Multi-scale TextTiling for Automatic Story Segmentation in Chinese Broadcast News

Lei Xie[1], Jia Zeng[2], and Wei Feng[3]

[1] Audio, Speech & Language Processing Group (ASLP),
School of Computer Science,
Northwestern Polytechnical University, Xi'an, China
[2] Department of Electronic Engineering,
City University of Hong Kong, Hong Kong SAR
[3] School of Creative Media,
City University of Hong Kong, Hong Kong SAR
lxie@nwpu.edu.cn, {j.zeng, wfeng}@ieee.org

Abstract. This paper applies Chinese subword representations, namely character and syllable n-grams, into the TextTiling-based automatic story segmentation of Chinese broadcast news. We show the robustness of Chinese subwords against speech recognition errors, out-of-vocabulary (OOV) words and versatility in word segmentation in lexical matching on errorful Chinese speech recognition transcripts. We propose a multi-scale TextTiling approach that integrates both the specificity of words and the robustness of subwords in lexical similarity measure for story boundary identification. Experiments on the TDT2 Mandarin corpus show that subword bigrams achieve the best performance among all scales with relative f-measure improvement of 8.84% (character bigram) and 7.11% (syllable bigram) over words. Multi-scale fusion of subword bigrams with words can bring further improvement. It is promising that the integration of syllable bigram with syllable sequence of word achieves an f-measure gain of 2.66% over the syllable bigram alone.

Key words: story segmentation, topic segmentation, spoken document segmentation, TextTiling, multi-scale fusion, spoken document retrieval, multimedia retrieval

1 Introduction

Story segmentation is the task of dividing text, audio or video into homogenous regions, each addressing a single central topic. It is a necessary pre-processing step for a wide range of speech and language processing tasks, namely topic tracking, summarization, information extraction, indexing and retrieval. These tasks usually assume the presence of individual 'documents'. For broadcast news (BN), a major media channel in the information era, the objective of story segmentation is to segment continuous audio/video streams into news stories. Manual segmentation requires annotators to work through the whole audio/video,

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 345–355, 2008.

which takes a large amount of time that makes it an intractable task. To perform automatic segmentation, three kinds of cues have been explored, namely acoustic/prosodic cues from audio, lexical cues from speech recognition transcripts or video captions and video cues such as anchor face and color histograms.

Borrowed from traditional text segmentation techniques, lexical-based story segmentation in BN is mainly performed on errorful text transcribed from audio stream via a large vocabulary continuous speech recognizer (LVCSR). Main approaches include word cohesiveness [1], [2], use of cue phrases and modeling (e.g. hidden Markov model [3]). *TextTiling* [1], a classical word-cohesiveness-based approach, has been recently introduced to segmenting spoken documents such as BN [4] and meetings [5] due to its simplicity and efficiency. This approach is based on a straightforward argument that different topic usually employs different set of words. Shifts in word usage are indicative of changes in topic. Therefore, lexical similarity measure between consecutive word chunks is performed across the text and a local similarity minimum indicates a possible topic shift.

Despite of receiving a considerable amount of attention from TREC SDR, TDT and TRECVID evaluation programs, state-of-the-art story segmentation error rates on BN transcripts remain fairly high. This is largely because of the high level of speech recognition error rates, e.g. about 30% for English and about 40% for Chinese Mandarin and Arabic in terms of the word error rate (WER) reported in TRECVID 2006. Besides the errors caused by harsh acoustic conditions (especially for 'field' speech) and diverse speaking styles from various speakers (anchors, reporters and interviewees), the out-of-vocabulary (OOV) problem remains the major obstacle for broadcast news segmentation. New words are introduced continuously from growing multimedia collections and words outside the speech recognizer vocabulary cannot be correctly recognized. Specifically for Chinese BN, the OOV words are largely named entities (e.g. Chinese person names and transliterated foreign names) that are highly related to topics. These OOV words induce incorrect lexical similarity measures across the word stream and thus decrease the segmentation performance greatly.

Recently, the use of *subword* indexing units (e.g. phonemes, syllables and sub-phonetic segments) has been shown to be very helpful to alleviate problems of speech recognition errors and OOV words in the spoken document retrieval (SDR) task [6]. Especially for Chinese, retrieval based on character or syllable indexing units is superior to words due to the special features of Chinese, namely character-based, monosyllabic and flexible wording structure [7], [8]. In this paper, we propose to apply subword units (characters and syllables) in automatic story segmentation of Chinese broadcast news. We present a *multi-scale* TextTiling approach, in which lexical similarities are measured in multiple scales (word and subword scales) and integrated in two schemes, namely representation fusion and score fusion. We aim at fusing the specificity of words and the robustness of subwords to improve story segmentation performance on errorful speech recognition transcripts.

