[1994] associates a penalty score to each node in a rhetorical structure 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 2 show in the style of Ono et al. (1994) the scope of the penalties that are associated with the corresponding spans. For example, span [4,15] has associated a penalty of 1, because it is one satellite away from the root. The penalty score of each unit, which is shown in bold italics, is given by the penalty score associated with the closest boundary.

The algorithm proposed by Marcu [1997,2000] exploits the salient units (promotion sets) associated with each node in a tree. By default, the salient units associated with the leaves are the leaves themselves. The salient units (promotion set) associated with each internal node are given by the union of the salient units of the children nodes that are nuclei. In Figure 3, the salients units associated with each node are shown in bold.

As one can see, the salient units induce a partial ordering on the importance of the units in a text : the salient units found closer to the root of the tree are considered to be more important than the salient units found farther. For example, units 3, 16, and 24 which are the promotion units of the root, are considered the most important units in the text whose discourse structure is shown in Figure 3. Marcu [1998] has shown that his method yields better results than Ono et al.'s. Yet, when we tried it on large texts, we obtained disappointing results (see Section 4).

2.2 Explanation

Both Ono et al.'s [1994] and Marcu's [1997, 2000] algorithms assume that the importance of textual units is determined by their distance to the root of the corresponding rhetorical structure tree.[1] Although this is a reasonable assumption, it is clearly not the only factor that needs to be considered.

Consider, for example, the discourse tree sketched out in Figure 1, in which the root node has three children, the first one subsuming 50 elementary discourse units (*edu*s), the second one 3, and the third one 40. Intuitively, we would be inclined to believe that since the author dedicated so much text to the first and third topics, these are more important than the second topic, which was described in only 3 edus. Yet, the algorithms described by Ono et al. [1994] and Marcu [1997] are not sensitive to the size of the spans.

Another shortcoming of the algorithms proposed by Ono et al. [1994] and Marcu [1997] is that they are fairly "un-localized". In our experiments, we have noticed that the units considered to be important by human judges are not uniformly distributed over the text. Rather, if a human judge considers a certain unit to be important, then it seems to be more likely that other units found in the neighborhood of the selected unit are also considered important.

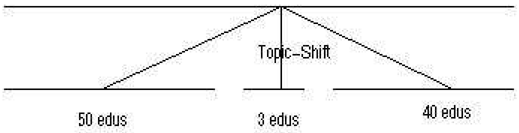

Figure 1: Example of unbalanced rhetorical structure tree.

And probably the most important deficiency, Ono et al.'s [1994] and Marcu's [1997] approaches are insensitive to the semantics of the rhetorical relations. It seems reasonable to expect, for instance, that the satellites of *EXAMPLE* relations are considered important less frequently than the satellites of *ELABORATION* relations. Yet, none of the extraction algorithms proposed so far exploits this kind of information.

3 Experiment

In order to enable the development of algorithms that address the shortcomings enumerated in Section 2.2, we took an empirical approach. That is, we manually annotated a corpus of 380 articles with rhetorical structures in the framework of Rhetorical Structure Theory. The leaves (*edus*) of the trees were clauses and clausal constructs. The agreement between annotators on the discourse annotation task was higher than the agreement reported by Marcu et al. [1999] – the kappa statistics computed over trees was 0.72 (see Carlson et al. [2001] for details). Thirty of the discourse annotated texts were used in one summarization experiment, while 150 in another experiment. In all summarization experiments, recall and precision figures are reported at the *edu* level.

3.1 Corpora used in the experiment

Corpus A consisted of 30 articles from the Penn Treebank collection, totaling 27,905 words. The articles ranged in size from 187 to 2124 words, with an average length of 930 words. Each of these articles was paired with:

- An informative abstract, built by a professional abstractor. The abstractor was instructed to produce an abstract that would convey the essential information covered in the article, in no more than 25% of the original length. The average size of the abstract was 20.3% of the original.

[1] The methods differ only in the way they compute this distance.

