Discourse trees are good indicators of importance in text

Daniel Marcu
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
marcu@isi.edu

Abstract

Researchers in computational linguistics have long speculated that the nuclei of the rhetorical structure tree of a text form an adequate "summary" of the text for which that tree was built. However, to my knowledge, there has been no experiment to confirm how valid this speculation really is.

In this paper, I describe a psycholinguistic experiment that shows that the concepts of discourse structure and nuclearity *can* be used effectively in text summarization. More precisely, I show that there is a strong correlation between the nuclei of the discourse structure of a text and what readers perceive to be the most important units in that text. In addition, I propose and evaluate the quality of an automatic, discourse-based summarization system that implements the methods that were validated by the psycholinguistic experiment. The evaluation indicates that although the system does not match yet the results that would be obtained if discourse trees had been built manually, it still significantly outperforms both a baseline algorithm and Microsoft's Office97 summarizer.

1 Motivation

Traditionally, previous approaches to automatic text summarization have assumed that the salient parts of a text can be determined by applying one or more of the following assumptions:

- important sentences in a text contain words that are used frequently (Luhn 1958; Edmundson 1968);
- important sentences contain words that are used in the title and section headings (Edmundson 1968);
- important sentences are located at the beginning or end of paragraphs (Baxendale 1958);
- important sentences are located at positions in a text that are genre dependent, and these positions can be determined automatically, through training techniques (Kupiec, Pedersen, & Chen 1995; Lin & Hovy 1997; Teufel & Moens 1997);
- important sentences use *bonus words* such as "greatest" and "significant" or *indicator phrases* such as "the main aim of this paper" and "the purpose of this article", while unimportant sentences use *stigma words* such as "hardly" and "impossible" (Edmundson 1968; Rush, Salvador, & Zamora 1971; Kupiec, Pedersen, & Chen 1995; Teufel & Moens 1997);
- important sentences and concepts are the highest connected entities in elaborate semantic structures (Skorochodko 1971; Hoey 1991; Lin 1995; Barzilay & Elhadad 1997; Mani & Bloedorn 1997);
- important and unimportant sentences are derivable from a discourse representation of the text (Sparck Jones 1993b; Ono, Sumita, & Miike 1994).

In determining the words that occur most frequently in a text or the sentences that use words that occur in the headings of sections, computers are accurate tools. Therefore, in testing the validity of using these indicators for determining the most important units in a text, it is adequate to compare the direct output of a summarization program that implements the assumption(s) under scrutiny with a human-made summary or to use human subjects to assess the quality of the generated summaries or their usefulness for carrying out specific tasks. However, in determining the concepts that are semantically related or the discourse structure of a text, computers are no longer so accurate; rather, they are highly dependent on the coverage of the linguistic resources that they use and the quality of the algorithms that they implement. Although it is plausible that elaborate cohesion- and coherence-based structures can be used effectively in summarization, I believe that we should distinguish between the adequacy for summarization of a method that we choose to implement and the adequacy of a particular implementation of that method.

Hence, the position that I advocate in this paper is that, in order to build high-quality summarization programs, we need to evaluate not only a representative set of automatically generated outputs (a highly difficult problem by itself), but also the adequacy of the assumptions that these programs use. That way, we are able to distinguish the problems that arise from a particular implementation from those that arise from the underlying theoretical framework and explore new

Relation name:	EVIDENCE
Constraints on N:	The reader R might not believe the information that is conveyed by the nucleus N to a degree satisfactory to the writer W.
Constraints on S:	The reader believes the information that is conveyed by the satellite S or will find it credible.
Constraints on N + S combination:	R's comprehending S increases R's belief of N.
The effect:	R's belief of N is increased.
Locus of the effect:	N.
Example:	[The truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life:[1]] [we know that 3,000 teens start smoking each day.[2]]

Figure 1: The definition of the EVIDENCE relation in Rhetorical Structure Theory [Mann and Thompson, 1988].

ways to improve each.

To this end, I first review briefly the Rhetorical Structure Theory (Mann & Thompson 1988) and the *rhetorical parsing algorithm* proposed by Marcu (1997a), which takes as input an unrestricted text and derives its discourse structure (see (Marcu 1997b) for details). I then show how one can use discourse structures in order to assign to each textual unit an importance score and to determine the most important units of the corresponding text. In section 3, I describe a psycholinguistic experiment that shows that the mapping between discourse structures and importance scores *can* be used effectively for determining the most important units in a text. More precisely, I show that there is a strong correlation between the nuclei of a discourse structure of a text and what readers perceive to be the most important units in a text. I end the paper with an evaluation of an implemented summarization system that uses the discourse structures derived by the rhetorical parser (Marcu 1997a) and with a broader analysis of the summarization and evaluation methodologies that I employed.

2 From discourse structures to text summaries

A short review of Rhetorical Structure Theory

Driven mostly by research in natural language generation, Rhetorical Structure Theory (RST) (Mann & Thompson 1988) has become one of the most popular discourse theories of the last decade. Central to the theory is the notion of *rhetorical relation*, which is a relation that holds between two non-overlapping text spans called NUCLEUS and SATELLITE. (There are a few exceptions to this rule: some relations, such as CONTRAST, are multinuclear.) The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's purpose than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa.

Text coherence in RST is assumed to arise from a set of constraints and an overall effect that are associated with each relation. The constraints operate on the nucleus, on the satellite, and on the combination of nucleus and satellite. For example, an EVIDENCE relation (see figure 1) holds between the nucleus (labelled as 1 in the example) and the satellite (labelled as 2 in the example), because the nucleus presents some information that the writer believes to be insufficiently supported to be accepted by the reader; the satellite presents some information that is thought to be believed by the reader or that is credible to her; and the comprehension of the satellite increases the reader's belief in the nucleus. The effect of the relation is that the reader's belief in the information presented in the nucleus is increased. Rhetorical relations can be assembled into rhetorical structure trees (RS-trees) by recursively applying individual relations to spans that range in size from one clause-like unit to the whole text.

