TextTiling: A Quantitative Approach to Discourse Segmentation

Marti A. Hearst
Computer Science Division, 571 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
marti@cs.berkeley.edu

Abstract

This paper presents TextTiling, a method for partitioning full-length text documents into coherent multi-paragraph units. The layout of text tiles is meant to reflect the pattern of subtopics contained in an expository text. The approach uses lexical analyses based on tf.idf, an information retrieval measurement, to determine the extent of the tiles, incorporating thesaural information via a statistical disambiguation algorithm. The tiles have been found to correspond well to human judgements of the major subtopic boundaries of science magazine articles.

1 Introduction

Expository texts such as science magazine articles and environmental impact reports can be viewed as being composed of a few main topics and a series of short, sometimes densely discussed, subtopics. For example, consider a 23-paragraph article from *Discover* magazine whose main topic is the exploration of Venus by the Magellan space probe. A reader divided this text into the following segments, with the labels shown, where the numbers indicate paragraph numbers:

1-2	*Intro to Magellan space probe*
3-4	*Intro to Venus*
5-7	*Lack of craters*
8-11	*Evidence of volcanic action*
12-15	*River Styx*
16-18	*Crustal spreading*
19-21	*Recent volcanism*
22-23	*Future of Magellan*

The capability to automate the recognition of this kind of structure in a full-text document should be useful for improving a variety of computational tasks, e.g., hypertext, text summarization and information retrieval. Toward this end, this paper describes TextTiling, a computational approach to segmenting written expository text into contiguous, non-overlapping discourse units that correspond to the pattern of subtopics in a text.[1]

(Skorochod'ko 1972) has suggested discovering a text's structure by dividing it up into sentences and seeing how much word overlap appears among the sentences. The overlap forms a kind of intra-structure; fully connected graphs might indicate dense discussions of a topic, while long spindly chains of connectivity might indicate a sequential account. The crucial idea is that of defining the structure of a text as a function of the connectivity patterns of the terms that comprise it. This is in contrast with segmenting guided primarily by fine-grained discourse cues such as register change, focus shift, and cue words. From a computational viewpoint, deducing textual topic structure from lexical connectivity alone is appealing, both because it is easy to compute, and also because discourse cues are sometimes misleading with respect to the topic structure (Brown & Yule 1983)(ch. 3).

Following Skorochod'ko, TextTiling attempts to discover coherent, interrelated subdiscussions by analyz-

[1]The use of 'topic' here is meant to signify pieces of text 'about' something, as opposed to the topic/comment distinction found within individual sentences. The intended sense is that described by (Brown & Yule 1983)(p. 69): "In order to divide up a lengthy recording of conversational data into chunks which can be investigated in detail, the analyst is often forced to depend on intuitive notions about where one part of a conversation ends and another begins . . . The notion of 'topic' is clearly an intuitively satisfactory way of describing the unifying principle which makes one stretch of discourse 'about' something and the next stretch 'about' something else, for it is appealed to very frequently in the discourse analysis literature. Yet the basis for the identification of 'topic' is rarely made explicit."

ing the "connectivity" of the terms. The simplest evidence to look for is repetition of words; repetition has been shown to be a coherence enhancer (Tannen 1989), (Walker 1991). Terms that are closely related in meaning also indicate coherence (Halliday & Hasan 1976),(Morris & Hirst 1991).[2] For example, evidence that a dense discussion of "volcanic activity" is taking place in the fourth segment of the example above could be the observation of words related to volcanism, such as *lava* and *eruption*. A third type of coherence evidence is the co-occurrence of multiple simultaneous themes. If the discussion of volcanism mentions its effects on the appearance of a planet's surface, it might be the case that terms related in meaning to "surface", but not semantically similar to "volcanism", occur in the same stretch of text. The fact that several threads of discussion occur contemporaneously should be used as evidence for a coherent subtopic. In other words, often it is the case that a writer discusses the relationship of one thing with respect to another (e.g., volcanic activity and where it takes place, volcanic activity and its effects on crops, or volcanic activity and Roman history) and when the discussion of one topic ends, so does discussion of the others.

Unlike standard discourse analysis approaches, TextTiling breaks the text into simple, contiguous 'tiles' that are meant to reflect only topical loci, and not the interrelations among the topics. Although there are many valid second-order structures that a text can take on – two prominent ones in expository text are hierarchical and sequential (as in a chronological biography) – for the purposes of this task the tiles are considered to be disjoint and no attempt is made to determine how they are related to one another. Higher level structural or functional roles (such as causation, elaboration, etc., found in theories like RST (Mann & Thompson 1987) and comprehensively categorized in (Hovy 1990)) might be determined in subsequent passes.