- A short, indicative abstract of 2-3 sentences, built by a professional abstractor, with an average length totaling 6.7% of the original document. This abstract was written so as to identify the main topic of the article.
- Two "derived extracts", $E_{d1}^{A\_long}$ and $E_{d2}^{A\_long}$, produced by two different analysts who were asked to identify the text fragments (*edu*s) whose semantics was reflected in the informative abstracts.
- Two "derived extracts", $E_{d1}^{A\_short}$ and $E_{d2}^{A\_short}$, produced by two different analysts who were asked to identify the text fragments (*edu*s) whose semantics was reflected in the indicative abstracts.
- An independent extract E^{A}, produced from scratch by a third analyst, by identifying the important *edu*s in the document, with no knowledge of the abstracts. As in the case of the informative abstract, the extract was to convey the essential information of the article in no more than 25% of the original length.

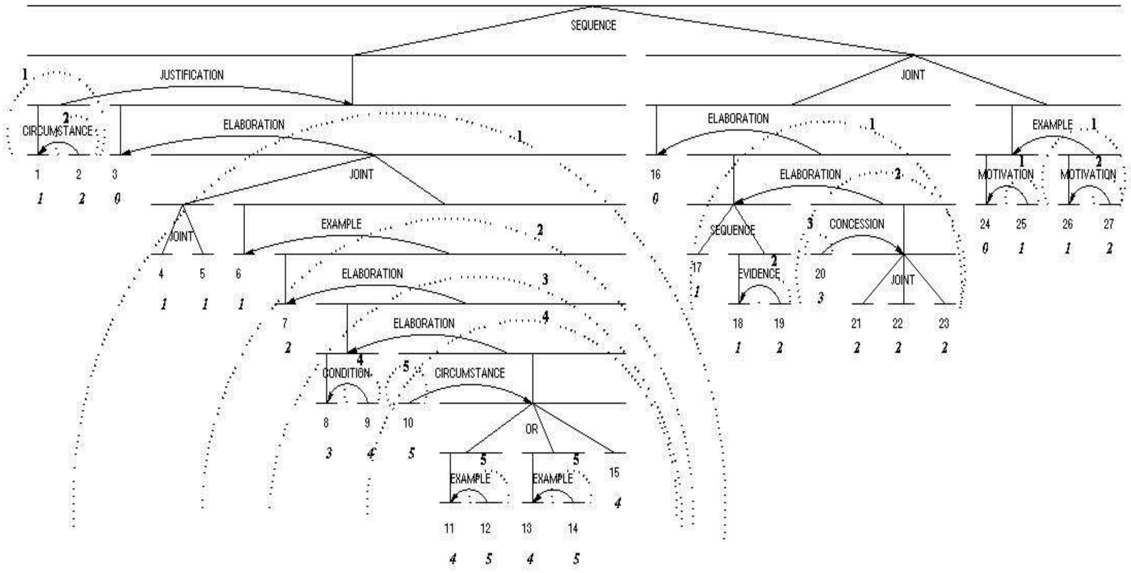

Figure 2: Assigning importance to textual units using Ono et al.'s method [1994].