Recent developments in computational linguistics have created the means for deriving the rhetorical structure of unrestricted texts. For example, when the text shown in (1), below, is given as input to the *rhetorical parsing algorithm* that is discussed in detail in (Marcu 1997a; 1997b), it is broken into ten elementary units (those surrounded by square brackets), and two parenthetical units (those surrounded by curly brackets).[1] The rhetorical parsing algorithm uses the cue phrases shown in italics in (1) in order to hypothesize rhetorical relations among the elementary units. Eventually, the algorithm derives the rhetorical structure tree shown in figure 2.

(1) [*With* its distant orbit {— 50 percent farther from the sun than Earth —^{P1}} and slim atmospheric blanket,[1]] [Mars experiences frigid weather conditions.[2]] [Surface temperatures typically average about −60 degrees Celsius (−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

[1] Parenthetical units are related only to the elementary units that they belong to; their deletion does not affect the coherence of the text.

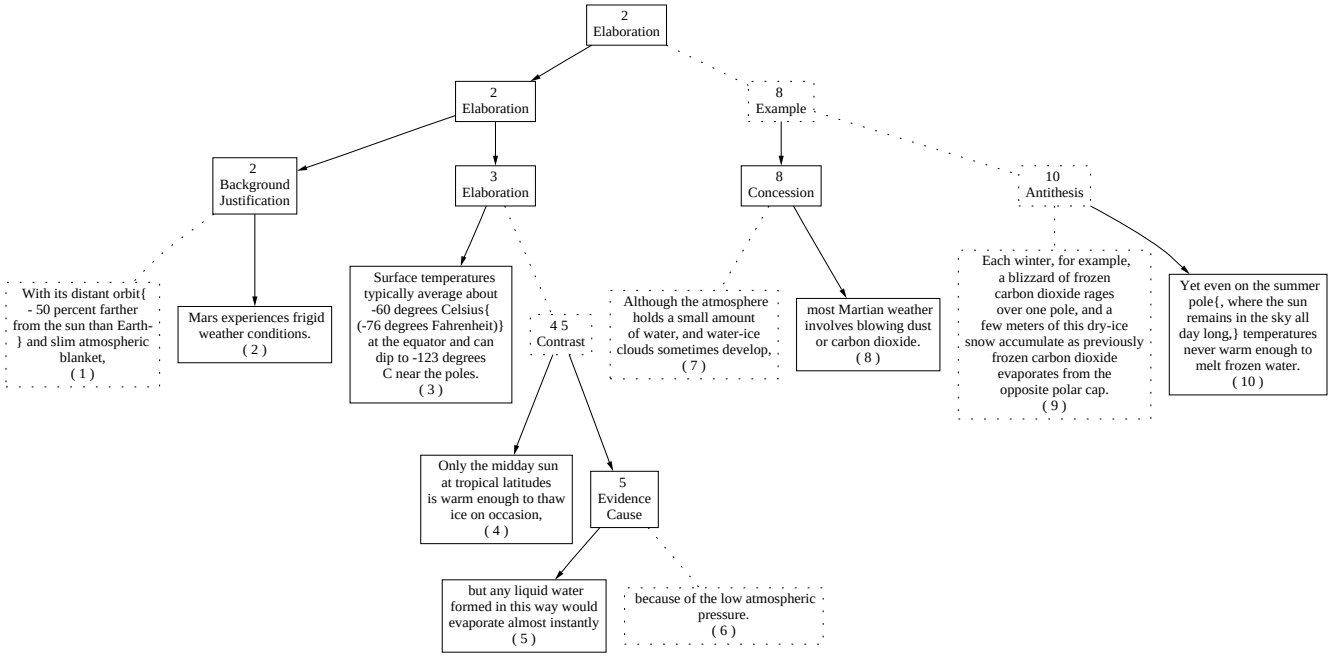

Figure 2: The discourse tree for text (1) that is built by the rhetorical parsing algorithm.

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,^{P10}} temperatures never warm enough to melt frozen water.[10]]

This discourse structure obeys the constraints put forth by Mann and Thompson (1988) and Marcu (1996). It is a binary tree whose leaves are the elementary textual units in (1). Each node in the tree plays either the role of nucleus or satellite. In figure 2, nuclei are represented by solid boxes, while satellites are represented by dotted boxes. The internal nodes of the discourse structure are labelled with names of rhetorical relations: for example, according to figure 2, a rhetorical relation of CONCESSION holds between units 7 and 8; and a rhetorical relation of ELABORATION holds between the text span that contains units [1–2] and the text span that contains units [3–6]. In addition to the names of rhetorical relations, each internal node has a *promotion set* that is given by the *salient* or *promotion* units of that node. The salient units are the most important units in the corresponding text span. They are determined in a bottom-up fashion, as follows: The salient unit of a leaf is the leaf itself; the salient units of an internal node are given by the union of the salient units of its immediate nuclear children. For example, the node that spans units [4–6] has salient units 4 and 5 because the immediate children of the node labelled with relation CONTRAST are both nuclei, which have promotion units 4 and 5 respectively; the root node, which spans units [1–10] has 2 as its salient unit because only the node that corresponds to span [1–6] is a nucleus, whose salient unit is 2. In figure 2, parent nodes are linked to subordinated nuclei by solid arrows; parent nodes are linked to subordinated satellites by dotted lines.[2]

From discourse structures to importance scores

Researchers in computational linguistics have long speculated that the nuclei of a rhetorical structure tree, such as that shown in figure 2, constitute an adequate summarization of the text for which that tree was built (Mann & Thompson 1988; Matthiessen & Thompson 1988; Hobbs 1993; Polanyi 1993; Sparck Jones 1993a; 1993b). As we have discussed in the previous subsection, the elementary units in the promotion set of a node of a tree structure, which depend on the nuclear statuses of its immediate children, denote the most important units of the textual span that is dominated by that node. A simple inspection of the structure in figure 2, for example, allows us to determine that

[2]For a first-order formalization of the mathematical properties of the discourse structure shown in figure 2, see (Marcu 1996; 1997b).

