TEXTUAL INFORMATION SEGMENTATION BY COHESIVE TIES

Samuel W.K. Chan, Benjamin K. T'sou and C.F. Choy

Language Information Sciences Research Centre
City University of Hong Kong
Hong Kong

ABSTRACT

This paper proposes a novel approach in clustering texts automatically into coherent segments. A set of mutual linguistic constraints that largely determines the similarity of meaning among lexical items is used and a weight function is devised to incorporate the diversity of linguistic bonds among the text. A computational method of extracting the gist from a higher order structure representing the tremendous diversity of interrelationship among items is presented. Topic boundaries between segments in a text are identified. Our text segmentation is regarded as a process of identifying the shifts from one segment cluster to another. The experimental results show that the combination of these constraints is capable to address the topic shifts of texts.

1. INTRODUCTION

Recent research in textual information science has increasingly turned to text processing beyond sentence level, partly because text analysis is manifestly necessary, and partly through implicit or explicit endorsement that negotiation of meaning in verbal transactions is achieved within the framework of text. Text has a rich structure in which sentences are grouped and related to one another in a variety of ways. A text is usually taken to be a piece of connected and meaningful sentences, rather than just a collection or sequence of sentences, connected only by contiguity. It is clear that text cannot simply be defined as *language above the sentence* (Winograd, 1972:65). Nor can we assume that stretches of language use which are isomorphic with sentences are the sole business of the grammarian, for these too are realizations of the text process. To understand a text, one must understand the relations between its parts and determine how the various pieces fit together. Clearly, these parts are not just the individual sentences: rather, sentences are joined together to form larger units, which in their turn may be the building blocks of yet larger units.

In information science, there is much ongoing research in finding textual regularities on the basis of empirical investigations. Analyzing text structure often calls for more than understanding individual sentences; it usually involves a process of making connections between sentences and sustaining that process as the text progresses (Grosz & Sidner, 1986). Current work also shows that centrality of text should no longer be defined in terms of any simplistic linguistic rules, but rather in terms of linguistic ties which exist among text segments (Stoddard, 1991). Among other kinds of interrelationship which a text may exhibit, *cohesion* has a prominent role in the understanding of its structure. As advocated by Halliday and Hasan (1976), cohesion can be a unity-creating device in texts. *Lexical cohesion*, which is a special kind of cohesion, investigates the repetitive sequences of lexically linked (co-articulated) items and their relations to the core sentences. Lexical cohesion has been identified as an important feature which underlines the structure of a wholesome text, distinguishing it from a non-text. Lexical cohesion, including lexical repetition as well as rhetorical continuity in conjunction with other related overt and covert linguistic markers, contributes to textual coherence by creating cohesive ties within the text. If the lexical items in a text can be related to the preceding or to following items, obviously, the text is seen more closely *knit together* than a text where such relationships do not exist. It has been ascertained that sentences with the greatest degree of lexical cohesion could be used to reflect the textual structure (Morris & Hirst, 1991).

Moreover, it is commonly believed that the recurrence of semantically related lexical items across the sentences could be used as an aid to identifying the core sentences which are characterized by their centrality and their expressive power. At the same time, while it is important for readers to be able to trace continuities in the entities under discussion, it is equally important to locate and understand the breaks in continuity. However, little research has demonstrated the functions of lexical cohesion in text segmentation and no computational theory or objective measure has as yet emerged in analyzing text structure on this basis. Given the increasing recognition of text structure in the fields of information retrieval in un-partitioned text, lexical cohesion also reveals the textual segmentability which means how texts are seen not as a continuous whole but as a complex grouping of larger pieces. There is a mounting demand for the in-depth study of an implementable quantitative model on lexical cohesion in text segmentation.