2 Corpus

We experiment with the TDT2 Mandarin corpus[1] that contains about 53 hours of Voice of America (VOA) Mandarin Chinese broadcast news with time span from February to June, 1998. The 177 audio files are accompanied with manually annotated story boundary files and word-level speech recognition transcripts. In TDT2, speech recognition was performed by the Dragon LVCSR with word, character and syllable error rates of 37%, 20% and 15%, respectively. We adopt a home-grown Pinyin lexicon to get the syllable sequences of words. The corpus covers about 2,907 news stories, and the average story duration is 65 seconds (142 words, 248 characters). We separate the corpus into three parts: a quarter as the training set, a quarter as the development set and another half as testing set for evaluation. According to the definition of TDT2, a detected story boundary is considered correct if it lies within a fifteen-second tolerant window on each side of a hand-annotated reference boundary (ground truth).

3 Robustness of Chinese Subwords

Robustness to Speech Recognition Errors. Chinese language is fundamentally different from western languages. Chinese is *character-based* while English is alphabetical. Each Chinese character is pronounced as a tonal syllable (known as the *monosyllabic* feature) and almost every Chinese character is a morpheme with its own meaning. A 'word' is made up of one or several characters. Chinese syllables are tonal because syllables with different lexical tones convey different meanings[2]. There are four lexical tones and a neutral tone in Mandarin Chinese. About 1200 phonologically allowed tonal syllables correspond to over 6500 commonly used simplified Chinese characters. When tones are disregarded, the 1200 tonal syllables are reduced to only about 400 base syllables. This means that the small number of syllables implies large number of *homonym* characters sharing the same syllable. Therefore in errorful Chinese ASR transcripts, it is common that a word is substituted by another character sequence with the same or similar pronunciations, in which homophone characters are the probable substitutions. There are considerable possibilities that a word (or part of a word) is substituted with their homophones in the ASR transcripts. Matching at syllable scale is robust to this kind of recognition errors.

Table 1 shows some word matching failures extracted from the TDT2 corpus. In Table 1, (a) illustrates the case that the foreign person name 哈里斯 is substituted by another person name 哈里森 in its neighborhood in an ASR transcript. Although the first two characters (哈 and 里) are correctly recognized, a single mis-recognized character (the third character 斯 was replaced by 森) will make the word level match fail. In story segmentation approaches based

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

[2] Lately, we have shown that the tonality of Mandarin Chinese affects the pitch reset cue for prosody-based story segmentation in Chinese broadcast news. Please refer to [9] for details.

Table 1. Samples of word matching failures due to speech recognition errors

#	Character sequence	Syllable sequence	English translation
(a)	哈里斯	/ha-li-si/	*Harris*
	哈里森	/ha-li-sen/	*Harrison*
(b)	阿尔及利亚	/a-er-ji-li-ya/	*Algeria*
	鲍尔 激励 要	/bao-er ji-li yao/	*Bauer inspire want*
(c)	股市	/gu-shi/	*stock exchange*
	故事	/gu-shi/	*story*

on word cohesiveness, e.g. TextTiling, this failure will induce incorrect lexical similarity measures. However, if character or syllable matching is adopted, the first two characters 哈 and 里 in the two different words still can be matched. Another failure in word matching is shown in (b), where the country name 阿尔及利亚 was mis-recognized as a sequence of three distinct words {鲍尔 激励 要} with similar pronunciations as word 阿尔及利亚. In this case, syllable matching can recall largely of the country name because character sequence 尔激励 has the same syllable representation with 尔及利 in the correct word 阿尔及利亚. (c) shows another failure, where the word 股市(*stock exchange*) was substituted by another word 故事(*story*) with character homophones. In this case, syllable matching can still link the two words together.

Robustness to OOV words. The limited number of Chinese characters with different meanings can be combined to produce unlimited Chinese words. As a result, there does not exist a commonly accepted lexicon for Chinese since new words are born everyday. Therefore, the monosyllabic feature makes the OOV problem more pronounced in Chinese spoken document segmentation, especially in the broadcast news domain. An OOV word distributed in different places of a spoken document may be substituted by several totally different character strings with the same (or partially same) syllable sequence. For example, foreign proper names are common OOV words in Chinese spoken documents as they are transliterated to Chinese character sequences based on the pronunciations (i.e. phonetic transliteration). As a result, speech recognizer may return different character sequences with the same or similar pronunciations.

