The Rhetorical Parsing of Natural Language Texts

Daniel Marcu
Department of Computer Science
University of Toronto
Toronto, Ontario
Canada  M5S 3G4
marcu@cs.toronto.edu

Abstract
We derive the rhetorical structures of texts by means of two new, surface-form-based algorithms: one that identifies discourse usages of cue phrases and breaks sentences into clauses, and one that produces valid rhetorical structure trees for unrestricted natural language texts. The algorithms use information that was derived from a corpus analysis of cue phrases.

1 Introduction
Researchers of natural language have repeatedly acknowledged that texts are not just a sequence of words nor even a sequence of clauses and sentences. However, despite the impressive number of discourse-related theories that have been proposed so far, there have emerged no algorithms capable of deriving the discourse structure of an unrestricted text. On one hand, efforts such as those described by Asher (1993), Lascarides, Asher, and Oberlander (1992), Kamp and Reyle (1993), Grover et al. (1994), and Prüst, Scha, and van den Berg (1994) take the position that discourse structures can be built only in conjunction with fully specified clause and sentence structures. And Hobbs's theory (1990) assumes that sophisticated knowledge bases and inference mechanisms are needed for determining the relations between discourse units. Despite the formal elegance of these approaches, they are very domain dependent and, therefore, unable to handle more than a few restricted examples. On the other hand, although the theories described by Groz and Sidner (1986), Polanyi (1988), and Mann and Thompson (1988) are successfully applied manually, they are too informal to support an automatic approach to discourse analysis.

In contrast with this previous work, the rhetorical parser that we present builds discourse trees for unrestricted texts. We first discuss the key concepts on which our approach relies (section 2) and the corpus analysis (section 3) that provides the empirical data for our rhetorical parsing algorithm. We then discuss an algorithm that recognizes discourse usages of cue phrases and that determines clause boundaries within sentences. Lastly, we present the rhetorical parser and an example of its operation (section 4).

2 Foundation
The mathematical foundations of the rhetorical parsing algorithm rely on a first-order formalization of valid text structures (Marcu, 1997). The assumptions of the formalization are the following. 1. The elementary units of complex text structures are non-overlapping spans of text. 2. Rhetorical, coherence, and cohesive relations hold between textual units of various sizes. 3. Relations can be partitioned into two classes: paratactic and hypotactic. Paratactic relations are those that hold between spans of equal importance. Hypotactic relations are those that hold between a span that is essential for the writer's purpose, i.e., a nucleus, and a span that increases the understanding of the nucleus but is not essential for the writer's purpose, i.e., a satellite. 4. The abstract structure of most texts is a binary, tree-like structure. 5. If a relation holds between two textual spans of the tree structure of a text, that relation also holds between the most important units of the constituent subspans. The most important units of a textual span are determined recursively: they correspond to the most important units of the immediate subspans when the relation that holds between these subspans is paratactic, and to the most important units of the nucleus subspan when the relation that holds between the immediate subspans is hypotactic.

In our previous work (Marcu, 1996), we presented a complete axiomatization of these principles in the context of Rhetorical Structure Theory (Mann and Thompson, 1988) and we described an algorithm that, starting from the set of textual units that make up a text and the set of elementary rhetorical relations that hold between these units, can derive all the valid discourse trees of that text. Consequently, if one is to build discourse trees for unrestricted texts, the problems that remain to be solved are the automatic determination of the textual units and the rhetorical relations that hold between them. In this paper, we show how one can find and exploit approximate solutions for both of these problems by capitalizing on the occurrences of certain lexicogrammatical constructs. Such constructs can include tense
Empirical studies on the disambiguation of cue phrases (Hirschberg and Litman, 1993) have shown that just by considering the orthographic environment in which a discourse marker occurs, one can distinguish between sentential and discourse usages in about 80% of cases. We have taken Hirschberg and Litman's research one step further and designed a comprehensive corpus analysis that enabled us to improve their results and coverage. The method, procedure, and results of our corpus analysis are discussed in section 3.

