Improving summarization through rhetorical parsing tuning

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

Abstract

We study the relationship between the structure of discourse and a set of summarization heuristics that are employed by current systems. A tight coupling of the two enables us to learn genre-specific combinations of heuristics that can be used for disambiguation during discourse parsing. The same coupling enables us to construct discourse structures that yield summaries that contain textual units that are not only important according to a variety of position-, title-, and lexical-similarity-based heuristics, but also central to the main claims of texts. A careful analysis of our results enables us to shed some new light on issues related to summary evaluation and learning.

1 Motivation

Current approaches to automatic summarization employ techniques that assume that textual salience correlates with a wide range of linguistic phenomena. Some of these approaches assume that important textual units contain words that are used frequently (Luhn, 1958; Edmundson, 1968) or words that are used in the title and section headings (Edmundson, 1968). Some of them assume that important sentences are located at the beginning or end of paragraphs (Baxendale, 1958) or at positions that can be determined through training for each particular text genre (Lin and Hovy, 1997). Other systems assume that important sentences in texts contain "bonus" words and phrases, such as *significant, important, in conclusion* and *In this paper we show,* while unimportant sentences contain "stigma" words such as *hardly* and *impossible* (Edmundson, 1968; Kupiec et al., 1995; Teufel and Moens, 1997). Other systems assume that important sentences and concepts are the highest connected entities in more or less elaborate semantic structures (Skorochodko, 1971; Hoey, 1991; Salton and Allan, 1995; Mani and Bloedorn, 1998; Barzilay and Elhadad, 1997). And, yet, others assume that important sentences and clauses are derivable from a discourse representation of texts (Ono et al., 1994; Marcu, 1997a; Marcu, 1997c).

A variety of systems (Edmundson, 1968; Kupiec et al., 1995; Teufel and Moens, 1997; Lin, 1998; Mani and Bloedorn, 1998) were designed to integrate subsets of the heuristics mentioned above. In these approaches, each individual heuristic yields a probability distribution that reflects the importance of sentences. A combination of the probability distributions defined by each heuristic yields the sentences that are most likely to be included in a summary.

What all these multiple heuristic-based systems have in common is that they treat texts as *flat sequences of sentences* — no such system employs discourse-based heuristics. As a consequence, it is possible, for example, a sentence to be assigned a high importance score on the basis of its position in the text and its semantic similarity with the title, although it is subsidiary to the main argument made in the text. In this paper, we remedy this shortcoming, by taking advantage of the structure of discourse.

More precisely, we study the relationship between the structure of discourse and a set of summarization heuristics that are employed by current systems. A tight coupling of the two, which is achieved by applying a simple learning mechanism, gives us two advantages over previous methods. First, two corpora of manually built summaries enable us to learn genre-specific combinations of heuristics that can be used for disambiguation during discourse parsing. Second, the discourse structures that we derive enable us to select textual units that are not only important according to a variety of position-, title-, and lexically-based heuristics, but also central to the main claims of texts.

In section 2, we review the discourse theory and the algorithms that constitute the foundation of our work. We explain then our approach to fusing various summarization heuristics in our discourse processing framework (section 3) and we review each of the heuristics from a discourse perspective. In section 4, we evaluate the appropriateness of using each individual heuristic for summarization and present an algorithm that finds combinations of heuristics that yield optimal summaries. We end the paper by assessing the strengths and limitations of our approach.

2 Background work

RST. The discourse theory that we are going to use is Rhetorical Structure Theory(RST) (Mann and Thompson, 1988). Central to RST is the notion of *rhetorical* *relation*, which is a relation that holds between two non-overlapping text spans called NUCLEUS and SATELLITE. (There are a few exceptions to this rule: some relations, such as CONTRAST, are multinuclear.) The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's purpose than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa.

Text coherence in RST is assumed to arise from a set of constraints. The constraints operate on the nucleus, on the satellite, and on the combination of nucleus and satellite. For example, an EVIDENCE relation holds between the nucleus (labelled as 5 in text (1), which is shown below) and the satellite (labelled as 6 in text (1)), because the nucleus presents some information that the writer believes to be insufficiently supported to be accepted by the reader; the satellite presents some information that is thought to be believed by the reader or that is credible to her; and the comprehension of the satellite increases the reader's belief in the nucleus. Rhetorical relations can be assembled into rhetorical structure trees (RS-trees) by recursively applying individual relations to spans that range in size from one clause-like unit to the whole text.

Rhetorical parsing. Recent developments in computational linguistics have created the means for the automatic derivation of rhetorical structures of unrestricted texts. For example, when the text shown in (1), below, is given as input to the *rhetorical parsing algorithm* that is discussed in detail by Marcu (1997b; 1997c), it is broken into ten elementary units (those surrounded by square brackets). The rhetorical parsing algorithm then uses cue phrases and a simple notion of semantic similarity in order to hypothesize rhetorical relations among the elementary units. Eventually, the algorithm derives the rhetorical structure tree shown in figure 1.

