Text Segmentation with Multiple Surface Linguistic Cues

MOCHIZUKI Hajime and HONDA Takeo and OKUMURA Manabu
School of Information Science
Japan Advanced Institute of Science and Technology
Tatsunokuchi Ishikawa 923-1292 Japan
Tel:(+81-761)51-1216, Fax: (+81-761)51-1149
{motizuki,honda,oku}@jaist.ac.jp

Abstract

In general, a certain range of sentences in a text, is widely assumed to form a coherent unit which is called a discourse segment. Identifying the segment boundaries is a first step to recognize the structure of a text. In this paper, we describe a method for identifying segment boundaries of a Japanese text with the aid of multiple surface linguistic cues, though our experiments might be small-scale. We also present a method of training the weights for multiple linguistic cues automatically without the overfitting problem.

1 Introduction

A text consists of multiple sentences that have semantic relations with each other. They form semantic units which are usually called discourse segments. The global discourse structure of a text can be constructed by relating the discourse segments with each other. Therefore, identifying segment boundaries in a text is considered as a first step to construct the discourse structure(Grosz and Sidner, 1986).

The use of surface linguistic cues in a text for identification of segment boundaries has been extensively researched, since it is impractical to assume the use of world knowledge for discourse analysis of real texts. Among a variety of surface cues, lexical cohesion(Halliday and Hasan, 1976), the surface relationship among words that are semantically similar, has recently received much attention and has been widely used for text segmentation(Morris and Hirst, 1991; Kozima, 1993; Hearst, 1994; Okumura and Honda, 1994). Okumura and Honda (Okumura and Honda, 1994) found that the information of lexical cohesion is not enough and incorporation of other surface information may improve the accuracy.

In this paper, we describe a method for identifying segment boundaries of a Japanese text with the aid of multiple surface linguistic cues, such as conjunctives, ellipsis, types of sentences, and lexical cohesion.

There are a variety of methods for combining multiple knowledge sources (linguistic cues)(McRoy, 1992). Among them, a weighted sum of the scores for all cues that reflects their contribution to identifying the correct segment boundaries is often used as the overall measure to rank the possible segment boundaries. In the past researches (Kurohashi and Nagao, 1994; Cohen, 1987), the weights for each cue tend to be determined by intuition or trial and error. Since determining weights by hand is a labor-intensive task and the weights do not always to achieve optimal or even near-optimal performance(Rayner et al., 1994), we think it is better to determine the weights automatically in order to both avoid the need for expert hand tuning and achieve performance that is at least locally optimal. We begin by assuming the existence of training texts with the correct segment boundaries and use the method of multiple regression analysis for automatically training the weights. However, there is a well-known problem in the methods of automatically training the weights, that the weights tend to be overfitted to the training data. In such a case, the weights cause the degrade of the performance for other texts. It is considered that the overfitting problem is caused by the relatively large number of the parameters (linguistic cues) compared with the size of the training data. Furthermore, all of the linguistic cues are not always useful. Therefore, we optimize the use of cues for training the weights. We think if only the useful cues are selected from the entire set of cues, better weights can be obtained. Fortunately, since several methods for parameters selection are already developed in the multiple regression analysis, we use one of these methods called the stepwise method. Therefore we think we can obtain the weights only for the useful by the using the multiple regression analysis and the stepwise method.

To give the evidence for the above claims that are summarized below, we carry out some preliminary experiments to show the effectiveness of our approach, even though our experiments might be small-scale.

- Combining multiple surface cues is effective for text segmentation.
- The multiple regression analysis with the stepwise method is good for selecting the useful cues for text segmentation and weighting these cues automatically.

In section two we outline the surface linguistic cues that we use for text segmentation. In section three we describe a method for automatically determining the weights for multiple cues. In section four we describe a method for automatically selecting cues. In section five we describe the experiments with our approach.

2 Surface Linguistic Cues for Japanese Text Segmentation

There are many linguistic cues that are available for identifying segment boundaries (or non-boundaries) of a Japanese text. However, it is not clear which cues are useful to yield better results for text segmentation task. Therefore, we first enumerate all the linguistic cues. Then, we select the useful cues and combine the selected cues for text segmentation. We use the method that a weighted sum of the scores for all cues is used as the overall measure to rank the possible segmentation with multiple linguistic cues.