Discourse markers are ambiguous with respect to the rhetorical relations that they mark and the sizes of the units that they connect. When we began this research, no empirical data supported the extent to which this ambiguity characterizes natural language texts. To better understand this problem, the corpus analysis described in section 3 was designed so as to also provide information about the types of rhetorical relations, rhetorical statuses (nucleus or satellite), and sizes of textual spans that each marker can indicate. We knew from the beginning that it would be impossible to predict exactly the types of relations and the sizes of the spans that a given cue marks. However, given that the structure that we are trying to build is highly constrained, such a prediction proved to be unnecessary: the overall constraints on the structure of discourse that we enumerated in the beginning of this section cancel out most of the configurations of elementary constraints that do not yield correct discourse trees.

Consider, for example, the following text:

\[
\begin{align*}
(1) & \text{[Although discourse markers are ambiguous,}\quad 1] \\
& \text{[one can use them to build discourse trees for unrestricted texts]:}\quad 2 \\
& \text{[this will lead to many new applications in natural language processing]:}\quad 3
\end{align*}
\]

For the sake of the argument, assume that we are able to break text (1) into textual units as labelled above and that we are interested now in finding rhetorical relations between these units. Assume now that we can infer that Although marks a CONCLUSIVE relation between satellite 1 and nucleus either 2 or 3, and the colon, an ELABORATION between satellite 3 and nucleus either 1 or 2. If we use the convention that hypotactic relations are represented as first-order predicates having the form \( \text{rhet\_rel}(\text{NAME}, \text{satellite}, \text{nucleus}) \) and that paratactic relations are represented as predicates having the form \( \text{rhet\_rel}(\text{NAME}, \text{nucleus}1, \text{nucleus}2) \), a correct representation for text (1) is then the set of two disjunctions given in (2):

\[
\begin{align*}
(2) & \quad \{ \text{rhet\_rel}(\text{CONCESSION}, 1, 2) \lor \text{rhet\_rel}(\text{CONCESSION}, 1, 3) \\
& \quad \text{rhet\_rel}(\text{ELABORATION}, 3, 1) \lor \text{rhet\_rel}(\text{ELABORATION}, 3, 2) \}
\end{align*}
\]

Despite the ambiguity of the relations, the overall rhetorical structure constraints will associate only one discourse tree with text (1), namely the tree given in figure 1: any discourse tree configuration that uses relations \( \text{rhet\_rel}(\text{CONCESSION}, 1, 3) \) and \( \text{rhet\_rel}(\text{ELABORATION}, 3, 1) \) will be ruled out. For example, relation \( \text{rhet\_rel}(\text{ELABORATION}, 3, 1) \) will be ruled
out because unit 1 is not an important unit for span \([1,2]\)
and, as mentioned at the beginning of this section, a
rhetorical relation that holds between two spans of a valid
text structure must also hold between their most impor-
tant units: the important unit of span \([1,2]\) is unit 2, i.e.,
the nucleus of the relation \(rhet.rel(\text{CONCESSION}, 1, 2)\).

3 A corpus analysis of discourse markers

3.1 Materials

We used previous work on cue phrases (Halliday and Hasan, 1976; Grosz and Sidner, 1986; Martin, 1992;
Hirschberg and Litman, 1993; Knot, 1995; Fraser, 1996)
to create an initial set of more than 450 potential dis-
course markers. For each potential discourse marker, we
then used an automatic procedure that extracted from the
Brown corpus a set of text fragments. Each text fragment
contained a “window” of approximately 200 words and
an emphasized occurrence of a marker. On average, we
randomly selected approximately 19 text fragments per
marker, having few texts for the markers that do not occur
very often in the corpus and up to 60 text fragments for
markers such as \(\text{and}\), which we considered to be highly
ambiguous. Overall, we randomly selected more than
7900 texts.

