TEXT SEGMENTATION BY LATENT SEMANTIC INDEXING

T. Ishioka

NATIONAL CENTER FOR UNIVERSITY ENTRANCE EXAMINATIONS, JAPAN
tunenori@rd.dnc.ac.jp

A global discourse structure of a text can be constructed by relating the discourse segments with each other. Identifying the segment boundaries in a text is considered first step to construct the discourse structures. Several proposed approaches to the text segmentation problem have been adopted. These can be summarized as follows:

1.Approach based on lexical cohesion, e.g., TextTiling algorithm (Hearst 1997),

2.Combing features with a decision tree (Passoneau 1997),

3.Topic detection and tracking (TDT) plot study (Allan 1998).

Beeferman et al.(1999) examined the behavior of text segmentation by three typical approaches, and introduced a new statistical approach associated with maximum entropy modeling.

We try to apply another statistical technique to text segmentation by using *Latent Semantic Indexing* (LSI: Deerwester 1990) which has only been used for document or term retrieval. When we use LSI, potential discourse segments can be represented as k-dimensional vectors. The Euclidean distance between two potential segments becomes an index that indicates the difference in meaning between the segments of text on either side of potential boundary.

By using this method, we illustrate the text segmentation results applied to several well-known plays, such as *"The Prince* (English translation)" by Machiavelli, *"The Tragedy of Hamlet," "The Duke of Venice,"* and *"The Tragedies of Romeo and Juliet"* by Shakespeare, and compare the results to the actual chapter or section boundaries. A page is defined to consist of 50 lines including null lines; one page contains approximately 300 words. Potential segment(s) will be appear in the page. In Machiavelli's *The Prince*, we defined 64 "pages" of this text. If we can detect the every page contains the actual chapter or section boundaries by using the Euclidean distance criterion, it shows that LSI methods works well.

We found that the determination of the document boundary is possible by using Singular Value Decomposition (SVD) based on the idea of LSI, and found that the conditional entropy method seems to be replaceable by the SVD method; the distance between the segments is quite similar to an entropy measure under the condition that the prior probability of the potential boundary is given.

In addition, we refer the statistical criterion that we should identify the segment boundary. If we use this entropy model, the homoscedasticity test can be available to detect the document boundary, because the entropy depends on the variance of k-variables of LSI.

In section 2, we describe singular value decomposition associated with LSI. In section 3, we present a new statistical text segmentation method and the relation with the conditional entropy. In section 4, we show the text segmentation results applied to several famous plays, and compare them to the actual chapter or section boundaries. Section 5 is a summary.

Key words: maximum entropy, randomness, singular value decomposition, text boundary.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J. & Yang, Y.(1998). Topic Detection and Tracking Pilot Study: Final Report, *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.*

Beeferman, D., Berger, A., & Lafferty, J.(1999). Statistical Models for Text Segmentation, *Machine Learning: special issue on Natural Learning,* Cardie, C. & Mooney, R. (Eds.), *34*(1-3), 177–210.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. A.(1999): Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science, 41* (6), 391–407.

Passoneau, R. J. & Litman, D. J.(1997): Discourse Segmentation by human and automated means, *Computational Linguistics, 23* (1), 103–139.