First we explain this method used for text segmentation with multiple linguistic cues. Here, we represent a point between sentences n and $n+1$ as $p(n,n+1)$, where n ranges from 1 to the number of sentences in the text minus 1. Each point, $p(n,n+1)$, is a candidate for a segment boundary and has a score $scr(n,n+1)$ which is calculated by a weighted sum of the scores for each cue i, $scr_i(n,n+1)$, as follows:

$$scr(n,n+1) = \sum_i w_i \times scr_i(n,n+1) \qquad (1)$$

A point $p(n,n+1)$ with a high score $scr(n,n+1)$ becomes a candidate with higher plausibility. The points in the text are selected in the order of the score as the candidates of segment boundaries.

We use the following surface linguistic cues for Japanese text segmentation:

- Occurrence of topical markers ($i = 1..4$). If the topical marker 'wa' or the subjective postposition 'ga' appears either just before or after $p(n,n+1)$, add 1 to $scr_i(n,n+1)$.
- Occurrence of conjunctives ($i = 5..10$). If one of the six types of conjunctives [1] appears in the head of the sentence $n+1$, add 1 to $scr_i(n,n+1)$.
- Occurrence of anaphoric expressions ($i = 11..13$). If one of the three types of anaphoric expressions[2] appears in the head of the sentence $n+1$, add 1 to $scr_i(n,n+1)$.
- Omission of the subject (i=14). If the subject is omitted in the sentence $n+1$, add 1 to $scr_i(n,n+1)$.
- Succession of the sentence of the same type ($i = 15..18$). If both sentences n and $n+1$ are judged as one of the four types of sentences[3], add 1 to $scr_i(n,n+1)$.
- Occurrence of lexical chains ($i = 19..22$). Here we call a sequence of words which have lexical cohesion relation with each other a *lexical chain* like(Morris and Hirst, 1991). Like Morris and Hirst, we assume that lexical chains tend to indicate portions of a text that form a semantic unit. We use the information of the lexical chains and the gaps of lexical chains that are the parts of the chains with no words. The gap of a lexical chain can be considered to indicate a small digression of the topic. In the case that a lexical chain or a gap ends at sentence n, or begins at sentence $n+1$, add 1 to $scr_i(n,n+1)$. Here we assume that related words are the words in the same class on thesaurus[4].
- Change of the modifier of words in lexical chains ($i = 23$). If the modifier word of words in lexical chains changes in the sentence $n+1$, add 1 to $scr_i(n,n+1)$. This cue originates in the idea that it might indicate the different aspect of the topic becomes the new topic.

The above cues indicate both the plausibility and implausibility of the point as the segment boundary. Occurrence of the topical marker 'wa', for example, the indicates the segment boundary plausibility, while occurrence of anaphora, succession of the same type sentence indicate the implausibility. The weight for each cue reflects whether the cue is the positive or negative factor for the segment boundary. In the next section, we present our weighting method.

3 Automatically Weighting Multiple Linguistic Cues

We think it is better to determine the weights automatically, because it can avoid the need for expert hand tuning and can achieve performance that is at least locally optimal. We use the training texts that are tagged with the correct segment boundaries. For automatically training the weights, we use the method of the multiple regression analysis(Jobson, 1991). We think the method can yield a set of weights that are better than those derived by a labor-intensive hand-tuning effort. Considering the following equation $S(n,n+1)$, at each point $p(n,n+1)$ in the training texts,

$$S(n,n+1) = a + \sum_{i=1}^{p} w_i \times scr_i(n,n+1) \qquad (2)$$

where a is a constant, p is the number of the cues, and w_i is the estimated weight for the i-th cue, we can obtain the above equations in the number of the points in the training texts. Therefore, giving some value to S, we can calculate the weights w_i for each cue automatically by the method of least squares.

The higher values should be given to $S(n,n+1)$ at the segment boundary points than non-boundary

[1] The classification of conjunctives is based on the work in Japanese linguistics(Tokoro, 1987), which can be considered to be equivalent to Schiffren's(Schiffren, 1987) in English.

[2] The classification of anaphoric expressions in Japanese arises from the difference of the characteristics of their referents from the viewpoint of the mutual knowledge between the speaker/writer and hearer/reader(Seiho, 1992).

[3] The classification of types of sentences originates in the work in Japanese linguistics(Nagano, 1986).

