Syntactic Identification of Attribution in the RST Treebank

Peter Rossen Skadhauge and **Daniel Hardt**
CMOL/Department of Computational Linguistics
Copenhagen Business School
DENMARK
{prs,dh}@id.cbs.dk

Abstract

We present a system that automatically identifies Attribution, an intra-sentential relation in the RST Treebank. The system uses uses syntactic information from Penn Treebank parse trees. It identifies Attributions as structures in which a verb takes an SBAR complement, and achieves a f-score of .92. This supports our claim that the Attribution relation should be eliminated from a discourse treebank, since it represents information that is already present in the Penn Treebank, in a different form. More generally, we suggest that intra-sentential relations in the RST Treebank might all be eliminable in this way.

1 Introduction

There has been a growing interest in recent years in Discourse Structure. A prominent example of this is the RST Treebank (Carlson et al., 2002), which imposes hierarchical structures on multi-sentence discourses. Since the texts in the RST Treebank are taken from the syntactically annotated Penn Treebank (Marcus et al., 1993), it is natural to ask what the relation is between the discourse structures in the RST Treebank and the syntactic structures of the Penn Treebank.

In our view, the most natural relationship would be that discourse structures always relate well-formed syntactic expressions, typically sentences. Discourse trees would then be seen as elaborations of syntactic trees, adding relations between sentential nodes that are not linked by syntactic relations. This would allow discourse structures and syntactic structures to coexist in a combined hierarchical structure.

Surprisingly, this is not what we have found in examining the syntax-discourse relation in the RST Treebank. A large proportion of relations apply to subsentential spans of text;[1] spans that may or may not correspond to nodes in the syntax tree. Is this complicated relation between syntax and discourse necessary? Our hypothesis is that the subsentential relations in the RST Treebank are in fact redundant; if this is true it should be possible to automatically infer these relations based solely on Penn Treebank syntactic information.

In this paper, we present the results of an initial study that strongly supports our hypothesis. We examine the Attribution relation, which is of particular interest for the following reasons:

- It appears quite frequently in the RST Treebank (15% of all relations, according to Marcu et al. (1999))

- It always appears within, rather than across, sentence boundaries

- It conflicts with Penn Treebank syntax, always relating text spans that do not correspond to nodes in the syntax tree

We describe a system that identifies Attributions by simple, clearly defined syntactic features.

---

[1] In the TRAINING portion of the RST Treebank, we found 17213 Elementary Discourse Units (EDU's). Of these only 6068 occurred at sentence boundaries.

This system identifies RST Attributions within precision and recall over 90%. In our view, this strongly supports the view that Attribution is in fact a syntactic relation. The system performs dramatically better than the results reported in (Soricut and Marcu, 2003) for automatic identification of such relations, where the precision and recall were reported at below .76. Furthermore, human annotator agreement reported in the RST Treebank project is also well below our results, with reported f-scores no higher than .77. (Soricut and Marcu, 2003)

In what follows, we first describe Attributions as they are understood in the RST Treebank project. Next we present the Attribution identification procedure, followed by a presentation of results. We compare these results with related work, as well as with inter-coder agreement reported in the RST Treebank project. Finally, we discuss plans for future work.

2 Attributions in the RST Treebank

The RST coding manual (Carlson and Marcu, 2001) gives the following definition of Attribution:

> Instances of reported speech, both direct and indirect, should be marked for the rhetorical relation of ATTRIBUTION. The satellite is the source of the attribution (a clause con- taining a reporting verb, or a phrase beginning with according to), and the nucleus is the content of the reported message (which must be in a separate clause). The ATTRIBUTION relation is also used with cognitive predicates, to include feelings, thoughts, hopes, etc.

The following is an example cited in the coding manual:

> [The legendary GM chairman declared] [that his company would make "a car for every purse and purpose."]wsj_1377

According to the RST Treebank, the attribution verb is grouped with the subject into a single text span. This constitutes the Attribution Satellite,

while the Nucleus is the SBAR complement of the attribution verb, as shown below in Figure 1.
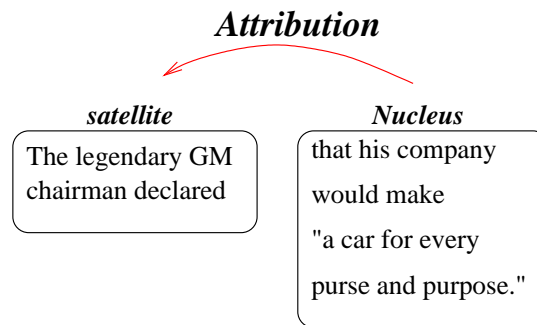

