From discourse structures to text summaries

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

Abstract

We describe experiments that show that the concepts of rhetorical analysis and nuclearity can be used effectively for determining the most important units in a text. We show how these concepts can be implemented and we discuss results that we obtained with a discourse-based summarization program.

1 Motivation

The evaluation of automatic summarizers has always been a thorny problem: most papers on summarization describe the approach that they use and give some "convincing" samples of the output. In very few cases, *the direct output of a summarization program* is compared with a human-made summary or evaluated with the help of human subjects; usually, the results are modest. Unfortunately, evaluating the results of a particular implementation does not enable one to determine what part of the failure is due to the implementation itself and what part to its underlying assumptions. The position that we take in this paper is that, in order to build high-quality summarization programs, one needs 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, one is able to distinguish the problems that pertain to a particular implementation from those that pertain to the underlying theoretical framework and explore new ways to improve each.

With few exceptions, automatic approaches to summarization have primarily addressed possible ways to determine the most important parts of a text (see Paice (1990) for an excellent overview). Determining the salient parts is considered to be achievable because one or more of the following assumptions hold: (i) important sentences in a text contain words that are used frequently (Luhn, 1958; Edmundson, 1968); (ii) important sentences contain words that are used in the title and section headings (Edmundson, 1968); (iii) important sentences are located at the beginning or end of paragraphs (Baxendale, 1958); (iv) important sentences are located at positions in a text that are genre dependent — these positions can be determined automatically, through training techniques (Lin and Hovy, 1997); (v) 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 non-important sentences use *stigma words* such as "hardly" and "impossible" (Edmundson, 1968; Rush, Salvador, and Zamora, 1971); (vi) important sentences and concepts are the highest connected entities in elaborate semantic structures (Skorochodko, 1971; Lin, 1995; Barzilay and Elhadad, 1997); and (vii) important and non-important sentences are derivable from a discourse representation of the text (Sparck Jones, 1993; Ono, Sumita, and 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. 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, we believe that before building summarization programs, we should determine the extent to which these assumptions hold.

In this paper, we describe experiments that show that the concepts of rhetorical analysis and nuclearity *can* be used effectively for determining the most important units in a text. We show how these concepts were implemented and discuss results that we obtained with a discourse-based summarization program.

2 From discourse trees to summaries — an empirical view

2.1 Introduction

Researchers in computational linguistics (Mann and Thompson, 1988; Matthiessen and Thompson, 1988; Sparck Jones, 1993) have long speculated that the nuclei that pertain to a rhetorical structure tree (RS-tree) (Mann and Thompson, 1988) constitute an adequate summariza-

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	3
18	2	1	1	0	1	0	1	0	2	0	1	1	2	4	3	4

Table 1: The scores assigned by the judges, analysts, and our program to the textual units in text 1.

tion of the text for which that RS-tree was built. However, to our knowledge, there was no experiment to confirm how valid this speculation really is. In what follows, we describe an experiment that shows that there exists a strong correlation between the nuclei of the RS-tree of a text and what readers perceive to be the most important units in a text.

2.2 Experiment

2.2.1 Materials and methods

We know from the results reported in the psychological literature on summarization (Johnson, 1970; Chou Hare and 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, we selected for our experiment five texts from *Scientific American* that we considered to be well-written. The texts ranged in size from 161 to 725 words. We used square brackets to enclose the minimal textual units (essentially the clauses) of each text. Overall, the five texts were broken into 160 textual units with the shortest text being broken into 18 textual units, and the longest into 70. The shortest text is given in (1), below (here, for the purpose of reference, the minimal units are not only enclosed by square brackets, but also are numbered):

(1) [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]]

We followed 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; we assumed that they had developed adequate comprehension and summarization skills on their own, so no training session was carried out. Table 1 presents the scores that were assigned by each judge to the units in text (1).

