Text Segmentation by Latent Semantic Indexing

Tsunenori Ishioka[1]

[1]Research Division, National Center for University Entrance Examinations, 2-19-23 Komaba, Meguro, Tokyo 153-8501, Japan

Summary. We point out that the determination of the document boundary is possible by using Singular Value Decomposition (SVD) based on the idea of LSI (Latent Semantic Indexing), and show that the conditional entropy method is replaceable by the SVD method using illustrations from several well-known plays. 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.

Key words. entropy model, natural language processing, singular value decomposition

1 Introduction

In general, a certain range of sentences in a text, is widely assumed to form a coherent unit which is called a discourse segment. 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(Grosz and Sidner 1986).

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 and Litman 1997),
3. Topic detection and tracking (TDT) plot study (Allan *et al.* 1989).

Beeferman *et al.* (1999) examined the behavior of those 3 approaches, and introduced a new statistical approach using maximum entropy modeling.

We try to apply another statistical technique to text segmentation by using *Latent Semantic Indexing* (LSI) which has only been used for document or term retrieval. Our method may be called a statistical TextTiling algorithm. We also illustrate the text segmentation results applied to several famous dramas.

If we use LSI, potential discourse segment boundaries can be represent as k-dimensional vectors. Thus, we can easily obtain the Euclidean distance between two potential boundaries. The distance is an index that indicates the difference in meaning between the segments of text on either side of potential boundary.

We show the distance is 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.

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. We show the text segmentation results applied to several famous dramas in section 4, and compare them to the actual chapter or section boundaries in these dramas.

2 Latent Semantic Indexing

2.1 Singular Value Decomposition Model

Latent semantic structure analysis starts with a matrix of terms by documents. This matrix is then analyzed by singular value decomposition (SVD) to derive our particular latent semantic structure.

Any rectangular matrix, for example a $t \times d$ matrix of terms and documents, X, can be decomposed into the product of three other matrices:

$$X = T_0 S_0 D_0' \tag{1}$$

so that T_0 and D_0 have orthonormal columns and S_0 are diagonal. This is called the singular value decomposition of X. T_0 and D_0 are the matrices of left and right singular vectors and S_0 is the diagonal matrix of singular values.

SVD is unique up to certain row, column and sign permutation and by convention the diagonal elements of S_0 are constructed to be all positive and ordered in decreasing magnitude.

2.2 Reduced Model

If the singular values in S_0 are ordered by size, the first k largest may be kept and the remaining smaller ones set to zero. The product of the resulting matrices is a matrix $\widehat{X}$ which is only approximately equal to X, and is of rank k. It can be shown that the new matrix $\widehat{X}$ is the matrix of rank k, which is closest in the least squares sense to X.

Since zeros were replaced into S_0, the representation can be simplified by deleting the zero rows and columns of S_0 to obtain a new diagonal matrix S, and then deleting the corresponding columns of T_0 and D_0 to obtain T and D respectively. The result is a reduced model:

$$X \approx \widehat{X} = TSD', \tag{2}$$

which is the rank-k model with the best possible least-squares-fit to X.

2.3 Comparing Two Documents

The dot product between two column vectors of X reflects the extent to which two documents have a similar pattern of occurrence across the set of documents. The matrix of $\widehat{X}'\widehat{X}$ is the square symmetric matrix containing all these document-to-document dot products. Since S is diagonal and D is orthonormal, it is easy to verify that:

$$\widehat{X}'\widehat{X} = DS^2D'. \tag{3}$$

Note that this means that the i, j cell of $\widehat{X}'\widehat{X}$ can be obtained by taking the dot product between the i and j columns of the matrix DS. Since S is diagonal, the DS space is just a stretched or shrunken version of the D space.

3 A New Statistical Approach

3.1 Text Segmentation based on LSI

If we use the reduced model, the coordinate position of a document is expressed by a vector whose number of components is not t (number of terms) but k, which is much smaller than t, because the vector corresponds to the i-th row of the matrix DS.

That means document i can be specified by one point in k-dimensional space. Thus, when we chain the coordinate positions of the documents, the magnitude of the linkage distance shows the dissimilarity between the two target documents, although LSI uses the cosine of the angle at which two target documents meet as the similarity measure. The cosine similarity corresponds to the correlation coefficient under the condition that the analyzed vector is composed of individual standardized variables.