Unit	1	$P1$	2	3	4	5	6	7	8	9	10	$P10$
Score	3	2	6	4	3	3	1	3	5	3	4	2

Table 1: The importance scores of the textual units in text (1).

unit 2 is the most important textual unit in text (1) because it is the only promotion unit associated with the root node. Similarly, we can determine that unit 3 is the most important unit of span [3–6] and that units 4 and 5 are the most important units of span [4–6].

A more general way of exploiting the ideas of nuclearity, discourse structure, and promotion units that are associated with a discourse tree is from the perspective of text summarization. If we repeatedly apply the concept of salience to each of the nodes of a discourse structure, we can induce a partial ordering on the importance of all the units of a text. The intuition behind this approach is that the textual units that are in the promotion sets of the top nodes of a discourse tree are more important than the units that are salient in the nodes found at the bottom. A very simple way to induce such an ordering is by computing a score for each elementary unit of a text on the basis of the depth in the tree structure of the node where the unit occurs first as a promotion unit. The larger the score of a unit, the more important that unit is considered to be in a text. Formula (2), below, provides a recursive definition for computing the importance score $s(u, D, d)$ of a unit u in a discourse structure D that has depth d.

$$\textbf{(2)}\quad s(u,D,d) = \begin{cases} 0 & \text{if } D \text{ is NIL,} \\ d & \text{if } u \in prom(D), \\ d-1 & \text{if } u \in paren(D), \\ max(s(u, & \\ \qquad C(D), & \\ \qquad d-1)) & \text{otherwise.} \end{cases}$$

The formula assumes that the discourse structure is a tree and that the functions $prom(D)$, $paren(D)$, and $C(D)$ return the promotion set, parenthetical units, and the child subtrees of each node respectively. If a unit is among the promotion set of a node, its score is given by the current value of d. If a unit is among the parenthetical units of a node, which can happen only in the case of a leaf node, the score assigned to that unit is $d-1$ because the parenthetical unit can be represented as a direct child of the elementary unit to which is related. For example, when we apply formula (2) to the tree in figure 2, which has depth 6, we obtain the scores in table 1 for each of the elementary and parenthetical units of text (1). Because unit 2 is among the promotion units of the root, it gets a score of 6. Unit 3 is among the promotion units of a node found two levels below the root, so it gets a score of 4. Unit 6 is among the promotion units of a leaf found 5 levels below the root, so it gets a score of 1. Unit $P1$ is a parenthetical unit of elementary unit 1, so its score is $s(1, D, 6) - 1 = 3 - 1 = 2$ because the elementary unit to which it belongs is found 3 levels below the root.

If we consider now the importance scores that are induced on the textual units by the discourse structure and formula (2), we can see that they correspond to a partial ordering on the importance of these units in a text. This ordering enables the construction of text summaries with various degrees of granularity: for example, the partial ordering shown in (3) was induced on the textual units of text (1) by the discourse structure in figure 2 and formula (2).

$$\textbf{(3)}\quad 2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > P1, P10 > 6$$

If we are interested in generating a very short summary of text (1), we can create a text with only one unit, which is unit 2. A longer summary can contain units 2 and 8; a longer one, units 2, 8, 3, and 10; and so on.

A discourse-based summarizer

Given that we can use the rhetorical parser described by Marcu (1997a; 1997b) to build the discourse structure of any text and that we can use formula (2) to determine the partial ordering that is consistent with the idea that the nuclei of a discourse structure constitute a good summary of a text, it is trivial now to implement a summarization program.

The summarization algorithm in figure 3 takes two arguments: a text and a number p between 1 and 100. It first uses the rhetorical parsing algorithm in order to determine the discourse structure of the text given as input. It then applies formula (2) and determines a partial ordering on the elementary and parenthetical units of the text. It then uses the partial ordering in order to select the p% most important textual units of the text.

The idea that I emphasized in the introductory section was that if we would take now the most important p% textual units and ascertain their suitability for a text summary, we would evaluate only the quality of one particular implementation of the discourse-based summarization algorithm. However, since this algorithm constructs trees that are not always correct (Marcu 1997a), such an evaluation would not assess the appropriateness of the discourse-based method for text summarization. In order to distinguish between the quality of the method and the quality of the implementation, I designed a psycholinguistic experiment that shows that the *theoretical concepts* of discourse structure and nuclearity *can* be used effectively for determining the most important units in a text. Once the suitability of using discourse structures for text summarization is established (see section 3), I turn back to the evaluation of

Input: A text T
A number p, such that $1 \leq p \leq 100$.
Output: The most important $p\%$ of the elementary units of T.

1. I. Determine the discourse structure DS of T by means of the rhetorical
2. parsing algorithm (Marcu 1997a; 1997b).
3. II. Determine a partial ordering on the elementary and parenthetical
4. units of DS by means of formula (2).
5. III. Select the first $p\%$ units of the ordering.

Figure 3: The discourse-based summarization algorithm

the algorithm and discuss its strengths and weaknesses. As will become apparent in section 4, although the implementation does not generate summaries of the quality predicted by the evaluation of the method, it still significantly outperforms both a baseline algorithm and Microsoft's Office97 summarizer.

3 From discourse structure to text summaries — an empirical view

Materials and methods of the experiment

We know from the results reported in the psychological literature on summarization (Johnson 1970; Chou Hare & Borchardt 1984; Sherrard 1989) that there exists a certain degree of disagreement between readers with respect to the importance that they assign to various textual units and that the disagreement is dependent on the quality of the text and the comprehension and summarization skills of the readers (Winograd 1984). In an attempt to produce an adequate reference set of data, I selected for my experiment five short texts from *Scientific American* that I considered to be well-written. The texts ranged in size from 161 to 725 words. The shortest text was that shown in (1).

Because my intention was to evaluate the adequacy for summarizing text not only of a particular implementation but also of discourse-based methods in general, I first determined manually the minimal textual units of each text. Overall, I broke the five texts into 160 textual units with the shortest text being broken into 18 textual units, and the longest into 70. Each textual unit was enclosed within square brackets and numbered. For example, when the text on Mars was manually broken into elementary units, I obtained not 10 units, as in the case when the rhetorical parsing algorithm was applied (see text (1)), but 18. The text whose minimal units were obtained manually is given in (4), below.