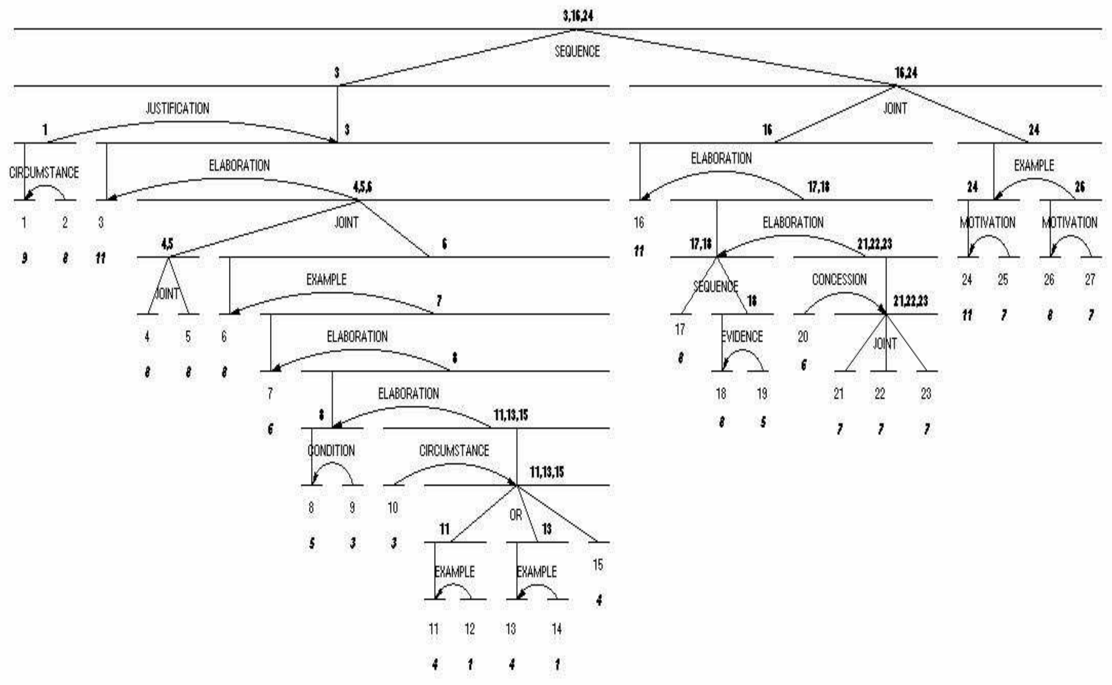

Figure 3: Assigning importance to textual units using Marcu's method [1997, 2000].

Corpus B consisted of 150 articles from the Penn Treebank collection, totaling 125,975 words. This set included the smaller Corpus A, and the range in size was the same. The average number of words per article was 840. Each article in this corpus was paired with:

- Two informative extracts, E_1^B and E_2^B, produced from scratch by two analysts, by identifying the important *edu*s in each document. For this experiment, a target number of *edu*s was specified, based on the square root of the number of *edu*s in each document. Analysts were allowed to deviate from this slightly, if necessary to produce a coherent extract. The average compression rate for these extracts was 13.30%.

3.2 Agreement on summary annotations

We have found that given an abstract and a text, humans can identify the corresponding extract, i.e., the important text fragments (*edu*s) that were used to write the abstract, at high levels of agreement. The average inter-annotator recall and precision figures computed over the edus of the derived extracts were higher than 80% (see the first two rows in Table 1).

Table 1: Inter-annotator agreements on various summarization tasks.

Agreement between	Judges	Rec	Prec	F-val
Extracts derived from informative abstracts	$E_{d1}^{A\_long}$ - $E_{d2}^{A\_long}$	85.71	83.18	84.43
Extracts derived from indicative abstracts	$E_{d1}^{A\_short}$ - $E_{d2}^{A\_short}$	84.12	79.93	81.97
Extracts created from scratch	E_1^B - E_2^B	45.51	45.58	45.54
Derived extracts vs. extracts created from scratch	$E_{d1}^{A\_long}$ - E^A $E_{d2}^{A\_long}$ - E^A	28.15 28.93	51.34 52.47	36.36 37.30

Building an extract from scratch proved though to be a much more difficult task : on Corpus B, for example, the average inter-annotator recall and precision figures computed over the edus in the extracts created from scratch were 45.51% and 45.58% respectively (see row 3, Table 1). This would seem to suggest that to enforce consistency, it is better to have a professional abstractor produce an abstract for a summary and then ask a human to identify the extract, i.e., the most important text fragments that were used to write the abstract. However, if one measures the agreement between the derived extracts and the extracts built from scratch, one obtains figures that are even lower than those that reflect the agreement between judges that build extracts from scratch. The inter-annotator recall and precision figures computed over *edu*s of the derived extracts and *edu*s of the extracts built from scratch by one judge were 28.15% and 51.34%, while those computed for the other judge were 28.93% and 52.47% respectively (see row 4, Table 1). The difference between the recall and precision figures is explained by the fact that the extracts built from scratch are shorter than those derived from the abstract.

These figures show that consistently annotating texts for text summarization is a difficult enterprise if one seeks to build generic summaries. We suspect this is due to the complex cognitive nature of the tasks and the nature of the texts.

Nature of the cognitive tasks

Annotating texts with abstracts and extracts are extremely complicated cognitive tasks, each involving its own set of inherent challenges.