The same texts were also given to two computational linguists with solid knowledge of rhetorical structure theory (RST). The analysts were asked to build one RS-tree

Text	1	2	3	4	5	All
All units	73	73	69	70	70	71
Very important units	88	63	65	64	67	66
Less important units	51	73	54	46	–	58
Unimportant units	75	83	73	73	71	74

Table 2: Percent agreement with the majority opinion.

for each text. We took then the RS-trees built by the analysts and used our formalization of RST (Marcu, 1996; Marcu, 1997b) to associate with each node in a tree its salient units. The salient units were computed recursively, associating with each leaf in an RS-tree the leaf itself, and to each internal node the salient units of the nucleus or nuclei of the rhetorical relation corresponding to that node. We then computed for each textual unit a score, depending on the depth in the tree where it occurred as a salient unit: the textual units that were salient units of the top nodes in a tree had a higher score than those that were salient units of the nodes found at the bottom of a tree. Essentially, from a rhetorical structure tree, we derived an importance score for each textual unit: the importance scores ranged from 0 to n where n was the depth of the RS-tree.[1] Table 1 presents the scores that were derived from the RS-trees that were built by each analyst for text (1).

2.2.2 Results

Overall agreement among judges. We measured the ability of judges to agree with one another, using the notion of *percent agreement* that was defined by Gale (1992) and used extensively in discourse segmentation studies (Passonneau and Litman, 1993; Hearst, 1994). Percent agreement reflects the ratio of observed agreements with the majority opinion 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 2. The agreements among judges for our 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. We considered a textual unit to be labeled consistently if a simple majority of the judges (≥ 7) assigned the same score to that unit. Overall, the judges labeled 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 labeled 36 of the units as very important, 8 as less important, and 96 as unimportant. They were inconsistent with respect to 20 textual units. For example, for text (1), the judges consistently labeled 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 and 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, we followed the same methodology as Passonneau and Litman (1993) and Hearst (1994) and applied the Cochran's Q summary statistics to our 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 our main goal was to determine a reliability figure for the agreement among judges with respect to what they believe to be important, we evaluated two versions of the data that reflected only one importance level. In the first version we 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, we considered as being important the judgments with a score of 2 and 1 and unimportant the judgments with a score of 0. Essentially, we 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, we computed for each version and each text the Cochran statistics Q, which approximates the χ^2 distribution with $n-1$ degrees of freedom, where n is the number of elements in the dataset. In all cases we 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

[1] Section 3.2 gives an example of how the importance scores were computed.

assigned a score of 2 by a majority of the judges.

As we have already mentioned, our 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, we decided to consider the set of 36 textual units labeled 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 (1) belong to this reference set.

Agreement between analysts. Once we determined the set of textual units that the judges believed to be important, we needed to determine the agreement between the analysts who built the discourse trees for the five texts. Because we did not know the distribution of the importance scores derived from the discourse trees, we computed the correlation between the analysts by applying Spearman's correlation coefficient on the scores associated to each textual unit. We 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, 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 very high: 0.798, at the $p < 0.0001$ level of significance. The differences between the two analysts came mainly from their interpretations of two of the texts: the RS-trees of one analyst mirrored the paragraph structure of the texts, while the RS-trees of the other mirrored a logical organization of the text, which that analyst believed to be important.

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, we selected, from each of the five texts, the set of textual units that were labeled as "very important" by a majority of the judges. For example, for text (1), we 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 a text. The percentages of important units for the five texts were 11, 36, 35, 17, and 22 respectively.

We 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, we selected a percentage of units that was closest to the one computed for the judges. For example, we selected units 4 and 12, which represented the most important 11% of units as induced from the RS-tree built by the first analyst. However, we selected only unit 4, which represented 6% of the most important units as induced from the RS-tree built by the second analyst. The reason for selecting only unit 4 for the second analyst was that units 10, 11, and 12 have the same score — 4 (see table 1). If we had selected units 10, 11 and 12 as well, we would have ended up selecting 22% of the units in text (1), which is farther from 11 than 6. Hence, we determined for each text the set of important units as labeled by judges and as derived from the RS-trees of those texts.

We calculated for each text the recall and precision of the important units derived from the RS-trees, with respect to the units labeled important by the judges. The overall recall and precision was the same for both analysts: 56% recall and 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%, $\sigma = 0.059$.

In summarizing text, it is often useful to consider not only clauses, but full sentences. To account for this, we considered to be important all the textual units that pertained to a sentence that was characterized by at least one important textual unit. For example, we labeled as important textual units 1 to 4 in text (1), because they make up a full sentence and because unit 4 was labeled as important. For the adjusted data, we determined again the percentages of important units for the five texts and we re-calculated the recall and precision for both analysts: the recall was 69% and 66% and the precision 82% and 75% respectively. In contrast, 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 we 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 RST by itself is not enough 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 enable one to derive 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 the one between the judges and the second analyst. It follows that most 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 seems adequate for the program to take advantage of the paragraph structure of the texts that it analyzes.

Currently, the rank assignment for each textual unit in an RS-tree is done entirely on the basis of the maximal depth in the tree where that unit is salient (Marcu, 1996). Our data seem to support the fact that there exists a correlation also between the types of relations that are used to connect various textual units and the importance of those units in a text. We plan to design other experiments that can provide clearcut evidence on the nature of this correlation.

3 An RST-based summarization program

3.1 Implementation

Our summarization program relies on a rhetorical parser that builds RS-trees for unrestricted texts. The mathematical foundations of the rhetorical parsing algorithm rely on a first-order formalization of valid text structures (Marcu, 1997b). 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.

The rhetorical parsing algorithm, which is outlined in figure 1, is based on a comprehensive corpus analysis of more than 450 discourse markers and 7900 text fragments (see (Marcu, 1997b) for details). When given a text, the rhetorical parser determines first the discourse markers and the elementary units that make up that text. The parser uses then the information derived from the corpus analysis in order to hypothesize rhetorical relations among the elementary units. In the end, the parser applies a constraint-satisfaction procedure to determine the text structures that are valid. If more than one valid structure is found, the parser chooses one that is the "best" according to a given metric. The details of the algorithms that

INPUT: a text T.

1. Determine the set D of all discourse markers in T 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. Determine the set *ValTrees* of all *valid* RS-trees of T that can be built using relations from R.
4. Determine the "best" RS-tree in *ValTrees* on the basis of a metric that assigns higher weights to the trees that are more skewed to the right.

Figure 1: An outline of the rhetorical parsing algorithm

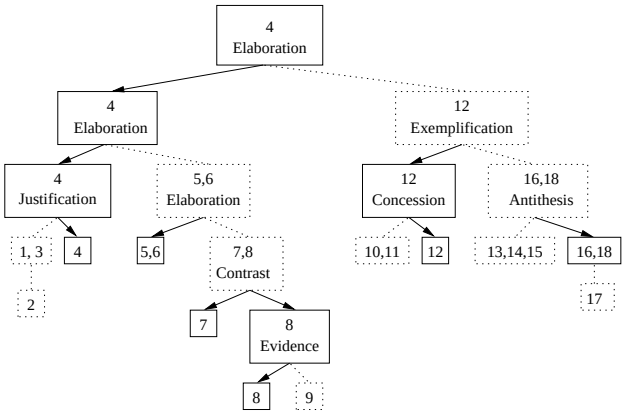

Figure 2: The RS-tree of maximal weight built by the rhetorical parser for text (1).

are used by the rethorical parser are discussed at length in (Marcu, 1997a; Marcu, 1997b).

When the rhetorical parser takes text (1) as input, it produces the RS-tree in figure 2. The convention that we use is that nuclei are surrounded by solid boxes and satellites by dotted boxes; the links between a node and a subordinate nucleus or nuclei are represented by solid arrows, and the links between a node and a subordinate satellite by dotted lines. The nodes with only one satellite denote occurrences of parenthetical information: for example, textual unit 2 is labeled as parenthetical to the textual unit that results from juxtaposing 1 and 3. The numbers associated with each leaf correspond to the numerical labels in text (1). The numbers associated with each internal node correspond to the salient units of that node and are explicitly represented in the RS-tree.

By inspecting the RS-tree in figure 2, one can notice that the trees that are built by the program do not have the same granularity as the trees constructed by the analysts. For example, the program treats units 13, 14, and 15 as one elementary unit. However, as we argue in (Marcu, 1997b), the corpus analysis on which our parser is built supports the observation that, in most cases, the global structure of the RS-tree is not affected by the inability of the rhetorical parser to uncover all clauses in a text —

most of the clauses that are not uncovered are nuclei of JOINT relations.

