

# Can readers recognize unit of summarization for reading?: an analysis of text segmentation task

Miho Fuyama (P)<sup>1</sup>, and Shohei Hidaka<sup>1</sup>

<sup>1</sup> Japan Advanced Institute of Science and Technology  
E-mail: miho02@sj9.so-net.ne.jp

## Abstract—

Viewing text reading as on-line summarization of a series of “text units” with coherent interpretation, we investigated human capability of segmentation texts into units of meaning. We hypothesize that readers detect the boundary of units of meaning by degree of meaningfulness of text within each unit. Our hypothesis predicts that the reader can identify latent units without reading the following sentence after the end-of-unit one. Our experiment showed a piece of evidence for this prediction.

**Keywords**— Comprehension, On-line summarization, Paragraph

## 1 On-line summarization and recognition of meaning units

How can we comprehend texts such as novels, reports, essays, and these articles? Readers cannot memorize all words and sentences of a text, therefore, they are supposed to hold the memory in a summarized form of the original text. We shall call this view of the text reading process *online summarization*. In this on-line summarization view, the reader is supposed to construct a unit which effectively compresses information on the text of a certain range. We call it *unit of summarization*, which is likely to match paragraph-level segmentation of text in practice.

## 2 Prediction of online summarization hypothesis

In the previous studies on text reading, unlike this view of reading as online summarization, the reader is hypothesized to analyze the sub-text based on some kind of similarity and dissimilarity [1, 2]. In this paper, we call this *text similarity hypothesis* of reading process.

What is critically different between the online summarization and similarity hypotheses of reading? Two views give different predictions if the reader summarize a certain range of text during reading it as being supposed in the online summarization hypothesis, the reader can predict the end of a paragraph or unit of meaningful sub-text, *before reading the next paragraph*. On the other hand, if the reader compare similarity or dissimilarity of sub-texts as being supposed in the text similarity hypothesis, the reader cannot predict the end of a paragraph, *until reading the new sentence in the next paragraph*.

In this study, we tested these prediction by two preliminary experiments of the paragraph segmentation tasks.

## 3 Experiment1

The goal of Experiment 1 is to confirm whether readers can detect the boundary from one paragraph to the other, when the boundary is not presented explicitly in the given text by the linguistic marker. In this paragraph segmentation task, each subject was shown a text, in which every end-of-line marker was removed, and was asked to answer whether or not it is the end of paragraph for each end of sentence. The prediction accuracy of the readers was analyzed by taking the end-of-paragraph markers in the original text as the ground truth.

### 3.1 Participants and Procedure

We employed seven participants. Their mother tongue are Japanese. We used openings of 10 Japanese texts (seven novels, three essays) each contains about 900 to 2200 letters. The original line breaks were take away from texts, so subjects did not show the original termination of paragraphs. Subjects were handed printed 10 texts and read them one by one with a natural speed. They marked sentences which they thought it was the last sentence of paragraphs.

### 3.2 Results

We analyzed participants’ responses using the signal detection theory [3]. The probability of “hit”  $P(\text{hit})$  is defined by the proportion of the number of sentences that the subject declared detection of end-of-paragraph out of the number of the end-of-paragraph sentences in the original text. The probability of “false alarm”  $P(\text{false alarm})$  is defined by the proportion of the number of sentences that the subject declared detection of end-of-paragraph out of the number of not-end-of-paragraph sentences in the original text.

Then the *d-prime* which can be interpreted as the degree of discriminability between the signal (presence of end-of-paragraph) and noise (absence of it) distributions in terms of distance normalized by the standard deviation, which is interpreted similarly as z-score. The d-primes across subjects ranges from 0.83 to 2.43, these result suggested all the subjects could detect the ends of paragraphs in the original text significantly better than the chance level.

Next, we analyzed the correlation coefficients between subject’s responses and variables reflecting the presense/absence of end-of-paragraph in the original text, where both two variables are coded as binary variables. The correlation coefficients range from 0.22 to 0.56, and showed that all pairs of the variables were significantly correlated ( $p < 0.01$ ).

Together with the first and second analyses, these

Table 1: The result of logistic regression analysis for experiments 2. The double and triple asterisk indicate  $p < 0.01$  and  $p < 0.001$ , respectively.