When humans produce an abstract, they create new language by synthesizing elements from disparate parts of the document. When the analysts produced derived extracts from these abstracts, the mapping from the text in the abstracts to *edu*s in documents was often one-to-many, rather than one-to-one. As a result, the *edu*s selected for these derived extracts tended to be distributed more broadly across the document than those selected for a pure extract. In spite of these difficulties, it appears that the intuitive notion of semantic similarity that analysts used in constructing the derived extracts was consistent enough across analysts to yield high levels of agreement.

When analysts produce "pure extracts", the task is much less well-defined. In building a pure extract, not only is an analyst constrained by the exact wording of the document, but also, what is selected at any given point limits what else can be selected from that point forward, in a linear fashion. As a result, the *edu*s selected for the pure extracts tended to cluster more than those selected for the derived extracts. The lower levels of agreement between human judges that constructed "pure extracts" show that the intuitive notion of "importance" is less well-defined than the notion of semantic similarity.

Nature of the texts

As Table 1 shows, for the 150 documents in Corpus B, the inter-annotator agreement between human judges on the task of building extracts from scratch was at the 45% level. (This level of agreement is low compared with that reported in previous experiments by Marcu [1997], who observed a 71% inter-annotator agreement between 13 human judges who labeled for importance five scientific texts that were, on average, 334 words long.) We suspect the following reasons explain our relatively low level of agreement:

- Human judges were asked to create informative extracts, rather than indicative ones. This meant that the number of units to be selected was larger than in the case of a high-level indicative summary. While there was general agreement on most of the main points, the analysts differed in their interpretation of what supporting information should be included, one tending to pick more general points, the other selecting more details.
- The length of the documents affected the scores, with agreement on shorter documents greater overall than on longer documents.
- The genre of the documents was a factor. Although these documents were all from the Wall Street Journal, and were generally expository in nature, a number of sub-genres were represented.
- The average size of an *edu* was quite small – 8 words/*edu*. At this fine level of granularity, it is difficult to achieve high levels of agreement.

We analyzed more closely the analysts' performance on creating extracts from scratch for a subset of this set that contained the same 30 documents as those contained in Corpus A.

This subset contained 10 short documents averaging 345 words; 10 medium documents averaging 832 words; and 10 long documents averaging 1614 words. The overall F measure for the short documents was 0.62; for the medium, 0.45, and for the long, 0.47. For the long documents, the results were slightly higher than the medium length ones because of an F score of 0.98 on one document with a well-defined discourse structure, consisting of a single introductory statement followed by a list of examples. For documents like these, the analysts were allowed to select only the introductory statement, rather than the pre-designated number of *edu*s. Excluding this document, the agreement for long documents was 0.41.

When the 30 documents were broken down by sub-genre, the corresponding F-scores were as follows (for two documents an error occurred and the F score was not computed):

- simple news events, single theme (9 articles) : 0.68
- financial market reports and trend analysis (5 articles) : 0.48 (excluding the one article that was an exception, the F measure was 0.36)
- narrative mixed with expository (8 articles) : 0.47
- complex or multiple news events, with analysis (3 articles) : 0.40

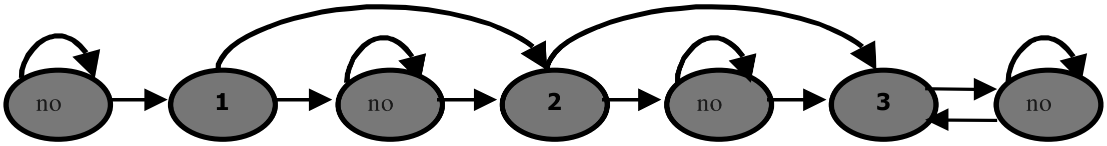

Figure 4: Example of summarization specific HMM chain.

- editorials/letters to the editor (3 articles) : 0.34

These scores suggest that genre does have an affect on how well analysts agree on what is relevant to an informative summary. In general, we have observed that the clearer the discourse structure of a text was, the more likely the same units were selected as important.

