Maximum Lexical Cohesion for Fine-Grained News Story Segmentation

Zihan Liu[1], *Lei Xie*[1], *Wei Feng*[2]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]School of Creative Media, City University of Hong Kong, Hong Kong SAR
zhliu81@gmail.com, lxie@nwpu.edu.cn, wfeng@ieee.org

Abstract

We propose a maximum lexical cohesion (MLC) approach to news story segmentation. Unlike sentence-dependent lexical methods, our approach is able to detect story boundaries at finer word/subword granularity, and thus is more suitable for speech recognition transcripts which have no sentence delimiters. The proposed segmentation goodness measure takes account of both lexical cohesion and a prior preference of story length. We measure the lexical cohesion of a segment by the KL-divergence from its word distribution to an associated piecewise uniform distribution. Taking account of the uneven contributions of different words to a story, the cohesion measure is further refined by two word weighting schemes, i.e. the inverse document frequency (IDF) and a new weighting method called difference from expectation (DFE). We then propose a dynamic programming solution to exactly maximize the segmentation goodness and efficiently locate story boundaries in polynomial time. Experimental results show that our MLC approach outperforms several state-of-the-art lexical methods.

Index Terms: story segmentation, KL-divergence, lexical cohesion, word weighting, dynamic programming, spoken document segmentation, spoken document retrieval

1. Introduction

The task of story segmentation is to partition a text, audio and/or video stream into a set of continuous segments, each addressing a single central topic. With the proliferation of multimedia contents, automatic story segmentation is highly in demand as a necessary pre-processing step for a variety of multimedia content processing tasks [1]. Lexical cues are of great interest for news story segmentation, since they reveal topic shift via semantic variation, which is independent of editorial rules.

Previous lexical story segmentation methods can be divided into two main categories. The first category is based on topic modeling, e.g. [2]. They treat word sequences as observations of some latent topics. By some optimal criterion, a sequence of topic labels is assigned to the input text or speech transcript. Then the segmentation is obtained simply by marking boundaries between every pair of adjacent parts with different topic labels. In contrast, the second category directly investigates word usage and segments the input stream into lexically cohesive parts. A typical method is TextTiling [3]. Based on an intuitive idea that different topics usually employ different sets of words, it scans the text and marks a boundary when lexical similarity of two adjacent sentences is lower than a tuned threshold. Some other methods aim at finding an optimal segmentation under some global criteria, rather than merely detecting local shifts [4, 5]. For instance, Malioutov *et al.* [5] formulated story segmentation as a sentence-level graph partitioning problem by optimizing the normalized cuts (NCuts) criterion.

Many approaches mentioned above are initially designed for segmenting pure text materials, although some of them have been borrowed to segment speech recognition transcripts of spoken documents, e.g. [6, 7]. In contrast to pure text documents, speech recognition transcripts have no sentence delimiters available. As a result, some widely-used inter-sentence similarity measures, e.g. cosine similarity, are not directly applicable, unless *pseudo*-sentences (i.e. word blocks of fixed length) are extracted first. However, using pseudo-sentence results in two major problems. First, the fact that a story boundary always appears at the end of a sentence does not hold for pseudo-sentence. Consequently, the boundaries lying in the middle of a pseudo-sentence can never be detected. Second, the appropriate length of pseudo-sentence is highly corpus or even document dependent, and has to be tuned carefully since it considerably influences segmentation accuracy [7]. Therefore, we would much prefer a fine-grained approach that does not rely on pseudo-sentences.

In this paper, we propose a maximum lexical cohesion (MLC) approach to news story segmentation. We measure the lexical cohesion of a segment by the KL-divergence from its word distribution to an associated piecewise uniform distribution. Since our measurement is at finer word/subword granularity, all possible story boundaries are detectable. Furthermore, we propose a word-weighted lexical cohesion measure to reflect the uneven contributions of different words to a story. We particularly employ two weighting schemes. One is inverse document frequency (IDF) [8], a popular weighting method in information retrieval and text mining; the other is a new weighting scheme, namely *difference from expectation* (DFE), aiming to diminish the effect of stop/prevalent words in the corpus and enhance the contribution of the words discriminating a segment from its complement in the document. By regularizing the weighted lexical cohesion with an empirical story length prior function, we finally obtain a general fine-grained segmentation goodness measure. We further show how to optimize the new criterion using a dynamic programming procedure in polynomial time. Experiments on the TDT2 Mandarin corpus show that the proposed MLC approach outperforms several state-of-the-art lexical methods based on local features detection and graph-theoretic sentence-level segmentation.