Subject No.	Coefficient	Std. Error	z value	Pr(> z )
1	-0.51	0.34	-1.50	0.1341
2	0.02	0.22	0.10	0.9197
3	0.81	0.29	2.82	0.0048 **
4	0.33	0.24	1.39	0.1648
5	0.60	0.22	2.77	0.0056 **
6	0.05	0.23	0.21	0.8373
7	0.13	0.37	0.34	0.7350
8	0.11	0.20	0.52	0.6053
9	-0.26	0.30	-0.89	0.3738
(Intercept)	-5.13	1.27	-4.05	0.0001 ***

results suggest that readers could detect the original paragraph significantly.

## 4 Experiment2

The goal of Experiment 2 is to test the prediction of the online summarization hypothesis relative to the text similarity hypothesis. In this experiment, the subject was shown one sentence at a time and was asked to respond whether it was the end-of-paragraph sentence or not by a five-point scale. The online summarization hypothesis predicts that the reader can detect end-of-paragraph line, under this circumstance, even when the subject cannot read the new sentence in the next paragraph in rating the likelihood of the end-of-paragraph. We analyzed the predictive accuracy of subject’s ratings under this situation.

### 4.1 Participants and Procedure

We employed nine participants. Their mother tongue is Japanese. We used the first 172 sentences of one Japanese novel “Kokoro” by Soseki Natsume as the text. Each subject was presented one new line at a time, and was asked to rate the likelihood for the new line to be the end of paragraph. In this way, the subjects cannot compare two lines across the paragraph boundary, when they rate it. Each of them was given a printed copy of the text, in which each sentence is printed as one line. The subject was asked to cover all the sentences but the first one by another piece of paper, and open the cover for the next sentence if they answered the rating. There is not time limit for each line, and the subject proceeded line by line as they wished.

### 4.2 Results and Discussion

First, the correlation between subjects scores and original paragraph was calculated (the end-of-paragraph sentences in the original text is coded by 1, and 0 otherwise.). The correlation coefficients range from 0.07 to 0.34 across subjects (five out of nine pairs showed significant correlation with  $p < 0.05$ ). This result suggested that about half of subjects should detect original paragraph significantly.

Second, we performed a logistic regression analysis on the binary variable indicating line of the original

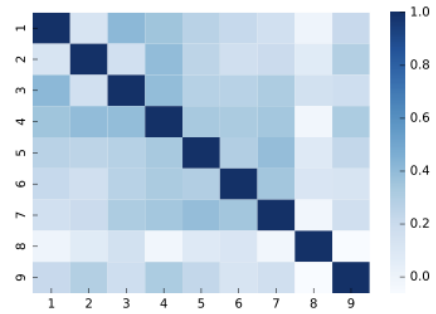


Figure 1: The correlation matrix among subjects on experiment 2.

end-of-paragraph by letting the subjects ratings as its predictors. The estimated regression coefficients and related statistics are summarized in Table 1. The ratings of the two subjects, Subject 3 and 5, were significant predictors. We also performed the model selection with AIC [4], the best estimated model was with ratings of the Subject 1, 3 and 5 as predictors.

Third, we also calculated the correlation coefficients for all the pairs of subjects to analyze the consistency of ratings across subjects. These correlation coefficients are  $-0.06 \sim 0.42$  (Figure 1), 22 pairs’ p-values were less than 0.01 and the other 5 pairs’ are less than 0.05. This result shows that 75% of pairs of correlation coefficients were significantly above 0 by the 5% criterion.

## 5 General Discussion

In this article, we conducted two experiments to test the online summarization hypothesis by its prediction that the reader can segment a text by summarization of sub-text rather than discrimination of difference between sub-texts. Experiment 1 suggested that readers can detect the ends of paragraphs, if the end-of-paragraph linguistic marker is not present in the text. Experiment 2, at least for some number of subjects, suggested that readers also can detect it, even before reading a new sentence in the next paragraph. In sum, these experimental results support the prediction of the online summarization hypothesis, rather than that of the text similarity hypothesis.

### Acknowledgments

This study was supported by the Grant-in-Aid for Young Scientists JP 16H05860.

### References

- [1] Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, 138(2), 307–327.
- [2] Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1), 33–64.
- [3] Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1), 137–149.
- [4] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.