What follows is a description of the TextTiling algorithm, first using only repetition of terms, and then incorporating terms that are closely related in meaning (and in both cases using theme overlap). This is followed by a discussion of the relationship of this work to that of (Morris & Hirst 1991), and others, and by a comparison of the algorithm's performance human judgement data. The paper concludes with a discussion of how this work will be extended.

[2](Raskin & Weiser 1987), following (Halliday & Hasan 1976), distinguishes between cohesion and coherence; cohesion relations often act to indicate coherence in a passage. They also differentiate between lexical cohesion and grammatical cohesion relations; an example of the latter is pronomial reference. Only lexical cohesion relations are used in this algorithm, although in future, grammatical relations may be added.

2 TextTiling

2.1 The Basic Algorithm

The algorithm is a two step process; first, all pairs of adjacent blocks of text (where blocks are usually 3-5 sentences long) are compared and assigned a similarity value, and then the resulting sequence of similarity values, after being graphed and smoothed, is examined for peaks and valleys. High similarity values, implying that the adjacent blocks cohere well, tend to form peaks, whereas low similarity values, indicating a potential boundary between tiles, create valleys. Figure 1 shows such a graph; the vertical lines indicate where human judges thought the topic boundaries should be placed. Note that a valley is meant to indicate where a discussion of interwoven themes ends, as opposed to monitoring for the ends of discussions of individual themes.

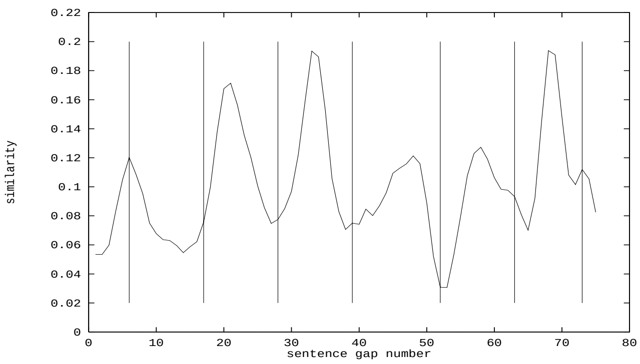

Figure 1: Results of TextTiling a 77-sentence popular science article ("Magellan"). Vertical lines indicate actual topic boundaries as determined by human judges, and the graph indicates computed similarity of adjacent blocks of text. Peaks indicate coherency, and valleys indicate potential breaks between coherent segments.

The one adjustable parameter is the size of the block used for comparison. This value, labeled k, varies slightly from text to text; as a heuristic it is assigned the average paragraph length (in sentences), although the block size that best matches the human judgement data is sometimes one sentence greater or fewer. Actual paragraphs are not used because their lengths can be highly irregular, leading to unbalanced comparisons.

Similarity is measured by putting a twist on tf.idf, a standard information retrieval measurement. The tf.idf value of a term is its frequency within a document divided by its frequency throughout a document collection as a whole (Salton 1988). Terms that are frequent in an individual document but relatively infrequent throughout the corpus are considered to be good distinguishers of the contents of the individual document. In TextTiling, each block of k sentences is treated as a unit unto itself, and the frequency of a term within each block is compared to its frequency in the entire document.[3] This helps bring out a distinction between local and global extent of terms; if a term is discussed frequently but within a localized cluster (thus indicating a cohesive passage), then it will be weighted more heavily than if it appears frequently but scattered evenly throughout the entire document, or infrequently within one block. Thus if adjacent blocks share many terms, and those shared terms are weighted heavily, there is strong evidence that the adjacent blocks cohere with one another.

Similarity between blocks is calculated by a cosine measure: given two text blocks $b1$ and $b2$,

$$cos(b1, b2) = \frac{\sum_{t=1}^{n} w_{t,b1} w_{t,b2}}{\sqrt{\sum_{t=1}^{n} w_{t,b1}^2 \sum_{t=1}^{n} w_{t,b2}^2}}$$

where t ranges over all the terms in the document and $w_{t,b1}$ is the tf.idf weight assigned to term t in block $b1$. Thus if the similarity score between two blocks is high, then not only do the blocks have terms in common, but the terms they have in common are relatively rare with respect to the rest of the document. The evidence in the reverse is not as conclusive: if adjacent blocks have a low similarity measure, this does not necessarily mean they don't cohere; however, in practice this negative evidence is often justified.

If similarity is measured between blocks b and $b+1$, where b spans sentences i through $i+k-1$ and $b+1$ spans $i+k$ to $i+2k-1$, then the measurement's x-axis coordinate would fall between sentences $i+k-1$ and $i+k$. Instead of graphing sentence number $i+k$ on the x-axis, we graph sentence *gap* number $i+k-1$. The straightforward way to use the similarity information is to plot, for each sentence gap, the similarity value measured there. This yields a very jagged graph which is smoothed using an intuitive algorithm. One measurement is made every k sentence gaps. This result is plotted and for the sentence gap numbers where no measurement was made, their values are filled in by piecewise linear interpolation. There are k different starting points, and thus k graphs are plotted along the same axes. We then compute the average similarity value at each sentence gap number, giving equal weight to input from each of the k measurements that cross that point.