This research is to devise a quantitative model of text segmentation based on the study of lexical cohesion as well as other related linguistic features. What distinguishes it from previous studies is that attention is not primarily focused on itemizing cohesive features across a text but on investigating how they combine with other linguistic features to organize text into a coherent whole. We propose a novel approach in textual information science. The approach will identify discoursally salient text segment boundaries. The main objectives of this research are (i) to investigate patterns of cohesion in expository texts in order to test hypothesis about the textual continuity; (ii) to devise a measure in order to analyze the interrelations between each segment; (iii) to formulate a computational model and an objective measure in analyzing textual continuity; (iv) to propose and implement a method for the segmentation of texts into thematically coherent units. Demonstrations will be focused on Chinese expository and argumentative texts which usually consist of long sentences with little punctuation and textual demarcation. Section 2 describes the bonding analysis among the text. Our algorithm in textual segmentation identification is described in Section 3. The experimental results are presented in Section 4, followed by a conclusion.

2. BONDING ANALYSIS

A text is composed of a number of paragraphs, each of which is made up of a number of segments. Given that our intention is to explore the means by which various linguistic factors link segments, it is necessary to have a formalism for representing the links that will accurately reflect the non-linear complexity of a discourse and, at the same time, permit us to handle and interpret them conveniently. In our consideration of how discourse structure is expressed, we have already established a discourse network that is employed to represent the inter-sentential relationships existing among the sentences.

[DEFINITION 1]

> A discourse network $\boldsymbol{D}$ is defined by a set of discourse segments, which stands in functional relations to each sentence in the discourse. The discourse network is represented as a graph characterized by a 5-tuple (Chan & Franklin, 1996).
>
> $$\boldsymbol{D} = \langle G, T, A, E, W \rangle \quad \text{where}$$
>
> - G is a finite set of the discourse segments composing the discourse.
> - T is a finite set of lexical items (hereafter, called *token*) composing the discourse segments.
> - A is a set of arcs representing the inter-sentential relations amongst the discourse segments.
> - E is a set of weights of the arcs, lies between [0,1].
> - W is a function $W: A \rightarrow E$ which assigns lateral weights to arcs.

In our discourse network, the lateral weights between the arcs among the discourse segments are defined by linguistic clues. Let g_i, $g_j \in G$ be two discourse segments in the discourse network $\boldsymbol{D}$, each representing a different segment. If both of these segments are interrelated, the connection between them, i.e., W_{ij}, is assigned a large positive weight. On the other hand, it is reasonable to assume that syntactic function words do not denote new topics, whereas new semantic content words (nouns, main verbs, adjectives, and some adverbs) do. Given this assumption in our segmentation, a segment could be generated for a document simply by removing all function words from those tokens. Our bonding analysis for text segmentation is shown in the algorithm as follows.

```
Partition the text to elementary segments. A segment is a sentence with all the
        function words removed.
While more that one segment left Do
        Identify the possible links between every discourse segments $g_i$
        Assign the lateral weights among the segments in the discourse network $D$
                under the three principles of lexical cohesion as shown below.
End While
Compute text boundary using cluster identification technique.
```

One aspect of world knowledge essential to constructing the network is to identify when two lexical items in the segments are related. Several major types of relationships provide a text with cohesiveness by relating its lexical items to one another: (i) *identical lexical items* [幣値(currency) & 幣値(currency)]; (ii) *synonym* [幣値(currency) & 匯價(exchange rate)]; (iii) *association* [幣値(currency) & 出口(export)]; (iv) *antonymy* [反升(surge) & 降低(slump)]; (v) *superordinate* or *subordinate* [金融(financial sector) & 銀行(bank)]. In addition to the above lexical reiteration, we also adopt the term saliency factor which takes into consideration the frequency of occurrence of the processing token in the database (Salton, 1989). One of the objectives of indexing an item is to discriminate the semantics of that item from other items in the database. It can be seen as an associate meaning relationship between regularly co-occurring lexical items in the text. In the following sections, we will describe how these can be utilized in building up the discourse network ***D***.

2.1 Lexical Repetition

Constant repetition of lexical items would make easier for any reader to match strings in a sequence of sentences, construct the appropriate entities, relate the individual segments, and make inferences from them easier than they would be from a text with pronouns and lexical replacement. Word repetitions in lexical cohesion are identified by same word matches and matches on inflections derived from the same stem.