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

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

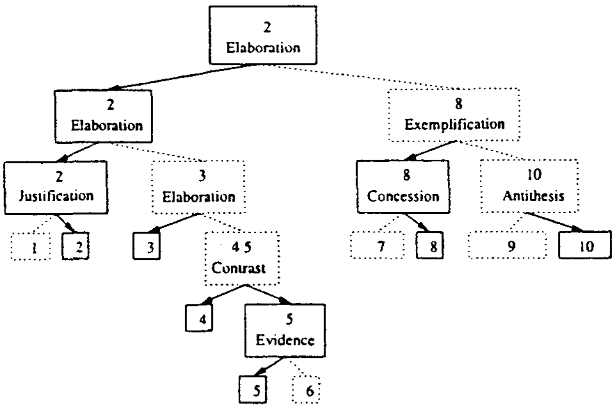

Figure 1: The discourse tree built by the rhetorical parser (Marcu, 1997c) for text (1).

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

Discourse-based summarization. Once a discourse structure such as that shown in figure 1 is created, we can derive a partial ordering of the important units in the original text by considering that the units that are promoted closer to the root are more important than those that are promoted less close. By applying this criterion to tree 1, we obtain the partial ordering shown in (2), below, because unit 2 is the only promotion unit associated with the root, unit 8 is the only unit found one level below the root, units 3 and 10 are the only units found two levels below the root, and so on.

(2) $2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > 6$

Using partial ordering (2) we can obtain a summary that contains k% of the original text by selecting the first k%

units in the partial ordering.

By applying this algorithm, Marcu (1997a; 1997c) has built a summarization system that recalled 52.77% (with precision 50.00%) of the clause-like units that were considered important by human judges in a collection of five texts.

3 An enhanced discourse-based framework for text summarization

3.1 Introduction

There are two ways in which one can integrate a discourse-based measure of textual saliency, such as that described above, with measures of saliency that are based on cohesion, position, similarity with the title, etc. The simplest way is to compute a probability distribution of the importance of textual units according to the discourse method and to combine it with all probability distributions produced by the other heuristics. In such an approach, the discourse heuristic is just one of the n heuristics that are employed by a system. Obtaining good summaries amounts then to determining a good way of combining the implemented heuristics.

Overall, a summarization system that works along the lines described above still treats texts as flat sequences of textual units, although the discourse method internally uses a more sophisticated representation. The shortcoming of such an approach is that it still permits the selection of textual units that do not play a central role in discourse. For example, if the text to be summarized consists only of units 7 and 8 in text (1), it may be possible that the combination of the position, title, and discourse heuristics will yield a higher score for unit 7 than for unit 8, although unit 8 is the nucleus of the text and expresses what is important. Unfortunately, if we interpret text as a flat sequence of units, the rhetorical relation and the nuclearity assignments with respect to these units cannot be appropriately exploited.

A more complex way to integrate discourse, cohesion, position, and other summarization-based methods is to consider that the structure of discourse is the most important factor in determining saliency, an assumption supported by experiments done by Mani et al. (1998). In such an approach, we no longer interpret texts as flat sequences of textual units, but as tree structures that reflect the nuclearity and rhetorical relations that characterize each textual span. When discourse is taken to be central to the interpretation of text, obtaining good summaries amounts to finding the "best" discourse interpretations. In the rest of the paper, we explore this approach.

3.2 Criteria for measuring the "goodness" of discourse structures

In order to find the 'best' discourse interpretations, i.e., the interpretations that yield summaries that are most similar to summaries generated manually, we considered seven metrics, which we discuss below.

The clustering-based metric. A common assumption in the majority of current text theories is that good texts exhibit a well-defined topical structure. In our approach, we assume that a discourse tree is "better" if it exhibits a high-level structure that matches as much as possible the topical boundaries of the text for which that structure is built.

In order to capture this intuition, when we build discourse trees, we associate with each node of a tree a clustering score. For the leaves, this score is 0; for the internal nodes, the score is given by the similarity between the immediate children. The similarity is computed using a traditional cosine metric, in the style of Hearst (1997). We consider that a discourse tree A is "better" than another discourse tree B if the sum of the clustering scores associated with the nodes of A is higher than the sum of the clustering scores associated with the nodes of B.

The marker-based metric. Naturally occurring texts use a wide range of discourse markers, which signal coherence relations between textual spans of various sizes. We assume that a discourse structure should reflect explicitly as many of the discourse relations that are signaled by discourse markers. In other words, we assume that a discourse structure A is better than a discourse structure B if A uses more rhetorical relations that are explicitly signaled than B.