[4] We use the Kadokawa Ruigo Shin Jiten(Oono and Hamanishi, 1981) as Japanese thesaurus.

points in the multiple regression analysis. If we can give the better value to $S(n, n+1)$ that reflects the real phenomena in the texts more precisely, we think we can expect the better performance. However, since we have only the correct segment boundaries that are tagged to the training texts, we decide to give 10 each $S(n, n+1)$ of the segment boundary point and -1 to the non-boundary point. These values were decided by the results of the preliminary experiment with four types of S.

Watanabe(Watanabe, 1996) can be considered as a related work. He describes a system which automatically creates an abstract of a newspaper article by selecting important sentences of a given text. He applies the multiple regression analysis for weighting the surface features of a sentence in order to determine the importance of sentences. Each S of a sentence in training texts is given a score that the number of human subjects who judge the sentence as important, divided by the number of all subjects. We do not adopt the same method for giving a value to S, because we think that such a task by human subjects is labor-intensive.

4 Automatically Selecting Useful Cues

It is not clear which cues are useful in the linguistic cues listed in section 2. Useless cues might cause a bad effect on calculating weights in the multiple regression model. Furthermore, the overfitting problem is caused by the use of too many linguistic cues compared with the size of training data.

If we can select only the useful cues from the entire set of cues, we can obtain better weights and improve the performance. However, we need an objective criteria for selecting useful cues. Fortunately, many parameter selecting methods have already been developed in the multiple regression analysis. We adopt one of these methods called the stepwise method which is very popular for parameter selection(Jobson, 1991).

The most commonly used criterion for the addition and deletion of variables in the stepwise method is based on the partial F-statistic. The partial F-statistic is given by

$$F = \frac{(SSR - SSR_R)/q}{SSE/(N-p-1)} \qquad (3)$$

where SSR denotes the regression sum of squares, SSE denotes the error sum of squares, p is the number of linguistic cues, N is the number of training data, and q is the number of cues in the model at each selection step. SSR and SSE refer to the larger model with p cues plus an intercept, and SSR_R refers to the reduced model with $(p-q)$ cues and an intercept(Jobson, 1991).

The stepwise method begins with a model that contains no cues. Next, the most significant cue is selected, and added to the model to form a new model(A) if and only if the partial F-statistic of the new model(A) is greater than F_{in}. After adding the cue, some cues may be eliminated from the model(A) and a new model(B) is constructed if and only if the partial F-statistic of the model(B) is less than F_{out}. These two processes occur repetitively until a certain termination condition is detected. F_{in} and F_{out} are some prescribed the partial F-statistic limits.

Although there are other popular methods for cue selection (for example, the forward selection method and the backward selection method), we use the stepwise method, because the stepwise method is expected to be superior to the other methods.

5 The Experiments

To give the evidence for the claims that are mentioned in the previous sections and are summarized below, we carry out some preliminary experiments to show the effectiveness of our approach.

- Combining multiple surface cues is effective for text segmentation.
- The multiple regression analysis with the stepwise method is good for selecting the useful cues and weighting these cues automatically.

We pick out 14 texts, which are from the exam questions of the Japanese language that ask us to partition the texts into a given number of segments. The question is like "Answer 3 points which partition the following text into semantic units." The system's performance is evaluated by comparing the system's outputs with the model answer attached to the above exam question.

In our 14 texts, the average number of points (boundary candidates) is 20 (the range from 12 to 47). The average number of correct answers boundaries from the model answer is 3.4 (the range from 2 to 6). Here we do not take into account the information of paragraph boundaries (such as the indentation) at all due to the following two reasons: Many of the exam question texts have no marks of paragraph boundaries; In case of Japanese texts, it is pointed out that paragraph boundaries and segment boundaries do not always coincide with each other(Tokoro, 1987).

In our experiments, the system generates the outputs in the order of the score $scr(n, n+1)$. We evaluate the performance in the cases where the system outputs 10%,20%,30%, and 40% of the number of boundary candidates. We use two measures, *Recall* and *Precision* for the evaluation: *Recall* is the quotient of the number of correctly identified boundaries by the total number of correct boundaries. *Precision* is the quotient of the number of correctly identified boundaries by the number of generated boundaries.

The experiments are made on the following cases:

1. Use the information of except for lexical cohesion (cues from 1 to 18 and 23).
2. Use the information of lexical cohesion(cues from 19 to 22).