Table 2 shows two OOV words from the TDT2 corpus. In (a), the OOV word, 尼姆佐夫 (*Nimzov*, a Russian person name), was substituted by four different character sequences (mainly come as singletons) within a news story. Lexical similarity measure at syllable level can partially recover this highly-topical-related OOV word because the last two syllables are the same for the four recognition outputs (all /*zuo fu*/). In (b), the former Korean president 金大中 (*Kim Dae-Jung*) was an OOV word and mis-recognized as three singleton sequences. Lexical similarity measure at syllable scale will successfully recover this OOV word since the three recognition outputs have the same syllable sub-sequence /da zhong/.

Table 2. Samples of word matching failures due to the OOV problem

#	Character sequence		Syllable sequence
(a)	OOV word	尼姆佐夫 (Nimzov)	/ni-mu-zuo-fu/
	Recognizer output	名 模 作 福	/ming mo zuo fu/
		你 目 作 赋	/ni mu zuo fu/
		英国 作 赋	/ying-guo zuo fu/
		你 没 坐 夫	/ni mei zuo fu/
(b)	OOV word	金大中 (Kim Dae-Jung)	/jin-da-zhong/
	Recognizer output	金 大 中	/jin da zhong/
		竞 达 中	/jing da zhong/
		近 大众	/jin da-zhong/

Robustness to Versatility in Word Segmentation. Words and sentences in Chinese text appear as a sequence of characters. Different from English, there are no blanks in Chinese text serving as word boundaries. As a result, 'word' is not clearly defined in Chinese. Consequently, word segmentation in Chinese texts is an ambiguous process and definitely not unique. For LVCSR, the same character sequence in different places might be recognized as several different word sequences that both syntactically valid and semantically meaningful. Therefore, the flexible wording structure of Chinese may contribute considerably to the word recognition error. For example, 北京市领导 (Translation: *leaders of the Beijing city*) can be segmented to words by the speech recognizer as:

a). 北京市 领导 (Translation: *Beijing city leaders*);
b). 北京 市 领导 (Translation: *Beijing city leaders*);
c). 北京 市领导 (Translation: *Beijing city leaders*).

Although they are indicating the same, word matching cannot link the three together. Specifically, matching between 北京市 (Translation: *Beijing city*) in a) and 北京 (Translation: *Beijing*) in b) leads to a failure. This problem can be solved easily by character matching.

4 Multi-scale TextTiling

We have demonstrated that Chinese subword units are robust to speech recognition errors, OOV words and versatility in word segmentation in imperfect Chinese ASR transcripts. This motivates the use of characters and syllables in TextTiling for story segmentation of Chinese BN. In this section, we first describe our TextTiling-based story segmentation algorithm, and then define the subword overlapping n-grams used in lexical similarity measure in subword scales. Finally, we explore the fusion schemes of multiple lexical representations (words and subwords n-grams) for further improving the story segmentation performance.

4.1 TextTiling-Based Story Segmentation

The classical TextTiling algorithm is composed of three steps: tokenization, lexical score determination and boundary identification [1]. The tokenization step splits a character stream into words, i.e. word segmentation. As the speech recognition outputs are word-level text transcripts in the TDT2 corpus, we thus bypass the tokenization step and jump to the rest two steps.

Lexical Score Determination. The Text Tiling algorithm first divides the text document into sentences (or pseudo-sentences). At each inter-sentence gap along the text stream, adjacent windows of fixed number of sentences are compared in terms of lexical similarity. For ASR transcripts of spoken documents, sentence boundaries are not readily available. A possible way for story boundary investigation is to perform lexical similarity measure at places with significant pauses (a useful prosodic cue). To evaluate story segmentation performance *purely* based on the lexical information, we divide the ASR transcripts into blocks/windows of fixed number of words (W). The *lexical score* between adjacent windows (at inter-window gap g) is calculated by cosine similarity:

$$lexscore(g) = \cos(\mathbf{v}_l, \mathbf{v}_r) = \frac{\sum_{i=1}^{I} v_{i,l} v_{i,r}}{\sqrt{\sum_{i=1}^{I} v_{i,l}^2 \sum_{i=1}^{I} v_{i,r}^2}} \quad (1)$$

where $\mathbf{v}_l$ and $\mathbf{v}_r$ are the term frequency vectors for the two adjacent windows (left and right windows to the inter-window gap). $v_{i,l}$ is the i^{th} element of $\mathbf{v}_l$, i.e., the term frequency of word i registered in the vocabulary (with size of I).