The summarization program takes the RS-tree produced by the rhetorical parser and selects the textual units that are most salient in that text. If the aim of the program is to produce just a very short summary, only the salient units associated with the internal nodes found closer to the root are selected. The longer the summary one wants to generate, the farther the selected salient units will be from the root. In fact, one can see that the RS-trees built by the rhetorical parser induce a partial order on the importance of the textual units. For text (1), the most important unit is 4. The textual units that are salient in the nodes found one level below represent the next level of importance (in this case, unit 12 — unit 4 was already accounted for). The next level contains units 5, 6, 16, and 18, and so on.

3.2 Evaluation

To evaluate our program, we associated with each textual unit in the RS-trees built by the rhetorical parser a score in the same way we did for the RS-trees built by the analysts. For example, the RS-tree in figure 2 has a depth of 6. Because unit 4 is salient for the root, it gets a score of 6. Units 5, 6 are salient for an internal node found two levels below the root: therefore, their score is 4. Unit 9 is salient for a leaf found five levels below the root: therefore, its score is 1. Table 1 presents the scores associated by our summarization program to each unit in text (1).

We used the importance scores assigned by our program to compute statistics similar to those discussed in the previous section. 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 53% 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 66% and the precision 68%. The lower recall and precision scores associated with clauses 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 significantly.

4 Comparison with other work

We are not aware of any RST-based summarization program for English. However, Ono et al. (1994) discuss a summarization program for Japanese whose minimal textual units are sentences. Due to the differences between English and Japanese, it was impossible for us to compare Ono's summarizer with ours. Fundamental differences concerning the assumptions that underlie Ono's work and ours are discussed at length in (Marcu, 1997b).

Unit type		Recall	Precision
Clauses	Random	25.7	25.7
	Microsoft Summarizer	28	26
	Our summarizer	**53**	**50**
	Analysts	56	66
Sentences	Random	38.4	38.4
	Microsoft Summarizer	41	39
	Our summarizer	**66**	**68**
	Analysts	67.5	78.5

Table 3: An evaluation of our summarization program.

We were able to obtain only one other program that summarizes English text — the one included in the Microsoft Office97 package. We run 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 we selected percentages of text that corresponded only to the clauses considered important by the judges, the Microsoft program recalled 28% of the units, with a precision of 26%. When we selected percentages of text that corresponded to sentences considered important by the judges, the Microsoft program recalled 41% of the units, with a precision of 39%. All Microsoft figures are only slightly above those that correspond to the baseline algorithms that select important units randomly. It follows that our program outperforms significantly the one found in the Office97 package.

We are not aware of any other summarization program that can build summaries with granularity as fine as a clause (as our program can).

5 Conclusions

We described the first experiment that shows that the concepts of rhetorical analysis and nuclearity can be used effectively for summarizing text. 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%. We showed how the concepts of rhetorical analysis and nuclearity can be treated algorithmically and we 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 we propose outperforms both the baseline and the commercial summarizer (see table 3). However, since its results 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 will result in better performance of subsequent implemetations.

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 reasearch was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Barzilay, Regina and Michael Elhadad. 1997. Using Lexical Chains for Text Summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*.

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

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

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

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

Gale, William, Kenneth W. Church, and David Yarowsky. 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)*, pages 249–256.

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

Hearst, Marti. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, June 27–30.

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

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

Lin, Chin-Yew. 1995. Knowledge-based automatic topic identification. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL–95)*, pages 308–310, Cambridge, Massachusetts, June 26–30.

Lin, Chin-Yew and Eduard Hovy. 1997. Identifying topics by position. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP–97)*, pages 283–290, Washington, DC, March 31 – April 3.

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

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

Marcu, Daniel. 1996. Building up rhetorical structure trees. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI–96)*, volume 2, pages 1069–1074, Portland, Oregon, August 4–8,.

Marcu, Daniel. 1997a. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL–97)*, Madrid, Spain, July 7-10.

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

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

Ono, Kenji, Kazuo Sumita, and Seiji Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the International Conference on Computational Linguistics (Coling–94)*, pages 344–348, Japan.

Paice, Chris D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26(1):171–186.

Passonneau, Rebecca J. and Diane J. Litman. 1993. Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 148–155, Ohio, June 22-26.

Rush, J.E., R. Salvador, and A. Zamora. 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, Carol. 1989. Teaching students to summarize: Applying textlinguistics. *System*, 17(1).

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

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

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