3. Use all linguistic cues mentioned in section 2. The weights are manually determined by one of the authors.
4. Use all linguistic cues mentioned in section 2. The weights are automatically determined by the multiple regression analysis. We divide 14 texts into 7 groups each consisting of 2 texts and use 6 groups for training and the remaining group for test. Changing the group for the test, we evaluate the performance by the cross validation(Weiss and Kulikowski, 1991).
5. Use only selected cues by applying the stepwise method. As mentioned in section 4, we use the stepwise method for selecting useful cues for training sets. The condition is the same as for the case 4 except for the cue selection.
6. Answer from five human subjects. By this experiment, we try to clarify the upper bound of the performance of the text segmentation task, which can be considered to indicate the degree of the difficulty of the task(Passonneau and Litman, 1993; Gale et al., 1992).

Figure 1,2 and table 1 show the results of the experiments. Two figures show the system's mean performance of 14 texts. Table 1 shows the 5 subjects' mean performance of 14 texts (experiment 6). We think table 1 shows the upper bound of the performance of the text segmentation task. We also calculate the lower bound of the performance of the task("lowerbound" in figure 2). It can be calculated by considering the case where the system selects boundary candidates at random. In the case, the precision equals to the mean probability that each candidate will be a correct boundary. The recall is equal to the ratio of outputs. In figure 1, comparing the performance among the case without lexical chains("ex.1"), the one only with lexical chains("ex.2"), and the one with multiple linguistic cues("ex.3"), the results show that better performance can be yielded by using the whole set of the cues. In figure 2, comparing the performance of the case where the hand-tuned weights are used for multiple linguistic cues("ex.3") and the one where the automatic weights are determined with the training texts("ex.4.test"), the results show that better performance can be yielded by automatically training the weights in general. Furthermore, since it can avoid the labor-intensive work and yield objective weights, automatic weighting is better than hand-tuning.

Comparing the performance of the case where the automatic weights are calculated with the entire set of cues("ex.4.test" in figure 2) and the one where the automatic weights are calculated with selected cues("ex.5.test"), the results show that better performance can be yielded by the selected cues. The result also shows that our cue selection method can avoid the overfitting problem in that the results for training and test data have less difference. The difference between "ex.5.training" and "ex.5.test" is less than the one between "ex.4.training" and "ex.4.test". In our cue selection, the average number of selected cues is 7.4, though same cues are not always selected. The cues that are always selected are the contrastive conjunctives(cue 9 in section 2) and the lexical chains(cues 19 and 20 in section 2).

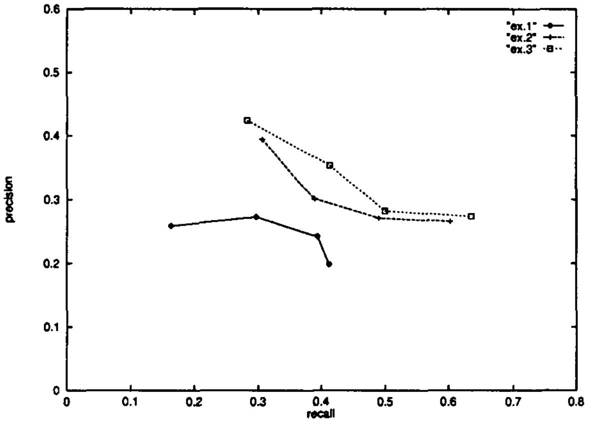

Figure 1: Hand tuning

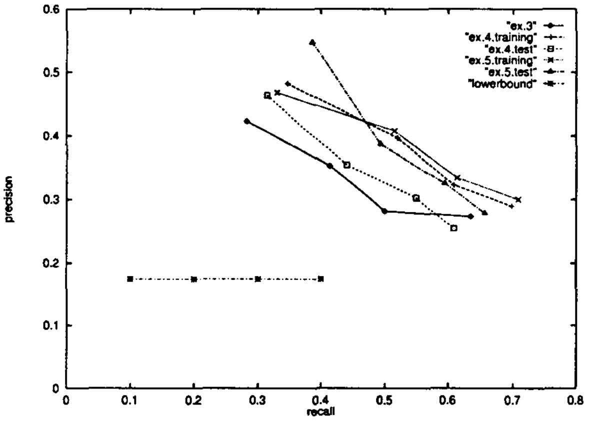

Figure 2: Automatic tuning

Table 1: The result of the human subjects

recall	precision
0.630714	0.571718

We also make an experiment with another answer, where we use points in a text that 3 or more human subjects among five judged as segment boundaries. The average number of correct answers is 3.5 (the range from 2 to 6). As a result, our system can yield similar results as the one mentioned above.