[DEFINITION 2]

The connection weights in the discourse network, due to the principle of lexical repetition, are assigned between word pairs, $\sigma_r(w, w') \in [0, 1]$ where $w \in g_i$, $w' \in g_j$, $i \neq j$, and $s(\cdot)$ is a length function

$$\sigma_r(w, w') = \frac{s(w \cap w')}{\max(s(w), s(w'))} \qquad \text{(Eqn. 1)}$$

The following Chinese examples suggest the feature of similarity σ_r of the lexical item 恒生指數(Heng Seng Index):

σ_r(恒生指數, 恒生指數) = 1.000

σ_r(恒生指數, 恒生銀行) = 0.727

σ_r(恒生指數, 恒基兆業) = 0.155

Another sense of lexical repetition is collocation. Collocation is a technique that originates from a particular distributional feature of lexical cohesion, namely that the number of links shared by segment pairs tends to increase as the distance between segments decrease. In other words, lexical items that occur together regularly are said to be collocated. This can happen in closely associated pairs like 恒生指數 (Heng Seng Index) and 上升(surge). In fact, this habitual association is largely independent of grammatical and semantic structures, being a relationship between lexical items and not between classes of words. In our approach, the frequency of co-occurrences for each pair of lexical items is collected and the collocation measure between them is calculated.

[Definition 3]

Let w and w' be two lexical items, the collocation measure $\sigma_c(w, w')$ is defined by

$$\sigma_c(w, w') = \frac{cf(w, w')}{f(w) + f(w') - cf(w, w')} \qquad \text{Eqn. (2)}$$

where $f(w)$ and $f(w')$ are the numbers of occurrence of lexical items w and w' respectively while $cf(w,w')$ is the number of co-occurrence of both lexical items in a predefined size of window. This collocation measure gives each pair of lexical items with a range [0,1].

2.2 Semantic Overlapping

Lexical preference is crucial in solving many natural language processing tasks. Whittemore and his colleagues (1990) find lexical preferences to be the key to resolve ambiguity. They echo Taraban and McClelland (1988) who have shown that the structural models of language analysis are not in fact good predictors of human behavior in semantic interpretation. Within the domain defined by this approach to devising semantic overlapping in text segmentation, the choice of thesaurus as one type of knowledge structure to better understand is indicated by certain properties which thesauri have as sources for the formation of computerized knowledge representations. Thesauri have the merit of being already semi-formalized. They also, by implication at least, embrace substantial subsets of any given natural language as a whole, inasmuch as they have the interesting property of serving an infinite multiplicity of functions as knowledge bases. Faceted thesauri are a special kind of semantic network with *is-a* and are being used in indexing and retrieving documents. Each occurrence of the same token under different categories of the thesaurus represents different senses of that word, i.e., the categories correspond roughly to word senses. A set of words in the same category is semantically related. In our approach, a Chinese thesaurus which defines a *is-a* hierarchy is employed. We make use the shortest path in the *is-a* hierarchy to measure the conceptual similarity between the tokens in the discourse segments (Rada *et al.*, 1989). That is, given tokens w and w' in the *is-a* hierarchy, the distance between the tokens is as follows:

[DEFINITION 4]

distance(w , w') = minimal number of *is-a* relationships between w and w'

An *is-a* hierarchy in the Chinese thesaurus, as in other thesaurus such as WordNet (Miller, 1985), can be viewed as a directed acyclic graph with single root. Figure 1 shows the structure of an *is-a* hierarchy.