All the text fragments associated with a potential cue
phrase were paired with a set of slots in which an ana-
lyst described the following. 1. The orthographic en-
vironment that characterizes the usage of the potential
discourse marker. This included occurrences of periods,
commas, colons, semicolons, etc. 2. The type of usage:
Sentential, Discourse, or Both. 3. The position of the
marker in the textual unit to which it belonged: Begin-
ning, Medial, or End. 4. The right boundary of the textual
unit associated with the marker. 5. The relative position
of the textual unit that the unit containing the marker
was connected to: Before or After. 6. The rhetorical rela-
tions that the cue phrase signaled. 7. The textual types of
the units connected by the discourse marker: from Clause
to Multiple.Paragraph. 8. The rhetorical status of each
textual unit involved in the relation: Nucleus or Satel-
lite. The algorithms described in this paper rely on the
results derived from the analysis of 1600 of the 7900 text
fragments.

3.2 Procedure

After the slots for each text fragment were filled, the
results were automatically exported into a relational
database. The database was then examined semi-
automatically with the purpose of deriving procedures
that a shallow analyzer could use to identify discourse
usages of cue phrases, break sentences into clauses, and
hypothesize rhetorical relations between textual units.
For each discourse usage of a cue phrase, we derived
the following:

- A regular expression that contains an unambigu-
ous cue phrase instantiation and its orthographic
environment. A cue phrase is assigned a regular
expression if, in the corpus, it has a discourse
usage in most of its occurrences and if a shallow
analyzer can detect it and the boundaries of the
textual units that it connects. For example, the
regular expression “[.] although” identifies such
discourse usage.

- A procedure that can be used by a shallow ana-
lyzer to determine the boundaries of the textual
unit to which the cue phrase belongs. For exam-
ple, the procedure associated with “[.] although”
instructions the analyzer that the textual unit that
pertains to this cue phrase starts at the marker and
ends at the end of the sentence or at a position to
be determined by the procedure associated with
the subsequent discourse marker that occurs in
that sentence.

- A procedure that can be used by a shallow ana-
lyzer to hypothesize the sizes of the textual units
that the cue phrase relates and the rhetorical rela-
tions that may hold between these units. For
example, the procedure associated with “[.] al-
though” will hypothesize that there exists a CON-
CESSION between the clause to which it belongs
and the clause(s) that went before in the same
sentence. For most markers this procedure makes
dischjunctive hypotheses of the kind shown in (2)
above.

3.3 Results

At the time of writing, we have identified 1253 occur-
cences of cue phrases that exhibit discourse usages and
associated with each of them procedures that instruct
a shallow analyzer how the surrounding text should be
broken into textual units. This information is used by an
algorithm that concurrently identifies discourse usages of
cue phrases and determines the clauses that a text is made
of. The algorithm examines a text sentence by sentence
and determines a set of potential discourse markers that
occur in each sentence. It then applies left to right the
procedures that are associated with each potential marker.
These procedures have the following possible effects:

- They can cause an immediate breaking of the cur-
rent sentence into clauses. For example, when
an “[.] although” marker is found, a new clause,
whose right boundary is just before the occur-
rence of the marker, is created. The algorithm is
then recursively applied on the text that is found
between the occurrence of "[,] although" and the end of the sentence.

- They can cause the setting of a flag. For example, when an "Although " marker is found, a flag is set to instruct the analyzer to break the current sentence at the first occurrence of a comma.

- They can cause a cue phrase to be identified as having a discourse usage. For example, when the cue phrase "Although" is identified, it is also assigned a discourse usage. The decision of whether a cue phrase is considered to have a discourse usage is sometimes based on the context in which that phrase occurs, i.e., it depends on the occurrence of other cue phrases. For example, an "and" will not be assigned a discourse usage in most of the cases; however, when it occurs in conjunction with "although", i.e., "and although", it will be assigned such a role.