4 Empirical grounded algorithms for discourse-based summarizers

We estimated the utility of discourse structure for summarization using three classes of algorithms : one class of algorithms employed probabilistic methods specific to Hidden Markov and Bayesian Models; one class employed decision-tree methods; and one class, used as a baseline, employed the algorithm proposed by Marcu [1997], which we discussed in Section 2. All these classes were compared against a simple position-based summarizer, which assumes that the most important units in a

text always occur at the beginning of that text; and against a human-based upper-bound. If we are able to produce a discourse-based summarization algorithm that agree with a gold standard as often as two human judges agree between themselves, that algorithm would be indistinguishable from a human.

4.1 Using Hiden Markov Models for Discourse-Based Summarization

In this section we present two probabilistic models for automatically extracting *edu*s to generate a summary: a hidden Markov model (HMM) and a Bayesian model.

The HMM for discovering *edu*s to extract for a summary uses the same approach as the sentence extraction model discussed by Conroy and O'Leary [2001]. The hidden Markov chain of the model consists of *k* summary states and *k*+1 non-summary states. The chain is ''hidden'' since we do not know which *edu*s are to be included in the summary. Figure 4 illustrates the Markov model for three such summary states, where the states correspond to edus.

The Markov model is used to model the positional dependence of the *edu*s that are extracted and the fact that if an *edu* in the *i*-th position is included in an extract then the prior probability to include in the extract the *edu* in the (*i*+1)-th position is higher than it would be if unit *i* was not included in the extract. The second part of the model concerns the initial state distribution, which is non-zero only for the first summary and non-summary states. The third piece of the HMM concerns the observations and the probabilistic mapping from states to observations. For this application we chose to use two observations for each *edu*: the original height in the discourse tree of the *edu* and its final height after promotion, where promotion units are determined as discussed in Section 2. The probabilistic mapping we use is a bi-variant normal model with a 2-long mean vector for each state in the chain and a common co-variance matrix. The unknown parameters for the model are determined by maximum likelihood estimation on the training data.

The Bayesian model is quite similar to the hidden Markov model except that the Markov chain is replaced by a prior probability of an *edu* to be contained in a summary. This prior is computed based on the position of each *edu* in a document, so that *edu*s that occur in the beginning of a document have a higher prior probability of being included in an extract than edus that occur towards the end. The prior probabilities for being included in a summary for *r*-1 leading *edu*s and a prior probability for subsequent *edu*s are estimated from the training data. The posterior probability for each *edu* being included in a summary is computed using the same bi-variant normal models used in the HMM. In particular, we have *r* bi-variant models corresponding to the quantitization of the prior probabilities.

4.2 Using Decision Trees for Discourse-Based Summarization

As we discussed in Section 2.2, the important units are rarely chosen uniformly from all over the text. To account for this, we decided to devise a dynamic selection model. The dynamic model assumes that a discourse tree is traversed in a top-down fashion, starting from the root. At each node, the traversal algorithm chooses between three possible actions, which have the following effects :

- **Select :** If the current node is a leaf, the corresponding text span is selected for summarization.
- **GoIn :** If the current node is an internal node, then the selection algorithm is applied recursively on all children nodes.

- **GiveUp :** The selection process is stopped; i.e., all textual units subsumed by the current node are considered to be unimportant.

Assume, for example, that a text has 9 edus, the rhetorical structure shown in Figure 5, and assume that units 1, 2, 8, and 9 were labeled as important by the human annotators. These units can be selected by the top-down traversal algorithm if starting from the root, the algorithm chooses at every level the actions shown in bold.

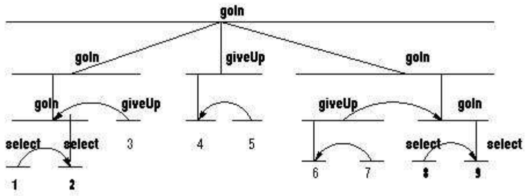

Figure 5: The top-down, dynamic selection algorithm.

To learn what actions to perform in conjunction with each node configuration, we have experimented with a range of features. We obtained the best results when we used the following features :

- An integer denoting the distance from the root of the node under scrutiny.
- An integer denoting the distance from the node to the farthest leaf.
- A boolean specifying whether the node under scrutiny is a leaf or not.
- Three integers denoting the number of edus in the span under consideration and the number of edus in the sibling spans to the left and right of the span under consideration.
- Three categorial variables denoting the nuclearity status of the node under scrutiny and the sibling nodes found immediately to its left and right.
- Three categorial variables denoting the rhetorical labels of the node under scrutiny and the sibling nodes found immediately to the left and right.

Using the corpora of extracts and discourse trees, we traversed each discourse tree top-down and generated automatically learning cases using the features and actions discussed above. This yielded a total of 1600 learning cases for corpus A and a total of 7687 learning cases for corpus B. We used C4.5 [Quinlan, 1993] to learn a decision tree classifier, which yielded an accuracy of 70.5% when cross-validated ten-fold on corpus A and 77.0% when cross-validated ten-fold on corpus B.

To summarize a text, a discourse tree is traversed top-down. At every node, the learned classifier decides to continue the top-down traversal (GoIn), abadon the traversal of all children nodes (GiveUp) or select the text subsumed by a given node for extraction (Select).

5 Evaluation of the discourse-based summarizers

To evaluate our extraction engines we applied a ten-fold cross-validation procedure. That is, we partitioned the discourse and extract files into ten sets. We trained our summarizers 10 times on the files in 9 sets (27 texts for corpus A, and 135 texts for corpus B) and then tested the summarizers on the files on the remaining set (3 texts for corpus A and 15 texts for corpus B). We compared the performance of our summarizers against two baselines : a position-based baseline, which assumes that important units always occur at the beginning of a text, and the algorithm proposed by Marcu [1997], which select important units according to their distance from the root in the corresponding discourse tree. Both baselines were given the extra advantage of selecting the same number of units as the humans. The HMM, Bayes, and Decision-based algorithms automatically learned from the corpus how many units to select. The Hidden Markov and Bayes models were tested only on Corpus B because Corpus A did not provide sufficient data for learning the parameters of these models.

For Corpus A, we trained and tested our decision-based summarization algorithm on all types of extracts, for all analysts : extracts derived from the informative abstracts, $E_{d1}^{A\_long}$ and $E_{d2}^{A\_long}$, extracts derived from the indicative abstracts, $E_{d1}^{A\_short}$ and $E_{d2}^{A\_short}$, and extracts built from scratch, E^{A}. Table 2 summarizes the results using traditional precision and recall evalutation metrics.

Table 2: Evaluation results on corpus A.

Method	Rec	Prec	F-val
Position-based Baseline	26.00	26.00	26.00
Marcu's [1997] selection algorithm	34.00	33.00	33.50
The dynamic, decision-based algorithm			
$E_{d1}^{A\_short}$	45.78	25.69	32.91
$E_{d1}^{A\_long}$	79.63	28.36	41.82
$E_{d2}^{A\_short}$	52.51	28.72	37.13
$E_{d2}^{A\_long}$	85.61	30.25	44.70
E^{A}	50.33	30.08	37.66
Agreement between human annotators (extracts created from	45.51	45.58	45.54

scratch: E_1^B - E_2^B)			

As one can see, the best results are obtained when the summarizer is trained on extracts derived from the informative abstracts.

Table 3 summarizes the evaluation results obtained on corpus B. The evaluation results in Tables 2 and 3 show that the relation between RST trees and the extracts produced by the second analyst was much tighter than the relation between the RST trees and the extracts produced by the first analyst. As a consequence, our algorithms were in a better position to learn how to use discourse structures in order to summarize text in the style of the second analyst. In general, all three algorithms produced good results, which show that discourse structures *can* be used successfuly for text summarization even in conjunction with large texts and different summarization styles. More experiments are needed though in order to determine what types of extracts are best suited for training discourse-based summarizers (informative, indicative, extracts built from scratch, extracts derived from the abstracts, or extracts built according to other protocols).

Table 3: Evaluation results on corpus B.

