To build text summaries of high quality, nuclearity is not sufficient

Daniel Marcu
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292-6601
marcu@isi.edu

Abstract

Researchers in discourse have long hypothesized that the nuclei of a rhetorical structure tree provide a good summary of the text for which that tree was built. In this paper, I discuss a psycholinguistic experiment that validates this hypothesis, but that also shows that the distinction between nuclei and satellites is not sufficient if we want to build summaries of very high quality. I empirically compare various techniques for mapping discourse trees into partial orders that reflect the importance of the elementary textual units in texts and I discuss both their strengths and weaknesses.

Motivation

Researchers in discourse have long hypothesized that the nuclei of a rhetorical structure tree provide a good summary of the text for which that tree was built (Mann & Thompson 1988; Matthiessen & Thompson 1988; Hobbs 1993; Polanyi 1993; Sparck Jones 1993; Ono, Sumita, & Miike 1994). And a psycholinguistic experiment that was designed to check the validity of this hypothesis (Marcu 1997a) showed that discourse trees *can* be used effectively in text summarization. More precisely, the experiment showed that the importance scores obtained on the basis of both manually and automatically built discourse trees yielded a partial order with respect to the importance of clause-like units in a text that matched the units considered important by human judges significantly better than random.

Although these results confirm the adequacy of using discourse trees for text summarization, they are still far from perfect. In this paper, I use data from the same experiment to investigate the reasons for the imperfect results. On the basis of this investigation, I suggest a number of techniques that can be used to improve the results of discourse-based summarizers.

Previous work — review of the experiment

In the experiment on discourse trees and summaries (Marcu 1997a; 1997c), I used five texts from *Scientific American* that ranged in size from 161 to 725 words. I broke the texts into 160 elementary textual units and asked 13 independent judges to rate each textual unit according to its importance to a potential summary. The judges assigned a score of 2 to the units that they believed to be very important, 1 to those that they considered somewhat important, and 0 to those that they considered unimportant and which should not appear in any summary. The percent agreement among judges was high (71%) and statistically significant. Overall, a majority of judges labelled 36 of the 160 units as very important.

In addition, two computational linguists built rhetorical structure trees (RS-trees) for each of the five texts. Using a nuclearity-based mapping, which I will describe in detail in section 3, I assigned an importance score to each unit in each discourse tree. The score of each unit reflected its importance for a summary of the text in which it occurred. The agreement between the two analysts, which was computed on the basis of these scores, was high and statistically significant.

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 for each of the five texts. Because the scores computed from the RS-trees yielded a partial order, I could not select in all cases the same number of units as the judges. In the cases in which there were ties, I selected a percentage of units that was closest to the one computed by judges. The overall recall and precision figures of the important units derived from the RS-trees with respect to the units labelled important by the judges were 56% and 66% respectively. 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%, $\sigma = 0.059$. When I considered sentences, the recall and precision results were 67.5% and 78% respectively, which contrasted the average recall and precision that were obtained when units were selected randomly: 38.4%, $\sigma = 0.048$.

Obviously, the above results provide enough evidence that discourse trees can be used effectively for text summarization. However, they are still far from the idealistic goal of 100% recall and precision.

Ways to improve discourse-based summarization programs

Build better rhetorical trees

In order to understand how the recall and precision figures can be improved, consider the text shown in (1), below, in which the elementary units are numbered from 1 to 27 and the units that a majority of the judges agreed to be important are shown in bold.

(1) [**Smart cards are becoming more attractive**[1]] [as the price of microcomputing power and storage continues to drop.[2]] [**They have two main advantages over magnetic-stripe cards.**[3]] [**First, they can carry 10 or even 100 times as much information**[4]] [— and hold it much more robustly.[5]] [**Second, they can execute complex tasks in conjunction with a terminal.**[6]] [For example, a smart card can engage in a sequence of questions and answers that verifies the validity of information stored on the card and the identity of the card-reading terminal.[7]] [A card using such an algorithm might be able to convince a local terminal that its owner had enough money to pay for a transaction[8]] [without revealing the actual balance or the account number.[9]] [Depending on the importance of the information involved,[10]] [security might rely on a personal identification number[11]] [such as those used with automated teller machines,[12]] [a midrange encipherment system,[13]] [such as the Data Encryption Standard (DES),[14]] [or a highly secure public-key scheme.[15]]