Litman and Passonneau(Litman and Passonneau, 1995)'s work can be considered to be a related research, because they presented a method for text segmentation that uses multiple knowledge sources. The model is trained with a corpus of spoken narratives using machine learning tools. The exact comparison is difficult. However, since the slightly lower upper bound for our task shows that our task is a bit more difficult than theirs, our performance is not inferior to theirs.

In fact, our experiments might be small-scale with a few texts to show the correctness of our claims and the effectiveness of our approach. However, we think the initial results described here are encouraging.

6 Conclusion

In this paper, we described a method for identifying segment boundaries of a Japanese text with the aid of multiple surface linguistic cues. We made the claim that automatically training the weights that are used for combining multiple linguistic cues is an effective method for text segmentation. Furthermore, we presented the multiple regression analysis with the stepwise method as a method of automatically training the weights without causing the overfitting problem. Though our experiments might be small-scale, they showed that our claims and our approach are promising. We think that we should experiment with large datasets.

As a future work, we now plan to calculate the weights for a subset of the texts by clustering the training texts. Since there may be some differences among real texts which reflect the differences of their author, their style, their genre, etc., we think that clustering a set of the training texts and calculating the weights for each cluster, rather than calculating the weights for the entire set of texts, might improve the accuracy. In the area of speech recognition, to improve the accuracy of the language models, clustering the training data is considered to be a promising method for automatic training(Carter, 1994; Iyer et al., 1994). Carter presents a method for clustering the sentences in a training corpus automatically into some subcorpora on the criterion of entropy reduction and calculating separate language model parameters for each cluster. He asserts that this kind of clustering offers a way to improve the performance of a model significantly.

Acknowledgments

The authors would like to express our gratitude to Kadokawa publisher for allowing us to use their thesaurus, and Dr.Shigenobu Aoki of Gunma Univ. and Dr.Teruo Matsuzawa of JAIST for their suggestions of statistical analysis, and Dr.Thanaruk Theeramunkong of JAIST for his suggestions of improvements to this paper.

References

D. Carter. 1994. Improving Language Models by Clustering Training Sentences. *Proc. of the 4th Conference on Applied Natural Language Processing*, pages 59–64.

R. Cohen. 1987. Analyzing the structure of argumentative discourse. *Computational Linguistics*, 13:11–24.

W.A. Gale, K.W. Church, and D. Yarowsky. 1992. Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs. In *Proc. of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 249–256.

B.J. Grosz and C.L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

H.A.K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman.

M.A. Hearst. 1994. Multi-Paragraph Segmentation of Expository Texts. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

R. Iyer, M. Ostendorf, and J.R. Rohlicek. 1994. Language modeling with sentence-level mixtures. In *Proc. of the Human Language Technology Workshop 1994*, pages 82–87.

J.D. Jobson. 1991. *Applied Multivariate Data Analysis Volume I: Regression and Experimental Design*. Springer-Verlag.

H. Kozima. 1993. Text segmentation based on similarity between words'. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 286–288.

S. Kurohashi and M. Nagao. 1994. Automatic Detection of Discourse Structure by Checking Surfce Information in Sentence. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 1123–1127.

D.J. Litman and R.J. Passonneau. 1995. Combining Multiple Knowledge Sources for Discourse. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*.

S.W. McRoy. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

J. Morris and G. Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21–48.

K. Nagano. 1986. *Bunsho-ron Sousetsu*. Asakura. in Japanese.

M. Okumura and T. Honda. 1994. Word sense disambiguation and text segmentation based on lexical cohesion. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 755–761.

Y. Oono and M. Hamanishi. 1981. *Kadokawa Ruigo Shin Jiten*. Kadokawa. in Japanese.

R.J. Passonneau and D.J. Litman. 1993. Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 148–155.

M. Rayner, D. Carter, V. Digalakis, and P. Price. 1994. Combining knowledge sources to reorder n-best speech hypothesis lists. In *Proc. of the Human Language technology Workshop 1994*, pages 271–221.

D. Schiffren. 1987. *Discourse Markers*. Cambridge University Press.

I. Seiho, 1992. *Kosoa no taikei*, pages 51–122. National Language Research Institute.

K. Tokoro. 1987. *Gendaibun Rhetoric Dokukaihou*. Takumi. in Japanese.

H Watanabe. 1996. A Method for Abstracting Newspaper Articles by Using Surface Clues. In *Proc. of the 16th International Conference on Computational Linguistics*, pages 974–979.

S.M. Weiss and C. Kulikowski. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann.