EXPLOITING TEXT STRUCTURE FOR TOPIC IDENTIFICATION

Tadashi Nomoto
Advanced Research Laboratory
Hitachi Ltd.
2520 Hatoyama Saitama, 350-03 Japan
nomoto@harl.hitachi.co.jp

Yuji Matsumoto
Nara Institute of Science and Technology
8916-5 Takayama Ikoma Nara, 630-01 Japan
matsu@is.aist-nara.ac.jp

Summary

The paper demonstrates how information on text structure can be used to improve the performance on the identification of topical words in texts, which is based on a probabilistic model of text categorization. We use texts which are not explicitly structured. A text structure is identified by measuring the similarity between segments comprising the text and its title. It is shown that a text structure thus identified gives a good clue to finding out parts of the text most relevant to its content. The significance of exploiting information on the structure for topic identification is demonstrated by a set of experiments conducted on the 19Mb of Japanese newspaper articles. The paper also brings concepts from the rhetorical structure theory (RST) to the statistical analysis of a text structure. Finally, it is shown that information on text structure is more effective for large documents than for small documents.

1. INTRODUCTION

Topic identification concerns a problem of predicting terms in text which indicate its subject or theme. In the past, the problem has been addressed mostly by computational linguists in relation to issues like coreference (Hobbs, 1978), anaphora resolution (Grosz and Sidner, 1986; Lappin and Leass, 1994), or discourse center (Joshi and Weinstein, 1981; Walker et al., 1994). In information retrieval, predicting important terms in document is crucial for an effective retrieval of relevant documents(Salton et al., 1993), though they do not necessarily correspond to the subject or the theme. Predicting important terms involves numerical weighting of terms in document. Terms with top weights are judged important and representative of document.

A spin-off of information retrieval, known as text categorization, shares a similar research interest. Text categorization concerns associating documents with their classification terms or categories (Lewis, 1992). Since in text categorization, categories are determined beforehand in such a way as to meet the user's specific tastes or needs, they may not serve as a topic or a theme in that they need not have a semantic relevance to the contents of documents.

Technically, however, it is straightforward to move from text categorization to topic identification, provided that we are able to somehow isolate themes in texts and use them as categories to be assigned to texts. But the problem with using text categorization for topic identification, is that categories are arbitrarily given by humans, with no regard for documents that are to be classified. There is thus always a danger of misrepresenting documents. One possible way out is to choose categories not from outside of the documents but from within. The feasibility of the idea is explored in the paper.

The use of text structure in information retrieval was motivated by the need for dealing with large documents, whose breadth of vocabulary may easily mislead the retrieval system into making a wrong

judgement about their relevancy to the query. Indeed, a new area of research known as *passage retrieval* has emerged to explore methods for using information from various levels of a document's structure, e.g. sentences, sections, paragraphs, and other semantically or rhetorically motivated textual units. Wilkinson (1994) describes weighting methods that combine the similarity measure with various textual categories like *abstract*, *purpose* and *supplementary*, etc. Salton (1993) compares the full-text retrieval with the passage retrieval based on sections and paragraphs, and reports that the latter form of retrieval led to an increased effectiveness. Allan (1995) examines the usefulness of passage for relevance feedback, which concerns deriving or *learning* useful query terms from retrieved documents. Hearst (1993) is an interesting attempt to enhance the retrieval performance by using what they call a *text tile*, a discourse unit determined on the basis of the subject or content of the text. Callan (1994) proposes a hybrid approach of using both passage and document.

Section 2 introduces the idea of bringing an IR technique to the topic identification task. Section 3 discusses a problem that the proposed method shows poor performance on large documents. Section 4 is a response to the problem: we propose the use of information on text structure to reduce irrelevancy in the document and increase effectiveness. In Section 5, we conduct a set of experiments to determine whether the use of text structure has a positive effect on the performance.

2. APPROACH TO TOPIC IDENTIFICATION

Topic identification concerns the problem of identifying a topic of text with no information being given on the text's title or keywords whatsoever. In this paper, we propose a probabilistic approach to the problem. We start with the definition of topic. Here we simply assume that a title term, or a noun that appears in the title, counts as a topic for the text. The reason is that it gives a clear cut, if not simple-minded, definition of topic and lends itself easily to a statistical treatment. Thus what is involved in identifying a topic is a task of locating an instance of a title term. One might ask how we might locate an instance of topic without any information on the title or keywords. We take a probabilistic approach where you estimate the likelihood that a noun appears as a title term and choose among the best scoring nouns. Although it is quite possible to expand the notion of 'topic' to include nouns semantically related to title terms, the possibility is not explored here.

The method we use for topic identification is basically one that is standardly used in research on text categorization, which aims at finding an effective way of classifying documents with predefined categories (Lewis, 1992). Roughly speaking, text categorization proceeds in two steps; first, for each of the given categories, estimate the likelihood that it is a correct category of a document, and second, decide whether to assign the category to the document based on the estimate; the rule is to use a suitable cutoff point to determine a choice.