The most important criterion for using a cue phrase in the marker identification procedure is that the cue phrase (together with its orthographic neighborhood) is used as a discourse marker in at least 90% of the examples that were extracted from the corpus. The enforcement of this criterion reduces on one hand the recall of the discourse markers that can be detected, but on the other hand, increases significantly the precision. We chose this deliberately because, during the corpus analysis, we noticed that most of the markers that connect large textual units can be identified by a shallow analyzer. In fact, the discourse marker that is responsible for most of our algorithm recall failures is and. Since a shallow analyzer cannot identify with sufficient precision whether an occurrence of and has a discourse or a sentential usage, most of its occurrences are therefore ignored. It is true that, in this way, the discourse structures that we build lose some potential finer granularity, but fortunately, from a rhetorical analysis perspective, the loss has insignificant global repercussions: the vast majority of the relations that we miss due to recall failures of and are Joint and Sequence relations that hold between adjacent clauses.

**Evaluation.** To evaluate our algorithm, we randomly selected three texts, each belonging to a different genre:

1. an expository text of 5036 words from *Scientific American*;
2. a magazine article of 1588 words from *Time*;
3. a narration of 583 words from the Brown Corpus.

Three independent judges, graduate students in computational linguistics, broke the texts into clauses. The judges were given no instructions about the criteria that they had to apply in order to determine the clause boundaries; rather, they were supposed to rely on their intuition and preferred definition of clause. The locations in texts that were labelled as clause boundaries by at least two of the three judges were considered to be "valid clause boundaries". We used the valid clause boundaries assigned by judges as indicators of discourse usages of cue phrases and we determined manually the cue phrases that signalled a discourse relation. For example, if an "and" was used in a sentence and if the judges agreed that a clause boundary existed just before the "and", we assigned that "and" a discourse usage. Otherwise, we assigned it a sentential usage. Hence, we manually determined all discourse usages of cue phrases and all discourse boundaries between elementary units.

We then applied our marker and clause identification algorithm on the same texts. Our algorithm found 80.8% of the discourse markers with a precision of 89.5% (see...
INPUT: a text \( T \).
1. Determine the set \( D \) of all discourse markers and the set \( U_T \) of elementary textual units in \( T \).
2. Hypothesize a set of relations \( R \) between the elements of \( U_T \).
3. Use a constraint satisfaction procedure to determine all the discourse trees of \( T \).
4. Assign a weight to each of the discourse trees and determine the tree(s) with maximal weight.

Figure 2: Outline of the rhetorical parsing algorithm

Table 1), a result that outperforms Hirschberg and Litman's (1993). The same algorithm identified correctly 81.3% of the clause boundaries, with a precision of 90.3% (see Table 2). We are not aware of any surface-form-based algorithms that achieve similar results.

4 Building up discourse trees

4.1 The rhetorical parsing algorithm

The rhetorical parsing algorithm is outlined in figure 2. In the first step, the marker and clause identification algorithm is applied. Once the textual units are determined, the rhetorical parser uses the procedures derived from the corpus analysis to hypothesize rhetorical relations between the textual units. A constraint-satisfaction procedure similar to that described in (Marcu, 1996) then determines all the valid discourse trees (see (Marcu, 1997) for details). The rhetorical parsing algorithm has been fully implemented in C++.

Discourse is ambiguous the same way sentences are: more than one discourse structure is usually produced for a text. In our experiments, we noticed, at least for English, that the "best" discourse trees are usually those that are skewed to the right. We believe that the explanation of this observation is that text processing is, essentially, a left-to-right process. Usually, people write texts so that the most important ideas go first, both at the paragraph and at the text level. The more text writers add, the more they elaborate on the text that went before: as a consequence, incremental discourse building consists mostly of expansion of the right branches. In order to deal with the ambiguity of discourse, the rhetorical parser computes a weight for each valid discourse tree and retains only those that are maximal. The weight function reflects how skewed to the right a tree is.