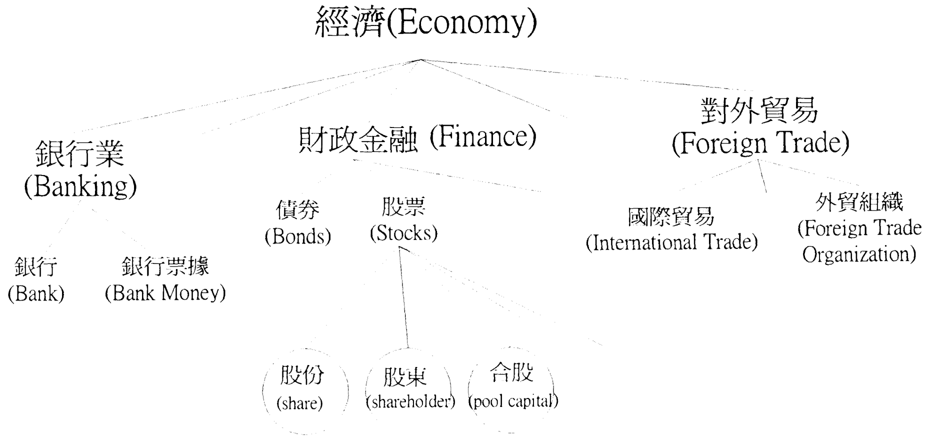

Figure 1: *is-a* hierarchy of a Chinese thesaurus

In order to compute the conceptual distances between each segment, all pairwise combination between tokens in one segment g_1 and tokens in every other segment g_2 are generated. For each pairwise combination, the following definition is used to create a metric over the segments in the discourse network.

[DEFINITION 5]

Let $g_1 = \{w_{x1}, w_{x2}, ..., w_{xm}\}$ and $g_2 = \{w_{y1}, w_{y2}, ..., w_{yn}\}$ be the two segments, the similarity $\sigma_S(.)$ and dissimilarity $\sigma_{DS}(.)$ components due to the semantic overlapping are defined by:

$$\sigma_S(g_1, g_2) = \frac{1}{n} \sum_{w_{xi} \in g_1} \min_{w_{yi} \in g_2} \{d(w_{xi}, w_{yi})\} \qquad \text{Eqn. (3)}$$

$$\sigma_{DS}(g_1, g_2) = \left(\frac{1}{mn} \sum_{w_{xi} \in g_1} \sum_{w_{yi} \in g_2} \sqrt{d(w_{xi}, w_{yi})} \right)^{-1} \qquad \text{Eqn. (4)}$$

where m, n are the cardinality of g_1 , g_2 respectively.

In Eqn.(3), the similarity measure is defined by the average of the distances between each token in one segment and the closest token in another segment. The dissimilarity measure in Eqn.(4) is on the basis of the fact that comparing distances between nodes that are close to each other seems more significant than comparing distances between nodes that are far from each other, a square root function is applied to the distance instead. Obviously, the distance satisfies the properties of a metric which are the zero, symmetric, positive and triangular inequality properties (Rada *et al.*, 1989).

2.3 Term Saliency Factor

Term saliency factor is a weight function which computes the number of times of each item occurs in a document *tf* times the inverse logarithm factor of the number of documents that the word occurs in a large collection *idf* (Salton, 1989; 1997). The term saliency factor algorithm is using a corpus-based knowledge. The advantage of using term saliency factor is the boundary of topics in a document can be distinguished by the coherence values of each segment pairs. The weight for each token in a document is defined as

$$r(w_i) = tf(w_i) \times idf(w_i) \qquad \text{Eqn.(5)}$$

where token frequency *tf* is the number of occurrences of a token w_i in a document $D_{i.}$ Document frequency *df* is the number of documents in a collection of N documents in which token w_i occurs. The saliency factor $r(w_i)$ is the product of *tf* and the inverse of *df* factor $log\ N/df_i$. When N is large and df_i is small, the token w_i is considered to be more important than other tokens. But, when the N is large and the df_i is large too, the token w_i is considered to be less important among the other tokens of documents. The frequent occurrences of tokens that are concentrated in particular documents are considered to be more important than the other tokens that are frequent but occur evenly over the entire document collection. This shows that saliency factor favors rare words than common words. Tokens that commonly occur throughout a collection are not necessarily good indicators of saliency because they are so common, and so their importance is down-weighted. Adjacent segments of text are compared to see how similar they are according to the number of tokens the adjacent segments have in common as defined in the following definition.

[DEFINITION 6]

Let $g_1 = \{w_{x1}, w_{x2}, ..., w_{xm}\}$ and $g_2 = \{w_{y1}, w_{y2}, ..., w_{yn}\}$ be the two segments, coherence value for the similarity between segments is calculated by a normalized inner product of the two text segments g_1 and g_2, the similarity component due to the semantic overlapping is defined by:

$$\sigma_{coh}(g_i, g_j) = \frac{2\sum_{i,j} r(w_i) \times r(w_j)}{\sum_i r(w_i)^2 + \sum_j r(w_j)^2} \qquad \text{Eqn. (6)}$$