Now to adapt text categorization for use in topic identification requires a slight change in the former. This is because in topic identification, we want to assign documents to nouns which occur in the document, rather than to categories given *a priori*.

As mentioned above, topic identification is a two-part process; estimating and assigning. Let us explain a bit more about how the estimating works. We will talk about the assigning part later in the paper. We begin with some terminology. We call a word or an expression which classifies the text its *potential topic* and those that appear in the title *actual topics*. Let $\mathcal{L}(c \mid d)$ denote the likelihood that a document d is assigned to, or classified by a category c, $W(d)$ a set of words or expressions comprising a text d, and $S(d)$ a set of potential topics for d. Then the process of estimating consists in computing the likelihood value $\mathcal{L}(c \mid d)$ for each c in $S(d)$ such that $S(d) \subseteq W(d)$. Later, we will be concerned with whether a particular choice of the set $S(d)$ will in any way affect the performance on topic identification.

Following Iwayama *et al* (1994) and Fuhr (1989), we define the likelihood function by:

$$\mathcal{L}(c \mid d) = \sum_t P(c \mid t)P(t \mid d) \tag{1}$$

which is meant to be a relativization of the relationship between c and d to some index t (Fuhr, 1989); The index t could be anything from simple units such as a word, a bigram, or a trigram to more complex forms such as a phrase or a sentence. Put simply, the equation above says that the greater the number of indices associated with both c and d is, the more likely d is to predict c. In text categorization, a set of indices is said to *represent* a text. Assume that every index t will be assigned to some category. Then by Bayes' theorem, equation 1 can be transformed into:

$$\mathcal{L}(c \mid d) = P(c) \sum_{t \in S(d)} \frac{P(t \mid c)P(t \mid d)}{P(t)} \tag{2}$$

We estimate the component probabilities by:

$$\begin{aligned} P(c) &= doc\_f(D_c)/doc\_f(D) \\ P(t \mid c) &= F_c(t)\,/token\_f(D_c) \\ P(t \mid d) &= F_d(t)\,/token\_f(d) \\ P(t) &= F_D(t)\,/token\_f(D) \end{aligned}$$

D is a collection of texts (news articles) found in the training corpus. D_c is a collection of texts in D whose title contains a term c. $doc\_f(D)$ is the count of texts in D. Similarly, $doc\_f(D_c)$ refers to the count of texts which have a term c in the title. $F_c(t)$ is the frequency of an index t in D_c. $token\_f(D_c)$ is the total count of word tokens in D_c, and similarly for $token\_f(d)$ and $token\_f(D)$. $F_d(t)$ is the frequency of an index t in d. Finally, $F_D(t)$ is the frequency of an index t in D.

(Futsu) (gin) -ga (Kiev)-ni (chuuzai) (in) (jimusho).
french bank SBJ Kiev at resident staff office

(Futsu) (gin) (Oote) -no (Societé) (General) -wa (15-nichi), (U*) (kura*) (ina*) -no (shuto)
french bank big-name which is Societé General as for on 15th U- kra- ine whose capital
(Kiev) -ni (chuzai) (in) (jimusho)-wo (kaisetsu)-ϕ suruto (happyo)-ϕ shita. Sude-ni (Kiev) (shi)
Kiev at resident staff office OBJ open plan disclose did Already Kiev city
(tookyoku)-no (kyoka) -mo eta to-iwu.
authority whose permission as well obtained sources say

MAJOR FRENCH BANK OPENS OFFICE IN KIEV
Societé General, a major french bank, disclosed on the 15th a plan to open a resident office in Kiev, capital of Ukraine. The bank has already obtained a permission from the city authority, sources say.

Figure 1: A sample news story

Shown in Fig. 1 is a sample news article from the corpus we used. Nouns marked with parentheses are automatically extracted by using a Japanese tokenizer program called JUMAN (Matsumoto et al., 1994). The star indicates that the relevant items are wrongly tokenized. What we have in Fig. 2 is sets

$T(d)$ = {French, bank, Kiev, resident, staff, office}

$S(d)$ = {French, bank, big-name, Societé, General, on 15th, U-, kra-, ine, capital, Kiev, resident, staff, office, open, disclose, city, authority, permission}

$W(d)$ = {French, bank, big-name, Societé, General, on 15th, U-, kra-, ine, capital, Kiev, resident, staff, office, open, disclose, city, authority, permission}

Figure 2: Potential and actual topics.

of actual and potential topics plus a set of indices that are used to represent the text. Notice that simple nouns are used as indices here.

3. PROBLEM

This section briefly discusses some of the problems with the present approach to topic identification (Nomoto, 1995). A most serious problem is that as the length of a story increases, the model's performance quickly degrades. A cause of the problem appears to be an assumption we made that $S(d)$ is equal to $W(d)$, that is, that every noun in the text counts as a potential topic of text. As a consequence, an increase in text length results in a larger set of potential topics.