2. Segmentation Goodness Measure

2.1. Measuring Lexical Cohesion

From the statistical viewpoint and the "bag-of-words" assumption, a story or a word sequence can be viewed as a random variable, with the words in it as its observations. The number of times (i.e. the frequency) of a word occurring in the sequence indicates the probability of the word.

Different from a sequence of randomly chosen words, a news story has a central topic. The words relevant to this central topic tend to have more occurrences in the story. Thus, word repetition is a strong lexical cohesion indicator [3]. This is also the fundamental of many existing lexical story segmentation methods. Based on the fact that if a word-sequence has more cohesive word usage (i.e. more word repetitions), the uncertainty of the sequence (as a potential story) is lower. Hence, it is intuitive to measure the repetition-based lexical cohesion of a word sequence by the opposite of its entropy. However, entropy is a biased measure, since it undesirably favors shorter sequences. Instead, to model the lexical cohesion of a sequence, we use the Kullback-Leibler divergence (i.e. relative entropy) from the word usage distribution of the sequence to an associated piecewise uniform distribution.

Formally, let $V = \{w_1, w_2, \ldots, w_n\}$ be the vocabulary of n words. A word sequence $\mathbf{s}$ can be represented by a probability vector $\mathbf{P_s}$. The probability of word w_i in $\mathbf{P_s}$ is calculated as

$$\mathbf{P_s}(w_i) = \frac{\text{freq}(w_i|\mathbf{s})}{\text{len}(\mathbf{s})}, \tag{1}$$

where $\text{freq}(w_i|\mathbf{s})$ is the number of times w_i occurs in $\mathbf{s}$ and $\text{len}(\mathbf{s})$ is the length of $\mathbf{s}$. A piecewise uniform probability vector associated with $\mathbf{s}$, $\mathbf{Q_s}$, is defined as

$$\mathbf{Q_s}(w_i) = \begin{cases} \frac{1}{\text{len}(\mathbf{s})} & \text{if } w_i \text{ appears in } \mathbf{s} \\ \lambda_{\mathbf{s}} & \text{otherwise} \end{cases}, \tag{2}$$

where $\lambda_{\mathbf{s}}$ is a normalizing parameter making $\sum_{i=1}^{n} \mathbf{Q_s}(w_i) = 1.0$.

The cohesion function of $\mathbf{s}$, $\text{Co}(\mathbf{s})$, is defined as the KL-divergence from $\mathbf{P_s}$ to $\mathbf{Q_s}$, i.e.,

$$\text{Co}(\mathbf{s}) = \sum_{i=1}^{n} \mathbf{P_s}(w_i) \log \frac{\mathbf{P_s}(w_i)}{\mathbf{Q_s}(w_i)}. \tag{3}$$

Clearly, $\text{Co}(\mathbf{s})$ has a number of desirable properties as a measure of repetition-based lexical cohesion:

- The words occurring only once in segment $\mathbf{s}$ have no contribution to $\text{Co}(\mathbf{s})$;
- $\text{Co}(\mathbf{s})$ increases when $\mathbf{s}$ has sparser word usage;
- $\text{Co}(\mathbf{s})$ achieves the minimum when all words in $\mathbf{s}$ have equal occurrence frequency.

These properties make $\text{Co}(\mathbf{s})$ a promising measure of lexical cohesion. Moreover, since the sentence structure is unnecessary in it, we believe it is more flexible than those measures based on inter-sentence similarity and particularly suitable for recognition transcript of spoken document.

2.2. Word-Weighted Lexical Cohesion

For a word sequence, different words are usually not equally significant, even if they have the same frequency. We measure this phenomenon by weighting the words.

Inverse Document Frequency (IDF). IDF is a popular word weighting scheme in information retrieval and text mining. The idea is that a word occurring in many documents is not a good discriminator and should be assigned with a lower weight, and vice versa. Assume there are N_d documents in the collection and that word w occurs in n_w of them. The IDF of w is commonly given by

$$\text{IDF}(w) = \log \frac{N_d}{n_w}. \tag{4}$$

In our experiments, N_d is replaced by the number of news stories in the development set, and n_w is also counted by story, because it is story that is of relevance to the segmentation task, rather than document.