Figure 1: Attribution in the RST Treebank

This conflicts with the syntactic structure in the Penn Treebank. As shown in Figure 2, the attribution verb is grouped with its SBAR complement, forming a VP, which is related to the subject.
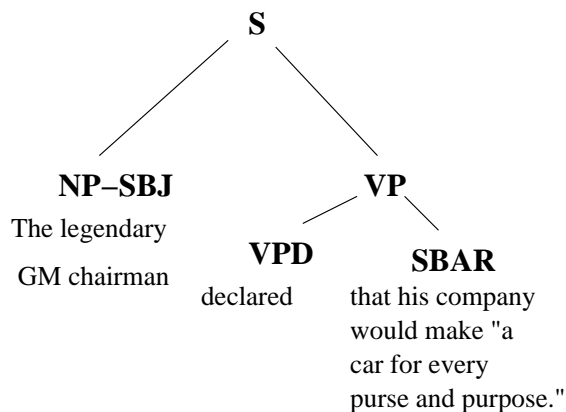

Figure 2: Attribution in the Penn Treebank

The main difference in the two structures regards the position of the verb; in the RST Treebank, the verb is grouped with the subject, while in the Penn Treebank, it is grouped with the SBAR complement. In the following section, we describe our method for identifying RST Attributions, based on the Penn Treebank syntactic structure.

3 Identifying Attributions

We define three forms of Attribution relations:

- **Basic:** A verb is followed by a sentential complement position

- **Backwards:** The sentential complement precedes the verb. In these cases, a trace ap-

pears as complement to the verb, and is coindexed with the sentential complement

- **According-To:** the phrase "according to" occurs

3.1 Basic Attributions

In this form, a sentential object immediately follows a verb.

Consider the example

(1) Now, the firm says it's at a turning point.

In PTB, the sentence is annotated as in :

(2)
```
( (S
    (ADVP-TMP (RB Now) )
    (, ,)
    (NP-SBJ (DT the) (NN firm) )
    (VP (VBZ says)
      (SBAR (-NONE- 0)
        (S
          (NP-SBJ (PRP it) )
          (VP (VBZ 's)
            (PP-LOC-PRD (IN at)
              (NP (DT a) (NN turning)
                  (NN point) ))))))
    (. .) ))
```

Sentential objects are annotated as SBAR regardless of the presence of complementizers. Thus, the subroutine searches the corpus for structures matching the template (3), which matches verb phrases in which a verb is followed by an SBAR.

(3) `(VP ... (V.. ...) (SBAR ...) ... )`

The SBAR must follow immediately after the verb, which may be the last verb in a verbal cluster. This represents a simplification, since adverbials may occur between the verb and its SBAR complement. Our implementation correctly identifies 1497 occurrences, and incorrectly identifies 215 occurrences of attributions, corresponding to a contribution to the total recall of 0.615 with a precision of 0.874.

3.2 Backwards Attributions

Where a sentential object does not immediately follow its corresponding verb, it is represented as a trace which is coindexed with the S. In the following example, the sentential complement precedes the sentence:

(4) "I believe that any good lawyer should be able to figure out and understand patent law",$_i$ Judge Mayer says t_i

The example is represented as follows in PTB:

(5)
```
((S-6 ('' '')
  (NP-SBJ-2 (PRP I) )
  (VP (VBP believe)
   (SBAR (IN that)
    (S
     (NP-SBJ-4 (DT any) (JJ good)
                        (NN lawyer) )
      (VP (MD should)
       (VP (VB be)
        (ADJP-PRD (JJ able)
         (S
          (NP-SBJ (-NONE- *-4) )
          (VP (TO to)
           (VP
            (VP (VB figure)
             (PRT (RP out) )
             (NP (-NONE- *RNR*-5) ))
            (CC and)
            (VP (VB understand)
             (NP (-NONE- *RNR*-5) ))
           NP-5 (NN patent)
                 (NN law) )))))))))
  (PRN
   (, ,)
   ('' '')
   (S
    (NP-SBJ (NNP Judge)
            (NNP Mayer) )
    (VP (VBZ says)
        (S (-NONE- *T*-6) ))))
```

The sentential object of "says" is represented by the trace (`(S (-NONE- *T*-6) )))`), which is coindexed with the outer sentence (`(S-6 )`).

The procedure searches for sentences of the types `S`, `S/SBAR`, and `VP/S-TPC` which are linked to a trace in the surrounding sentence. Thus, it covers cases of topicalization and sentence inversion which are the most frequent reasons for sentential objects not occurring immediately after the verb.

The subroutine covering sentential objects linked by traces make 700 correct and 4 incorrect predictions, corresponding to a recall contribution of 0.287 with a precision of 0.994.

3.3 According-To Attributions

Also categorized as attributions are "according to" expressions. These are identified with a separate subroutine which simply identifies occur-

rences of the two words "according" and "to" in sequence.

Example:

(6) Now, <u>according to</u> <u>a Kidder World story about Mr. Megargel</u>, all the firm has to do is "position ourselves more in the deal flow."

(7)

```
( (S
   (ADVP-TMP (RB Now) )
   (, ,)
   (PP (VBG according)
     (PP (TO to)
       (NP
         (NP (DT a) (NNP Kidder)
             (NNP World) (NN story) )
         (PP (IN about) (NP (NNP Mr.)
                 (NNP Megargel) )))))
   (, ,)
   (NP-SBJ
     (NP (DT all) )
     (SBAR
       (WHNP-1 (-NONE- 0) )
       (S
         (NP-SBJ-2 (DT the)
                   (NN firm) )
         (VP (VBZ has)
           (S
             (NP-SBJ (-NONE- *-2) )
             (VP (TO to)
               (VP (VB do)
               (NP (-NONE- *T*-1)
                   )))))))) 
   (VP (VBZ is) (`` ``)
     (VP (VB position)
       (NP (PRP ourselves) )
       (ADVP-MNR (RBR more)
         (PP (IN in)
           (NP (DT the)
           (NN deal) (NN flow) )))))
   (. .) ('' '') ))
```

The subroutine identifies 87 "according to" expressions correctly, and 1 incorrectly.

4 Discussion of Results

Our system for recognizing Attributions is a quite direct implementation of the description of Attribution given in the RST Tagging Manual, relying on simple structural characteristics. In developing the system, we examined data in the Training portion of the RST Treebank. To ensure that our implementation was not tuned to any idiosyncrasies of the data we examined, we performed two tests of our system, on the Test portion of the

RST Treebank as well as the Training portion. We avoided any examination of data in the Test portion of the Treebank.

Given the general nature of the syntactic characteristics of our system, it is not surprising that the results on the Training and Test portions of the Treebank our quite similar. We present the overall results on both portions of the Treebank, followed by more detailed results, giving the contributions of the main subparts of the system.

4.1 Overall Results

The following figure summarizes the results of executing the procedure on the two portions of the Treebank.

Corpus	Precision	Recall	F-score
Training	0.912	0.938	0.925
Test	0.897	0.944	0.920

Figure 3: Overall results

4.2 Subparts of the System

Next, we present the contribution of each of the three subparts of the system.

	+	−	Prec	Rec
Basic	1497	215	0.874	
Backwards	700	4	0.994	
According-to	87	1	0.989	
Total	2284	220	0.912	0.938

Figure 4: Breakdown of system results (Training corpus)

	+	−	Prec	Rec
Basic	193	33	0.854	
Backwards	90	0	1.000	
According-to	4	0	1.000	
Total	286	33	0.897	0.994

Figure 5: Breakdown of system results (Test corpus)

5 Related Work

Soricut and Marcu (2003) describe a *Discourse Parser* – a system that uses Penn Treebank syntax to identify intra-sentential discourse relations in the RST Treebank. Since this applies to *all* intra-sentential relations in the RST Treebank, while our system is limited to Attribution, the systems are not directly comparable. Still, the results and discussion from (Soricut and Marcu, 2003) provide some useful perspective on our results.

Soricut and Marcu (2003) evaluate their Discourse Parser under a variety of scenarios; the most favorable has human-corrected syntax trees and discourse segmentation. In this scenario, the system achieves an f-score of .703 with the full set of 110 Relation Labels, and 75.5 with the relation label set collapsed to 18 labels. Soricut and Marcu (2003) note that human annotator agreement receives comparable f-scores, of .719 and .77 respectively. In the light of these numbers, our Attribution system f-score of .92 is quite impressive. This provides some measure of support for our hypothesis that the intra-sentential relations in the RST Treebank are in fact properly viewed as alternative notations for syntactic information that is already present in the Penn Treebank.

Of course, it may well be that some of the other intra-sentential relations present much greater difficulties than Attribution. But these results suggest that it is worth pursuing our project of attempting to automatically derive the intra-sentential RST Treebank relations from specific syntactic features.

6 Conclusion and Future Work

We have shown that Attribution relations can be identified successfully by using the syntactic structure of the Penn Treebank. In a sense, then, notating Attribution relations in syntactically parsed texts adds no information. Our hypothesis is that all intra-sentential relations in the RST Treebank are of this character.

This is important for several reasons. First, it is clear that the relations *across* sentences in the RST Treebank are not directly derivable from syntax, at least not in any obvious way. Our approach to identifying Attributions is a direct implementation of the description in the RST Tree-

bank tagging manual. For inter-sentential relations such as CONTRAST or EXPLANATION-EVIDENCE, the situation is quite different. Syntactic criteria are relevant, but clearly not decisive, as can be observed in (Marcu and Echihabi, 2002). Finally, the elimination of intra-sentential relations like Attribution would appear to be more in line with the original vision behind RST; for example, according to Mann & Thompson (1988), the basic unit for RST relations is the clause.

References

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging manual. ISI Tech Report ISI-TR-545.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

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

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings, 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA.

Daniel Marcu, Magdalena Romera, and Estibaliz Amorrortu. 1999. Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In *Proceedings of the Workshop on Levels of Representation in Discourse*, pages 71–78, Edinburgh, Scotland.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2).

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edmonton, Canada.