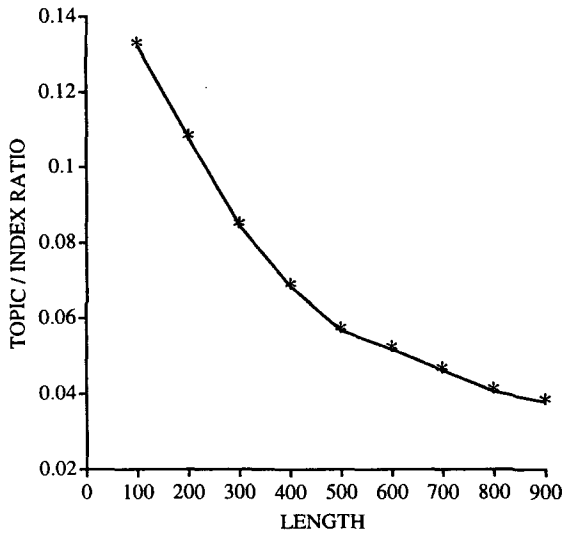

Figure 3: The proportion of words in text against words in title.

Fig. 8 shows how the proportion of actual topics to indices (that is, $\frac{T(d)}{W(d)}$) changes with the increase in text length. (Information is based on the news articles in the test set we used in the experiments later. '100' denotes a set of news articles between 100 to 200 character long, '200' means news articles between 200 to 300 character long, and similarly for others. The vertical dimension represents the proportion of words in text against words in title.) Since the title length stays rather constant over the test corpus, the possibility that an actual topic is identified by chance would be higher for short texts than for lengthy ones; we find 13% of indices to be actual at 100, while the rate goes down to 3% at 900. In the

following, we will investigate ways of reducing the size of $S(d)$ without hurting the performance of topic identification.

4. METHOD

Our approach to the reduction problem above is to use a text structure to demarcate between relevant and irrelevant parts of text. Since newspaper articles in general do not have formal structure indicators, we use some similarity function to discover the structure of text. One such function is proposed by Hearst (1994), where a text structure is determined by measuring the similarity between adjacent blocks of text. Rather than to use a measure for within document similarity, the present approach chose to use a similarity measure between the text and its title to determine the structure of text. Behind this is an assumption that parts of the text most similar to its title would best represent its content. Thus it is expected that discarding parts dissimilar to the title reduces irrelevancy in text, contributing to an improvement on the performance.

Let $\mathcal{F}(d)$ be a complete set of non-overlapping text segments comprising a document d. Let $d \in \mathcal{F}(d)$ and h be a title associated with the document d. Then the similarity between a headline and text segment is given by the usual $tf \cdot idf$ measurement (Wilkinson, 1994):

$$SIM(h, d) = \sum_{i=0}^{N} ntf_{td} \cdot idf_t$$

N is the number of words that appear in h. ntf_{td} is a normalized term frequency of t in d, which is given by:

$$ntf_{td} = \frac{tf_{td}}{max\_tf_d}$$

where tf_{td} denotes a frequency of the term t in d and max_tf_d the frequency of the most frequent term in d.

$$idf_t = \frac{\log \frac{df}{tdf}}{\log df}$$

tdf is the number of segments which have an occurrence of t. df is the total number of segments in d. $\log df$ is a normalization factor. Further, any repetitions of words are removed from the title. The length of a segment d is fixed at 10 words in the experiments. The idea of idf or the inverse document frequency is to give more points to words which have a localized distribution, that is, those that appear only in some of the documents and not others.

Fig. 4 shows similarity graphs for news articles published from January, 1992 through April, 1992 (Nihon-Keizai-Shimbun-Sha, 1992). Each graph corresponds to articles of a particular length; the one marked with '100-200' means that it is for articles from 100 to 200 character long. Since the test sets we used in the later experiments of topic identification ranges from 100 to 1000 characters in length, the title/block similarity is measured only for the relevant sizes. 't' denotes the number of articles considered.

We divided each news article into a set of 10-word segments ordered according to their appearance. For each of the 10-word segments contained in the text, we measured its similarity to the title. The horizontal dimension represents a position at which a segment appears. A segment at $x = 5$, for instance, spans the 40th to the 49th word of text. The vertical dimension gives the probability that a segment at a particular position is chosen as most similar to the title. It is clear from Fig. 4 that the initial portion of text is more likely to be chosen as most similar to the title than other parts of text; the later a segment appears in text, the less chance it has of being selected as most similar to the title.