Eqn. (6) yields a value between 0 and 1 representing the term saliency factors within the collection of documents. The weight generated from all these three major principles are combined to form an overall lateral matrix W which represents the connection across each segment.

$$W(g_1,g_2)=\sum\nolimits_{w\in g_1,w'\in g_2}\{\sigma_r(w,w')+\sigma_c(w,w')\}+\sigma_S(g_1,g_2)-\sigma_{DS}(g_1,g_2)+\sigma_{Coh}(g_1,g_2) \qquad \text{Eqn.(7)}$$

In representational terms, there is a greater concentration of weights indicating rich links near the diagonal of the lateral weight matrix. Since concentration of links are potential indicators of textual coherence, our textual information segmentation can be regarded as a process of identifying segment clusters through the lateral matrix as described in the following.

3. CLUSTER IDENTIFICATION AS TEXTUAL INFORMATION SEGMENTATION

Most of the discourse segmentation techniques are based on the premise that the coherence should be lower in areas of the discourse where the discourse topic changes. Our approach turns to identify clusters which the discourse segments belong to. Boundaries are detected through the shifts of discourse segments from one cluster to another. We make use of an orthogonal decomposition known as the Singular Value Decomposition (SVD), which is a generalization of the well-known eigenvalue decomposition. SVD is a technique closely related to eigenvector decomposition and factor analysis. It is usually used in the solution of unconstrained linear least squares problems, matrix rank estimation and canonical correlation analysis. First, we shall define what is SVD in the remaining section and then explain how it can be used to identify the discourse boundaries.

[Theorem 1]

Given a matrix $A \in \mathbf{R}^{mxn}$, without loss of generality $m \geq n$ and rank $(A) = r$, then there exist orthogonal matrices $U \in \mathbf{R}^{mxn}$ and $V \in \mathbf{R}^{mxn}$ such that $A = U\Sigma V^T$

where $$\Sigma = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}$$

and $\Sigma = \mathrm{diag}(\lambda_1, \lambda_2, .., \lambda_n)$, $\lambda_i > 0$ for $1 \leq i \leq r$, $\lambda_j = 0$ for $j \geq r + 1$ and

$U^T U = V^T V = I$

The first r columns of the orthogonal matrices U ($m \times p$) and V ($n \times p$) define the orthogonal eigenvectors associated with the r nonzero eigenvalues of AA^T and A^TA, respectively. The columns of U and V are referred to as the left and right singular vectors, respectively, and the singular values of A are defined as the diagonal elements of Σ, which are the nonnegative square roots of the n eigenvalues of AA^T. These matrices reflect a breakdown of the original relationships into linearly independent vectors or factor values. In our application, the first step is to represent the inter-relationships among the tokens in the text, as defined in Eqn. (7) in Section 2, by an overall $m \times m$ matrix W in which each row and column stands for a unique segment. Each entry, say W_{ij}, contains the weight in which the segment i is related to segment j and the entry subsumes the contribution coming from the lexical repetition, semantic overlapping and term saliency factor. The SVD of the matrix W is then defined as the product of three matrices,

$$W = B\Sigma B^T$$

where the columns of B contains the eigenvectors of W and Σ is a diagonal matrix containing the eigenvalues in descending order:

$$\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_n$$

The eigenvectors are normalized to have length 1 and orthogonal, which means that they satisfy the following condition: $B^TB = I$. Decomposing a regular matrix into a product of three other matrices is not too interesting. However, if the first k ($<< m$) columns of the B matrix and the first (largest) k singular values of

W are used to construct a rank-k of W via W_k, such that

$$W_k = B_k \Sigma_k B_k^T$$

then the W_k, constructed from the k-largest singular triplets of W is the closest rank-k approximation in the least squares sense. The Singular Value Decomposition is truncated into a new segment-by-segment matrix by multiplying the first two singular values of diagonal matrix Σ with the first two columns of singular vectors of the orthogonal matrix of B.

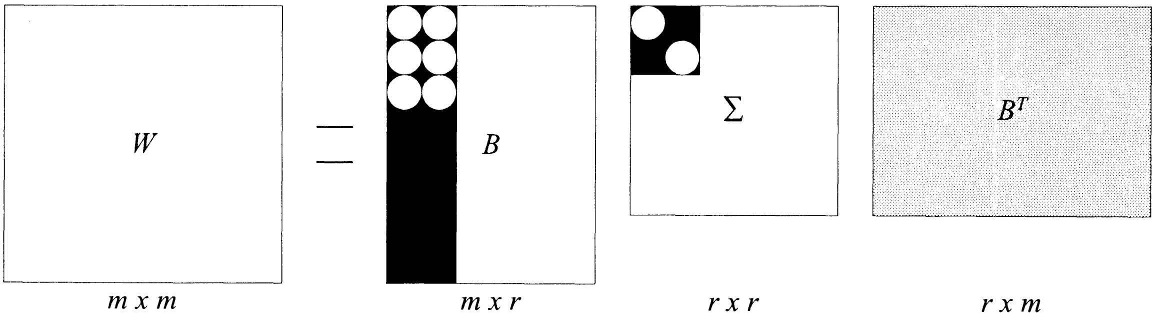

Figure 2: Mathematical representation of the matrix W.

The advantage of using SVD is that the higher-order structure in the association of segments within the document can be approximated by the truncated segment-by-segment matrix. The W_k is the best possible rank-k approximation of W in several senses, including the square root of the sum of squares of the elements. Another way to express this is that if we project onto the first k principal components, we have the most accurate rank-k reconstruction of the original data points in W_k. The truncated SVD matrix is used to show the high coherence relationship of the segments in the document, and to estimate the structure in segments across the document. It also captures the most important underlying structure in the association of segments in document and removes the noise or variability in segment usage that plagues segment-based cohesive ties. By reducing the dimensionality of W, much of the noise that indicates the less important bonds among the text can be eliminated. SVD allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important links among the segments. This dimension reduction step has collapsed our original piecewise and leads to an approximate model that contains many fewer dimensions. In summary, in this reduced model, the bonding are now approximated by values on this smaller number of dimensions. The result can still be represented geometrically by a spatial configuration in which the dot product or cosine between vectors representing two segments corresponds to their estimated similarity.

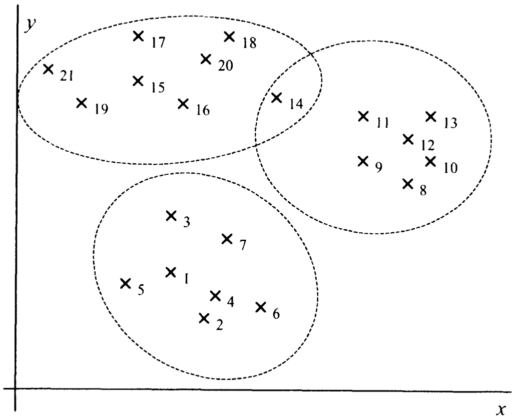

Figure 3: Boundaries are detected in discourse segments 8, 14 since the segments move from one cluster to another.

The result of Singular Value Decomposition is a k-dimensional vector space containing a vector for each segment. In order to visualize the clusters so formed that a text may exhibit, we use the first column of B

multiplied by the first singular value λ_1 for the x-coordinates and the second column of B multiplied by the second singular values λ_2 for the y-coordinates. The segments can then be represented on the Cartesian plane. The location of segment vectors reflects the correlation in their usage across segments. Segments within the same topic will cluster around while any topic shift in text can be detected by the shift of segments from one cluster to another as shown in Figure 3.

4. IMPLEMENTATION AND RESULTS

The goal of this experiment is to understand the inter-relationships among segments within documents and demonstrate how the principles influence textual segmentation. Fifteen documents are selected in our implementation with a total of 476 segments and about 10,000 tokens. The documents are extracted from several categories, including economic, health, education and sport. With the assumption that function words do not denote much important meaning while semantic content words do, our preprocessing first removes function words from the documents. At the same time, other relevant information, such as document ID, segment ID, segment-token number, token ID, token and thesaurus index are stored into a segmentation table. In Chinese language processing, lexical and semantic meanings, compared with grammatical linkages, play an extra-ordinarily dominant role because of little surface inflectional morphology in Chinese. It is therefore essential to identify and analyze the recurrent lexical and semantic patterns relevant to Chinese language understanding. In order to represent the sole effect of each principle as described in Section 2, we demonstrate their effects in textual segmentation. In lexical repetition, every pair of segments is compared to find the number of same or similar tokens as defined in Eqns. (1) and (2). As more repetition among tokens can be found between the segments, this segment pair will have a higher coherence value.

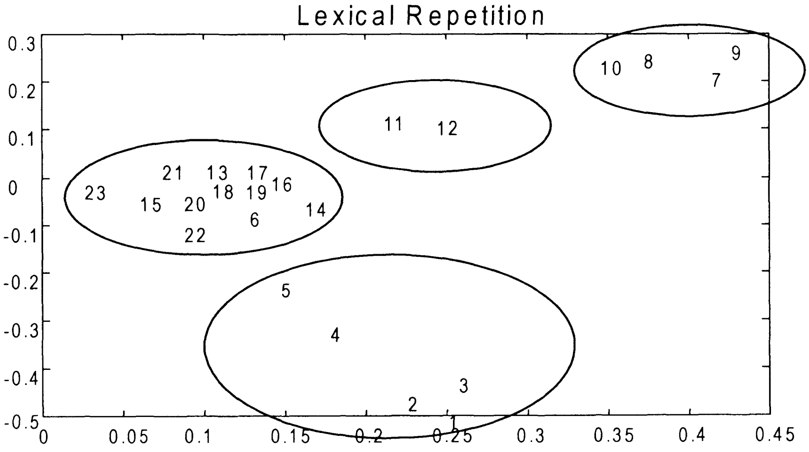

Figure 4: Segment cluster linked by lexical repetition after singular value decomposition with k =2

Figure 4 shows the segment cluster linked by lexical repetition after the singular value decomposition with k equal to 2. The number in the figure represents the corresponding segment in the document. The oval indicates those segments that are likely close together and may be considered as a group under the same topic. The distance between segments represents the segment similarity measure among them. The more closer the segments in the figure, the higher the similarity of segments. Similarly, in semantic overlapping, every pair of segments is compared in order to identify the semantic overlapping as defined in Eqns. (3) and (4) using a built-in thesaurus. When more semantic overlapping can be found between the segment pairs, the pairs will have a higher coherence value and it is most unlikely that the segment boundaries lie among them.

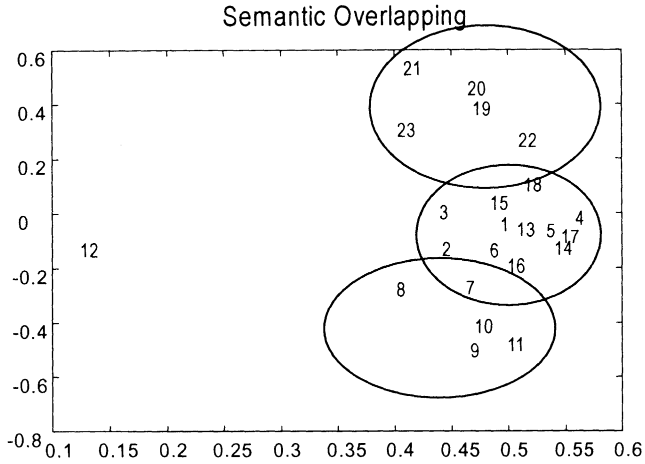

Figure 5: The segment cluster linked by semantic overlappping after the singular value decomposition

Similarly, Figure 5 shows the segment cluster linked by semantic overlapping after the singular value decomposition. It can be observed that segments 1-6, 7-11, 19-23 are under different topics. Chaos appears in segments 13-18 which seem to be overlapped with segments 1-6. The boundaries between the two clusters, 1-6 and 13-18 are totally unclear.

As shown in the preceding sections, the three potential constraining principles in textual continuity are described. The application of each is supported by empirical studies of text structure, and each is consistent with general assumptions about the nature of discourse. However, as shown in Figures 4 and 5, it is clear that none of these principles are by themselves sufficient to emulate humans' solution to discourse segmentation. In the process of constructing a coherent representation of discourse segments, the reader must make a number of bridging inferences that do not solely rely upon either one. Given this situation, our working hypothesis described here is that all three must be applied simultaneously. Figure 6 shows the combined effect of the three principles such as lexical repetition, semantic overlapping and term saliency factor.