[3]The algorithm uses a large "stop list"; i.e., closed class words and other very frequent terms are omitted from the calculation.

As can be seen in Figure 2, this tends to smooth the graph in a desireable way, eliminating sudden quick dips while at the same time preserving the general trends of the graph. This calculation can be shown[4] to be equivalent to performing a discrete convolution of the similarity function with the function $h_k(\cdot)$, where:

$$h_k(i) \equiv \begin{cases} \frac{1}{k^2}(k - |i|), & |i| \leq k-1 \\ 0, & otherwise \end{cases}$$

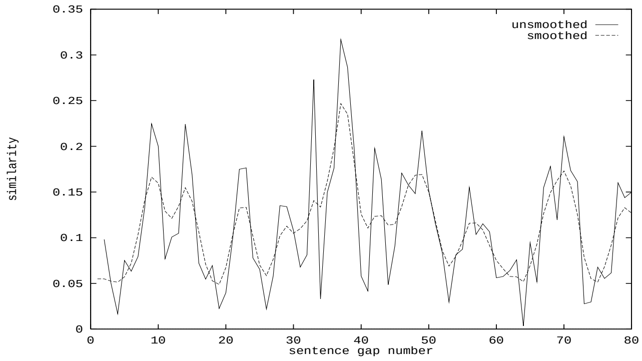

Figure 2: Smoothed and unsmoothed analyses of "Earth" (before median smoothing).

The result is smoothed further with a simple median smoothing algorithm (Rabiner & Schafer 1978), with a window of size three, to eliminate small local minima. Tile boundaries are determined by locating the lowermost portions of valleys in the resulting plot. The actual values of the similarity measures are not taken into account; the relative differences are what are of consequence.

This algorithm can be modified in several ways. Just as irregularity in paragraph length makes collections of sentences a more desireable unit of comparison, using same-sized collections of words, without regard for sentence boundaries, may improve results where overly short or long sentences appear. Additionally, other approaches to smoothing might be more appropriate.

Since the algorithm is to be measured against human judgement data, it is designed to group the sentences at the level of granularity that the judges displayed. Furthermore, from preliminary experiments this seems like a useful level of granularity for other text processing tasks (Hearst & Plaunt 1993). However, this is not necessarily the only level of granularity that is useful (although finer distinctions most likely are not accurately detectable with lexical relatedness measures

[4]I am grateful to Michael Braverman for finding the proof.

alone) and if alternative levels are desired, the algorithm will have to be adjusted.

This was devised as a baseline test, against which to compare the incorporation of thesaural information. As described in the next subsection, addition of thesaural information tends to strengthen the demarcation between the main bulk of the tiles and their boundaries, and seems to slightly improve the placement of the boundaries as well.

2.2 Incorporating Disambiguated Category Information

For some texts, the amount of word repetition is not sufficient to make a confident judgement of the topic boundary. To counter this, we would like to group together terms that are close in meaning, e.g., substituting for the terms *lava, eruption*, and *volcano* one category label, in effect treating these each of these different words as instances of the same word. Unfortunately, this kind of strategy usually encounters problems with polysemy; since many words can take on more than one meaning, they are erroneously linked up with words to which they should not be related.

To help alleviate the polysemy problem, a version of Yarowsky's statistical lexical disambiguation algorithm (Yarowsky 1992) was implemented. Yarowsky defines word senses as the categories listed for a word in *Roget's Thesaurus* (Fourth Edition), where a category is something like TOOLS/MACHINERY. For each category, the algorithm

1) Collects contexts that are representative of the category.
2) Identifies salient words in the collective contexts and determines the weight for each word.
3) Uses the resulting weights to predict the appropriate category for a word occurring in a novel context.

A strong advantage of this algorithm over others is that it does not require a hand-labeled training corpus. A disadvantage is that it does require a good thesaurus. Since Roget's Fourth Edition is not openly available online, a set of 740 categories was derived from the noun hierarchy of WordNet (Miller *et al.* 1990), a hand-built lexical thesaurus.

Sometimes the algorithm works quite well, as shown in the groupings below where each column shows terms taken from localized pieces of text:

hills 95	bubonic 206	motions 430
mounds 95	black 206	quivering 430
highlands 95	epidemics 206	shaking 430
mountain 95	aids 206	perturbations 430

For example, both *mound* and *black* have more than one sense, but the correct one is identified here. An evaluation of the performance of this version of the algorithm is still pending; Yarowsky's results were in the range of 90% correct, but the categories used in this version are inferior, most likely leading to less accurate results.