We divided a larger document into several potential segments by using the delimiter of paragraph, page or section. These can be regarded as documents in LSI. Then, we can recognize where the text segmentation should occur by finding the two documents whose distance is more than certain threshold. This distance represents the degree of the change in the meanings associated with the two documents.

The contents of the text are changing progressively, even if it was written by same author. Therefore we can detect the text boundary by finding the part where semantic structure is changing a lot.

3.2 Entropy Model

The entropy for a continuous distribution is defined as follows:

$$H(X) = -\int f(x) \log f(x) dx, \tag{4}$$

where X is a random variable, and $f(x)$ is the probability density function.

If we assume any k element in document i of the reduced model is under the normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{5}$$

the entropy becomes

$$\begin{aligned} H(X) &= -\int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \left\{\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\ &= \log(\sqrt{2\pi}\sigma) \underbrace{\int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx}_{=1} \\ &\quad + \int \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{(x-\mu)^2}{2\sigma^2} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx. \end{aligned} \tag{6}$$

We set

$$y = \frac{x-\mu}{\sqrt{2\pi}\sigma}, \tag{7}$$

so that to be $dx = \sqrt{2\pi}\sigma dy$, the entropy is rewritten to the following using the transformed variable:

$$H(X) = \log(\sqrt{2\pi}\sigma) + \pi \underbrace{\int y^2 \exp[-\pi y^2] dy}_{1/(2\pi)} = \log(\sqrt{2\pi}\sigma) + \frac{1}{2}. \tag{8}$$

If we assume that document $i(1 \leq i \leq d)$ is a potential segment, X_i obey a stochastic process. Under the assumption that the k-dimensional occurrence for document i is given as $X_i = r_i$, the conditional entropy for document j , where $j > i$, becomes

$$H(X_j | X_i = r_i) = \log(\sqrt{2\pi}\sigma_j) + \frac{1}{2}, \tag{9}$$

where the occurrence for j is r_j, and the variance of the random variable $X_{j|i} = r_j - r_i$ is σ_j^2 .

Note that the conditional entropy is defined only by variance of the potential segment. Thus we can utilize a statistical homoscedasticity test to determine whether the conditional entropy changes significantly; the test should be performed by setting

$$\text{null hypothesis } H_0 : \sigma_{j-1}^2 = \sigma_j^2 \tag{10}$$

and

$$\text{alternative hypothesis } H_1 : \sigma_{j-1}^2 \neq \sigma_j^2. \tag{11}$$

Now we use the ratio of two samples as the test statistic, $F = S_{j-1}^2/S_j^2$. F obeys the F distribution whose degree of freedom is $(k-1, k-1)$.

The critical region W of a significant level α is

$$W = \{S_{j-1}^2/S_j^2 > c_1\}, \tag{12}$$

where

$$c_1 = f(1 - \alpha/2, k-1, k-1). \tag{13}$$

By using the singular value decomposition model or the reduced model, we can reduce the t-dimensional term vector indicating whether each term is contained, to a k-dimensional vector showing a latent semantic meaning; k usually has a value from 50 to 100. An entropy model can further reduce to one statistic, which shows the randomness of the semantic meanings.

It is difficult to judge whether a k-dimensional coordinates position is changed significantly, even if we set the proper assumptions against the distribution of statistics. However by using the entropy model, the statistical testing becomes easier, because it allows a test of homoscedasticity of two samples.

The assumption of continuity and the further assumption of normality in this section may be difficult to be accepted without some proofs or justifications. However, when there is no prior information about the distribution, it is thought most appropriate to assume a normal distribution.

4 Illustrations

We consider the following classical famous dramas:

(a) Machiavelli's *"The Prince"* (English translation),
(b) Shakespeare's *"The Tragedy of Hamlet,"*
(c) Shakespeare's *"The Duke of Venice,"*
(d) Shakespeare's *"The Tragedies of Romeo and Juliet."*

These texts are available on the internet at `http://www.gutenberg.net/`. Table 1 summarizes the number of word types, word tokens, and pages in each text. 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.