Since story boundaries are searched at inter-window gaps, we increase the boundary hypotheses by sliding. The lexical scores are calculated at positions of $J, 2J, 3J \cdots$ word positions along the ASR transcript, where J is the sliding length and $J \leq W$.

Boundary Identification. Boundary identification can be carried out directly on the time trajectory of lexical score. However, TextTiling adopts the relative score information instead of the absolute values. The inter-window gaps whose similarity valleys represent a 'valley' are considered for boundary identification. Specifically, *depth score* is calculated for each valley point on the lexical score time trajectory. Denote the valley point as v, and the nearest left and right peaks, p_l and p_r, around the valley points, the depth score of valley v is

$$depthscore(v) = (lexscore(p_l) - lexscore(v)) + (lexscore(p_r) - lexscore(v)). \quad (2)$$

Note that every non-valley point is given a depth score of 0. The depth score considers a sharp drop in lexical similarity as more indicative of a story boundary than a gentle drop. This helps make decisions in the cases in which a inter-window gap's lexical score falls into the middle of the lexical score range, but is flanked by tall peaks on either side. This situation happens commonly enough to be possible story boundaries. Boundary identification is performed on the depth score time trajectory, in which a point whose depth score exceeds a pre-set threshold θ is considered as a story boundary.

4.2 Subword Overlapping *N*-grams

We perform lexical similarity measures in subword scales by subword overlapping n-grams. For a sequence of m subword units (characters or syllables) $\{S_1 S_2 S_3 \cdots S_m\}$, the subword overlapping bigram and trigram are formed as

$$\{S_1S_2\ S_2S_3\ S_3S_4 \cdots S_{m-1}S_m\} \text{ and} \tag{3}$$

$$\{S_1S_2S_3\ S_2S_3S_4\ S_3S_4S_5 \cdots S_{m-2}S_{m-1}S_m\} \tag{4}$$

respectively. Other n-grams can be listed accordingly. To reduce the possibility of missing any useful information embedded in the subword sequence, overlapping between subwords is used. Term frequency vectors, lexical scores and depth scores are calculated on sequences of subword overlapping n-gram units transformed from ASR word transcripts.

4.3 Fusion of Multi-scale Representations

We propose to combine multiple lexical scales (words and subwords) for improving story segmentation performance by taking advantage of different scales. We present two different fusion strategies: *representation fusion* and *score fusion*.

Representation Fusion. Representation fusion merges lexical representations from different scales before the lexical similarity measure. The term frequency vectors for all scales are combined to form a concatenated vector with $\sum_{k=1}^{K} I^k$ dimensions, where I^k denotes the dimension of scale k. The concatenated vector is formulated as

$$\mathbf{v} = [w_1 \cdot \mathbf{v}^1, w_2 \cdot \mathbf{v}^2, \cdots, w_K \cdot \mathbf{v}^K], \tag{5}$$

where w_k denotes the fusion weight for scale k and $\sum_{k=1}^{K} w_k = 1$. The fusion weight is used to reflect the importance of each lexical scale. The lexical score and the depth score are thus calculated on the concatenated vector, and boundary identification is performed on the time trajectory of depth score.

Score Fusion. In score fusion, the lexical scores are calculated for all scales separately, and then linearly integrated to a combined score by

$$lexscore = \sum_{k=1}^{K} w_k \cdot lexscore^k, \tag{6}$$

where $lexscore^k$ is the lexical score for scale k calculated on the term frequency vector $\mathbf{v}^k$ and $\sum_{k=1}^{K} w_k = 1$. Subsequently the depth scores are calculated and story boundary hypotheses are made.

5 Evaluations

We have carried out story segmentation experiments on the TDT2 corpus described in Section 2. We evaluate our story segmentation approach by comparing