In terms of the rhetorical structure theory (RST)(Fox, 1987), the results could be interpreted as indicating a stylistic norm particular to the newspaper domain, i.e., that the main claim is presented in

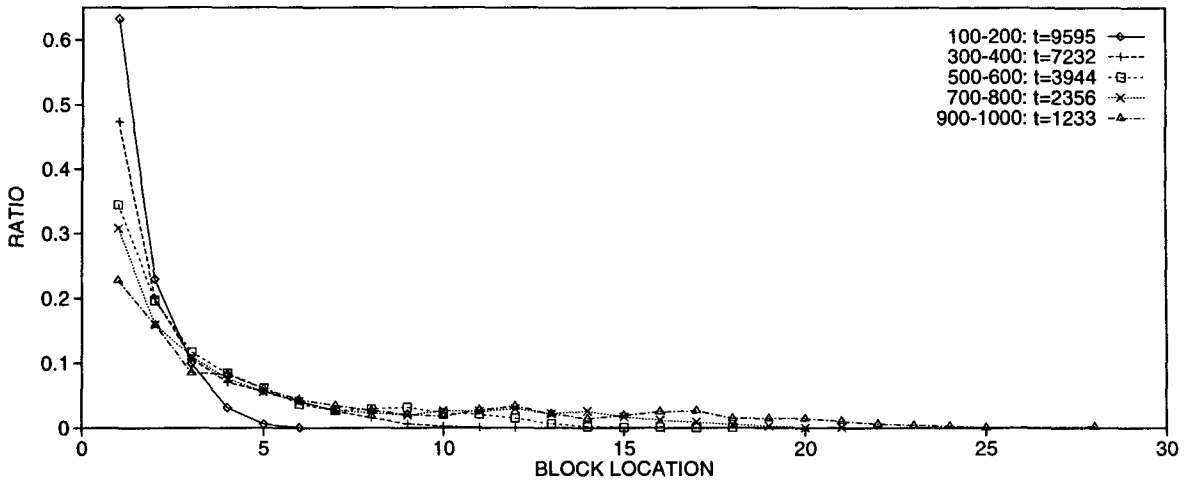

Figure 4: Text Structure as determined by the title/block similarity.

the beginning of the article, followed by supplemental materials. Consider, for instance, a sample news article given in Fig. 1. Since it does not affect points of the discussion, the English translation is used here for the convenience. The article consists of a title (1) plus two sentences(2,3).

(1) *Major French Bank opens office in Kiev.*

(2) *Societé General, a major French bank, disclosed on the 15th a plan to open a resident office in Kiev, capital of Ukraine.*

(3) *The bank has already obtained a permission from the city authority, sources say.*

Considering the title (1), it is fair to say that the first sentence constitutes a main news of the article, while the second is its elaboration, providing supplementary details about the news.

In RST, the article would be analyzed as having an Issue structure, which consists of one *nucleus* and some *adjuncts*. A nucleus is a set of clauses that presents a main claim of the text, something that makes the text newsworthy, while an adjunct supplements the main claim with some background or ancillary information. Fig. 5 shows a diagrammatic representation of the article along the lines of Fox (1987).

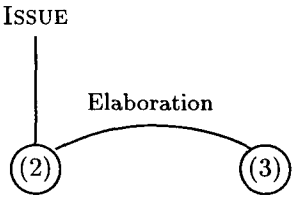

Figure 5: A rhetorical structure for the sample news story in Fig. 1

A node labelled with '(2)' represents a nucleus of the text and a node labelled with '(3)' an adjunct to the nucleus. The arc going from '(2)' to '(3)' is labelled with the type of relationship that holds between the relevant nodes.[1]

[1] 'Elaboration' is one of the three types that occur with an Issue structure; the other two are 'Evidence' and 'Background' (Fox, 1987).

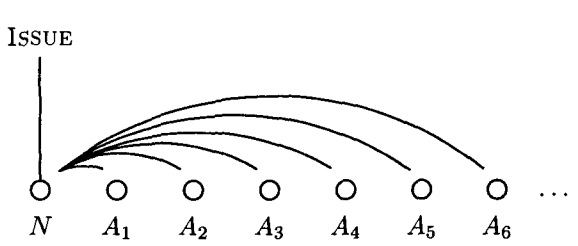

Figure 6: A rhetorical structure for the newswire texts.

In fact the results of the experiments shown in Fig. 4 give some ground for believing that texts from the newspaper domain, as a rule, take a rhetorical structure similar to Fig. 5 such as one in Fig. 6, where the nucleus appears at the beginning of the text, followed by any number of supplementary adjuncts.

In the light of this, we have conducted a series of experiments to determine whether discarding rear portions of text affects the performance of topic identification.

5. EXPERIMENTS

We have conducted a set of experiments to see how a full-text and "discard" model compare in terms of the performance on the topic identification task. Our experiments used the total of 43,253 full-text news articles from *Nihon Keizai Shimbun*, a Japanese business daily (Nihon-Keizai-Shimbun-Sha, 1992). All of the articles appeared in the first half of the year 1992. Of these, 40,553 articles, which appeared on May 31, 1992 and earlier, were used for training and the remaining 2,700 articles, which appeared on June 1, 1992 or later, were used for testing.

A training set and a test set were obtained by extracting nouns from the newspaper corpus, which involves as a sub-step tokenizing each article into a set of words. The procedure was carried out with the tokenizer program JUMAN. The resultant training set contained some 2.5 million words excluding stop words.