Based on IDF, the word-weighted lexical cohesion function $\text{Co}(\mathbf{s})$ is defined as

$$\text{Co}(\mathbf{s}) = \sum_{i=1}^{n} \text{IDF}(w_i)\mathbf{P_s}(w_i) \log \frac{\mathbf{P_s}(w_i)}{\mathbf{Q_s}(w_i)}. \tag{5}$$

Difference from Expectation. IDF is independent of document segmentation, hence can be pre-calculated. However, we believe that it is reasonable for segmentation algorithm to take the information of a segment into consideration when weighting the words in it. We propose another word weighting scheme which is specifically designed for the segmentation task, namely *difference from expectation* (DFE). Let $\mathcal{D}$ denote the document to be segmented and $\mathbf{s}$ be a segment in a given segmentation of $\mathcal{D}$. Assume that word w occurs t times in $\mathbf{s}$ and T times in the whole document $\mathcal{D}$. Then the weight of w in $\mathbf{s}$ is given by

$$\text{DFE}(w|\mathbf{s}) = \left| t - \frac{T}{\text{len}(\mathcal{D})} \text{len}(\mathbf{s}) \right|. \tag{6}$$

In (6), $\frac{T}{\text{len}(\mathcal{D})}$ is the density of w over $\mathcal{D}$ and $\frac{T}{\text{len}(\mathcal{D})} \text{len}(\mathbf{s})$ is the expected frequency of w.

For the words that occur frequently throughout the document (e.g. *the*, *that*) and the words that are less frequent but quite uniformly distributed (e.g. *today*, *reporter* in news programs), (6) tends to give a low weight. Further, when most occurrences of w are in $\mathbf{s}$ (i.e. $t \approx T$) and the length of $\mathbf{s}$, $\text{len}(\mathbf{s})$, is relatively short, $\text{DFE}(w|\mathbf{s})$ is likely to be large. This is reasonable because w effectively discriminates $\mathbf{s}$ from other parts of the document. Note that if $\text{len}(\mathbf{s})$ is approximately equal to the length of the entire document, i.e. $\text{len}(\mathcal{D})$, the DFE weights of all words in $\mathbf{s}$ will decrease to near zero. Thus, DFE can also avoid irregularly long segments. This character is highly desirable for the segmentation task. The corresponding weighted cohesion function using DFE is defined as

$$\text{Co}(\mathbf{s}) = \sum_{i=1}^{n} \text{DFE}(w_i|\mathbf{s})\mathbf{P_s}(w_i) \log \frac{\mathbf{P_s}(w_i)}{\mathbf{Q_s}(w_i)}. \tag{7}$$

2.3. Story Length Regularization

Besides the lexical cohesion, our measure of segmentation goodness also takes account of the general prior distribution of story length in real-world broadcast news programs. As shown in Figure 1, the news story length l approximately follows an exponential distribution:

$$\Pr(l) \propto \alpha^l \quad (0 < \alpha < 1), \tag{8}$$

where α is the suppression rate parameter. The exponential prior function reflects the similar fact described by the well-known power-law, i.e., very long news stories are rather rare, most stories of real-world news are relatively short. From our experiments, we empirically find that the exponential prior α^l is more suitable to the task of news story segmentation and has more appropriate suppression rate, as compared to the common power-law expression l^{α}.

Let $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ denote a hypothesis segmentation of document $\mathcal{D}$, which divides $\mathcal{D}$ into k segments. The goodness score of segment $\mathbf{s}_i$ is defined as

$$\text{Score}(\mathbf{s}_i) = \text{Co}(\mathbf{s}_i) \Pr[\text{len}(\mathbf{s}_i)]. \tag{9}$$

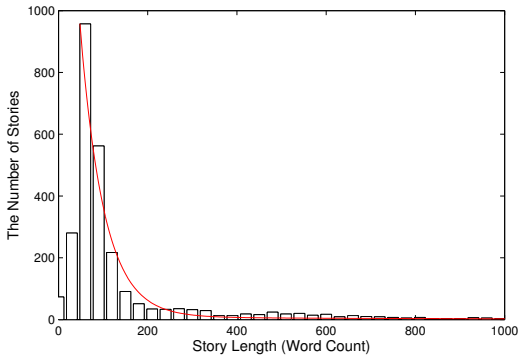

Figure 1: *The empirical histogram of the lengths of all 2,648 stories in the TDT2 Mandarin corpus and the fitted exponential distribution (red curve).*

We then define the goodness score of the whole segmentation S as the weighted average of the segment scores, i.e.,

$$\text{Score}(S) = \frac{\sum_{i=1}^{k} \text{len}(\mathbf{s}_i)\,\text{Score}(\mathbf{s}_i)}{\sum_{j=1}^{k} \text{len}(\mathbf{s}_j)}. \tag{10}$$

Therefore, seeking the optimal story segmentation $\hat{S}$ can be achieved by solving the following optimization problem

$$\hat{S} = \underset{S}{\text{argmax}}\ \text{Score}(S). \tag{11}$$

3. Dynamic Programming Solution

Due to the linear constraint of story segmentation [5], we can obtain the global maximum of (10) using the following dynamic programming (DP) algorithm in polynomial time:

$$A(j,i) = \frac{j-1}{i} F(j-1) + \big(1 - \frac{j-1}{i}\big)\,\text{Score}(\mathbf{s}_{j\leftrightarrow i})\,, \tag{12}$$

$$F(i) = \max_{1\le j\le i}\ A(j,i)\,,\ i \le \text{len}(\mathcal{D})\,, \tag{13}$$

$$B(i) = \underset{1\le j\le i}{\text{argmax}}\ A(j,i)\,,\ i \le \text{len}(\mathcal{D})\,, \tag{14}$$

$$\text{s.t. } F(0) = 0\,. \tag{15}$$

$\mathbf{s}_{j\leftrightarrow i}$ denotes the segment starting from the jth word to the ith word in document $\mathcal{D}$. $A(j,i)$ is the score of a particular segmentation of the first i words, with the last boundary lying between the $(j-1)$th and the jth word, and other boundaries forming the optimal segmentation of the first $j-1$ words. $F(i)$ is the score of the optimal segmentation of the first i words in $\mathcal{D}$. Clearly, $F(\text{len}(\mathcal{D})) = \text{Score}(\hat{S})$. $B(i)$ is used to recover the segment boundaries in $\hat{S}$.

According to the principle of Occam's razor, if there are multiple solutions to (11), we choose the one with the fewest segments. For the same reason, we merge the segments with zero scores into their respective previous neighbors.

3.1. Fast and Incremental Implementation

The time complexity of naively calculating $\text{Score}(\mathbf{s})$ is $O(|V|)$ or $O(l_\mathbf{s}\log(l_\mathbf{s}))$ depending on the implementation, where $|V|$ is the size of the vocabulary V and $l_\mathbf{s} = \text{len}(\mathbf{s})$ is the length of segment $\mathbf{s}$. Specifically, to calculate $\text{Score}(\mathbf{s})$, one needs to either maintain and go through a word frequency table of the whole vocabulary V, or sort the words in segment $\mathbf{s}$ to get the frequency vector over a subset vocabulary $V_\mathbf{s} = \{w_i | w_i \in V, \text{freq}(w_i|\mathbf{s}) > 0\}$. Therefore, calculating $\text{Score}(\mathbf{s}_{j\leftrightarrow i})$ for all possible pairs of (j,i) separately has time complexity of $O(|V|N^2)$ or $O(N^3\log(N))$, where $N = \text{len}(\mathcal{D})$ is the number of words in the document. We now introduce an incremental method to reduce the time complexity to $O(N^2)$.

We first consider the unweighted cohesion function (3), which can be rewritten as

$$\begin{aligned}\text{Co}(\mathbf{s}) &= \sum_{w_i\in V_\mathbf{s}} \frac{\text{freq}(w_i|\mathbf{s})}{\text{len}(\mathbf{s})} \log\left[\frac{\text{freq}(w_i|\mathbf{s})}{\text{len}(\mathbf{s})}\text{len}(\mathbf{s})\right] \\ &= \frac{1}{\text{len}(\mathbf{s})} \sum_{w_i\in V_\mathbf{s}} \text{freq}(w_i|\mathbf{s}) \log[\text{freq}(w_i|\mathbf{s})].\end{aligned} \tag{16}$$

Let $\tilde{w}_i$ be the ith word in $\mathcal{D}$ and $f_{\tilde{w}_i}^{j\leftrightarrow i} = \text{freq}(\tilde{w}_i|\mathbf{s}_{j\leftrightarrow i})$ denote the frequency of $\tilde{w}_i$ in segment $\mathbf{s}_{j\leftrightarrow i}$. From (16), we have

$$\begin{aligned}\text{Co}(\mathbf{s}_{j\leftrightarrow i}) = \frac{1}{i-j+1}\Big\{&(i-j)\,\text{Co}(\mathbf{s}_{j\leftrightarrow i-1}) \\ &+ f_{\tilde{w}_i}^{j\leftrightarrow i}\log(f_{\tilde{w}_i}^{j\leftrightarrow i}) - (f_{\tilde{w}_i}^{j\leftrightarrow i}-1)\log(f_{\tilde{w}_i}^{j\leftrightarrow i}-1)\Big\}.\end{aligned} \tag{17}$$

Using (17), we can calculate $\text{Co}(\mathbf{s}_{j\leftrightarrow i})$ given $\text{Co}(\mathbf{s}_{j\leftrightarrow i-1})$ in $O(1)$. Then $\text{Score}(\mathbf{s}_{j\leftrightarrow i})$ can be immediately obtained by (9). Thus, we can obtain $\text{Score}(\mathbf{s}_{j\leftrightarrow i})$ for all pairs of (j,i) in $O(N^2)$ time. Similarly, the above procedure also applies to the word-weighted cohesion functions (5) and (7). Therefore, the overall time complexity of the proposed DP solution (12)–(15) is $O(N^2)$.

4. Experiments

We carried out experiments on the TDT2 Mandarin BN corpus[1], which contains about 53 hours of VOA Mandarin Chinese broadcast news audio. Manually annotated story boundaries and word-level speech recognition transcripts of the audio recordings are provided. The TDT2 audio was transcribed by the Dragon LVCSR with word, character and base syllable error rates of 37%, 20% and 15%, respectively. The 177 audio recordings are divided into two non-overlapping sets: a development set of 90 recordings for parameter tuning and a set of 87 recordings for performance testing. According to TDT2, a detected story boundary is considered correct if it lies within a 15-second tolerant window on each side of a manually-annotated reference boundary.

Since subword n-gram units are robust to speech recognition errors and out-of-vocabulary (OOV) words in Mandarin broadcast news [6], we applied the proposed method on both word level and n-gram character/syllable levels. For the syllable level, we extracted syllable sequences of the word-level speech recognition transcripts with a home-grown Pinyin lexicon. The balanced F-measure, i.e. the harmonic mean of precision and recall, was adopted as the evaluation criterion.

4.1. Results and Analysis

Table 1 summarizes the experimental results of the proposed MLC approach with different weighting schemes, i.e., the unweighted cohesion function (3) and the two weighted cohesion functions (5) and (7). We observe that IDF performs best on the word and unigram levels and the proposed DFE shows superior performance on the bigram, trigram and quadgram levels. The MLC approach with the DFE weighting scheme on the character bigram level achieves the highest F1-measure of 0.7230.

[1] http://projects.ldc.upenn.edu/TDT2/

Table 1: *Story segmentation results (F1-measure) for the proposed MLC approach with different weighting schemes*

Weighting Scheme	Word		Unigram		Bigram		Trigram		Quadgram	
	Char.	Syl.	Char.	Syl.	Char.	Syl.	Char.	Syl.	Char.	Syl.
Unweighted	0.6090	0.6111	0.6233	0.6138	0.6777	0.6736	0.6647	0.6681	0.6305	0.6308
IDF	**0.6797**	**0.6821**	**0.6738**	**0.6527**	0.7106	0.7133	0.6546	0.6598	0.5823	0.5805
DFE	0.5003	0.5052	0.5683	0.5559	**0.7230**	**0.7200**	**0.7005**	**0.7015**	**0.6486**	**0.6495**

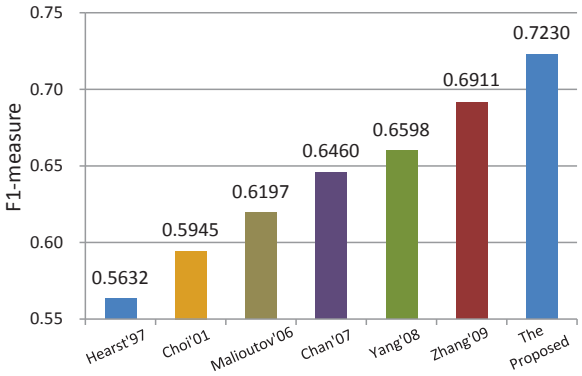

Figure 2: *Experimental results of several lexical-cue-based approaches on the TDT2 corpus.*

On the word level and the unigram subword level, the unweighted version and the DFE version of the proposed MLC method have inferior performances as compared with the IDF version. For the word level, we believe the inferior results are mainly due to the high word error rate of the LVCSR (37%). By contrast, the subword levels have relatively low error rates (20% for character and 15% for syllable). However, due to the low discriminative capacity of unigrams, it incurs many term repetitions that have nothing to do with semantic similarity. IDF is robust to these problems because it is a statistic of the whole development set, which only cares whether a term occurs in a certain document (story) or not and does not concern term repetition. Thus IDF is inherently robust to recognition errors and misleading unigram repetitions.

We observe that all three versions of the proposed method have superior performance on the bigram subword levels. This is mainly due to the fact that the most frequently used words in Chinese are bi-character and bigram subwords are robust to speech recognition errors and OOV words [6].

The DFE version shows superior performance on the trigram and quadgram subword levels as compared with the unweighted version and IDF version. This is because most of trigram and quadgram subwords are merely combinations of characters (syllables) from adjacent words. Recurrence of a trigram or quadgram is often due to two or more words that frequently recur in a particular story. As a result, recurrence of a trigram or quadgram tends to be a local phenomenon, and effectively discriminates the area it occurs in from other parts of the document. Nevertheless, IDF cannot capture this kind of feature because it does not care in-story recurrence. On the contrary, DFE can effectively capture it since DFE is very sensitive to local high density of a term that is rare in the rest of the document. This explains why the DFE version achieves clearly better results on the trigram and quadgram subword levels.

For performance comparison, we re-implemented several previous lexical-similarity-based segmentation approaches, including TextTiling [3] (Hearst'97), LSA-based TextTiling [9] (Choi'01), NCuts [5] (Malioutov'06), modeling lexical chains' statistical behavior [1] (Chan'07), subword-LSA-based TextTiling [6] (Yang'08) and subword-based NCuts [7] (Zhang'09). The performances of these approaches on the TDT2 corpus are shown in Figure 2. We can see that the proposed MLC approach outperforms the others. For instance, it achieves a relative improvement of 4.62% over the subword-based NCuts approach.

5. Conclusions

This paper has proposed a maximum lexical cohesion (MLC) approach to news story segmentation. Different from inter-sentence similarity based approaches, the proposed MLC approach works at finer word granularity. This makes it very suitable for spoken document segmentation since speech recognition transcripts have no sentence delimiters. Our contributions are: (1) a segmentation goodness measure that takes account of both lexical cohesion and a prior preference of story length; (2) a novel word weighting scheme called *difference from expectation* (DFE); (3) a dynamic programming algorithm for finding the optimal segmentation in polynomial time. Experiments show that the proposed MLC approach outperforms several state-of-the-art lexical methods; the two word-weighting schemes, i.e. the classical IDF and the proposed DFE, have respective advantages at different word/subword levels in broadcast news story segmentation.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (60802085), the Research Fund for the Doctoral Program of Higher Education (20070699015), the Program for New Century Excellent Talents in University and the NPU Foundation for Fundamental Research (W018103).

7. References

[1] S.-K. Chan, L. Xie, and H. M.-L. Meng, "Modeling the statistical behavior of lexical chains to capture word cohesiveness for automatic story segmentation," in *Interspeech*, 2007, pp. 2408–2411.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] M. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[4] P. Fragkou, V. Petridis, and A. Kehagias, "A dynamic programming algorithm for linear text segmentation," *Journal of Intelligent Information Systems*, vol. 23, no. 2, pp. 179–197, 2004.

[5] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *Proc. ACL*, 2006, pp. 25–32.

[6] Y. Yang and L. Xie, "Subword latent semantic analysis for TextTiling-based automatic story segmentation of Chinese broadcast news," in *Proc. ISCSLP*, 2008, pp. 358–361.

[7] J. Zhang, L. Xie, W. Feng, and Y. Zhang, "A subword normalized cut approach to automatic story segmentation of chinese broadcast news," in *Proc. AIRS*, 2009, pp. 136–148.

[8] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[9] F. Y. Y. Choi, P. Wiemer-hastings, and J. Moore, "Latent semantic analysis for text segmentation," in *Proc. EMNLP*, 2001.