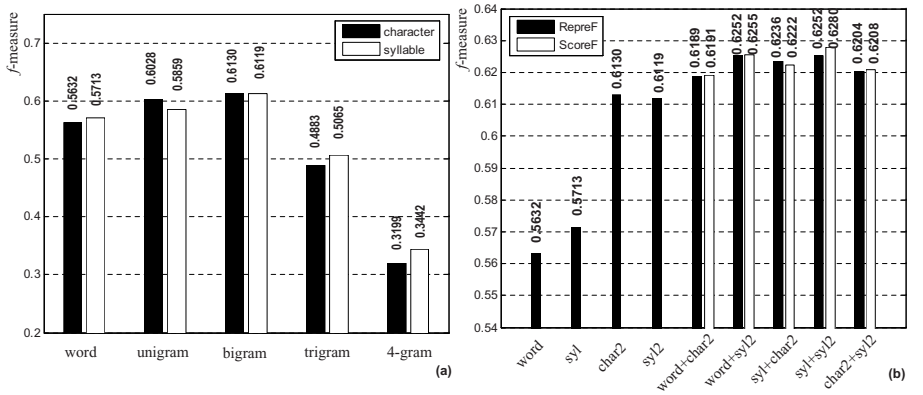

Fig. 1. Experimental results on (a) individual scales and (b) multi-scale fusion. Syl: syllable sequence of word; char2: character bigram; syl2: syllable bigram.

the automatically detected boundaries to the human-labeled boundaries. Recall, precision and their harmonic mean, f-measure, are adopted as the evaluation metrics. The training set is used to optimize the window size (W), the sliding length (J) and the boundary identification threshold (θ). An empirical selection procedure is adopted on the testing set to choose the best W, J and θ maximizing the f-measure for the word scale. Experiment shows that the combination of $W = 50$ and $J = 10$ achieves the best f-measure. The selected W and J are then fixed in the empirical searching for the best θ on each subword scale. For multi-scale fusion experiments, we use the development set to find the optimal fusion weights, w_k for each fusion form. We iterate through weights from 0 to 1.0 at intervals of 0.1 to seek the best weights that maximize the f-measure on the development set. Story segmentation results are reported for the testing set.

Experimental Results on Individual Scales. Story segmentation results for individual lexical scales are shown in Fig. 1 (a). We observe that subword bigrams perform the best as compared with other lexical scales in story segmentation of Chinese BN. The f-measure for character bigram and syllable bigram are 0.6130 and 0.6119, respectively. Promisingly, the relative improvement of subword bigrams over words are as high as 8.84% (characters) and 7.11% (syllables). This improvement is mainly because of the robustness of bigrams against speech recognition errors, OOV words and versatility in word segmentation, as illustrated in Section 3. This observation also accords with Chinese SDR using subword units, where subword bigrams are also superior to other scales [7], [8].

Another observation is that as n increases from one to four for the subword n-grams, performance reaches the peak at bigram and gradually drops from bigram to 4-gram. It can also be seen that subword unigrams also bring considerable improvement over words, performing the second best. The subword trigram and 4-gram perform worse even than the word scale. This can be explained by the fact that most frequently used words in Chinese are bi-character, and the proba-

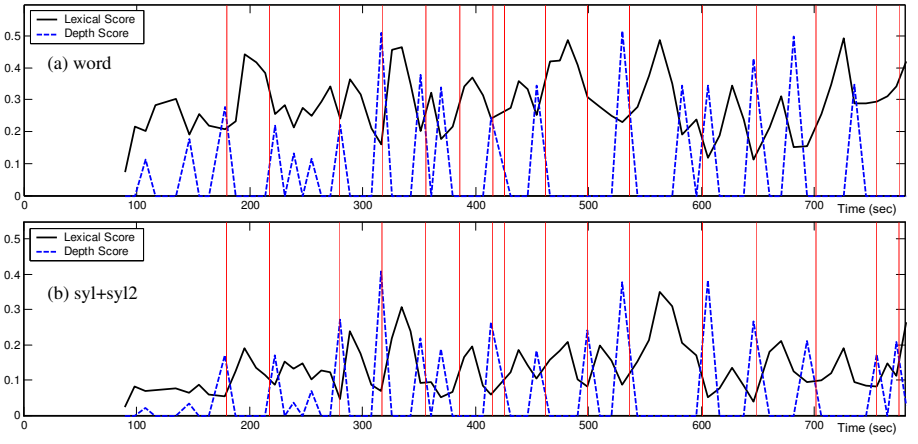

Fig. 2. Lexical score and depth score time trajectories for a broadcast news audio file (partial) in the testing set, calculated using (a) word scale alone and (b) fusion between syllable sequence of words and syllable bigrams (*syl+syl2*, score fusion). The vertical lines (in red) show the positions of human-labeled boundaries.

bility of long sequences with correctly recognized characters is smaller than two character units.