As an example, Table 1 shows that Machiavelli's *The Prince* consists of 32,331 words, of which 3,666 are unique, and that we defined 64 "pages" of this text.

Table 1. dramas with which we dealt

drama's name	number of words	number of unique words	number of pages
(a) *The Prince*	32,331	3,666	64
(b) *The Tragedy of Hamlet*	31,974	4,631	98
(c) *The Duke of Venice*	22,253	3,178	64
(d) *The Tragedies of Romeo and Juliet*	25,917	4,035	81

We get a singular value decomposition of a matrix of terms by documents for each drama; the number of singular values we required is 50. The distance between neighboring pages, when we use this reduced model, is shown upper section in each part of Figure 1. The vertical axis shows k-dimensional distance, and the horizontal axis the sequential number of pages. We can recognize the page in which the text meaning is changed. The letters (a)-(d) associated with each title in Table 1 correspond with same labels in Figure 1.

On the other hand, the lower sections of Figure 1 show the conditional entropy under the condition that the k-variables from the previous page are given. When comparing two sections, we find the state of affairs are quite similar.

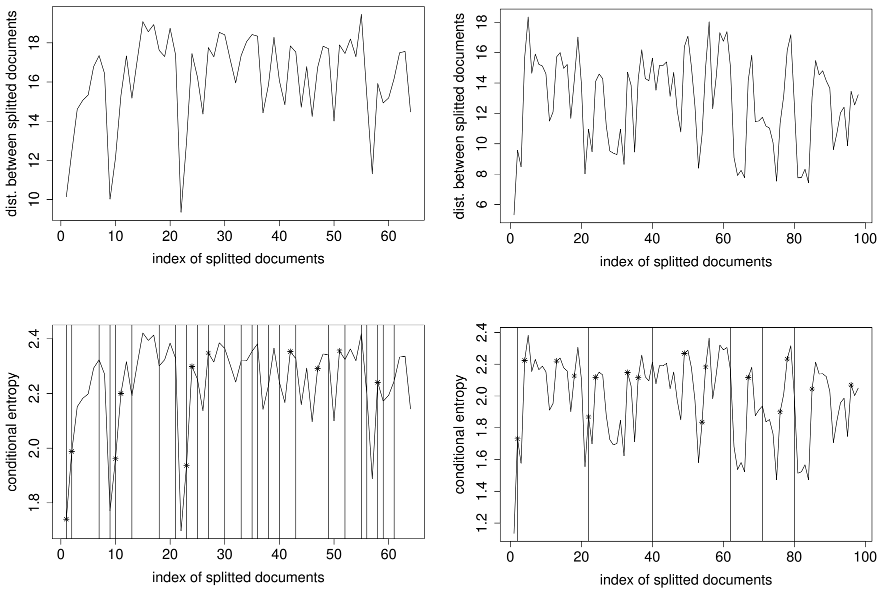
Euclidean linkage distance d_j, indeed, is shown as $d_j = \sqrt{(r_j - r_i)^2}$ and σ_j^2 in Eq. (9) as $\sigma_j^2 = (r_j - r_i)^2/k$. Thus, we can obtain the entropy for the document j by using d_j instead of σ_j^2; the formula is

$$H(X_j|X_i = r_i) = \log\sqrt{\frac{2\pi}{k}} + \log d_j + \frac{1}{2}. \tag{14}$$