[Smart cards are not a new phenomenon.[16]] [**They have been in development since the late 1970s**[17]] [and have found major applications in Europe,[18]] [with more than a quarter of a billion cards made so far.[19]] [The vast majority of chips have gone into prepaid, disposable telephone cards,[20]] [but even so the experience gained has reduced manufacturing costs,[21]] [improved reliability[22]] [and proved the viability of smart cards.[23]] [**International and national standards for smart cards are well under development**[24]] [to ensure that cards, readers and the software for the many different applications that may reside on them can work together seamlessly and securely.[25]] [Standards set by the International Organization for Standardization (ISO), for example, govern the placement of contacts on the face of a smart card[26]] [so that any card and reader will be able to connect.[27]]

The discourse structure shown in figure 1, which was built by one of the analysts in the experiment, yielded the lowest recall and precision figures at the clause-like unit level. In figure 1, the discourse tree is represented in the style of Mann and Thompson (1988), with each elementary unit in the structure being labelled with a number from 1 to 27.

One of the ways in which we can use the tree in figure 1 for text summarization is by exploiting the difference between nuclei and satellites; nuclei express what is more essential to the writer's purpose. For example, the system proposed by Ono et al. (1994) associates a penalty score to each node in a tree by assigning a score of 0 to the root and by increasing the penalty by 1 for each satellite node that is found on every path from the root to a leaf. The dotted arcs in figure 1 show in the style of Ono et al. (1994) the scope of the penalties that are associated to the corresponding text spans; the penalties are shown in bold. For example, span [4,15] has associated a penalty of 1 because it is the first satellite encountered on the path from the root to the corresponding node. Span [6,15] has associated a penalty of 2, because is the second satellite encountered on the path from the root to the corresponding node. And so on. The penalty score of each unit, which is shown in bold italics, is given by the penalty score associated with the closest boundary.

The penalties of the 27 units in text (1) induce a partial ordering on the importance of the units in the text, which is shown in (2), below.

$$\textbf{(2)} \quad \begin{aligned} &3, 16 > 1, 4, 5, 6, 17, 18, 21, 22, 23, 24 > 2, 7, 19, \\ &20, 25, 26 > 8, 27 > 9, 11, 13, 15 > 10, 12, 14 \end{aligned}$$

When judges analyzed text (1), they considered 6 units to be important. If we try to use partial ordering (2) in order to select approximately 6 units, we find that we can select either 2 units (3 and 16) or 12 (3, 16, 1, 4, 5, 6, 17, 18, 21, 22, 23, and 24). In such cases, I have chosen to select the number of units that comes closest to the number of units considered important by judges. For text (1) and partial ordering (2), we will hence select units 3 and 16, which corresponds to a 1/6=16.66% recall and 1/2=50% precision.

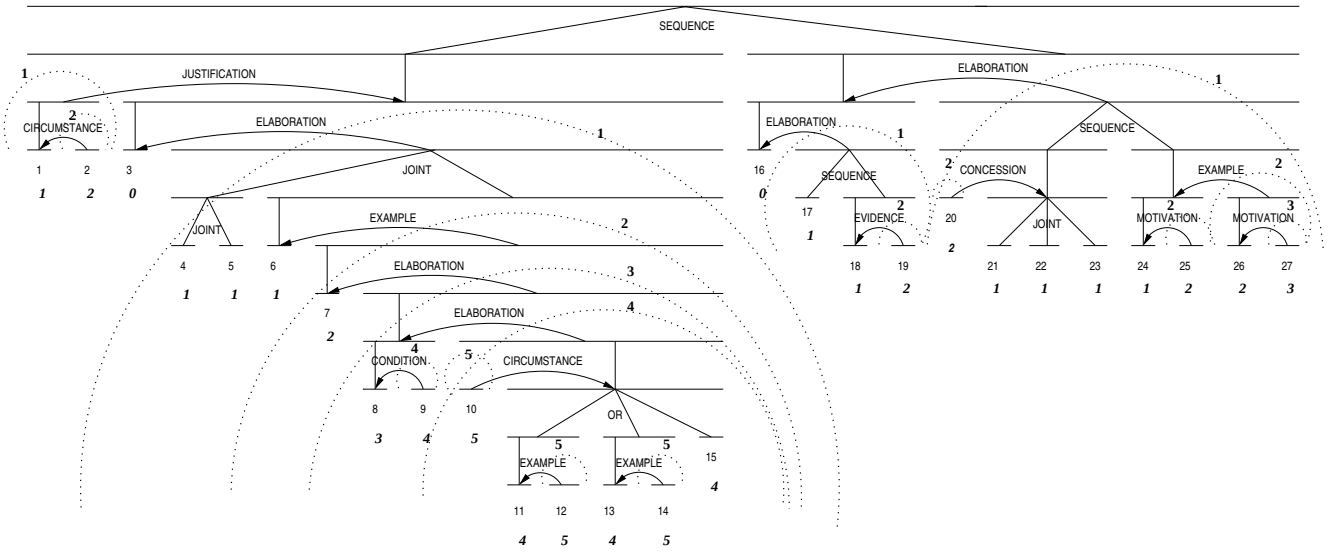

Figure 1: The discourse tree that was built for text (1) by one of the analysts.

Why are the results so poor? The first explanation that comes to one's mind is that maybe the discourse tree that was built by the analyst is incorrect. By examining it closely, we may dispute the correctness of the structure that corresponds to span [16,27], where the analyst considered spans [20,23] and [24,27] to be both ELABORATIONS of span [16,19]. However, if we examine the text closely, it would seem that a better tree is that shown in figure 2, where it is only span [20,23] that elaborates on span [17,19]. The fact that units 17–23 use the present perfect while units 16 and 24–27 use the present strongly supports the validity of the structure in figure 2. One may still argue about the appropriateness of using certain rhetorical relations, but if we ignore this, the nuclearity of tree 2 seems to be right. In spite of the correctness of the tree, the partial ordering that follows from the application of Ono et al.'s method increases the performance of the system only slightly. The new partial ordering, that shown in (3) recalls 33% (2 of the 6 units considered important by judges), with a precision of 66%.

(3) $3, 16, 24 > 1, 4, 5, 6, 17, 18, 25, 26 > 2, 7, 19, 21$
$22, 23, 27 > 8, 20 > 9, 11, 13, 15 > 10, 12, 14$

The example above suggests that the correctness of a discourse tree affects the quality of the summary that is derived from it. And the results reported by Marcu (1997a) confirm this hypothesis: the recall and precision figures that pertained to summaries that were obtained for the five texts using discourse trees that were derived automatically (Marcu 1997b; 1997c) were about 10% lower than the figures derived using the trees that were built manually by the two analysts.

Exploit better the nuclearity feature

However, the correctness of the discourse trees that are used in summarization are only one part of the problem. Even if we take the trees that were built manually by the two analysts for the five texts and apply Ono et al.'s scoring function, we will obtain recall and precision figures of only 45.83% and 52.38% at the clause level and 57.75% and 68.36% at the sentence level.

If we closely examine Ono et al.'s proposal, we will notice that it assigns a partial order over the importance scores that is relatively "flat". For example, although the tree in figure 2 has a depth of 11, there are only 6 levels of importance in the partial order defined by Ono et al. Because the partial ordering assigned by Ono et al. is very coarse, many units are assigned the same score. In addition, units that occur on a certain level in the tree are assigned the same scores as units that occur much lower, which seems counterintuitive. For example, although unit 10 occurs in the discourse tree in figure 2 two levels above units 12 and 14, all three units are assigned the same penalty score.

A better way to take advantage of the difference between nuclei and satellites is by exploiting the notions of *salient* and *parenthetical unit* that are associated with textual spans in the formalization proposed by Marcu (1997b; 1997c). By default, the salient units associated with the leaves are the leaves themselves. The salient units (promotion set) of each internal node

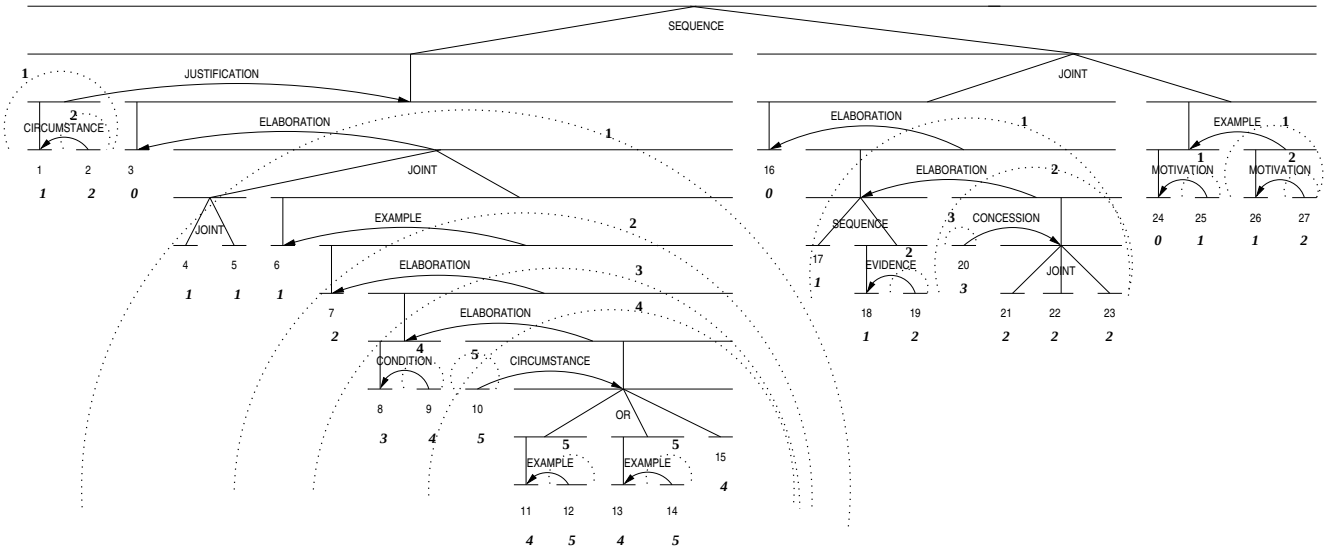

Figure 2: A modified version of the discourse tree that was built for text (1) by one of the analysts.

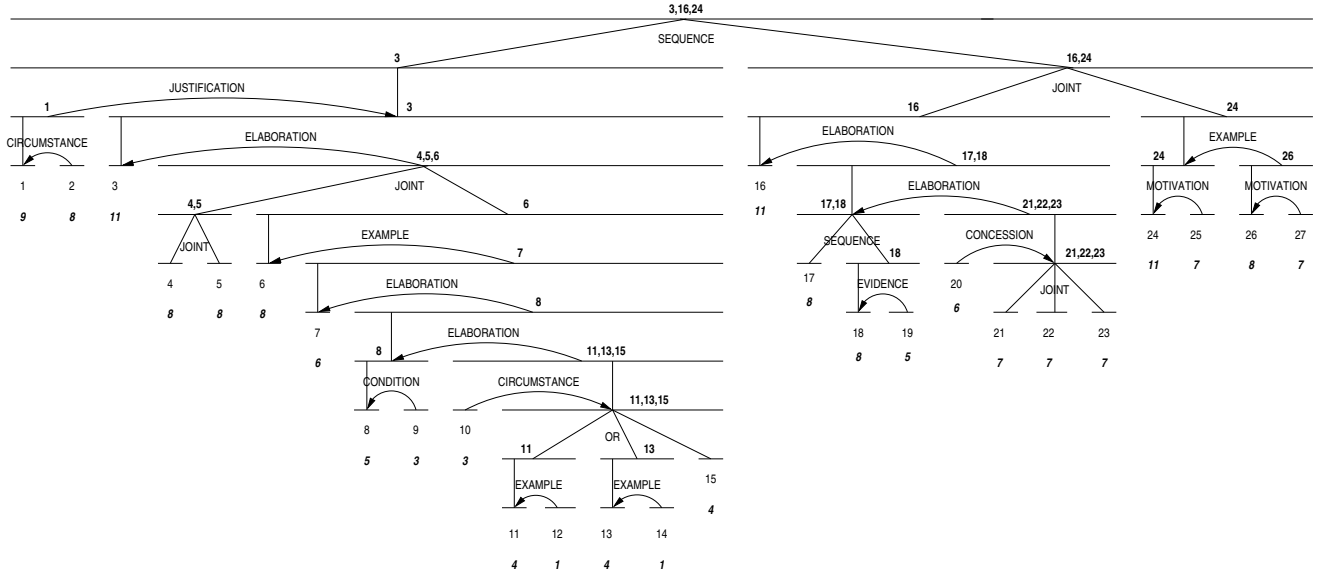

Figure 3: The discourse tree shown in figure 2, with the promotion units and importance scores assigned by formula (5).

are given by the salient units of the nucleus node in the case the relation is mononuclear and by the union of the salient units of the nuclei nodes in the case the relation is multinuclear. The set of salient units associated with the discourse structure in text (1) are explicitly shown in figure 3 in bold. The parenthetical units, such as that shown in italics in (4), are related only to the larger units that they belong to or to the units that immediately precede them.

(4) With its distant orbit — *50 percent farther from the sun than the Earth* — and slim atmospheric blanket, Mars experiences frigid weather conditions.

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. One 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 (5), which is given 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.

$$\textbf{(5)}\quad s(u,D,d) = \begin{cases} 0 & \text{if } D \text{ is NIL,} \\ d & \text{if } u \in prom(D), \\ d-1 & \text{if } u \in paren(D), \\ max(s(u, C(D), d-1)) & \text{otherwise.} \end{cases}$$

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 (5) to the tree in figure 3, which has depth 11, we obtain the scores shown in italics bold for each of the elementary and parenthetical units of text (1). Because units 3,16, and 24 are among the promotion units of the root, they get a score of 11. Units 4,5,6 are among the promotion units of a node found three levels below the root, so they get a score of 8. Unit 14 is among the promotion units of a leaf found 10 levels below the root, so it gets a score of 1.

The scores assigned by formula (5) induce a partial ordering on the importance of textual units in text (1) that is shown in (6), below.

$$\textbf{(6)}\quad 3, 16, 24 > 1 > 2, 4, 5, 6, 17, 18, 26 > 21, 22, 23, 27 > 7, 20 > 8, 19 > 11, 13, 15 > 9, 10 > 12, 14$$

In comparison with the partial orders induced by Ono et al. (1994), the partial orders induced by formula (5) are much finer grained. For example, for the tree in figure 3, formula (5) induced a partial ordering with 9 levels, while Ono et al.'s method induced a partial ordering of only 6 levels. If we use the new partial ordering to determine the most important units of text (1), we obtain a recall of 3/6=50% and a precision of 3/4=75%.

When we apply formula (5) to the trees that were built by the two analysts in order to determine the most important units in all five text in the experiment, we obtain partial orderings that yield recall and precision figures of 55.55% and 66.66% at the clause level and 67.24% and 78.00% at the sentence level. These recall and precision figures, which are about 10% higher than those obtained using Ono et al.'s method (1994), suggest that exploiting the promotion units that are associated with each node enables the derivation of text summaries that are better than the summaries that result from a straightforward application of the nuclearity principle.

An alternative method to computing importance scores is one that assumes that the importance of a textual unit is given not only by the distance from the root to the first node that has that unit in its promotion set, but also by the number of links between the unit and the closest node to the root that has that unit in its promotion set.[1] The number of links $upwards(u)$ that a unit u is projected upwards in a tree reflects its importance relative to the units found in its immediate neighborhood: the bigger the number of links a unit is projected upwards in a tree, the more important that unit is relative to its neighbors. For example, in the tree in figure 4, $upwards(2) = 0$ because unit 2 does not belong to the promotion set of any internal node. In contrast, $upwards(18) = 3$ because there are three links between the unit and the node that has 18 in its promotion set and that is closest to the root; and $upwards(3) = 3$ as well because unit 3 is three links away from the root, which has 3 in its promotion set.

[1] I am grateful to Eduard Hovy for suggesting this idea.

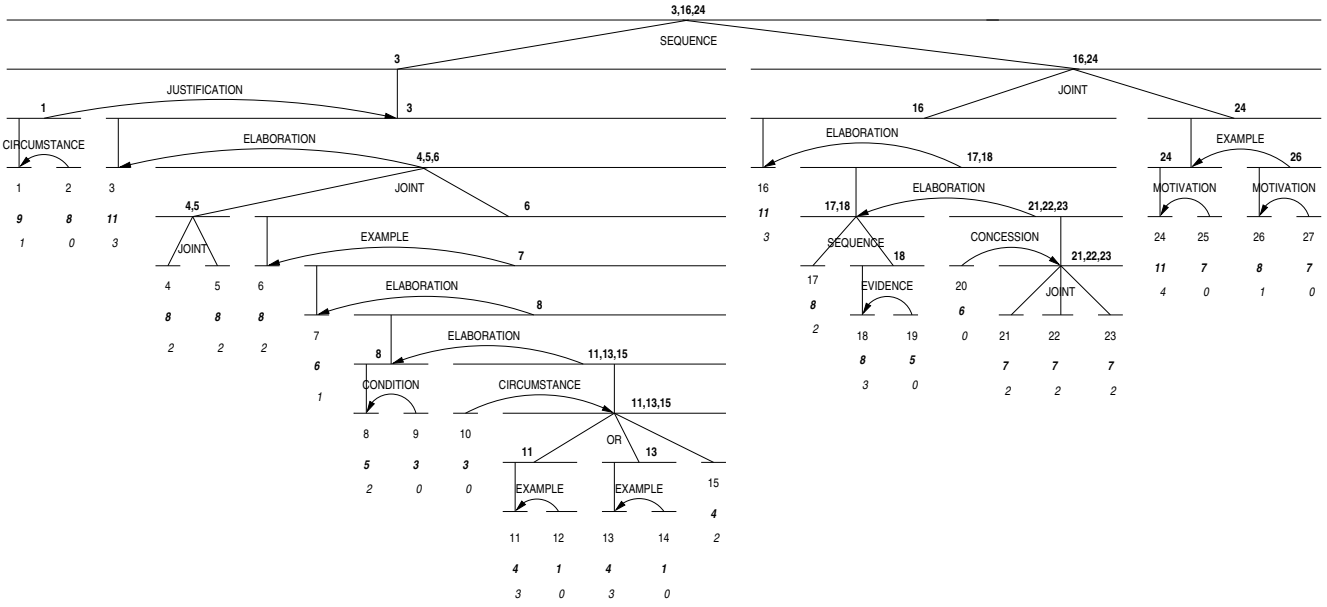

Figure 4: The discourse tree shown in figure 3, with the promotion units and importance scores assigned by formula (5) and the importance scores that reflect the relative significance of a unit among the units found in its neighborhood. The importance scores that reflect the relative significance are shown at the bottom of each unit in italics.

If we add the two scores for each unit of text (1), we obtain the partial ordering shown in (7), below.

$$\textbf{(7)} \quad \begin{aligned} &24 > 3, 16 > 1, 4, 5, 6, 17 > 21, 22, 23, 26 > 2 > 7, \\ &8, 11, 13, 25, 27 > 15, 20 > 19 > 9, 10 > 12, 14 \end{aligned}$$

If we use the new partial ordering to determine the most important units of text (1), we obtain a recall of 2/6=33.33% and a precision of 2/4=50%. When we apply this method to the trees that were built by the two analysts in order to determine the most important units in all five text in the experiment, we obtain partial orderings that yield recall and precision figures of 62.50% and 60.81% at the clause level and 68.10% and 69.91% at the sentence level. Hence, the new method improves recall at the expense of precision.

Exploit the semantics of rhetorical relations

If we examine the discourse structure in figure 3 and the units that judges perceived as being important, we notice a couple of very interesting facts. For example, a majority of the judges labelled units 3, 4, and 6 as important. The discourse structure in figure 3 shows that an ELABORATION relation holds between units 4 and 3 and between units 6 and 3. Because units 4 and 6 are the satellites of the ELABORATION relation, they are assigned a lower score than unit 3. However, if we examine the text closely, we also find it natural to include in the summary not only the information that "smart cards have two main advantages over magnetic-stripe cards" (unit 3), but also the advantages per se, which are given in units 4 and 6. Hence, for certain kinds of ELABORATION relations, it seems adequate to assign a higher score to their satellites than formula (5) currently does.

By examining the same discourse structure and the importance scores assigned by judges, we can see that none of the units in the span [7,15] were considered important. This observation seems to correlate with the fact that the whole span [7,15] is an EXEMPLIFICATION of the information given in unit 6. If the observation that satellites of EXAMPLE relations are not important generalizes, then it would be appropriate to account for this in the formula that computes importance scores. Also interesting is the fact that judges considered unit 24 important, which seems to correlate with a topic shift. Again, if this observation generalizes, it will have to be properly accounted for by the formula that computes importance scores.

The discussion above strongly suggests that the semantics of rhetorical relations should play a major role in determining the important units in a text. However, more data is needed in order to support the development of scoring functions that are sensitive not only to

the difference in rhetorical status between various textual units but also to the semantics of the rhetorical relations that hold among them.

Beyond discourse trees

So far, we have looked at phenomena in which there was a clear correlation between the structure of discourse and what human judges perceived as being important in the corresponding texts. However, such a correlation is not universal.

As an example, consider the following two cases, in which the judges considered important only the first nucleus of a multinuclear relation. Although a rhetorical relation of JOINT holds between units 4 and 5 and a rhetorical relation of SEQUENCE holds between units 17 and 18, a majority of the judges considered only units 4 and 17 important. According to formula (5), both pairs of units should be assigned the same score. Obviously, mechanisms that are not inherent to the rhetorical structure of text are needed in order to explain why only one nucleus of a multinuclear relation is considered important by humans. For example, in the case of units 4 and 5 it might be the case that the choice of the marker *"— and"* influenced the decision of the judges: the dash found at the beginning of unit 5 usually signals a parenthetical role. Such an explanation would be consistent with research in psycholinguistics that suggests that different connectives signal with different strengths the rhetorical relations between units. In three experiments, Deaton and Gernsbacher (1997; 1997) have shown that two-clause sentences that describe moderately causal events were read faster when the clauses were conjoined by *because* (Susan called the doctor for help *because* the baby cried in his playpen) than when they were conjoined by *and, then,* or *after*. In addition, when the clauses were conjoined by *because*, subjects recalled the second clauses more frequently when prompted with the first clause.

To make things even more complicated, in some cases, the textual units that are considered important by human judges correspond to satellites of rhetorical relations and not to nuclei. For example, 5 judges considered unit 16 to be very important and 2 to be somewhat important, while 10 judges considered unit 17 to be important and 2 to be somewhat important. This contradicts the nuclearity assignment in figure 3, which assigns a nucleus status to unit 16 and a satellite status to unit 17. One can argue that unit 16 is a satellite of a JUSTIFICATION relation that has span [17,23] as nucleus, but it is difficult to accept that: after all, unit 16 can be even considered to be by itself a good summary of the whole span [16,23]. So why did judges considered it less important than unit 17? If we are to provide adequate answers to the questions we raised in this section, it seems that we need go beyond the predictions that pertain to the rhetorical structure of text.

Conclusion

In this paper, I have examined the limitations of simplistic use of discourse structures in automatic summarization. On the basis of a psycholinguistic experiment, I have shown that the quality of discourse-based summaries depends on the quality of the discourse trees that are used; the mapping between discourse structures and importance scores; the semantics of the rhetorical relations that are used; and other factors that are not inherent to the rhetorical description of text.

Acknowledgements. I am grateful to Graeme Hirst for the invaluable help he gave me during every stage of this work.

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

Deaton, J., and Gernsbacher, M. 1997. Causal conjunctions and implicit causality cue mapping in sentence comprehension. *Journal of Memory and Language.*

Gernsbacher, M. 1997. Coherence cues mapping during comprehension. In Costermans, J., and Fayol, M., eds., *Processing Interclausal Relationships. Studies in the Production and Comprehension of Text.* Lawrence Erlbaum Associates. 3–22.

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

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

Marcu, D. 1997a. From discourse structures to text summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, 82–88.

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

Marcu, D. 1997c. *The rhetorical parsing, summarization, and generation of natural language texts.* Ph.D.

Dissertation, Department of Computer Science, University of Toronto.

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.

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

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