4.2 The rhetorical parser in operation

Consider the following text from the November 1996 issue of Scientific American (3). The words in italics denote the discourse markers, the square brackets denote the boundaries of elementary textual units, and the curly brackets denote the boundaries of parenthetical textual units that were determined by the rhetorical parser (see Marcu (1997) for details); the numbers associated with the square brackets are identification labels.

(3) [With its distant orbit \( \text{--} 50 \) percent farther from the sun than Earth \( \text{--} \) and slim atmospheric blanket,\(^1\)] Mars experiences frigid weather conditions.\(^2\) [Surface temperatures typically average about \(-60\) degrees Celsius \( \text{--76} \) degrees Fahrenheit] at the equator and can dip to \(-123\) degrees C near the poles.\(^3\) [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,\(^4\) but any liquid water formed in this way would evaporate almost instantly\(^5\) because of the low atmospheric pressure.\(^6\)]

[Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,\(^7\) most Martian weather involves blowing dust or carbon dioxide.\(^8\)] Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap.\(^9\) [Yet even on the summer pole, where the sun remains in the sky all day long, temperatures never warm enough to melt frozen water.\(^10\)]

Since parenthetical information is related only to the elementary unit that it belongs to, we do not assign it an elementary textual unit status. Such an assignment will only create problems at the formal level as well, because then discourse structures can no longer be represented as binary trees.

On the basis of the data derived from the corpus analysis, the algorithm hypothesizes the following set of relations between the textual units:

\[
\begin{align*}
\text{rhet\_rel(JUSTIFICATION,1,2)} & \lor \\
\text{rhet\_rel(CONDITION,1,2)} & \lor \\
\text{rhet\_rel(ELABORATION,3,[1,2])} & \lor \\
\text{rhet\_rel(ELABORATION,3,[1,2])} & \lor \\
\text{rhet\_rel(ELABORATION,4,6,3)} & \lor \\
\text{rhet\_rel(ELABORATION,4,6,1,3)} & \lor \\
\text{rhet\_rel(CONTRAST,4,5)} & \\
\text{rhet\_rel(EVIDENCE,6,5)} & \lor \\
\text{rhet\_rel(ELABORATION,7,10,[1,6])} & \lor \\
\text{rhet\_rel(CONCESSION,7,8)} & \lor \\
\text{rhet\_rel(EXAMPLE,9,[7,8])} & \lor \\
\text{rhet\_rel(EXAMPLE,9,[7,8])} & \lor \\
\text{rhet\_rel(ANTITHESIS,9,10)} & \lor \\
\text{rhet\_rel(ANTITHESIS,9,10)} & \lor \\
\end{align*}
\]

(4)

The algorithm then determines all the valid discourse trees that can be built for elementary units 1 to 10, given the constraints in (4). In this case, the algorithm constructs 8 different trees. The trees are ordered according to their weights. The "best" tree for text (3) has weight 3 and is fully represented in figure 3. The PostScript file corresponding to figure 3 was automatically generated by

\(^1\)In fact, journalists are trained to employ this "pyramid" approach to writing consciously (Cumming and McKercher, 1994).
Each winter, for example, a blizzard of frozen carbon dioxide rages over the summer pole. Temperatures there can dip to about 230 degrees Fahrenheit, and the wind can blow snow from the opposite polar cap, where the sun remains in the sky all day long. But any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure.

Figure 3: The discourse tree of maximal weight that can be associated with text (3).

4.3 Discussion and evaluation

We believe that there are two ways to evaluate the correctness of the discourse trees that an automatic process builds. One way is to compare the automatically derived trees with trees that have been built manually. Another way is to evaluate the impact that the discourse trees that we derive automatically have on the accuracy of other natural language processing tasks, such as anaphora resolution, intention recognition, or text summarization. In this paper, we describe evaluations that follow both these avenues.

Unfortunately, the linguistic community has not yet built a corpus of discourse trees against which our rhetorical parser can be evaluated with the effectiveness that traditional parsers are. To circumvent this problem, two analysts manually built the discourse trees for five texts that ranged from 161 to 725 words. Although there were some differences with respect to the names of the relations that the analysts used, the agreement with respect to the status assigned to various units (nuclei and satellites) and the overall shapes of the trees was significant.