(4) [With its distant orbit[1]] [— 50 percent farther from the sun than Earth —[2]] [and slim atmospheric blanket,[3]] [Mars experiences frigid weather conditions.[4]] [Surface temperatures typically average about –60 degrees Celsius (–76 degrees Fahrenheit) at the equator[5]] [and can dip to –123 degrees C near the poles.[6]] [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,[7]] [but any liquid water formed in this way would evaporate almost instantly[8]] [because of the low atmospheric pressure.[9]]

[Although the atmosphere holds a small amount of water,[10]] [and water-ice clouds sometimes develop,[11]] [most Martian weather involves blowing dust or carbon dioxide.[12]] [Each winter, for example, a blizzard of frozen carbon dioxide rages over one pole,[13]] [and a few meters of this dry-ice snow accumulate[14]] [as previously frozen carbon dioxide evaporates from the opposite polar cap.[15]] [Yet even on the summer pole,[16]] [where the sun remains in the sky all day long,[17]] [temperatures never warm enough to melt frozen water.[18]]

I followed Johnson's (1970) and Garner's (1982) strategy and asked 13 independent judges to rate each textual unit according to its importance to a potential summary. The judges used a three-point scale and assigned a score of 2 to the units that they believed to be very important and should appear in a concise summary, 1 to those they considered moderately important, which should appear in a long summary, and 0 to those they considered unimportant, which should not appear in any summary. The judges were instructed that there were no right or wrong answers and no upper or lower bounds with respect to the number of textual units that they should select as being important or moderately important. The judges were all graduate students in computer science; I assumed that they had developed adequate comprehension and summarization skills on their own, so no training session was carried out. Table 2 presents the scores that were assigned by each judge to the units in text (4).

The same texts were also given to two computational linguistics analysts with solid knowledge of Rhetorical Structure Theory. The analysts were asked to build one rhetorical structure tree (RS-tree) for each text. I took then the RS-trees built by the analysts and associated with each node in a tree its salient units. I then computed for each textual unit a score, by applying formula (2). Table 2 also presents the scores that were derived from the RS-trees that were built by each analyst for text (4) and the scores that were derived from

Unit	Judges													Analysts		Program
	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	
1	0	2	2	2	0	0	0	0	0	0	0	0	0	3	3	3
2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2
3	0	2	0	2	0	0	0	0	0	0	0	0	1	3	2	3
4	2	1	2	2	2	2	2	2	2	2	2	2	2	6	5	6
5	1	1	0	1	1	1	0	1	2	1	0	2	2	4	3	4
6	0	1	0	1	1	1	0	1	1	1	0	2	2	4	3	4
7	0	2	1	0	0	0	1	1	1	0	0	0	0	4	3	3
8	0	1	0	0	0	0	0	0	0	0	0	0	0	4	3	3
9	0	0	2	0	0	0	0	0	0	0	1	0	1	1	0	1
10	0	2	2	2	0	0	2	0	0	0	0	0	0	3	4	3
11	0	0	0	2	0	0	0	1	0	0	0	0	1	3	4	3
12	2	2	2	2	2	2	2	2	2	0	1	2	2	5	4	5
13	1	1	0	0	0	1	0	1	0	0	0	2	0	3	3	3
14	1	0	0	0	0	1	1	0	0	0	0	2	0	3	3	3
15	0	0	0	0	0	1	0	0	0	0	0	1	0	2	3	3
16	0	1	1	0	1	0	0	0	2	0	0	1	0	4	3	4
17	0	1	0	0	0	0	0	0	1	0	0	1	0	2	1	2
18	2	1	1	0	1	0	1	0	2	0	1	1	2	4	3	4

Table 2: The scores assigned by the judges, analysts, and the discourse-based summarizer to the textual units in text (4).

Text	T.1	T.2	T.3	T.4	T.5	Overall
All units	72.64	73.23	69.23	69.89	70.08	70.67
Very important units	88.46	63.07	64.83	63.73	67.30	65.66
Less important units	51.28	73.07	53.84	46.15	–	58.04
Unimportant units	75.14	82.51	73.07	72.85	71.25	73.86

Table 3: Percent agreement with the majority opinion.

the discourse tree that was built by the discourse-based summarizer.

Usually, the granularity of the trees that are built by the rhetorical parser is coarser than the granularity of those that are built manually. The last column in table 2 reflects this: all the units that were determined manually and that overlapped an elementary unit determined by the rhetorical parser were assigned the same score. For example, units 1 and 3 in text (4) correspond to unit 1 in text (1). Because the score of unit 1 in the discourse structure that is built by the rhetorical parser is 3, both units 1 and 3 in text (4) are assigned the score 3.

Agreement among judges

Overall agreement among judges. I measured the agreement of the judges with one another, by means of the notion of *percent agreement* that was defined by Gale (1992) and used extensively in discourse segmentation studies (Passonneau & Litman 1993; Hearst 1997). Percent agreement reflects the ratio of observed to possible agreements with the majority opinion. The percent agreements computed for each of the five texts and each level of importance are given in table 3. The agreements among judges for my experiment seem to follow the same pattern as those described by other researchers in summarization (Johnson 1970). That is, the judges are quite consistent with respect to what they perceive as being very important and unimportant, but less consistent with respect to what they perceive as being less important. In contrast with the agreement observed among judges, the percentage agreements computed for 1000 importance assignments that were randomly generated for the same texts followed a normal distribution with $\mu = 47.31, \sigma = 0.04$. These results suggest that the agreement among judges is significant.

Agreement among judges with respect to the importance of each textual unit. I considered a textual unit to be labelled consistently if a simple majority of the judges (≥ 7) assigned the same score to that unit. Overall, the judges labelled consistently 140 of the 160 textual units (87%). In contrast, a set of 1000 randomly generated importance scores showed agreement, on average, for only 50 of the 160 textual units (31%), $\sigma = 0.05$.