Experimental Results on Multi-Scale Fusion. Since subword bigrams show superior performance among other scales, we run experiments on multi-scale fusion that integrates both subword bigrams and words for TextTiling-based story segmentation. Results in Fig. 1 (b) shows that multi-scale fusion can bring further improvement to the story segmentation performance. All multi-scale integration forms under investigation outperform the single scales. The best performance is achieved when syllable bigrams are combined with syllable sequence of words (*syl+syl2*) in terms of score fusion. The f-measure is as high as 0.6280 and the performance gain of 2.66% is achieved over using the syllable bigram alone. The integration of words and syllable bigrams (*word+syl2*) offers the second best performance with f-measures of 0.6255 for score fusion and 0.6252 for representation fusion. When syllable bigrams and character bigrams are integrated (*char2+syl2*), the improvement is not as salient as the integration between words and subword bigrams. This is probably because the two subword bigram scales carry similar information (which can be seen from their very close performance in Fig. 1 (a)) and neither have the specificity of words. Results also show that score fusion marginally outperforms representation fusion.

Fig. 2 illustrates the lexical score and depth score time trajectories for a broadcast news clip in the testing set, where (a) is calculated using the word scale alone and (b) the fusion of syllable sequence of word and syllable bigrams (*syl+syl2*, score fusion). We can clearly see that some of the story boundary misses and false alarms in (a) are corrected in (b). This performance gain is mainly due to the robustness of the syllable bigram.

6 Conclusions and Future Work

In this paper, we have applied subword representations (characters and syllables) into story segmentation of Chinese broadcast news. We have shown the robustness of Chinese subwords in lexical matching in errorful speech recognition transcripts. This have motivated us to propose a multi-scale TextTiling story segmentation approach to integrate both the word and subword scales. This approach aims at combining the specificity of words and the robustness of subwords. Two different fusion schemes have been adopted, namely representation fusion and score fusion. Experimental results on the TDT2 Mandarin corpus show that subword bigrams achieve the best performance among all scales with relative f-measure improvement of 8.84% (character bigram) and 7.11% (syllable bigram) over words. Multi-scale fusion experiments demonstrate that integration of subword bigrams with words can bring further improvement for both fusion schemes. The integration between syllable bigram and syllable sequence of words achieves an f-measure gain of 2.66% over the syllable bigram alone.

There is still substantial work to be done. The overall story segmentation performance of the TextTiling-based approach is not high. This is mainly because the TextTiling algorithm is based on local lexical similarity measure, and thus sensitive to sub-topics changes. This motivates us to integrate both local and global measurements in the future work. In addition, we will experiment with the fusion between lexical cues and prosodic cues to further improve the story segmentation performance in Chinese broadcast news.

Acknowledgements

This work is supported by the Research Fund for the Doctoral Program of Higher Education in China (Program No. 20070699015), the Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2007F15), NPU Aoxiang Star Plan (Program No. 07XE0150) and NPU Foundation for Fundamental Research.

References

1. Hearst, M.A.: TexTiling: Segmenting text into multiparagraph subtopic passages. Computational Linguistics 23(1), 33–64 (1997)
2. Chan, S.K., Xie, L., Meng, H.: Modeling the statistical behavior of lexical chains to capture word cohesiveness for automatic story segmentation. In: Proc. Interspeech, pp. 2851–2854 (2007)
3. Yamron, J., Carp, I., Gillick, L., Lowe, S., van Mulbregt, P.: A hidden Markov model approach to text segmentation and event tracking. In: Proc. ICASSP, pp. 333–336 (1998)
4. Rosenberg, A., Hirschberg, J.: Story segmentation of broadcast news in English, Mandarin and Arabic. In: Proc. HLT-NAACL, pp. 125–128 (2006)
5. Banerjee, S., Rudnicky, I.A.: A TextTiling based approach to topic boundary detection in meetings. Proc. Interspeech (2006) 57–60

6. Ng, K.: Subword-based approaches for spoken document retrieval. Ph.D. Thesis of MIT (2000)
7. Chen, B., Wang, H.M., Lee, L.S.: Discriminating capabilites of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese. IEEE Transactions on Speech and Audio Processing 10(5), 202–314 (2002)
8. Lo, W.K., Meng, H., Ching, P.C.: Multi-scale spoken document retrieval for Cantonese broadcast news. International Journal of Speech Technology 7(2-3), 1381–2416 (2004)
9. Xie, L., Liu, C., Meng, H.: Combined Use of Speaker- and Tone-Normalized Pitch Reset with Pause Duration for Automatic Story Segmentation in Mandarin Broadcast News. In: Proc. HLT-NAACL, pp. 193–196 (2007)