In order to measure this agreement we associated an importance score to each textual unit in a tree and computed the Spearman correlation coefficients between the importance scores derived from the discourse trees built by each analyst. The Spearman correlation coefficient 2

2The Spearman rank correlation coefficient is an alternative to the usual correlation coefficient. It is based on the ranks of the data, and not on the data itself, and so is resistant to outliers. The null hypothesis tested by Spearman is that two variables
between the ranks assigned for each textual unit on the bases of the discourse trees built by the two analysts was very high: 0.798, at p < 0.0001 level of significance. The differences between the two analysts came mainly from their interpretations of two of the texts: the discourse trees of one analyst mirrored the paragraph structure of the texts, while the discourse trees of the other mirrored a logical organization of the text, which that analyst believed to be important.

The Spearman correlation coefficients with respect to the importance of textual units between the discourse trees built by our program and those built by each analyst were 0.480, p < 0.0001 and 0.449, p < 0.0001. These lower correlation values were due to the differences in the overall shape of the trees and to the fact that the granularity of the discourse trees built by the program was not as fine as that of the trees built by the analysts.

Besides directly comparing the trees built by the program with those built by analysts, we also evaluated the impact that our trees could have on the task of summarizing text. A summarization program that uses the rhetorical parser described here recalled 66% of the sentences considered important by 13 judges in the same five texts, with a precision of 68%. In contrast, a random procedure recalled, on average, only 38.4% of the sentences considered important by the judges, with a precision of 38.4%. And the Microsoft Office 97 summarizer recalled 41% of the important sentences with a precision of 39%.

We discuss at length the experiments from which the data presented above was derived in (Marcu, 1997).

The rhetorical parser presented in this paper uses only the structural constraints that were enumerated in section 2. Co-relational constraints, focus, theme, anaphoric links, and other syntactic, semantic, and pragmatic factors do not yet play a role in our system, but we nevertheless expect them to reduce the number of valid discourse trees that can be associated with a text. We also expect that other robust methods for determining coherence relations between textual units, such as those described by Harabagiu and Moldovan (1995), will improve the accuracy of the routines that hypothesize the rhetorical relations that hold between adjacent units.

We are not aware of the existence of any other rhetorical parser for English. However, Sumita et al. (1992) report on a discourse analyzer for Japanese. Even if one ignores some computational "bonuses" that can be easily exploited by a Japanese discourse analyzer (such as co-reference and topic identification), there are still some key differences between Sumita's work and ours. Particularly important is the fact that the theoretical foundations of Sumita et al.'s analyzer do not seem to be able to accommodate the ambiguity of discourse markers: in their system, discourse markers are considered unambiguous with respect to the relations that they signal. In contrast, our system uses a mathematical model in which this ambiguity is acknowledged and appropriately treated. Also, the discourse trees that we build are very constrained structures (see section 2): as a consequence, we do not overgenerate invalid trees as Sumita et al. do. Furthermore, we use only surface-based methods for determining the markers and textual units and use clauses as the minimal units of the discourse trees. In contrast, Sumita et al. use deep syntactic and semantic processing techniques for determining the markers and the textual units and use sentences as minimal units in the discourse structures that they build. A detailed comparison of our work with Sumita et al.'s and others' work is given in (Marcu, 1997).

5 Conclusion

We introduced the notion of rhetorical parsing, i.e., the process through which natural language texts are automatically mapped into discourse trees. In order to make rhetorical parsing work, we improved previous algorithms for cue phrase disambiguation, and proposed new algorithms for determining the elementary textual units and for computing the valid discourse trees of a text. The solution that we described is both general and robust.

Acknowledgements. This research would have not been possible without the help of Graeme Hirst; there are no right words to thank him for it. I am grateful to Melanie Baljko, Phil Edmonds, and Steve Green for their help with the corpus analysis. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References