The judges consistently labelled 36 of the units as very important, 8 as less important, and 96 as unimportant. They were inconsistent with respect

Text T.1	Text T.2	Text T.3	Text T.4	Text T.5	Overall
0.645	0.676	0.960	0.772	0.772	0.798

Table 4: The Spearman correlation coefficients between the ranks assigned to each textual unit on the basis of the RS-trees built by the two analysts.

to 20 textual units. For example, for text (4), the judges consistently labelled units 4 and 12 as very important, units 5 and 6 as less important, units $1, 2, 3, 7, 8, 9, 10, 11, 13, 14, 15, 17$ as unimportant, and were inconsistent in labeling unit 18. If we compute percent agreement figures only for the textual units for which at least 7 judges agreed, we get 69% for the units considered very important, 63% for those considered less important, and 77% for those considered unimportant. The overall percent agreement in this case is 75%.

Statistical significance. It has often been emphasized that agreement figures of the kinds computed above could be misleading (Krippendorff 1980; Passonneau & Litman 1993). Since the "true" set of important textual units cannot be independently known, we cannot compute how valid the importance assignments of the judges were. Moreover, although the agreement figures that would occur by chance offer a strong indication that our data are reliable, they do not provide a precise measurement of reliability.

To compute a reliability figure, I followed the same methodology as Passonneau and Litman (1993) and Hearst (1997) and applied Cochran's Q summary statistics to the data (Cochran 1950). Cochran's test assumes that a set of judges make binary decisions with respect to a dataset. The null hypothesis is that the number of judges that take the same decision is randomly distributed. Since Cochran's test is appropriate only for binary judgments and since my main goal was to determine a reliability figure for the agreement among judges with respect to what they believe to be important, I evaluated two versions of the data that reflected only one importance level. In the first version I considered as being important the judgments with a score of 2 and unimportant the judgments with a score of 0 and 1. In the second version, I considered as being important the judgments with a score of 2 and 1 and unimportant the judgments with a score of 0. Essentially, I mapped the judgment matrices of each of the five texts into matrices whose elements ranged over only two values: 0 and 1. After these modifications were made, I computed for each version and each text the Cochran Q statistics, which approximates the χ^2 distribution with $\mathrm{N} - 1$ degrees of freedom, where N is the number of elements in the dataset. In all cases I obtained probabilities that were very low: $p < 10^{-6}$. This means that the agreement among judges was extremely significant.

Although the probability was very low for both versions, it was lower for the first version of the modified data than for the second. This means that it is more reliable to consider as important only the units that were assigned a score of 2 by a majority of the judges.

As I have already mentioned, my ultimate goal was to determine whether there exists a correlation between the units that judges find important and the units that have nuclear status in the rhetorical structure trees of the same texts. Since the percentage agreement for the units that were considered very important was higher than the percentage agreement for the units that were considered less important, and since the Cochran's significance computed for the first version of the modified data was higher that the one computed for the second, I decided to consider the set of 36 textual units labelled by a majority of judges with 2 as a reliable reference set of importance units for the five texts. For example, units 4 and 12 from text (4) belong to this reference set.

Agreement between analysts

Once I determined the set of textual units that the judges believed to be important, I needed to determine the agreement between the analysts who built the discourse trees for the five texts. Because I did not know the distribution of the importance scores derived from the discourse trees, I computed the correlation between the analysts by applying Spearman's correlation coefficient on the scores associated to each textual unit. I interpreted these scores as ranks on a scale that measures the importance of the units in a text.

The 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 the Spearman coefficient is that two variables are independent of each other, against the alternative hypothesis that the rank of a variable is correlated with the rank of another variable. The value of the statistics ranges from -1, indicating that high ranks of one variable occur with low ranks of the other variable, through 0, indicating no correlation between the variables, to $+1$, indicating that high ranks of one variable occur with high ranks of the other variable.

The Spearman correlation coefficient between the ranks assigned for each textual unit on the bases of the RS-trees built by the two analysts was high for each of the five texts. It ranged from 0.645, for text T.1, to 0.960, for text T.3 at the $p < 0.0001$ level of significance. The Spearman correlation coefficient between the ranks assigned to the textual units of all five texts was 0.798, at the $p < 0.0001$ level of significance (see

Text	No. of units that were considered important by judges	First Analyst			
		No. of units that were labelled as important on the basis of the RS-tree built by the analyst	No. of units that were correctly labelled as important on the basis of the RS-tree built by the analyst	Recall	Precision
T.1	2	2	2	100.00	100.00
T.2	9	6	5	55.55	83.33
T.3	7	5	4	57.14	80.00
T.4	12	10	6	50.00	60.00
T.5	6	7	3	50.00	42.85
All	36	30	20	55.55	66.66

Table 5: Summarization results obtained by using the text structures built by the first analyst — the clause-like unit case.

Text	No. of units that were considered important by judges	Second Analyst			
		No. of units that were labelled as important on the basis of the RS-tree built by the analyst	No. of units that were correctly labelled as important on the basis of the RS-tree built by the analyst	Recall	Precision
T.1	2	1	1	50.00	50.00
T.2	9	8	6	66.66	75.00
T.3	7	5	4	57.14	80.00
T.4	12	7	5	41.66	71.42
T.5	6	9	4	66.66	44.44
All	36	30	20	55.55	66.66

Table 6: Summarization results obtained by using the text structures built by the second analyst — the clause-like unit case.

table 4).

Agreement between the analysts and the judges with respect to the most important textual units

In order to determine whether there exists any correspondence between what readers believe to be important and the nuclei of the RS-trees, I selected, from each of the five texts, the set of textual units that were labelled as "very important" by a majority of the judges. For example, for text (4), I selected units 4 and 12, i.e., 11% of the units. Overall, the judges selected 36 units as being very important, which is approximately 22% of the units in all the texts. The percentages of important units for the five texts were 11, 36, 35, 17, and 22 respectively.

I took the maximal scores computed for each textual unit from the RS-trees built by each analyst and selected a percentage of units that matched the percentage of important units selected by the judges. In the cases in which there were ties, I selected a percentage of units that was closest to the one computed for the judges. For example, I selected units 4 and 12, which represented the most important 11% of the units that were induced by formula (2) on the RS-tree built by the first analyst. However, I selected only unit 4, which represented 6% of the most important units that were induced on the RS-tree built by the second analyst, because units 10, 11, and 12 have the same score (see table 2). If I had selected units 10, 11 and 12 as well, I would have ended up selecting 22% of the units in text (4), which is farther from 11 than 6. Hence, I determined for each text the set of important units as labelled by judges and as derived from the RS-trees of those texts.

I calculated for each text the recall and precision of the important units derived from the RS-trees, with respect to the units labelled important by the judges. The overall recall and precision was the same for both analysts: 55.55% recall and 66.66% precision. In contrast, the average recall and precision for the same percentages of units selected randomly 1000 times from the same five texts were both 25.7%, $\sigma = 0.059$. Tables 5 and 6 show the recall and precision figures for each an-

Text	No. of units that were considered important by judges	First Analyst			
		No. of units that were labelled as important on the basis of the RS-tree built by the analyst	No. of units that were correctly labelled as important on the basis of the RS-tree built by the analyst	Recall	Precision
T.1	7	7	7	100.00	100.00
T.2	12	12	12	100.00	100.00
T.3	10	9	8	80.00	88.88
T.4	18	11	8	44.44	72.72
T.5	11	10	5	45.45	50.00
All	58	49	40	68.96	81.63

Table 7: Summarization results obtained by using the text structures built by the first analyst — the sentence case.

Text	No. of units that were considered important by judges	Second Analyst			
		No. of units that were labelled as important on the basis of the RS-tree built by the analyst	No. of units that were correctly labelled as important on the basis of the RS-tree built by the analyst	Recall	Precision
T.1	7	7	7	100.00	100.00
T.2	12	11	9	75.00	81.81
T.3	10	9	8	80.00	88.88
T.4	18	11	6	33.33	54.54
T.5	11	13	8	72.72	61.53
All	58	51	38	65.51	74.50

Table 8: Summarization results obtained by using the text structures built by the second analyst — the sentence case.

alyst and each of the five texts.

In summarizing text, it is often useful to consider not only clause-like units, but full sentences. To account for this, I considered as important all the textual units that pertained to a sentence that was characterized by at least one important textual unit. For example, I labelled as important textual units 1 to 4 in text (4), because they make up a full sentence and because unit 4 was labelled as important. For the adjusted data, I determined again the percentages of important units for the five texts and I recalculated the recall and precision for both analysts: the recall was 68.96% and 65.51% and the precision 81.63% and 74.50% respectively. Tables 7 and 8 show the sentence-related recall and precision figures for each analyst and each of the five texts.

In contrast with the results in tables 7 and 8, the average recall and precision for the same percentages of units selected randomly 1000 times from the same five texts were 38.4%, $\sigma = 0.048$. These results confirm that there exists a strong correlation between the nuclei of the RS-trees that pertain to a text and what readers perceive as being important in that text. Given the values of recall and precision that I obtained, it is plausible that an adequate computational treatment of discourse theories would provide most of what is needed for selecting accurately the important units in a text. However, the results also suggest that the discourse theory that was employed here is not enough by itself if one wants to strive for perfection.

The above results not only provide strong evidence that discourse theories can be used effectively for text summarization, but also suggest strategies that an automatic summarizer might follow. For example, the Spearman correlation coefficient between the judges and the first analyst, the one who did not follow the paragraph structure, was lower than that between the judges and the second analyst. This might suggest that human judges are inclined to use the paragraph breaks as valuable sources of information when they interpret discourse. If the aim of a summarization program is to mimic human behavior, it would then seem adequate for the program to take advantage of the paragraph structure of the texts that it analyzes.

Text	No. of units that were considered important by judges	Discourse-based Summarizer			
		No. of units that were labelled as important on the basis of the tree built by the rhetorical parser	No. of units that were correctly labelled as important on the basis of the tree built by the rhetorical parser	Recall	Precision
T.1	2	2	2	100.00	100.00
T.2	9	8	5	55.55	62.50
T.3	7	8	3	42.85	37.50
T.4	12	14	6	50.00	42.85
T.5	6	6	3	50.00	50.00
All	36	38	19	52.77	50.00

Table 9: Summarization results obtained by using the text structures built by the rhetorical parser — the clause-like unit case.

Text	No. of units that were considered important by judges	Discourse-based Summarizer			
		No. of units that were labelled as important on the basis of the tree built by the rhetorical parser	No. of units that were correctly labelled as important on the basis of the tree built by the rhetorical parser	Recall	Precision
T.1	7	7	7	100.00	100.00
T.2	12	14	12	100.00	85.71
T.3	10	9	6	60.00	66.66
T.4	18	20	10	55.55	50.00
T.5	11	6	5	45.45	83.33
All	58	56	38	65.51	67.85

Table 10: Summarization results obtained by using the text structures built by the rhetorical parser — the sentence case.

4 An evaluation of the discourse-based summarization program

Agreement between the results of the summarization program and the judges with respect to the most important textual units

To evaluate the summarization program, I followed the same method as in section 4. That is, I used the importance scores assigned by formula (2) to the units of the discourse trees built by the rhetorical parser in order to compute statistics similar to those discussed in conjunction with the manual analyses. Tables 9 and 10 summarize the results.

When the program selected only the textual units with the highest scores, in percentages that were equal to those of the judges, the recall was 52.77% and the precision was 50%. When the program selected the full sentences that were associated with the most important units, in percentages that were equal to those of the judges, the recall was 65.51% and the precision 67.85%. Tables 9 and 10 show recall and precision results for each of the five texts that were summarized. The lower recall and precision scores associated with clause-like units seem to be caused primarily by the difference in granularity with respect to the way the texts were broken into subunits: the program does not recover all minimal textual units, and as a consequence, its assignment of importance scores is coarser. When full sentences are considered, the judges and the program work at the same level of granularity, and as a consequence, the summarization results improve.

Comparison with other work

I am not aware of any other discourse-based summarization program for English. However, Ono et al. (1994) discuss a summarization program for Japanese that uses a discourse parser built by Sumita (1992) and that constructs trees whose minimal textual units are sentences. Due to the differences between English and

Text	No. of units considered important by judges	Microsoft Office97 Summarizer			
		No. of units identified	No. of units identified correctly	Recall	Precision
T.1	2	3	1	50.00	33.33
T.2	9	10	5	55.55	50.00
T.3	7	9	3	42.85	33.33
T.4	12	11	1	8.33	9.09
T.5	6	6	0	0.00	0.00
All	36	39	10	27.77	25.64

Table 11: Recall and precision figures obtained with the Microsoft Office97 summarizer — the clause-like unit case.

Text	No. of units considered important by judges	Microsoft Office97 Summarizer			
		No. of units identified	No. of units identified correctly	Recall	Precision
T.1	7	8	3	42.85	37.50
T.2	12	12	5	41.66	41.66
T.3	10	11	8	80.00	72.72
T.4	18	20	3	16.66	15.00
T.5	11	11	5	45.45	45.45
All	58	62	24	41.37	38.70

Table 12: Recall and precision figures obtained with the Microsoft Office97 summarizer — the sentence case.

Japanese, it was impossible to compare Ono's summarizer with ours. Fundamental differences concerning the assumptions that underlie Ono's work and the work described here are discussed at length in (Marcu 1997b; 1998b).

I was able to obtain one other program that summarizes English text — the program included in the Microsoft Office97 package. I ran the Microsoft summarization program on the five texts from *Scientific American* and selected the same percentages of textual units as those considered important by the judges. When I selected percentages of text that corresponded only to the clause-like units considered important by the judges, the Microsoft program recalled 27.77% of the units, with a precision of 25.64%. When I selected percentages of text that corresponded to sentences considered important by the judges, the Microsoft program recalled 41.37% of the units, with a precision of 38.70%. Tables 11 and 12 show the recall and precision figures for each of the five texts.

In order to provide a better understanding of the results in this section, I also considered a baseline algorithm that randomly selects from a text a number of units that matches the number of units that were considered important in that text by the human judges. Tables 13 and 14 show recall and precision results for the baseline, Microsoft Office97, and discourse-based summarizers, as well as the results that would have been obtained if we had applied the score function (2) on the discourse trees that were built manually. In tables 13 and 14, I use the term "Analyst-based Summarizer" as a name for a summarizer that identifies important units on the basis of discourse trees that are manually built. The recall and precision figures associated with the baseline algorithm that selects textual units randomly represent averages of 1000 runs. The recall and precision results associated with the "Analyst-based Summarizer" in tables 13 and 14 are averages of the results shown in tables 5 and 6, and 7 and 8 respectively.

Discussion

Discussion of the summarization methodology. Throughout this paper, I used words such as "summary" and "to summarize" in a very narrow way, which often equated summarization with the process of selecting the most important units in a text. Obviously, selecting the salient units is only part of the problem that a sophisticated summarization system needs to solve, because the information encoded in the selected units must be eventually mapped into coherent abstracts. From this perspective, the experiment described here showed only that discourse structures can be reliably used to *extract* salient textual units at a level that is comparable to that of humans. However, the experiment did not show that these units are actually useful

Text	Baseline Summarizer	Microsoft Summarizer		Discourse-based Summarizer		Analyst-based Summarizer	
	Recall & Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.
T.1	12.05	50.00	33.33	100.00	100.00	75.00	75.00
T.2	38.01	55.55	50.00	55.55	62.50	61.11	78.57
T.3	36.20	42.85	33.33	42.85	37.50	57.14	57.14
T.4	18.32	8.33	9.09	50.00	42.85	45.83	64.70
T.5	23.06	0.00	0.00	50.00	50.00	58.33	43.75
All	25.7	27.77	25.64	52.77	50.00	55.55	66.66

Table 13: Recall and precision figures obtained with the baseline, Microsoft Office97, discourse-based, and analyst-based summarizers — the clause-like unit case.

Text	Baseline Summarizer	Microsoft Summarizer		Discourse-based Summarizer		Analyst-based Summarizer	
	Recall & Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.
T.1	40.12	42.85	37.50	100.00	100.00	100.00	100.00
T.2	50.02	41.66	41.66	100.00	85.71	87.50	91.30
T.3	52.12	80.00	72.72	60.00	66.66	80.00	88.88
T.4	26.91	16.66	15.00	55.55	50.00	38.88	63.63
T.5	42.31	45.45	45.45	45.45	83.83	59.09	56.52
All	38.40	41.37	38.70	65.51	67.85	67.24	78.00

Table 14: Recall and precision figures obtained with the baseline, Microsoft Office97, discourse-based, and analyst-based summarizers — the sentence case.

to create *abstracts* of document content.

The simplest way to make use of a set of textual units that are extracted according to any salience-based method is to catenate them in their order of occurrence in the original text. When I applied this procedure to the texts used in this experiment, I found that the resulting extracts read well — after all, the discourse-based summarizer selects nuclei, which represent what is most essential for the writer's purpose and which can be understood independent of their satellites. Yet, I have not carried out any readability evaluation. Although it is plausible that the extracts resulted from catenating the units that are identified as salient by a discourse structurer read better than extracts resulted from catenating units identified using lexically-based, position-based, or other heuristics, a discourse-based summarizer still needs to solve the problem of dangling references: in some cases, the selected units use anaphoric expressions to referents that were not selected. It is likely that by exploiting the relationship between discourse structure and anaphora (Fox 1987), one can provide an elegant solution to this problem. Also, by taking advantage of the rhetorical relations that hold between the selected units, which are implicitly represented in the discourse structure of a text, one is in a good position to investigate ways of mapping the selected units into coherent abstracts. Dealing with these issues is, however, beyond the scope of this paper.

The results presented here confirm the suitability of using discourse structures for summarizing texts from the *Scientific American* genre. In (Marcu 1998a), I show that the same techniques can be applied successfully for summarizing texts from the newspaper genre as well and provide a methodology for integrating the discourse-based approach to summarization with position-, title-, and semantic-similarity-based approaches.

In spite of the promising results, in some cases, the recall and precision figures obtained with the discourse-based summarizer are still far from 100%. I believe that there are two possible explanations for this: either the rhetorical parser does not construct adequate discourse trees; or the mapping from discourse structures to importance scores is too simplistic. Examining the influence of these factors on summarization requires, however, more sophisticated experiments (see Marcu (1998b) for a discussion).

