented in Marcu (2000). GE1 was considered as the "gold" annotation. The precision and recall values, provided in Table 2, show relatively higher agreement for spans and nuclearity, but low agreement for relations. Precision and recall are the same, because there are equal numbers of false positives and false negatives.

	Precision	Recall
Span	0.65	0.65
Nuclearity	0.56	0.56
Relation	0.30	0.30

Table 2: Precision and recall for expert versus student annotation (GE1-GL1)

We also conducted various more detailed analyses, but for reasons of time, only a randomly chosen subset of five texts and their RST trees could be handled in this phase. In Table 3, we report the percent agreement results for all pairs of annotators.

	Span	Nuclearity	Relation
GE1 - GE2	63.6	43.8	27.0
GL1 - GL2	60.6	35.2	15.4
GE1 - GL1	56.6	38.8	13.2
GE1 - GL2	48.8	31.2	19.6
GE2 - GL1	63.4	44.2	23.8
GE2 - GL2	44.2	35.2	15.4

Table 3: Percent agreement of all annotator pairs (German study, 5 texts)

First of all, notice that the results for GL1-GL2 are considerably closer to those of GE1-GE2 than in the comparison of the full 10 texts; this indicates that the texts selected are "easy" ones. But the main insight to be gained from Table 3 is that the poor results of GL1-GL2 are mainly due to the performance of GL2, who consistently reaches low agreement with all three other annotators (the single exception being the Relation agreement with GE1), while GL1 does a fairly good job; in particular s/he agrees with GE2 essentially as much as GE1 does.

One other factor we investigated is the "difficulty" of individual RST relations. On the basis of the five texts, we computed how many pairs of annotators achieve at least one perfect agreement for a particular relation type. The results are given in Table 4. The second column gives the number of pairs of annotators that agree on the relation label (and also on spans and nuclearity) in at least one text.

Relation	Ann.pairs	Percent
Preparation Condition	6	100

annotated corpus, in this case the RST Discourse Treebank (Carlson et al., 2003), but we did not use the associated annotation guidelines, as explained earlier. To match the genre of “commentary”, we looked especially for argumentative text (which in general we expect to be more prone to competing analyses, since more interpretation and subjectivity is involved than in plain news text). In total we found 19 such documents in the RST-DT, which are letters to the editor, editorials, op-ed pieces, or reviews. For our present experiment, we selected four of the documents. One document contains multiple letters; we split it up and thus have a set of seven individual texts to work with. With an average length of 205 words per text, they are somewhat shorter than the German texts.

Also in line with the German study, we performed a pre-segmentation (following the rules mentioned in Section 3.2) of all the texts, so that annotators started from a basis that allows for a solid comparison of span, nuclearity and relation decisions. In terms of annotator teams, however, we could not exactly replicate the setting of the

	Precision	Recall
Span	0.88	0.88
Nuclearity	0.58	0.58
Relation	0.41	0.41

Table 7: Precision and recall for expert versus student annotation (EE1-EL1)

Qualitative analysis. We are also interested in a qualitative comparison: Which phenomena in the texts triggered discrepancies in the two anal-

5 The complex gold: Capturing ambiguity

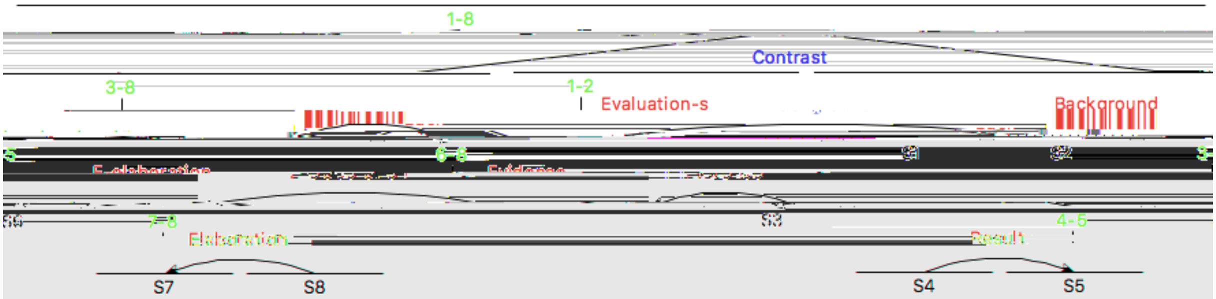

Figure 1: Annotation by EE1 for part of a corpus text (English study)

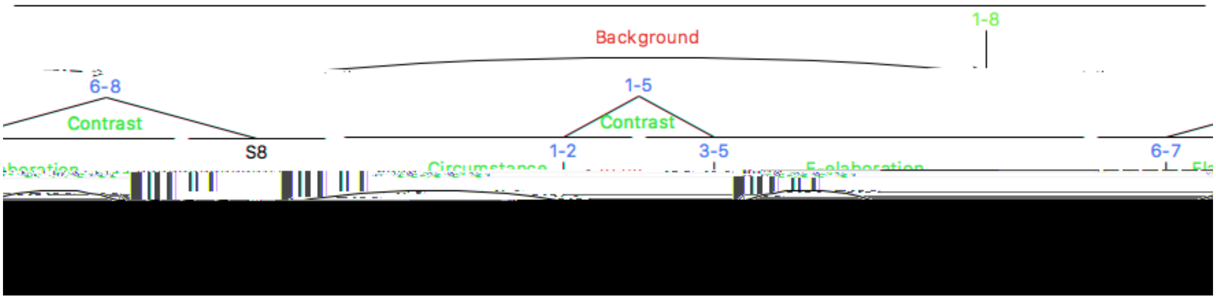

Figure 2: Annotation by EE2 for part of a corpus text (English study)

```
   <satellite id="N4">
</hypRelation>

<parRelation id="N4" type="Contrast"
     annotator="V2">
   <nucleus id="N2b">
   <nucleus id="N3">
</parRelation>
```

The declarations state that nodes N1a and N1b are alternative analyses provided by annotators EE1 and EE2. They are alternatives because they belong to the same group N1, and cover the same sequence of EDUs (S1–S8). In contrast, N4 does not belong to a group, i.e., it occurs only in EE-2's analysis. The first nucleus of both CONTRAST relations is an alternative of group N2 (not shown here), which represents the analyses for segments S1–S2.

In the same way, the other disagreements between EE1 and EE2 can be captured in the same URML representation, which thus plays the role of a "complex gold" annotation.

6 Summary

With two empirical studies, we demonstrated that annotator agreement depends on the amount of training and expertise the annotators have acquired. While this is hardly surprising, our next step is to differentiate between non-expert disagreement (some of which can arise from failure to adhere to the given guidelines, annotation flaws, or other human factors) and what we call "legitimate disagreement", i.e., that between expert annotators. Our proposal here is that competing expert analyses should be regarded as part of the "ground truth" in an annotated corpus. Besides differentiating between annotator expertise by means of quantitative measures, we undertook a first qualitative analysis of the types of disagreements encountered among experts. In future work, this needs to be elaborated.

The second point we made is that we can use the URML representation framework (which had originally been designed for a somewhat different purpose) to capture the disagreement in annotations in a single representation for a text. Our initial result is that the analyses used in the English study could all be mapped to URML and adequately represent the alternatives in the annotations. Here, the next step for us is to provide tools for automatic mapping (and merging) from the rs3 format of RSTTool to URML, and to devise ways of computing annotator agreement between a "new" annotator, or an RST parser for that matter, and the URML graph representing the "complex gold".

References

Nicholas Asher and Alex Lascarides. 2003. Logics of Conversation