The test set was then divided into nine subsets of news articles according to the length. Each subset contained 300 articles. In Table 1, the test set 1, for instance, consists of articles, each of which which contains from 100 to 200 characters. The test set 2, on the other hand, consists of larger articles, which are between 200 and 300 character long.

test set	length (in char.)	num. of doc.
1	100-200	300
2	200-300	300
3	300-400	300
4	400-500	300
5	500-600	300
6	600-700	300
7	700-800	300
8	800-900	300
9	900-1000	300

Table 1: Test Sets

In the experiments, we were interested in finding out the effectiveness of a segment model which considers a starting block of the article and ignores everything else. Here we tried two approaches; one is based on a fixed-length segment and the other on a proportional-length segment. The fixed-length approach uses the first i words of the text, i being constant across texts, whereas the proportional-length approach uses the first j% of words contained in the text, so that the actual length of segment is

Table 2: Fixed-Length Model (FLM): a summary

$l \setminus i$	10	20	30	40	50	60	70	80	90	100
100-200	.38(−.10)	.42(−.00)	.41	.41(−.02)	.41	.41(−.02)	.41	.41(−.02)	.41	.41(−.02)
200-300	.35(+.03)	.39(+.15)	.38	.36(+.06)	.35	.34(+.00)	.34	.34(+.00)	.34	.34(+.00)
300-400	.34(+.06)	.38(+.19)	.37	.36(+.12)	.35	.34(+.06)	.33	.33(+.03)	.33	.33(+.03)
400-500	*.34(.30)	.37(+.19)	.37	.36(+.16)	.35	.34(+.10)	.33	.33(+.06)	.32	.32(+.03)
500-600	*.34(.28)	.37(+.23)	.36	.36(+.20)	.36	.35(+.17)	.33	.34(+.13)	.33	.33(+.10)
600-700	*.36(.29)	.37(+.32)	.37	.36(+.29)	.35	.33(+.18)	.33	.32(+.14)	.31	.31(+.11)
700-800	*.32(.25)	.34(+.31)	.34	.34(+.31)	.33	.32(+.23)	.31	.30(+.15)	.30	.29(+.12)
800-900	*.32(.25)	.34(+.31)	.34	.33(+.27)	.33	.32(+.23)	.32	.31(+.19)	.30	.30(+.15)
900-1000	*.29(.23)	.32(+.23)	.31	.30(+.15)	.30	.29(+.12)	.29	.29(+.12)	.28	.28(+.08)

Table 3: Proportional-Length Model (PLM): a summary

$l \setminus j$	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
100-200	*.40(.22)	.35(−.17)	.39	.41(−.02)	.42	.42(−.00)	.41	.41(−.02)	.41	.42
200-300	*.38(.24)	.37(+.09)	.39	.38(+.12)	.36	.36(+.06)	.36	.35(+.03)	.35	.34
300-400	*.36(.30)	.38(+.19)	.38	.36(+.12)	.36	.35(+.09)	.34	.34(+.06)	.33	.32
400-500	*.34(.33)	.38(+.23)	.37	.35(+.13)	.34	.34(+.10)	.33	.33(+.06)	.32	.31
500-600	.34(+.13)	.36(+.20)	.36	.35(+.17)	.34	.33(+.10)	.33	.32(+.07)	.32	.30
600-700	.35(+.25)	.36(+.29)	.35	.33(+.18)	.32	.31(+.11)	.30	.29(+.04)	.29	.28
700-800	.36(+.38)	.34(+.31)	.32	.32(+.23)	.29	.28(+.08)	.28	.27(+.04)	.26	.26
800-900	.35(+.35)	.33(+.27)	.32	.31(+.19)	.30	.29(+.12)	.28	.27(+.04)	.27	.26
900-1000	.32(+.23)	.30(+.15)	.29	.28(+.08)	.27	.27(+.04)	.26	.26(+.00)	.26	.26

proportional to that of the whole text.

Table 2 and Table 3 show break even points of experiments using the fixed-length and proportional-length strategies, respectively. A break even point is a highest point where recall and precision is equal. It is meant as a summary figure of the performance. Precision and recall are determined for each text in the test set, by the formulae below:

$$\text{Precision} = \frac{\text{the number of words correctly identified as title words}}{\text{the number of words assigned}}$$

$$\text{Recall} = \frac{\text{the number of words correctly identified as title words}}{\text{the number of actual topics}}$$

We use a assigning strategy called *probabilistic thresholding* (Lewis, 1992) to decide what words to be assigned to the text as potential title indicators. Basically, what we do is to pick up a thresholding constant k and assign words whose probability of being a title word is greater than k. Typically, a large value of k gives high recall and low precision, while the opposite is the case with a small value of k. A break even point is obtained by varying the value of k.

Returning to Table 2 and Table 3, i indicates the size of segment, and l the length of text. The '+/−' figure next to each break even point indicates the improvement (or drop) as compared to a topic identification task using full texts. The asterisk '*' means that no break even point was found for the associated experiment and the precision at the highest recall is listed instead (the highest recall is given parenthetically). In case that the length of a text is smaller than that of the segment, the whole text is used.

The column labelled "10" in Table 2 is the result of applying a segment model which considers the starting 10-word block from a text. The table shows that at $i = 10$, there were no break even points found for texts with more than 400 characters ($l > 400$).

Both the FLM and PLM approaches produced an improvement over the full-text model. Discarding rear portions of a text turns out to be more effective for large texts ($l > 200$) than for short texts ($100 < l \leq 200$). However, the effectiveness of the "discard" strategy slowly declines as the text length increases. In Table 2, for instance, the effectiveness falls from .42 to .32 at $i = 20$. The distribution of similarity measurements for large texts in Fig. 4 suggests that the similarity distribution for large texts tend to be less *skewed* to the left than that for short texts. This would mean that title-indicating terms are scattered more evenly over the text, and thus it becomes all the more difficult to demarcate between relevant and irrelevant parts of the text.

A problem with the PLM approach is that a segment from which topical words are chosen is too small for short texts. Thus at 20%, for instance, its performance on 100-200 character texts drops by 17% compared to the full-text approach, but gradually improves as the value of j increases. (See Table 3 and Fig. 8). Interestingly enough, the situation turns around when l is large and j is small: thus at $j = 20$, there is a 20 % increase for $500 < l \leq 600$ but a 17 % decrease for $100 < l \leq 200$.

The results of experiments using paragraphs are shown in Fig. 4. The experiments used the first paragraph of a text as a segment. Though the use of paragraph achieved better results at some points (300-400 and 500-600) than other approaches, the overall performance is not outstanding compared to either FLM or PLM. In particular, the 'first paragraph' strategy is outperformed by the full-text method on texts with more than 900 characters.

Table 4: Results for using paragraphs. Figures are in break even point.

size	100-200	200-300	300-400	400-500	500-600	600-700	700-800	800-900	900-1000
b.e. pts	.396	.358	0.389	.371	.381	.338	.290	.283	.250

6. FINAL REMARKS

Two major benefits of using text structure in topic identification are an improvement in effectiveness and a considerable reduction of the text volume necessary for the correct identification of text topics. For instance, a fixed-length model using the first 20-word block requires only about one tenth of the words that are used in a full-text model and still performs significantly and consistently better than the latter. Contrary to our expectation, the results of the experiments cast some doubt as to the usefulness of paragraphs for topic identification.

Fig. 10 shows a particular implementation of the present method (full-text model), which operates in the Emacs environment. The article shown in Fig. 10 is from *Nihon Keizai Shimbun* (1992).

As a conclusion, let us mention a few points. The present paper demonstrated that evidence on text structure enhanced the performance on the identification of topical words in texts, which is based on a probabilistic model of text categorization. Importantly, we used texts which are not explicitly structured. A text structure is identified by measuring the similarity between segments comprising the text and its title. It was shown clearly that a text structure thus identified gives a good clue to finding out parts of the text most relevant to its content.

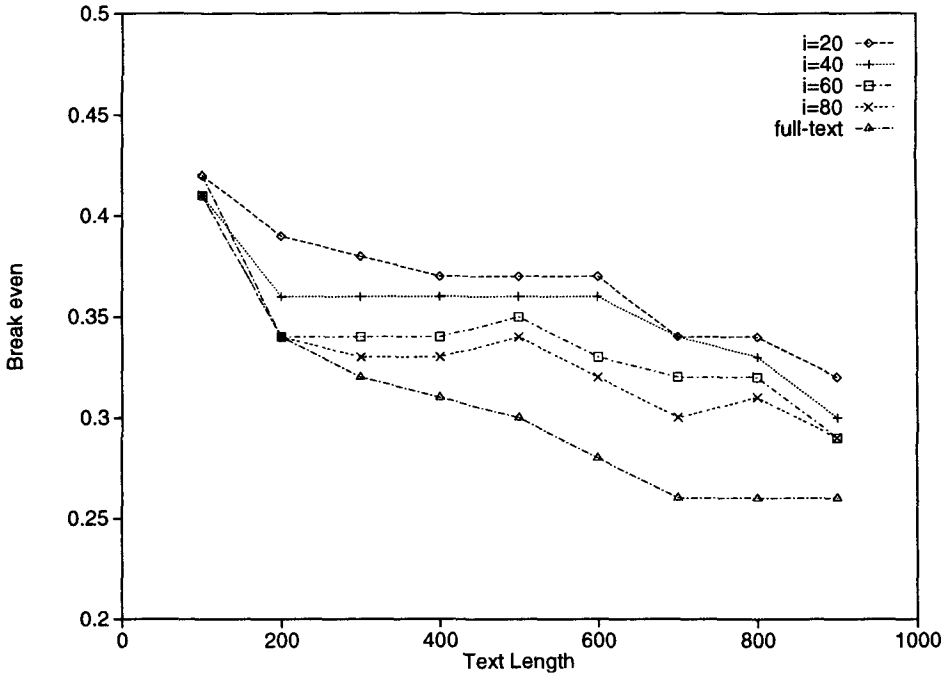

Figure 7: Fixed-Length Models

REFERENCES

James Allan. 1995. Relevance Feedback With Too Much Data. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 337–343. ACM Press.