Method	Rec	Prec	F-val
Position-based Baseline	30.60	30.60	30.60
Marcu's [1997] selection algorithm	31.94	31.94	31.94
HMM model HMM vs. E_1^B HMM vs. E_2^B	 30.00 37.00	 30.00 37.00	 29.00 37.00
Bayes model Bayes vs. E_1^B Bayes vs. E_2^B	 34.00 41.00	 34.00 40.00	 34.00 40.00
The dynamic, decision-based algorithm (DDB) DDB vs. E_1^B DDB vs. E_2^B	 53.96 57.66	 24.86 34.71	 34.03 43.43
Agreement between human annotators (extracts created from scratch: E_1^B - E_2^B)	45.51	45.58	45.54

6 Discussion

This paper shows that rhetorical structure trees can be successfuly used in the context of summarization to derive extracts even for large texts. The learning mechanisms we have proposed here manage to exploit correlations between rhetorical constructs and elementary discourse units that are selected as important by human judges. In spite of this, we believe RST is not capable of explaining all our data.

For example, RST does not differentiate between local and global levels of discourse. Yet, research in reading comprehension suggests that when people read, they often create a macro-structure of the document in their heads, in order to constrain the possible inferences that can be made at any given point (Rieger, 1975; Britton and Black, 1985). Even though we were able to achieve a statistically significant level of agreement on the discourse annotation task (Anonymous, 2001), we believe that investigating approaches that distinguish between local microstrategies and global macrostrategies (Meyer, 1985; Van Dijk and Kintsch, 1983) would help produce higher consistency in hierachical tagging, particularly at higher levels of the discourse structure, enabling us to exploit the discourse structure more effectively in creating text summaries.

For example, by manually examining the discourse tree for a document on which two analysts who created pure extracts had high agreement on selecting the important units (F score = 0.67), it could be seen that both analysts selected from the same sub-trees, both marked with an *elaboration-additional* relation. However, the rhetorical labels were insufficient to tell us why they chose these particular *elaboration-additional* sections over others that preceded or followed the ones they chose. The same phenomenon was observed in a number of other cases when comparing two different extracts against the corresponding discourse trees. We believe that an important next step in this work is to take a closer look at the topology of the trees, to see if there are macro-level generalizations that could help explain why certain sections get picked over others in the creation of extracts.

Another important direction is to use discourse structure in order to increase the inter-annotator agreement with respect to the task of identifying the most important information in a text. Our experiments suggest that the clearer the discourse structure of a text is, the higher the chance of agreement between human annotators who identify important edus in a text. We suspect that if human judges can visualize the discourse structure of a text, they are able to comprehend the text at a level of abstraction that may not be accessible immediately from the text, and produce better abstracts/extracts. Naturally, these are hypotheses that need further experiments in order to be tested.

References

Carlson Lynn, Daniel Marcu, and Mary Ellen Okurowsky. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. *Submitted for publication.*

Conroy, John M. and O'Leary, Dianne P., 2001. Text Summarization via Hidden Markov Models and Pivoted QR Decomposition. *Comp. Sci. Tech Rep. Univ. Of Maryland.*

Britton Bruce and John Black eds., 1985. *Understanding Expository Text*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Mann, William and Sandra Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text* 8 (3):243-281.

Marcu, Daniel. 1997. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. Dissertation, Dept. of Computer Science, University of Toronto.

Marcu, Daniel. 1998. 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.

Marcu, Daniel, Estibaliz Amorrortu, and Magdalena Romera (1999). Experiments in Constructing a Corpus of Discourse Trees. *The ACL'99 Workshop on Standards and Tools for Discourse Tagging,* pages 48-57, Maryland, June 1999.

Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge, MA: The MIT Press.

Matthiessen, Christian and Sandra Thompson. 1988. The Structure of Discourse and 'Subordination'. In Haiman, J. and Thompson, S., eds., *Clause Combining in Grammar and Discourse.* Amsterdam: John Benjamins Publishing Company, 275-329.

Meyer, Bonnie. 1985. Prose Analysis: Purposes, Procedures, and Problems. In Britton Bruce and John Black eds., *Understanding Expository Text.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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),* 344-348.

Polanyi, Livia. 1993. Linguistic Dimensions of Text Summarization. In *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication.*

Quinlan, Ross J. 1993. *C4.5: Programs for Machine Learning.*. San Mateo, CA: Morgan Kaufmann Publishers.

Rieger, C. 1975. Conceptual Memory. In Roger Schank, ed., *Conceptual Information Processing.* Amsterdam: North-Holland

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

VanDijk, Teun A. and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. New York: Academic Press.