The rhetorical-clustering-based metric. The clustering-based metric discussed above computes an overall similarity between two textual spans. However, in the discourse formalization proposed by Marcu (1996; 1997c), it is assumed that whenever a discourse relation holds between two textual spans, that relation also holds between the salient units (nuclei) associated with those spans. We extend this observation to similarity as well, by introducing the rhetorical-clustering-based metric, which measures the similarity between the salient units associated with two spans. For example, the clustering-based score associated with the root of the tree in figure 1 measures the similarity between spans [1,6] and [7,10]. In contrast, the rhetorical-clustering-based score associated with the root of the same tree measures the similarity between units 2 and 8, which are the salient units that pertain to spans [1,6] and [7,10] respectively. In the light of the rhetorical-clustering-based metric, we consider that a discourse tree A is "better" than another discourse tree B if the sum of the rhetorical-clustering scores associated with the nodes of A is higher than the sum of the rhetorical-clustering scores associated with the nodes of B.

The shape-based metric. The only disambiguation metric that we used in our previous work (Marcu, 1997b) was the shape-based metric, according to which the "best" trees are those that are skewed to the right. The explanation for this metric is that text processing is, essentially, a left-to-right process. In many genres, people write texts so that the most important ideas go first, both at the paragraph and at the text levels.[1] The more text writers add, the more they elaborate on the text that went before: as a consequence, incremental discourse building consists mostly of expansion of the right branches. According to the shape-based metric, we consider that a discourse tree A is "better" than another discourse tree B if A is more skewed to the right than B (see Marcu (1997c) for a mathematical formulation of the notion of skewedness).

The title-based metric. A variety of systems assume that important sentences in a text use words that occur in the title. We measure the similarity between each textual unit and the title by applying a traditional cosine metric. We compute a title-based score for each discourse structure by computing the similarity between the title and the units that are promoted as salient in that structure. The intuition that we capture in this way is that a discourse structure should be constructed so that it promotes as close to the root as possible the units that are similar with the title. According to the title-based metric, we consider that a discourse structure A is "better" that a discourse structure B if the title-based score of A is higher than the title-based score of B.

The position-based metric. Research in summarization (Baxendale, 1958; Edmundson, 1968; Kupiec et al., 1995; Lin and Hovy, 1997) has shown that, in genres with stereotypical structure, important sentences are often located at the beginning or end of paragraphs/documents. Our position-based metric captures this intuition by assigning a positive score to each textual unit that belongs to the first two or last sentences of the the first three or last two paragraphs. We compute a position-based score for each discourse structure by averaging the position-based scores of the units that are promoted as salient in that discourse structure. The intuition that we capture in this way is that a discourse structure should be constructed so that it promotes as close to the root as possible the units that are located at the beginning or end of a text. According to the position-based metric, we consider that a discourse structure A is "better" that a discourse structure B if the position-based score of A is higher than the position-based score of B.

The connectedness-based metric. A heuristic that is often employed by current summarization systems is that of considering important the highest connected entities in more or less elaborate semantic structures (Skorochodko, 1971; Hoey, 1991; Salton and Allan, 1995; Mani and Bloedorn, 1998; Barzilay and Elhadad, 1997). We implement this heuristic by computing the average cosine similarity of each textual unit in a text with respect to all the other units. We associate a connectedness-based score to each discourse structure by averaging the connectedness-based scores of the units that are promoted as salient in that discourse structure. As in the case of the other metrics, we consider that a discourse structure A is "better" that a discourse structure B if the connectedness-based score of A is higher than the connectedness-based score of B.

4 Combining heuristics

4.1 The approach

As we have already mentioned, discourse parsing is ambiguous the same way sentence parsing is: the rhetorical parsing algorithm often derives more than one discourse structure for a given text. Each of the seven metrics listed above favors a different discourse interpretation.

For the purpose of this paper, we assume that the "best" discourse structures are given by a linear combination of the seven metrics. Hence, along the lines described in section 3.2, we associate with each discourse structure a clustering-based score s_{clust}, a marker-based score s_{mark}, a rhetorical-clustering-based score s_{rhet_clust}, a shape-based score s_{shape}, a title-based score s_{title}, a position-based score s_{pos}, and a connectedness-based score s_{con}; and we assume that the best tree of a text is that that corresponds to the discourse structure D that has the highest score $s(D)$. The score $s(D)$ is computed as shown in (3), where $w_{clust}, \ldots, w_{con}$ are weights associated with each metric.

$$
\begin{aligned}
s(D) = {} & w_{clust} \times s_{clust}(D) + \\
& w_{mark} \times s_{mark}(D) + \\
& w_{rhet\_clust} \times s_{rhet\_clust}(D) + \\
& w_{shape} \times s_{shape}(D) + w_{title} \times s_{title}(D) + \\
& w_{pos} \times s_{pos}(D) + w_{con} \times s_{con}(D).
\end{aligned} \tag{3}
$$

To avoid data skewedness, the scores that correspond to each metric are normalized to values between 0 and 1.

Given the above formulation, our goal is to determine combinations of weights that yield discourse structures that, in turn, yield summaries that are as close as possible to those generated by humans. In discourse terms, this amounts to using empirical summarization data for discourse parsing disambiguation.

4.2 Corpora used in the study

In order to evaluate the appropriateness for summarization of each of the heuristics, we have used two corpora: a corpus of 40 newspaper articles from the TREC collection (Jing et al., 1998) and a corpus of five articles from *Scientific American* (Marcu, 1997a).