James P. Callan. 1994. Passage-Level Evidence in Document Retrieval. In Croft and van Rijsbergen (Croft and van Rijsbergen, 1994), pages 302–310.

W. Bruce Croft and C. J. van Rijsbergen, editors. 1994. *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Dublin City University, Springer-Verlag.

Barbara A. Fox. 1987. *Discourse structure and anaphora.* Cambridge Studies in Linguistics 48. Cambridge University Press, Cambridge, UK.

Norbert Fuhr. 1989. Models for Retrieval with Probabilistic Indexing. *Information Processing & Management*, 25(1):55–72.

Barbara Grosz and Candance Sidner. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

Marti A. Hearst and Christian Plaunt. 1993. Subtopic Structuring for Full-Length Document Access. In Korfhage et al. (Korfhage et al., 1993), pages 59–68.

Marti A. Hearst. 1994. Multi-Paragraph Segmentation of Expository Text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, New Mexico, USA.

Jerry Hobbs. 1978. Resolving pronoun references. *Lingua*, (44):311–338.

Makoto Iwayama and Takenobu Tokunaga. 1994. A Probabilistic Model for Text Categorization: Towards a Tool for Personal Knowledge Acquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, Institute for Computational Linguistics,University of Stuttgart, Germany. Association for Computational Linguistics.

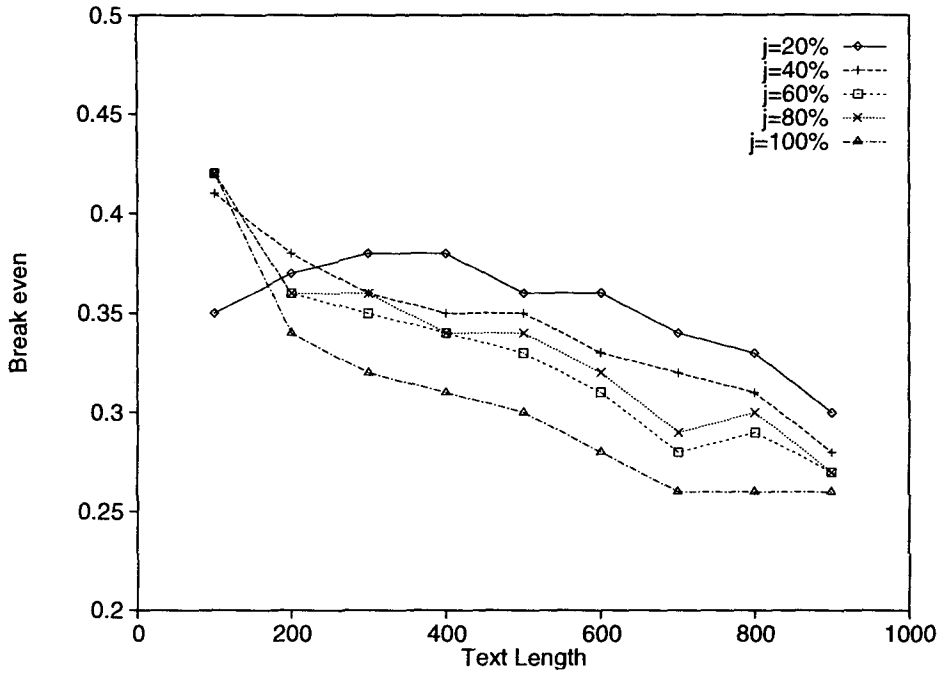

Figure 8: Proportional-Length Models

Aravind K. Joshi and Scott Weinstein. 1981. Control of Inference: Role of Some Aspects of Discourse Structure - Centering. In *Proceeding of International Joint Conference on Artificial Intelligence.*

Robert Korfhage, Edie Rasmussen, and Peter Willet, editors. 1993. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM Press.

Shalom Lappin and Herbert J. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):235–561.

David D. Lewis. 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50.

Yuji Matsumoto, Sadao Kurohashi, Takehito Utsuro, Yutaka Myoki, and Makoto Nagao. 1994. Japanese Morphological Analysis System JUMAN Manual. NAIST-IS-TR 94025, Nara Institute of Technology and Science, Nara, Japan.

Nihon-Keizai-Shimbun-Sha. 1992. Nihon Keizai Shimbun 92 nen CD-ROM ban. CD-ROM. Nihon Keizai Shimbun, Inc., Tokyo.