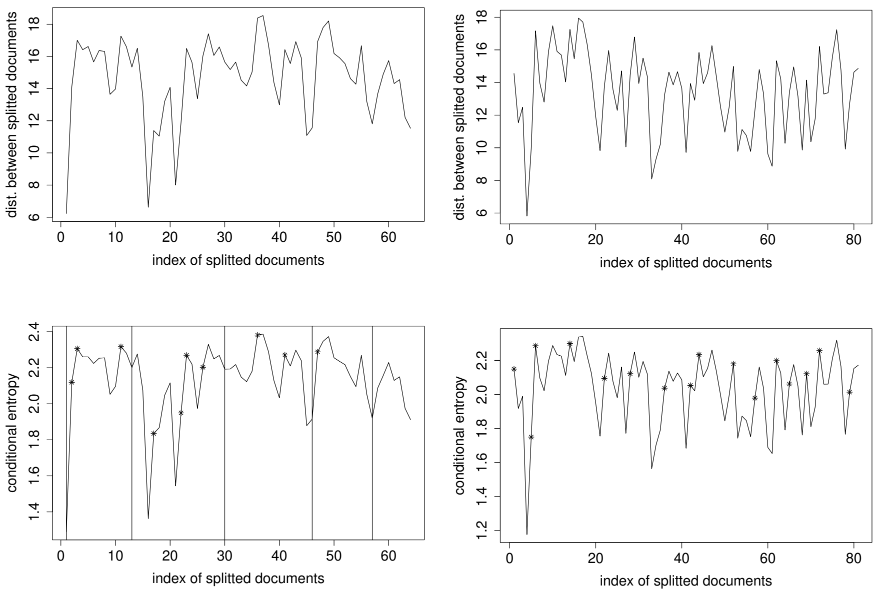
This equation shows that the linkage (k-dimensional) distance is replaceable by the conditional entropy (that is only one statistics), under the assumption that the distribution of $(r_j - r_i)$ has the normality.

For further comprehension, we add vertical lines as the formal punctuation, e.g., ACT, SCENE, or CHAPTER (depending on the drama) at the specified pages. In addition, we append marks of '$*$' where the conditional entropy increases significantly (significant level $\alpha = 0.20$); the conditional entropy depends on the variance of different vector from the previous page, so the homoscedasticity test is adopted.

Machavelli's "(a) The prince", which include 26 chapters, deals with political theory. The beginning five chapters are mentioned about followings: Chapter 1, how many kinds of principalities there are, and by what means they are acquired. Chapter 2, concerning hereditary principalities. Chapter 3, concerning mixed principalities. Chapter 4, why the kingdom of darius, conquered by Alexander, did not rebel against the successors of Alexander at his death. Chapter 5, concerning the way to govern cities or principalities which lived under their own laws before they were annexed. Chapter 6, concerning new principalities which are acquired by one's own arms and ability.



(a) Machiavelli's *"The Prince"* (b) *"The Tragedy of Hamlet"*



(c) *"The Duke in Venice"* (d) *"The Tragedies of Romeo and Juliet"*

Fig. 1. k-dimensional Distance from the Previous Page, and the Conditional Entropy

Thus, we find that the context is changing in chapter 1, 2, and 6. And $*$ marks actually put there.

In "(b) The Tragedy of Hamlet", $*$ marks are put on 16 places, that is page 3, 4, 13, 18, 22, 24, 33, 36, 49, 54, 55, 67, 76, 78, 85, and 96; they coincide with well the locations of "scene" in the drama on page 2, 3, 12, 15, 17, 24, 48, 54, 57, 63, 64, 66, 68, 74, 75 and 84; several vertical lines show "act" in the drama.

In "(c) The Duke in Venice", $*$ marks are put on 10 places in page 2, 3, 11, 17, 22, 23, 26, 36, 41, and 47; on the other hand, the "scene"s are located in page 2, 3, 11, 17, 21, 23, 25, 27, 30, 36, 41, 42, 44, and 56; these are also seemed to be coincident considerably.

The drama of (d) has no formal punctuation.

From these illustrations, we can obtain the following results:

– The sequential k-dimensional distances between two pages can be replaced by the conditional entropies; if we replace a k-dimensional vector with only one statistic, no lack of information is observed from the viewpoint of whether the semantic meaning of the document is changed or not.
– The pages in which the entropy is changed seem to correspond with those in the formal punctuation, e.g., ACT, SCENE and so on.

References

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

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

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

Berry, M. W., Dumais, S. T., and O'Brien, G. W.(1995) Using linear algebra for intelligent information retrieval *SIAM Review* **37**(4): 573–595.

Gous, A.(1999) Spherical Subfamily Models, submitted, *special issue on Natural Learning*, C. Cardie and R. Mooney eds., **34**(1-3): 177–210. `http://www-stat.stanford.edu/~gous/papers/ssm.ps`

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

Hearst, M.(1997) TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics, **23**(1), 33–64.

Hofmann, T.(1999) Probabilistic Latent Semantic Indexing, *ACM 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval*: 50–57.

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