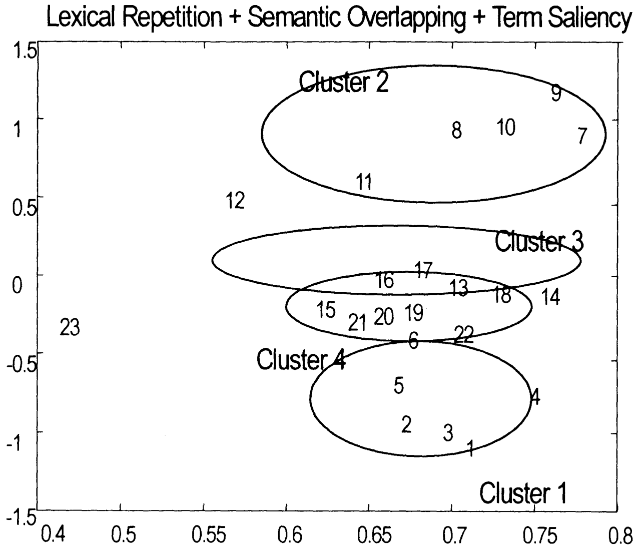

Figure 6: Segment cluster linked by combined effect after the singular value decomposition

The figure shows that segments 1-5, 7-11, 13-18, 19-23 are clearly under different topics and the corresponding topic shifts occur at segments 6-7, 12-13, 18-19. By investigating these results in turn, it is clear that the combined effect achieves the best result, although the semantic overlapping, among the three principles, shows better performance. One may expect that the performance will be deteriorated by the inappropriate links as more inter-relationships are added. However, the dimension reduction using singular value decomposition has demonstrated its capability by distilling the main gist or segment clusters in the noisy environment. This coarse segmentation provides the outlines and the gist of the text, omitting details and inconsistencies. The segmentation using lexical cohesion has obvious applications at the beginning of any summarisation processes.

5. CONCLUSION

In this research, the modeling we put forward is to employ a novel approach which establishes a network of relations among segments in the discourse. Lexical cohesion between linguistic items is reflected by using various linguistic clues modeled in our discourse network. The process of discourse segmentation, from a microscopic point of view, can be regarded as a process of assigning weights between the discourse segments. We have presented a method for segmenting texts into thematically coherent units using the techniques in matrix computation. In order to exaggerate the cohesive effect, our initial discourse network is subjected to a singular valued decomposition which is interpreted as a particular transformation of a given set of weights into a set of lexical clusters. This novel approach, different from any others, not only provides more sophisticated segmentation by reducing the noise but also provides a clear visual effect in the analysis.

ACKNOWLEDGEMENTS

This research was supported in part by a Competitive Earmarked Research Grant of the Research Grants Council of Hong Kong SAR (CityU 1196/99E) and a Strategic Research Grant of City University of Hong Kong under grant #7000868.

REFERENCES

Chan, S.W.K., & Franklin, J. (1996). A Brain-State-in-a-Box Network for Narrative Comprehension and Recall. *Proceedings of IEEE International Conference on Neural Networks* (ICNN'96), Vol.2, 694-699. Washington, D.C.

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

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Miller, G.A. (1985). Wordnet: A Dictionary Browser in Information in Data. *Proceedings of the First Conference of the UW Center for the New Oxford Dictionary*. Waterloo, Canada: University of Waterloo.

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

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19, 1, 17-30.

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley.

Salton, G., Singhal, A., Mitra, M. & Buckley, C. (1997). Automatic text structuring and summarization. *In Information Processing and Management, Elsevier Science*, 33, 193-207.

Stoddard, S. (1991). *Text and Texture: Patterns of Cohesion*. Advances in Discourse Processes, volume XL. Ablex.

Taraban, R., and McClelland, J.L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectation. *Journal of Memory and Language*, 27, 597-632.

Whittemore, G., Ferrara, K., & Brunner, H. (1990). Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. *Proceedings of 28th Annual Meeting of the Association for Computational Linguistics*, 23-30.

Winograd, T. (1972). *Understanding Natural Language*. Academic.