Tadashi Nomoto. 1995. Effects of Grammatical Annotation on the Topic Identification Task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

Gerald Salton, J. Allan, and C. Buckley. 1993. Approaches to Passage Retrieval in Full Text Information Systems. In Korfhage et al. (Korfhage et al., 1993), pages 49–58.

Marilyn Walker, Masayo Iida, and Sharon Cote. 1994. Japanese Discourse and the Process of Centering. *Computational Linguistics*, 20(2):193–232.

Ross Wilkinson. 1994. Effective Retrieval of Structured Documents. In Croft and van Rijsbergen (Croft and van Rijsbergen, 1994), pages 311–317.

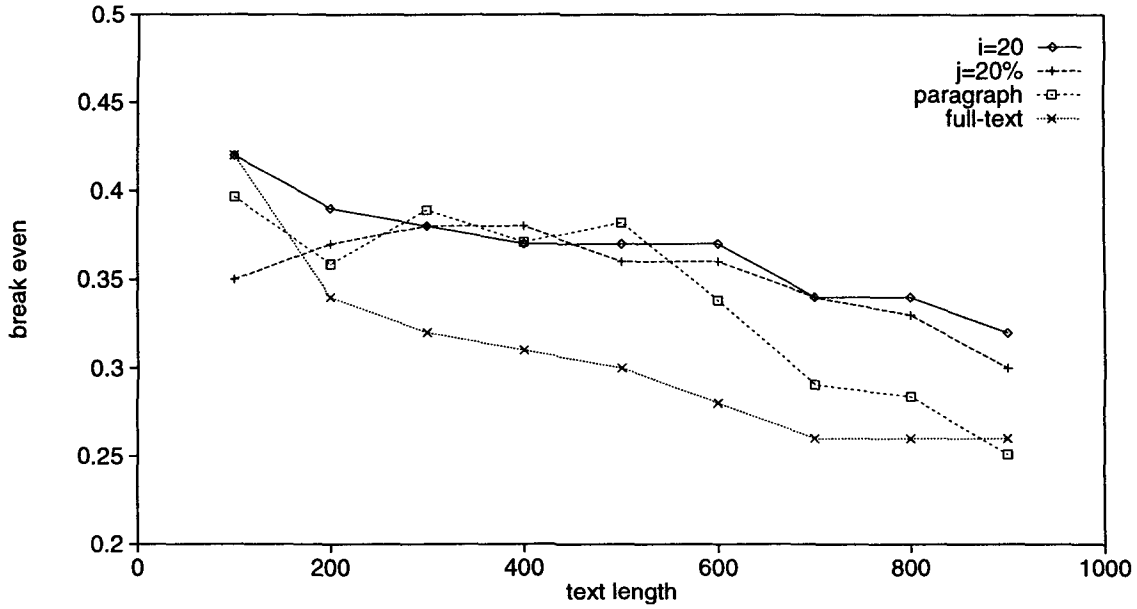

Figure 9: Comparing effectiveness of various strategies; FLM with $i = 20$, PLM with $j = 20\%$, 'first-paragraph' model, and the full-text model.

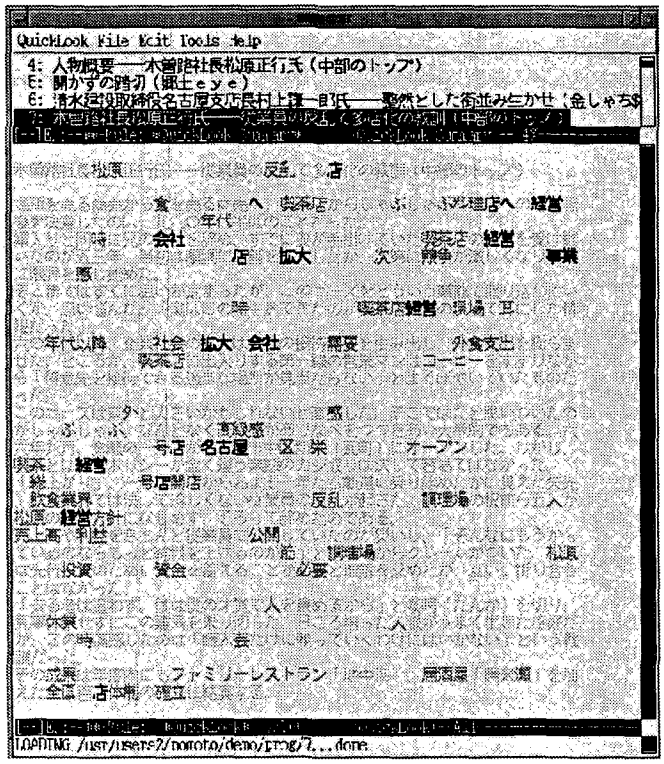

Figure 10: Topic Finder: What we see in the upper window is a list of headlines and corresponding news articles appear in the lower window. Each word in the article has a varying degree of visibility, reflecting the confidence in representing a topic of text; words with high visibility are more likely to be topical words of text.