Five human judges selected sentences to be included in 10% and 20% summaries of each of the articles in the TREC corpus (see (Jing et al., 1998) for details). For each of the 40 articles and for each cutoff figure (10% and 20%), we took the set of sentences selected by at least three human judges as the "gold standard" for summarization. In our initial experiments, we noticed that the rhetorical parsing algorithm needed more than 1 minute in order to automatically generate summaries for seven of the 40 articles in the TREC corpus, which were highly ambiguous from a discourse perspective. In order to enable a better employment of training techniques that are

[1] In fact, journalists are trained to employ this "pyramid" approach to writing consciously (Cumming and McKercher, 1994).

Metric	Recall	Precision
Humans	83.20%	75.95%
Clustering	48.08%	54.29%
Marker	38.63%	44.44%
Rhetorical clustering	26.26%	27.87%
Shape	44.04%	52.52%
Title	58.93%	67.67%
Position	52.87%	63.73%
Connectedness	35.35%	31.31%
Lead	82.91%	63.45%
Random	9.44%	9.44%

Table 1: The appropriateness of each of the seven metrics for text summarization in the TREC corpus — the 10% cutoff.

Metric	Recall	Precision
Humans	82.83%	64.93%
Clustering	40.99%	43.61%
Marker	37.91%	38.78%
Rhetorical clustering	23.10%	24.68%
Shape	46.54%	49.73%
Title	42.29%	40.17%
Position	37.48%	40.97%
Connectedness	29.87%	32.78%
Lead	70.91%	46.96%
Random	15.80%	15.80%

Table 2: The appropriateness of each of the seven metrics for text summarization in the TREC corpus — the 20% cutoff.

specific to machine learning, we partitioned the TREC collection into two subsets. The first subset contained 15 documents: this subset included the seven documents for which our summarization algorithm required extensive computation and eight other documents that were selected randomly. The second subset contained the rest of 25 documents for which our algorithm could generate summaries sufficiently fast. For the purpose of this paper, we will refer to the collection of 25 articles as the "training corpus" and to the collection of 15 articles as the "test corpus". However, the reader should not take the denotations associated with these referents literally, because the partitioning was not performed randomly. Rather, the reader should see the partitioning only as a means for accelerating the process that determines combinations of heuristics that yield the best summarization results for all the texts in the corpus.

The second corpus consisted of five *Scientific American* texts whose elementary textual units (clause-like units) were labelled by 13 human judges as being very important, somewhat important, or unimportant (see (Marcu, 1997c) for the details of the experiment). For each of the five texts, we took the set of textual units for which at least seven judges agreed to be very important as the gold standard for summarization.

We built automatically discourse structures for the texts in the two corpora using various combinations of weights and we compared the summaries that were derived from these structures with the gold standards. The comparison employed traditional recall and precision figures, which reflected the percent of textual units that were identified correctly by the program with respect to the gold standards and the percent of textual units that were identified correctly by the program with respect to the total number of units that were identified by the program.

For both corpora, we attempted to mimic as closely as possible the summarization tasks carried out by human judges. For the TREC corpus, we automatically extracted summaries at 10% and 20% cutoffs; for the *Scientific American* corpus, we automatically extracted summaries that reflected the lengths of the summaries on which human judges agreed.

4.3 Appropriateness for summarization of the individual heuristics

The TREC corpus. Initially, we evaluated the appropriateness for text summarization of each of the seven heuristics at both 10% and 20% cutoffs for the collection of texts in the TREC corpus. By assigning in turn value 1 to each of the seven weights, while the other six weights were assigned value 0, we estimated the appropriateness of using each individual metric for text summarization.

Tables 1 and 2 show the recall and precision figures that pertain to discourse structures that were built for the TREC corpus, in order to evaluate the appropriateness for text summarization of each of the seven metrics at 10% and 20% cutoffs, respectively. For a better understanding of the impact of each heuristic, tables 1 and 2 also show the recall and precision figures associated with the human judges and with two baseline algorithms. The recall and precision figures for the human judges were computed by taking the average recall and precision of the summaries built by each human judge individually when compared with the gold standard. These recall and precision figures can be interpreted as summarization upper-bounds for the collection of texts that they characterize. Since each judge contributed to the derivation of the gold standards, the recall and precision figures that pertain to human judges are biased: they are probably higher than the figures that would characterize an outsider to the experiment.

The recall and precision figures that pertain to the baseline algorithms are computed as follows: the lead-based algorithm assumes that important units are located at the beginning of texts; the random-based algorithm assumes that important units can be selected randomly.

The results in table 1 show that, for newspaper articles, the title- and position-based metrics are the best individual metrics for distinguishing between discourse trees that are appropriate for generating 10% summaries and discourse trees that are not. Interestingly, none of these heuristics taken in isolation is better than the lead-based algorithm. In fact, the results in table 1 show that there is almost no quantitative difference in terms of recall and precision between summaries generated by the lead-based algorithm and summaries generated by humans.

We were so puzzled by this finding that we investigated further this issue: by scanning the collection of 40 articles, we came to believe that since most of them are very short and simple, they are inappropriate as a testbed for summarization research. To estimate the validity of this belief, we focused our attention on a subset of 10 articles that seemed to use a more sophisticated writing style, that did not follow straightforwardly the pyramid-based approach; each of these 10 articles used at least once the word "computer". When we evaluated the performance of the lead-based algorithm on this subset, we obtained figures of 66.00% recall and 43.66% precision at the 10% cutoff. This result suggests that as soon more sophisticated texts are considered, the performance of the lead-based algorithm decreases significantly even within the newspaper genre.

The results in table 2 show that, for newspaper articles, the shape-based metric is the best individual metric for distinguishing between discourse trees that are appropriate for 20% summaries and discourse trees that are not. Still, the shape-based heuristic is not better than the lead-based algorithm.

The *Scientific American* corpus. When we evaluated the appropriateness for text summarization of the heuristics at both clause and sentence levels for the collection of texts in the *Scientific American* corpus, we obtained a totally different distribution of the configuration of weights that yielded the highest recall and precision figures.

A close analysis of the results in table 3 shows that, for *Scientific American* articles, the clustering-, rhetorical-clustering-, and shape-based metrics are the best individual metrics for distinguishing between discourse trees that are good for clause-based summarization and discourse trees that are not.

The results in table 4 show that, for *Scientific American* articles, the shape-based metric is the best individual metric for distinguishing between discourse trees that are appropriate for sentence-based summarization. Surprisingly, the title-, position-, and connectedness-based metrics underperform even the random-based metric.

In contrast with the results that pertain to the TREC corpus, the lead-based algorithm performs significantly worse than human judges for the texts in the *Scientific American* corpus, despite the *Scientific American* texts being shorter than those in the TREC collection.

Discussion. Overall, the recall and precision figures presented in this section suggest that no individual heuristic consistently guarantees success across different text genres. Moreover, the figures suggest that, even within the same genre, the granularity of the textual units that are selected for summarization and the overall length of the summary affect the appropriateness of a given heuristic.

Metric	Recall	Precision
Humans	72.66%	69.63%
Clustering	54.05%	66.66%
Marker	43.24%	55.17%
Rhetorical clustering	48.65%	62.07%
Shape	51.35%	63.33%
Title	40.54%	55.56%
Position	29.73%	47.83%
Connectedness	24.32%	40.91%
Lead	39.68%	39.68%
Random	25.70%	25.70%

Table 3: The appropriateness of each of the seven metrics for text summarization in the *Scientific American* corpus — the clause-like unit case.

Metric	Recall	Precision
Humans	78.11%	79.37%
Clustering	42.31%	42.31%
Marker	42.31%	42.31%
Rhetorical clustering	46.15%	40.00%
Shape	57.69%	51.72%
Title	30.77%	33.33%
Position	30.77%	38.10%
Connectedness	23.08%	25.00%
Lead	54.22%	54.22%
Random	38.40%	38.40%

Table 4: The appropriateness of each of the seven metrics for text summarization in the *Scientific American* corpus — the sentence case.

By focusing only on the human judgments, we notice that the newspaper genre yields a higher consistency than the *Scientific American* genre with respect to what humans believe to be important. Also, the results in this section show that humans agree better on important sentences than on important clauses; and that within the newspaper genre, they agree better on what is very important (the 10% summaries) than on what is somewhat important (the 20% summaries).

4.4 Learning the best combinations of heuristics

The individual applications of the metrics suggest what heuristics are appropriate for summarizing texts that belong to the text genres of the two corpora. In addition to this assessment, we were also interested in finding *combinations* of heuristics that yield good summaries. To this end, we employed a simple learning paradigm, which we describe below.

4.4.1 A GSAT-like algorithm

In the framework that we proposed in this paper, finding a combination of metrics that is best for summarization amounts to finding a combination of weights

$w_{clust}, \dots, w_{con}$ that maximizes the recall and precision figures associated with automatically built summaries. The algorithm shown in figure 2 performs a greedy search in the seven-dimensional space defined by the weights, using an approach that mirrors that proposed by Selman, Levesque, and Mitchell (1992) for solving propositional satisfiability problems.

The algorithm assigns initially to each member of the vector of weights $\vec{W}_{max}$ a random value in the interval $[0, 1]$. This assignment corresponds to a point in the n-dimensional space defined by the weights. The program then attempts *NoOfSteps* times to move incrementally, in the n-dimensional space, along a direction that maximizes the F-measure of the recall and precision figures that pertain to the automatically built summaries. The F-measure is computed as shown in (4), below.

$$(4) \quad F = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The F-measure always takes values between the values of recall and precision, and is higher when recall and precision are closer.

For each point $\vec{W}_t$ the program computes the F-value of the recall and precision figures of the summaries that correspond to all the points in the neighborhood of $\vec{W}_t$ that are at distance Δw along each of the seven axes (lines 6–10 in figure 2). From the set of 14 points that characterize the neighborhood of the current configuration $\vec{W}_t$, the algorithm selects randomly (line 12) one of the weight configurations that yielded the maximum F-value (line 11). In line 13, the algorithm moves in the n-dimensional space to the position that characterizes the configuration of weights that was selected on line 12. After *NoOfSteps* iterations, the algorithm updates the configuration of weights $\vec{W}_{max}$ such that it reflects the combination of weights that yielded the maximal F-value of the recall and precision figures (line 15 in figure 2). The algorithm repeats this process *noOfTries* times, in order to increase the chance of finding a maximum that is not local.

Since the lengths of the summaries that we automatically extracted was fixed in all cases, we chose to look for configurations of weights that maximized the F-value of the recall and precision figures. However, one can use the algorithm in figure 2 to find configurations of weights that maximize only the recall or only the precision figure as well.

4.4.2 Results

The TREC corpus. We have experimented with different values for *noOfTries*, *noOfSteps*, and Δw. When we ran the algorithm shown in figure 2 on the collection of 25 texts in our training TREC corpus, with *noOfTries* $= 50$, *noOfSteps* $= 60$, and $\Delta w = 0.4$, we obtained multiple configurations of weights that yielded maximal F-values of the recall and precision figures at both 10% and 20% cutoffs. Table 5 shows only the two best configurations for each cutoff. The best configuration of weights for the 10% cutoff recalls 68.33% of the sentences considered important by human judges in the whole TREC corpus with a precision of 84.16%. The F-value of the recall and precision figures for this configuration is 75.42%, which is approximately 4% lower than the F-value that pertains to human judges and 3.5% higher that the F-value that pertains to the lead-based algorithm. The results in table 5 show that at 10% cutoff there is not too much difference between summaries built by human judges, by the rhetorical parser, and by the lead-based algorithm. Since the lead-based algorithm performs so well at the 10% level, the only conclusion that we can draw is that *for short newspaper articles*, the lead-based algorithm is the most efficient solution.

At the 20% cutoff, the best configuration of weights recalls 59.51% of the sentences considered important by human judges in the whole corpus, with 72.11% precision. The F-value of the recall and precision figures for this configuration is 65.21%, which is about 7.5% lower than the F-value that pertains to human judges and 8.5% higher than the F-value that pertains to the lead-based algorithm. These results suggest that when we want to build longer summaries, the lead-based heuristic is no longer appropriate even within the newspaper genre (and even for simple articles).

The *Scientific American* corpus. We also ran the algorithm shown in figure 2 on the collection of five texts in our *Scientific American* corpus, with *noOfTries* $= 120$, *noOfSteps* $= 50$, and $\Delta w = 0.4$. Table 6 shows 2 configurations of weights that yielded maximal F-values of the recall and precision figures at the clause-like unit level and 4 configurations of weights that yielded maximal F-values at the sentence level. The best combination of weights for summarization at the clause-like unit level recalls 67.57% of the elementary units considered important by human judges, with a precision of 73.53%. The F-value of the recall and precision figures for this configuration is 70.42%, which is less than 1% lower than the F-value that pertains to human judges and about 30% higher than the F-value that pertains to the lead-based algorithm. This result outperforms significantly the previous 52.77% recall and 50.00% precision figures that were obtained by Marcu using only the shape-based heuristic (1997c; 1997a).

The best combination of weights for summarization at the sentence level recalls 69.23% of the sentences considered important by human judges, with a precision of 64.29%. The F-value of the recall and precision figures for this configuration is 66.67%, which is about 12% lower than the F-value that pertains to human judges but about 12% higher than the F-value that pertains to the lead-based algorithm. These results suggest that although *Scientific American* articles cannot be summarized properly by applying a simple, lead-based heuristic, they can be by applying the discourse-based algorithm.

Discussion. Because of the limited size of the corpora and because of the rhetorical ambiguity of some of the texts in the TREC corpus, carrying out cross-validation experiments was either meaningless or prohibitively ex-

```
Input: A corpus of texts $C_T$.
       The manually built summaries $S_T$ for the texts in $C_T$.
       $NoOfTries$, $NoOfSteps$, $\Delta w$.
Output: The weights $\vec{W}_{max} = \{w_{clust}, w_{mark}, \dots, w_{con}\}$ that yield the best summaries with respect to $C_T$ and $S_T$.

$\vec{W}_{max} = \{w_{clust}, w_{mark}, \dots, w_{con}\} = \{rand(0,1), rand(0,1), \dots, rand(0,1)\}$
for tries = 1 to $NoOfTries$
    $\vec{W}_t = \{w_{clust}, w_{mark}, \dots, w_{con}\} = \{rand(0,1), rand(0,1), \dots, rand(0,1)\}$
    $F_t$ = F_RecallAndPrecision($w_{clust}, w_{mark}, \dots, w_{con}$);
    for flips = 1 to $NoOfSteps$
        $F_1$ = F_RecallAndPrecision($w_{clust} + \Delta w, w_{mark}, \dots, w_{con}$);
        $F_2$ = F_RecallAndPrecision($w_{clust} - \Delta w, w_{mark}, \dots, w_{con}$);
        ...
        $F_{13}$ = F_RecallAndPrecision($w_{clust}, w_{mark}, \dots, w_{con} + \Delta w$);
        $F_{14}$ = F_RecallAndPrecision($w_{clust}, w_{mark}, \dots, w_{con} - \Delta w$);
        $F_{max} = max(F_t, F_1, F_2, \dots, F_{14})$;
        $F_t$ = randomOf($F_{max}$);
        $\vec{W}_t$ = weightsOf($F_t$);
    endfor
    $\vec{W}_{max} = max(\vec{W}_{max}, \vec{W}_t)$;
endfor
return $\vec{W}_{max}$
```

Figure 2: A GSAT-like algorithm for improving summarization.

Corpus	Method	w_{clust}	w_{mark}	w_{rhet_clust}	w_{shape}	w_{title}	w_{pos}	w_{con}	Recall	Precision	F-val
10%	Humans								83.20%	75.95%	79.41%
Training 10%	Program	0.69	0.22	−0.53	1.32	0.75	3.79	−0.78	75.33%	82.00%	78.52%
Testing 10%									56.66%	87.77%	68.86%
Both									68.33%	84.16%	**75.42%**
Training 10%	Program	1.71	0.62	0.58	0.99	1.27	1.20	−1.17	75.33%	84.66%	79.92%
Testing 10%									51.66%	81.66%	63.28%
Both									66.45%	83.53%	74.01%
10%	Lead								82.91%	63.45%	71.89%
20%	Humans								82.83%	64.93%	72.80%
Training 20%	Program	0.79	0.23	−0.09	1.33	0.40	0.45	−0.28	57.40%	70.31%	63.20%
Testing 20%									63.02%	75.11%	68.54%
Both									59.51%	72.11%	**65.21%**
Training 20%	Program	0.65	1.02	0.37	0.87	0.88	0.37	0.61	58.77%	58.26%	58.51%
Testing 20%									54.41%	54.66%	54.53%
Both									57.14%	56.91%	57.02%
20%	Lead								70.91%	46.96%	56.50%

Table 5: The combination of heuristics that yielded the best summaries for the texts in the TREC corpus.

pensive. As a consequence, the recall and precision results that are reported in this paper can be interpreted as being only suggestive of discourse-based summarization performance. However, the experiments do support conclusions that pertain to the integration of multiple heuristics.

The analysis of the patterns of weights in tables 5 and 6 shows that, for both corpora, no individual heuristic is a clear winner with respect to its contribution to obtaining good summaries. For the TREC corpus, with the exception of the rhetorical-clustering- and the connectedness-based heuristics, all other heuristics seem to contribute consistently to the improvement in summarization quality. For the *Scientific American* corpus, when combined with other heuristics, the marker-, rhetorical-clustering-, shape-, and title-based heuristics seem to contribute consistently to the improvement in recall and precision figures in almost all cases. In contrast, the clustering-, position-, and connectedness-based heuristics seem to be even detrimental with respect to the collection of texts that we considered.

However, the conclusion that seems to be supported by

Granularity	Method	w_{clust}	w_{mark}	w_{rhet_clust}	w_{shape}	w_{title}	w_{pos}	w_{con}	Recall	Precision	F-val
Clause-like unit	Humans								72.66%	69.93%	71.27%
	Program	0.37	0.04	1.27	0.58	0.54	−1.16	−1.24	67.57%	73.53%	**70.42%**
		1.34	0.27	0.69	0.58	−0.08	2.07	−0.53	62.16%	71.87%	66.66%
	Lead								39.68%	39.68%	39.68%
Sentence	Humans								78.11%	79.37%	78.73%
	Program	0.41	0.36	0.14	1.75	0.65	−0.46	−0.73	69.23%	64.29%	**66.67%**
		−1.51	0.13	0.65	0.84	0.52	−1.23	1.06	65.38%	68.00%	66.66%
		0.59	0.03	0.86	0.56	0.75	−0.79	0.50	61.54%	66.67%	64.00%
		−0.82	0.04	0.47	0.18	0.54	0.03	3.15	61.54%	66.67%	64.00%
	Lead								54.22%	54.22%	54.22%

Table 6: The combination of heuristics that yielded the best summaries for the texts in the *Scientific American* corpus.