Discussion of the evaluation methodology. Most of the current summarization systems have focused only on the task of extracting the salient units of a text (usually sentences or paragraphs). Although the evaluation methodology that I employed in this paper is adequate for assessing the quality of such systems, it cannot be probably used at a very large scale because it assumes the existence of a large corpus of texts whose units were manually annotated for salience. Also, it is not very

clear whether this methodology can be applied in the case in which the selected units are smaller than clauses, i.e., they are, for example, concepts, noun compounds, and verbal phrases (see Boguraev and Kennedy (1997) for an approach to summarization that outputs such constructs). A prerequisite for applying the same evaluation methodology in such a case is that human judges can agree on the important concepts of a text. To my knowledge, no experiments have been carried out to investigate this. If the agreement between humans proves to be low, a different evaluation methodology will have to be sought.

5 Conclusions

I described the first experiment that shows that the concepts of rhetorical analysis and nuclearity can be used effectively for summarizing texts. The experiment suggests that discourse-based methods can account for determining the most important units in a text with a recall and precision as high as 70%. I showed how the concepts of rhetorical analysis and nuclearity can be treated algorithmically and I compared recall and precision figures of a summarization program that implements these concepts with recall and precision figures that pertain to a baseline algorithm and to a commercial system, the Microsoft Office97 summarizer. The discourse-based summarization program that I propose outperforms both the baseline and the commercial summarizer. Since the results of the discourse-based summarizer do not match yet the recall and precision figures that pertain to the manual discourse analyses, it is likely that improvements of the rhetorical parser algorithm and more sophisticated mappings from discourse structures to importance scores will result in better performance of subsequent implementations.

Acknowledgements. I am grateful to Graeme Hirst for the invaluable help he gave me during every stage of this work and to Marilyn Mantei, David Mitchell, Kevin Schlueter, and Melanie Baljko for their advice on experimental design and statistics. I am also grateful to Marzena Makuta for her help with the RST analyses and to my colleagues and friends who volunteered to act as judges in the experiments described here.

This research was conducted while I was at the University of Toronto, and was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Barzilay, R., and Elhadad, M. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, 10–17.

Baxendale, P. 1958. Machine-made index for technical literature — an experiment. *IBM Journal of Research and Development* 2:354–361.

Boguraev, B., and Kennedy, C. 1997. Salience-based content characterisation of text documents. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, 2–9.

Chou Hare, V., and Borchardt, K. 1984. Direct instruction of summarization skills. *Reading Research Quarterly* 20(1):62–78.

Cochran, W. 1950. The comparison of percentages in matched samples. *Biometrika* 37:256–266.

Edmundson, H. 1968. New methods in automatic extracting. *Journal of the Association for Computing Machinery* 16(2):264–285.

Fox, B. 1987. *Discourse Structure and Anaphora.* Cambridge Studies in Linguistics; 48. Cambridge University Press.

Gale, W.; Church, K.; and Yarowsky, D. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL–92)*, 249–256.

Garner, R. 1982. Efficient text summarization: costs and benefits. *Journal of Educational Research* 75:275–279.

Hearst, M. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64.

Hobbs, J. 1993. Summaries from structure. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication.*

Hoey, M. 1991. *Patterns of Lexis in Text.* Oxford University Press.

Johnson, R. 1970. Recall of prose as a function of structural importance of linguistic units. *Journal of Verbal Learning and Verbal Behaviour* 9:12–20.

Krippendorff, K. 1980. *Content analysis: An Introduction to its Methodology.* Beverly Hills, CA: Sage Publications.

Kupiec, J.; Pedersen, J.; and Chen, F. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval*, 68–73.

Lin, C., and Hovy, E. 1997. Identifying topics by position. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP–97)*, 283–290.

Lin, C. 1995. Knowledge-based automatic topic identification. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL–95)*, 308–310.

Luhn, H. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2):159–165.

Mani, I., and Bloedorn, E. 1997. Multi-document summarization by graph search and matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI–97)*, 622–628.

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

Marcu, D. 1996. Building up rhetorical structure trees. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI–96)*, volume 2, 1069–1074.

Marcu, D. 1997a. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL–97)*, 96–103.

Marcu, D. 1997b. *The rhetorical parsing, summarization, and generation of natural language texts.* Ph.D. Dissertation, Department of Computer Science, University of Toronto.

Marcu, D. 1998a. Improving summarization through rhetorical parsing tuning. In preparation.

Marcu, D. 1998b. To build text summaries of high quality, nuclearity is not sufficient. In *Working Notes of the AAAI–98 Spring Symposium on Intelligent Text Summarization*, 1–8.

Matthiessen, C., and Thompson, S. 1988. The structure of discourse and 'subordination'. In Haiman, J., and Thompson, S., eds., *Clause combining in grammar and discourse*, volume 18 of *Typological Studies in Language*. John Benjamins Publishing Company. 275–329.

Ono, K.; Sumita, K.; and Miike, S. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the International Conference on Computational Linguistics (Coling–94)*, 344–348.

Passonneau, R., and Litman, D. 1993. Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL–93)*, 148–155.

Polanyi, L. 1993. Linguistic dimensions of text summarization. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication.*

Rush, J.; Salvador, R.; and Zamora, A. 1971. Automatic abstracting and indexing. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of American Society for Information Sciences* 22(4):260–274.

Sherrard, C. 1989. Teaching students to summarize: Applying textlinguistics. *System* 17(1).

Skorochodko, E. 1971. Adaptive method of automatic abstracting and indexing. In *Information Processing*, volume 2, 1179–1182. North-Holland Publishing Company.

Sparck Jones, K. 1993a. Summarising: analytic framework, key component, experimental method. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication.*

Sparck Jones, K. 1993b. What might be in a summary? In *Information Retrieval 93: Von der Modellierung zur Anwendung*, 9–26.

Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; and Amano, S. 1992. A discourse structure analyzer for Japanese text. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, volume 2, 1133–1140.

Teufel, S., and Moens, M. 1997. Sentence extraction as a classification task. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, 58–65.

Winograd, P. 1984. Strategic difficulties in summarizing texts. *Reading Research Quaterly* 19(4):404–425.