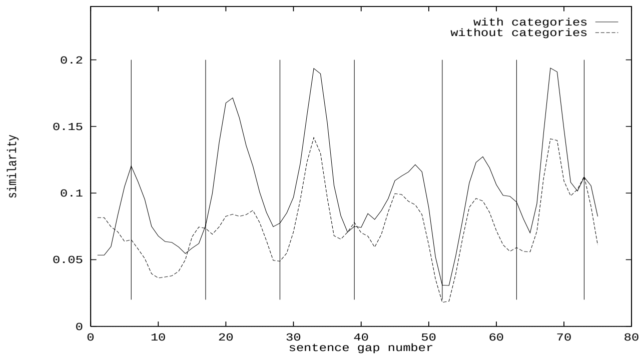

Figure 3: The results of TextTiling the Magellan text with and without category information.

Incorporating category information improved the results in several cases, and did not adversely affect the results in any of the cases. Often the net effect is to make the evidence for the tiles stronger. Figure 3 shows the effects of incorporating category information for the "Magellan" text.

3 Relation to Other Work

3.1 Morris and Hirst

Morris and Hirst's pioneering work on computing discourse structure from lexical relations (Morris & Hirst 1991; Morris 1988) is a precursor to the work reported on here. Morris, influenced by Halliday and Hasan's theory of lexical coherence (Halliday & Hasan 1976), developed an algorithm that finds chains of related terms via a comprehensive thesaurus (again, Roget's Fourth Edition). For example, the words *residential* and *apartment* both index the same thesaural category and can thus be considered to be in a coherence relation with one another. The chains are used to structure texts according to Grosz and Sidner's theory of intentional discourse structure (Grosz & Sidner 1986). This theory is intrinsically hierarchical and fine-grained, and so its goals differ somewhat from those of this paper. Morris provides five short example texts for which she has determined the intentional structure, and states that the lexical chains generated by her algorithm provide a good indication of the segment boundaries that Grosz and Sidner's theory assumes.

Morris' algorithm was executed by hand because, as mentioned above, the thesaurus is not generally available. However, Project Gutenberg has donated an online copy of Roget's 1911 thesaurus which, although smaller and less structured than the thesaurus used by Morris, can be used for an implementation of the algorithm. Aside from the fact that using such a thesaurus lowers the quality of the connections found among terms, an implementation of the Morris algorithm using it brought forth some unforeseen difficulties. First, although ambiguous chain links were rare in Morris's texts, the texts analyzed here had many ambiguous links, even when connections were restricted to being made between terms in the same category. Second, when looking at somewhat longer texts, the chains' extent tended to overlap so much that it wasn't possible to determine structure with them, i.e., many chains would end at a particular paragraph while at the same time many other chains would extend past it. With longer texts, the chains did not neatly delineate the intentional structure.

To get around this difficulty, the Morris algorithm can be extended to create graphs similar to those shown in the previous section, by plotting the number of active chains against paragraph or sentence numbers. In future experimentation will be done comparing the results of modifying Morris' algorithm to the results presented here.

3.2 Other Approaches

The text generation work of (Mooney *et al.* 1990) is related to that presented here, in that they assert that the high level structure of extended explanations is determined by processes separate from those which organize text at lower levels. They present a scheme for text generation that is centered around the notion of Basic Blocks: multi-paragraph units of text, each of which consists of (1) an organizational focus such as a person or a location, and (2) a set of concepts related to that focus. Thus their scheme emphasizes the importance of organizing the high level structure of a text according to its topical content, and afterwards incorporating the necessary relatedness information, as reflected in discourse cues, in a finer-grained pass.

The approach advocated here is similar in that it does not make *a priori* assumptions about how the text is structured. Many texts are best characterized as having a sequential structure; a hierarchical assumption could be misleading ((Sibun 1992) discusses this point in the context of generating spoken text). Separating tiling from a subsequent functional role analysis stage might make discerning these roles easier.

(Givon 1983) presents a quantitative approach to discourse analysis which focuses on identification of syntactic clues, such as pronoun anaphora and cleft/focus constructions, to discover the continuity patterns of the discourse. Many other researchers have studied the effects of trails of pronominal reference and other discourse cues, although usually not in such quantified terms.

Significant work has been done in the segmentation of speech signals, e.g. (Hirschberg & Pierrehumbert 1986). (Glass & Zue 1987) suggest an algorithm, originally intended for segmentation of speech into phonemic units, in which adjacent units are grouped hierarchically, forming a dendogram. This is a modification of scale space filtering (Witkin 1984) and is also similar to a technique suggested by (Rotondo 1984), and can represent potential segments at many different levels of granularity simultaneously.

4 Evaluation

This algorithm is most useful for texts that lack copious orthographic structure, as opposed to technical texts which tend to be highly structured by the author. Lacking a standard test set, [5] the algorithm was evaluated on two data sources: three expository articles (length 77 to 160 sentences), and the five short general interest articles from (Morris 1988) (length 23 to 44 sentences).

Human judgement data was gathered via an informal study in which several readers were given each text, with all section markers and diagrams removed, and asked to draw a line beneath a sentence wherever they perceived a change in topic. After a pilot study in which all but one reader chose to place boundaries between paragraphs, it was decided to require the boundaries to lie between paragraphs, in order to facilitate evaluation of the algorithm. All articles were in English, and they consisted of two general interest science articles, (nicknamed "Magellan" and "Earth"),

[5] Experiments on how readers determine various aspects of text structure have been done (Britton *et al.* 1986), (Rotondo 1984), (Vipond 1980), (Mandler 1987). However, the texts used in these studies are too short for the purposes of this work.

and one environmental impact study ("Sequoia").

As to be expected, the readers' judgements do not agree with one another completely. In order to use the judgements to evaluate the algorithm's performance, the dominant trends must be captured. A "consensus" value is computed: the majority opinion for each potential boundary point is designated to be the correct judgement for that position. To evaluate the agreement quality of the data, how often each reader's judgements match the consensus is computed. The agreement on average should be 90 per cent or higher per reader in order to assure that the consensus judgement is representative.[6] In the test set, the agreement for two of the documents was greater than 80% but less than 90%, but their judgements are presented here pending more testing.

Figure 4 demonstrates the results of the algorithm on the three test texts. For Magellan and Earth most of the correspondences were correct although there are many off-by-one-or-two sentences errors. This is to be expected as the readers were required to place boundaries only at paragraph breaks, whereas the algorithm can place a boundary at any sentence gap. If the algorithm were to choose the nearest paragraph boundary instead, these errors would be eliminated; it may also benefit from the use of discourse cue heuristics. The Sequoia text, especially the first 70 sentences, was more difficult to divide correctly, since it was an informative report that referred to diagrams and tables (which were removed), thus making it less free-flowing. In most cases, when the algorithm missed one boundary, it was correct on the following boundary, and after adding a gratuitous boundary, it still usually managed to recover and place the following boundary correctly.

(Bachenko *et al.* 1992), in evaluating a system for placing prosodic phrase boundaries in spoken text, point out that one can't use as a measure of success how many times a boundary is not marked. Since there are relatively few boundaries as compared with non-boundaries, a strategy that places no boundaries at all can be around 80% correct by this measure. Therefore, what should be measured is the number of *errors* the algorithm makes. Using this criterion, and assuming that being off by one or two sentences is not considered an error, we obtain the following summary results:

text	deleted boundary	inserted boundary	off by $>$ 2 sentences
Magellan	1	1	1
Earth	0	1	2
Sequoia	1	2	2

[6] I am grateful to Bill Gale for this suggestion.

In a few rare cases, the algorithm's results differ strongly from the reader judgements; an example of this can be seen in the first segment of Figure 3 from sentence gaps 1 - 17. This disagreement after sentence 6 seems to be caused by a strong local discourse cue; a quotation ending a paragraph is followed by a reintroduction of a previously-discussed entity, leading the reader to believe that a new topic is being introduced ((Stark 1988) calls this phenomenon over-reference). Further investigation into the interaction of the global and local cues is warranted.

Figure 5 demonstrates the relationship between Morris's results with the chaining algorithm and the intentional structure for two of the texts she analyzed (recall that the intentional structure is hierarchical, so different chains are meant to correspond to different levels of the analysis). Also shown are the results of TextTiling these documents. Note that quite a few sentences were very short, only three or four words long.

As mentioned above, this algorithm is not designed to recognize the most fine-grained segments of the intentional structure; rather it should find a structure midway up the hierarchical structure (thus making evaluation more difficult). The tiles almost always lined up with one of the boundaries indicated by the intentional structure; for the runs without category information there were only four instances of boundaries being off by more than one sentence, and for the runs with category information, there were only five (and these were only off by two sentences). However, the layering pattern of the intentional structure was not always respected. For example, in Figure 5 under Morris Text 1, tiling found a grouping from sentences 23-30, which subsumed two of the intentional structure's segments exactly; however these segments were judged to be part of two different supersegments. Figure 6 shows the results for tiling the remaining three texts.

5 Discussion

This paper has described an algorithm for the segmentation of expository texts into discourse units that are meant to reflect the topic flow of the text. The algorithm is fully implemented and can be run without benefit of inference mechanisms or a large knowledge base. The structure it obtains is coarse-grained but generally reflects human judgement data.

To be more useful the tiles should be labeled according to what subtopic discussions they contain. Experiments are underway in which using the categorizaton algorithm to classify the terms related to those with the highest tf.idf weights in each tile. This would al-low, for example, the classification of the fourth tile in the example in Section 1 with the label "geological activity," although multiple labels should be associated with a tile when more than one theme runs through it. TextTiling lengthy documents also opens the door to interesting new information retrieval paradigms, as discussed in detail in (Hearst & Plaunt 1993). Also, techniques for text summarization and for determining good index terms, as explored in (Evans *et al.* 1991), could be improved by providing information about the roles individual terms play with respect to what subtopic tile they appear in.

Acknowledgments

I would like to thank Anne Fontaine for inspiration and for running the reader experiments, Chris Plaunt for technical assistance, and Michael Braverman, Ken Church, Bill Gale, David Yarowsky, and John Maxwell for discussion that improved the quality of this work. I would also like to thank Narciso Jaramillo, Penni Sibun, and Robert Wilensky for helpful suggestions to improve an earlier draft of this paper. This research was sponsored in part by the University of California and Digital Equipment Corporation under Digital's flagship research project Sequoia 2000: Large Capacity Object Servers to Support Global Change Research.

References

Bachenko, J., J. Daugherty, & E. Fitzpatrick (1992). A parser for real-time speech synthesis of conversational texts. In *Proceedings of the ACL Conference on Applied Natural Language Processing*, Trento, Italy.

Britton, B. K., K. D. Muth, & S. M. Glynn (1986). Effects of text organization on memory; test of a cognitive effort hypothesis with limited exposure time. *Discourse Processes*, 9:475–487.

Brown, G. & G. Yule (1983). *Discourse Analysis*. Cambridge Textbooks in Linguistics Series. Cambridge University Press.

Evans, D. A., K. Ginther-Webster, M. Hart, R. G. Lefferts, & I. A. Monarch (1991). Automatic indexing using selective nlp and first-order thesauri. In *Proceedings of the RIAO*, volume 2, pages 624–643.

Givon, T., editor (1983). *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. John Benjamins Publishing Company, Philadelphia.

Glass, J. R. & V. W. Zue (1987). Acoustic segmentation and classification. Technical Report SAIC-87/1644, DARPA.

Grosz, B. J. & C. L. Sidner (1986). Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):172–204.

Halliday, M. A. K. & R. Hasan (1976). *Cohesion in English*. Longman, London.

Hearst, M. A. & C. Plaunt (1993). Subtopic structuring for full-length document access. In *Proceedings of SIGIR*. to appear.

Hirschberg, J. & J. Pierrehumbert (1986). The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 136–144.

Hovy, E. (1990). Parsimonious and profligate approaches to the question of discourse structure relations. In *5th ACL Workshop on Natural Language Generation*, Dawson, Pennsylvania.

Mandler, J. M. (1987). On the psychological reality of story structure. *Discourse Processes*, 10(1):1–29.

Mann, W. C. & S. A. Thompson (1987). Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS 87-190, ISI.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, & K. J. Miller (1990). Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

Mooney, D. J., M. S. Carberry, & K. F. McCoy (1990). The generation of high-level structure for extended explanations. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, volume 2, pages 276–281, Helsinki.

Morris, J. (1988). Lexical cohesion, the thesaurus, and the structure of text. Technical Report CSRI-219, Computer Systems Research Institute, University of Toronto.

Morris, J. & G. Hirst (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Rabiner, L. R. & R. W. Schafer (1978). *Digital Processing of Speech Signals*. Prentice-Hall, Inc, New Jersey.

Raskin, V. & I. Weiser (1987). *Language and Writing: Applications of Linguistics to Rhetoric and Composition*. ABLEX Publishing Corporation, Norwood, New Jersey.

Rotondo, J. A. (1984). Clustering analysis of subjective partitions of text. *Discourse Processes*, 7:69–88.

Salton, G. (1988). *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading, MA.

Sibun, P. (1992). Generating text without trees. *Computational Intelligence: Special Issue on Natural Language Generation*, 8(1):102–122.

Skorochod'ko, E. (1972). Adaptive method of automatic abstracting and indexing. In C. Freiman, editor, *Information Processing 71: Proceedings of the IFIP Congress 71*, pages 1179–1182. North-Holland Publishing Company.

Stark, H. (1988). What do paragraph markers do? *Discourse Processes*, 11(3):275–304.

Tannen, D. (1989). *Talking Voices: Repetition, dialogue, and imagery in conversational discourse*. Studies in Interactional Sociolinguistics 6. Cambridge University Press.

Vipond, D. (1980). Micro- and macroprocesses in text comprehension. *Journal of Verbal Learning and Verbal Behavior*, 19:276–296.

Walker, M. (1991). Redundancy in collaborative dialogue. In J. Hirschberg, D. Litman, K. McCoy, & C. Sidner, editors, *AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation*, Pacific Grove, CA.

Witkin, A. P. (1984). Scale space filtering: A new approach to multi-scale description. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Yarowsky, D. (1992). Word sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 454–460, Nantes, France.

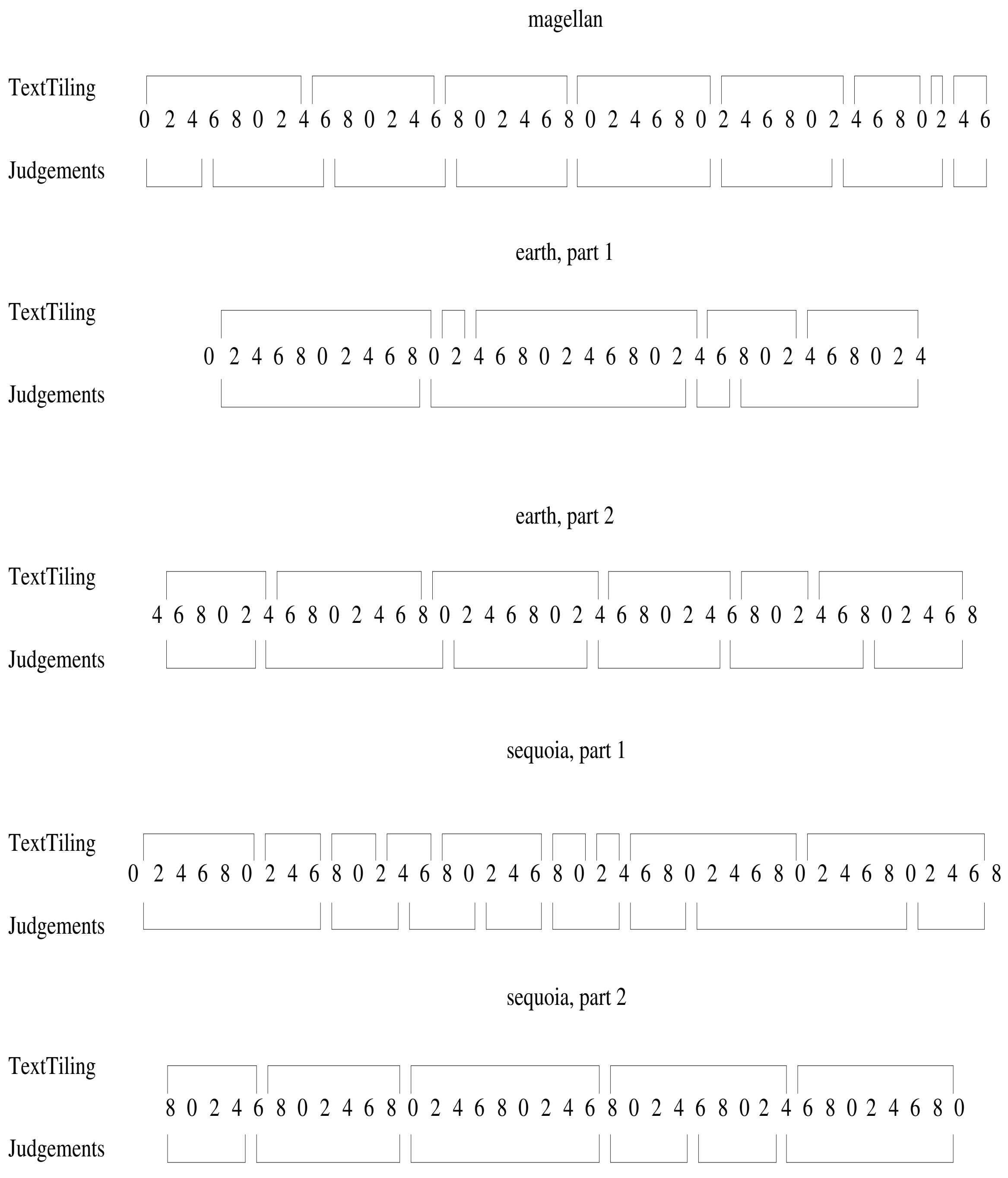

Figure 4: Results of TextTiling on two science magazine articles and one environmental impact report. The numbers represent sentence gap numbers, below the numbers appear consensus judgements, and above them are the results of tiling with categories.

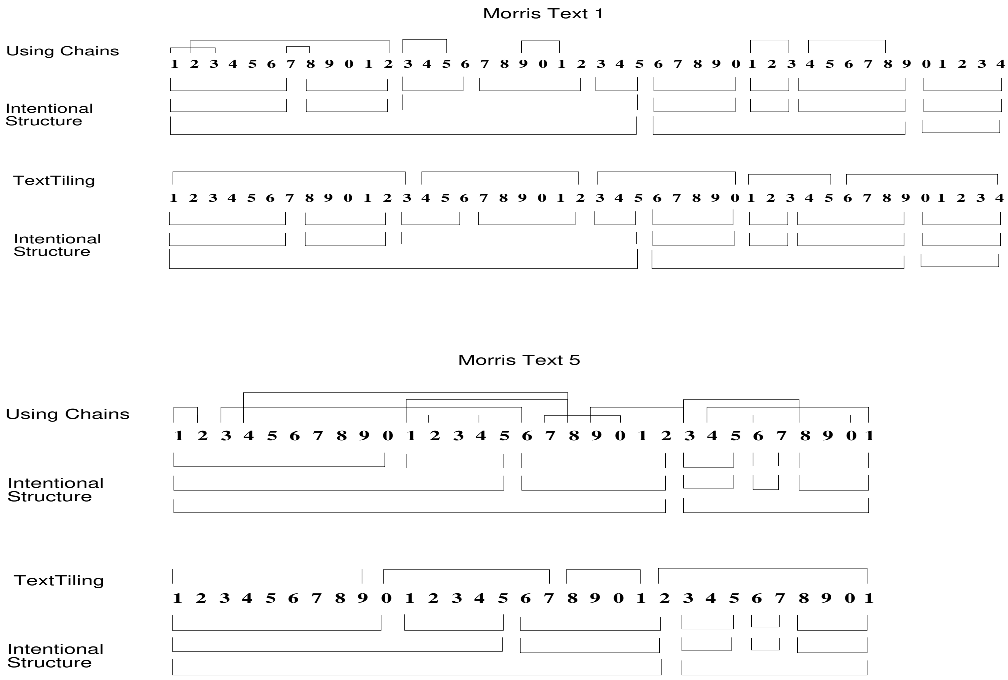

Figure 5: Chaining's and TextTiling's results on two texts from Morris 1988.

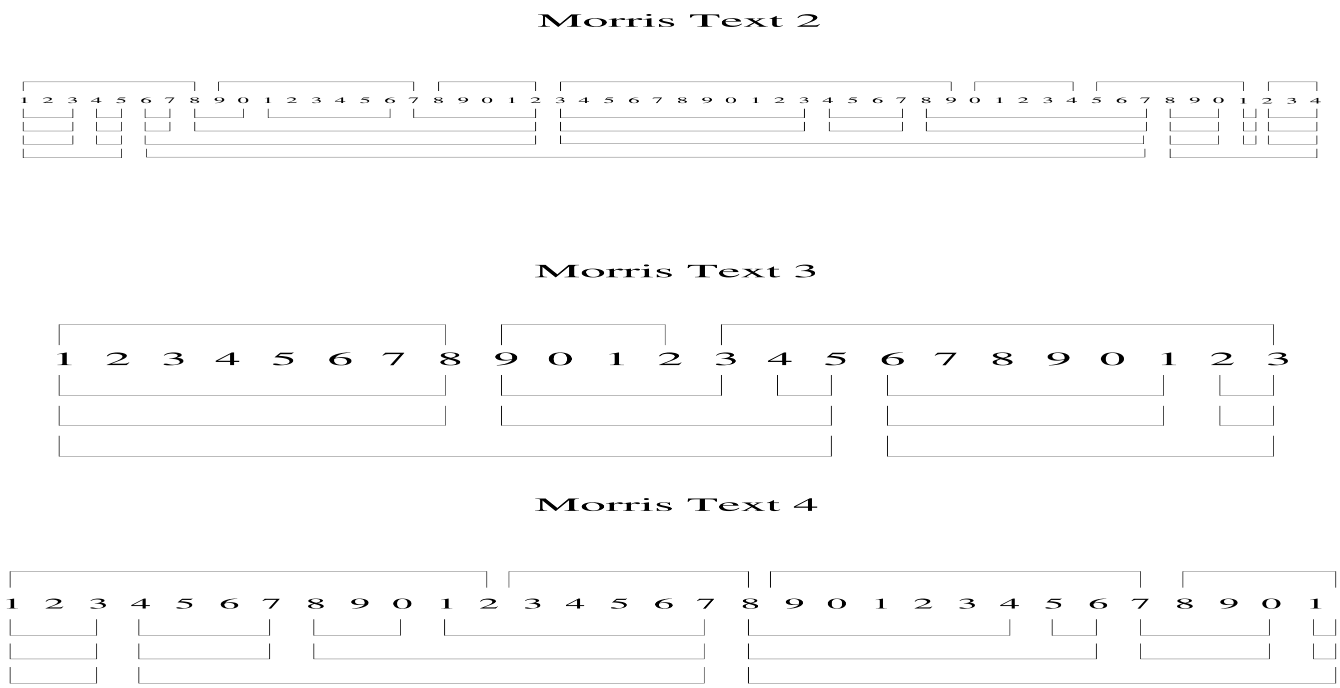

Figure 6: Results of TextTiling on the remaining three texts from Morris 1988. The numbers represent sentence gap numbers, below the numbers appear the intentional structure and them are the results of tiling with categories.