Corpus	Method	w_{clust}	w_{mark}	w_{rhet_clust}	w_{shape}	w_{title}	w_{pos}	w_{con}	Recall	Precision	F-val
10%	Humans								83.20%	75.95%	79.58%
	Program (best at 10%)								68.33%	84.16%	75.42%
	Program (best at 20%)	0.79 0.65	0.23 1.02	−0.09 0.37	1.33 0.87	0.40 0.88	0.45 0.37	−0.28 0.61	64.58% 60.83%	81.67% 69.99%	72.13% 65.09%
	Lead								82.91%	63.45%	71.89%
20%	Humans								82.83%	64.93%	72.80%
	Program (best at 20%)								59.51%	72.11%	65.21%
	Program (best at 10%)	0.69 1.71	0.22 0.62	−0.53 0.58	1.32 0.99	0.75 1.27	3.79 1.20	−0.78 −1.17	58.13% 54.11%	62.23& 59.78%	60.18% 56.80%
	Lead								70.91%	46.96%	56.50%

Table 7: A cross-analysis of summarization results in the TREC corpus.

the data in tables 5 and 6 is that the strength of a summarization system does not depend so much on the use of one heuristic, but rather on the ability of the system to use an optimal combination of heuristics. The data also shows that the optimal combinations need not necessarily follow a common pattern: for example, the combinations of heuristics that yield the highest F-values of the recall and precision figures for the 10% cutoff in the TREC corpus differ dramatically: one combination relies almost entirely on the position-based heuristic, while the other combination uses a much more balanced combination of heuristics that is slightly biased towards assigning more importance to the clustering-based heuristic.

In addition to the analysis of the patterns of weights that yielded optimal summaries in the two corpora, we also examined the appropriateness of using combinations of weights that were optimal for a given summary cutoff in order to summarize texts at a different cutoff. Table 7 shows the recall and precision figures that are obtained when the patterns of weights that yielded optimal summaries at 10% cutoff are used to summarize the texts in the TREC corpus at 20% cutoff; and the recall and precision figures that are obtained when the patterns of weights that yielded optimal summaries at 20% cutoff are used to summarize the same texts at 10% cutoff. As the figures in table 7 show, the combinations of heuristics that yielded optimal summaries at a particular cutoff do not yield optimal summaries at other cutoffs, although we can still find combinations of heuristics that outperform the lead-based algorithm at both cutoffs. The results in table 7 suggest that there are at least two ways in which one can train a summarization system. If the system is going to be used frequently to summarize texts at a given cutoff, then it makes sense to train it to produce good summaries at that cutoff. However, if the system is going to be used to generate summaries of various lengths, then a different training methodology should be adopted, one that would ensure optimality across the whole cutoff spectrum, from 1 to 99%.

5 Conclusions

The empirical and computational experiments that we described in this paper support at least the following conclusions. 1. For extracting 10% summaries of short articles of the news story genre, a simple lead-based algorithm is the most efficient solution. 2. For extracting longer summaries of short newspaper articles and for extracting any size summaries of complex (not necessarily news stories) newspaper articles, a simple lead-based algorithm does not provide a satisfactory solution. This assertion holds for other text genres, such as that of *Scientific American*, as well. 3. There is no magic key (heuristic) for obtaining good summarization results; rather the strength of a summarization system seems to come from its ability to combine a multitude of heuristics. 4. Combinations of heuristics that yield "optimal" results for certain summary extract lengths might not yield optimal results for different lengths. 5. Incorporating various heuristics into a discourse-based summarization framework yields good results. 6. In order to assess confidently the effectiveness of the summarization methodology that was introduced here, much larger corpora are required.

Acknowledgements. I am grateful to Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad for sharing their TREC-based summarization corpus and to Chin-Yew Lin for fruitful discussions of previous versions of this paper. Special thanks go to Eduard Hovy for comments, discussions, and stimulative enthusiasm.

References

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, July 11.

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

Carmen Cumming and Catherine McKercher. 1994. *The Canadian Reporter: News writing and reporting*. Hartcourt Brace.

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

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

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

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *Proceedings of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 60–68, Stanford, March 23–25.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM/SIGIR Annual Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington.

Chin-Yew Lin 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.

Chin-Yew Lin. 1998. Assembly of topic extraction modules in SUMMARIST. In *Proceedings of the AAAI Spring Symposium on Intelligent Text Summarization*, Stanford, CA, March 23–25.

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

Inderjeet Mani and Eric Bloedorn. 1998. Machine learning of generic and user-focused summarization. In *Proceedings of Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin, July 26–30.

Inderjeet Mani, Eric Bloedorn, and Barbara Gates. 1998. Using cohesion and coherence models for text summarization. In *Proceedings of the AAAI'98 Spring Symposium on Intelligent Text Summarization*, Stanford, CA, March 23–25.

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

Daniel Marcu. 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.

Daniel Marcu. 1997a. From discourse structures to text summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 11.

Daniel Marcu. 1997b. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 96–103, Madrid, Spain, July 7-12.

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

Kenji Ono, 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.

Gerard Salton and James Allan. 1995. Selective text utilization and text traversal. *International Journal of Human-Computer Studies*, 43:483–497.

Bart Selman, Hector Levesque, and David Mitchell. 1992. A new method for solving hard satisfiability problems. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, pages 440–446, San Jose, California.

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

Simone Teufel and Marc Moens. 1997. Sentence extraction as a classification task. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 58–65, Madrid, Spain, July 11.