SUBWORD LATENT SEMANTIC ANALYSIS FOR TEXTTILING-BASED AUTOMATIC STORY SEGMENTATION OF CHINESE BROADCAST NEWS

Yulian Yang, Lei Xie

Audio, Speech and Language Processing Group, School of Computer Science,
Northwestern Polytechnical University, Xi'an
lxie@nwpu.edu.cn, ylyang@nwpu-aslp.org

ABSTRACT

This paper proposes to perform latent semantic analysis (LSA) on character/syllable n-gram sequences of automatic speech recognition (ASR) transcripts, namely subword LSA, as an extension of our previous work on subword TextTiling for automatic story segmentation of Chinese broadcast news. LSA represents the 'meaning' of a lexical term by a feature vector conveying the term's relations with other terms. We apply subword LSA vectors to the measurement of inter-sentence lexical score in TextTiling-based story segmentation. Subword n-grams are robust to speech recognition errors, especially out-of-vocabulary (OOV) words, in lexical matching on Chinese ASR transcripts. This work combines the concept matching merit of LSA and the robustness of subwords. Experimental results on the TDT2 Mandarin corpus show that subword-LSA-based TextTiling can effectively improve the story segmentation performance. Character-bigram-LSA-based TextTiling achieves the best F1-measure of 0.6598 with relative improvement of 17.4% over the conventional word-based TextTiling and 6.5% over our previous syllable-bigram-based TextTiling.

Index Terms— latent semantic analysis, TextTiling, story segmentation, topic segmentation, spoken document retrieval, subword

1. INTRODUCTION

Automatic story segmentation aims at segmenting a text, audio or video stream into individual stories, each addressing a single central topic. It serves as a preprocessing step for subsequent tasks such as topic tracking, summarization, information extraction, indexing and retrieval, because these tasks usually assume the presence of individual topical 'documents'. Specifically for a broadcast news (BN) retrieval task, continuous audio/video streams have to be divided into distinct news stories before retrieval since users are expecting short clips of relevant news stories rather than an entire news stream.

To perform automatic story segmentation, acoustic/prosodic cues [1], video cues [2] and lexical/texual cues have been extensively explored. Main lexical approaches involve lexical cohesion [3], use of cue phrases [4] and modeling [5]. Lexical-cohesion-based approaches make use of the inter-word semantic relations, e.g. repetition, synonymy, and part-whole relation, which 'hang words together' within a topic. TextTiling [6] is a classical lexical-cohesion-based text segmentation approach that has been recently introduced to segmenting spoken documents such as BN [7] and meetings [8] because of its simplicity and efficiency. This approach assumes that different topic often employs different set of words, and a shift of word usage may signal a topic boundary. It identifies story boundaries by detecting lexical similarity minima across the text stream. Choi *et. al.* have introduced latent semantic analysis (LSA) [9] in formulating the similarity measure of lexical-cohesion-based text segmentation. Conventional lexical cohesion relies on rigid word/term matching, while LSA is capable of simulating human cognitive ability on *concept matching*. LSA-based story segmentation employs the contextual-usage meaning of words by principle components analysis (PCA) and improves separability among different topics over conventional lexical approaches [10].

Prior research on lexical-based story segmentation has been applied to clean texts, while segmentation on spoken documents (e.g. BN) has to be performed on errorful texts transcribed from audio via a large vocabulary continuous speech recognizer (LVCSR). Speech recognition errors induce noises on texts and break lexical cohesion, causing lexical similarity measures and resulting in both word and concept matching failures. The word error rate of the state-of-the-art transcription system is fairly high, for example, about 30% for English and 40% for Mandarin reported in TRECVID 2006. Besides errors caused by adverse acoustic conditions and diverse speaking styles, the out-of-vocabulary (OOV) words for words outside the vocabulary of the speech recognizer account for a large part of the error rate. The OOV problem is especially serious for Chinese because of the open vocabulary nature and the feasible wording structure of the Chinese language. Chinese OOV words are largely named entities that are key to topics, and their mis-recognition remains the major obstacle for BN segmentation.

Motivated by the successful use of subword indexing units in spoken document retrieval [11], we have recently applied Chinese subword units (characters and syllables) in TextTiling-based automatic story segmentation of Chinese BN [7]. Subwords have the advantages of partial matching, because the incorrectly recognized words may include several subword units correctly recognized. Especially for Chinese, lexical similarity measure on character or syllable level is superior to words due to the special features of Chinese [7]. In this paper, we propose to measure lexical similarities by subword LSA, as an extension of our previous work. We believe that employing latent semantics on subword level is more effective than word level due to the partial matching merit of subwords. We demonstrate the advantages of subword matching in errorful speech recognition transcripts and perform story boundary detection by lexical similarities measured on subword LSA vectors.

2. CORPUS

We experiment with the TDT2 Mandarin BN Corpus that contains about 53 hours of VOA Mandarin Chinese BN. The 177 audio

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 in China (20070699015), the Natural Science Basic Research Plan of Shaanxi Province (2007F15), the NPU Aoxiang Star Plan (07XE0150), the NPU Foundation for Fundamental Research (W018103).

978-1-4244-2942-4/08/$25.00 (c) 2008 IEEE

recordings are accompanied with manually annotated story boundaries and word-level speech recognition transcripts with silence annotations. TDT2 audio was transcribed by the Dragon LVCSR with word, character and base syllable error rates of 37%, 20% and 15%, respectively. We adopted a home-grown Pinyin lexicon to get the syllable sequences of words. We separated the corpus into two non-overlapping parts: a development set of 90 recordings (1390 story boundaries) for parameter tuning and a set of 87 recordings (1400 story boundaries) for story segmentation 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.

3. LATENT SEMANTIC ANALYSIS

Latent semantic analysis (LSA) was originally used in information retrieval [12]. Humans typically retrieve documents according to the conceptual content, while individual words provide unreliable evidence about the conceptual topic or meaning of a document. Also, there are usually many words to express a given concept. Hence rigid literal term matching may not lead to the relevant documents. LSA aims to discover the conceptual relations among words via measuring the contextual use of words by PCA [10] . In LSA, the 'meaning' of a word can be represented by a 'feature vector' that embodies its relations to other words, and words occurring in similar contexts have the similar 'feature vectors'.

Given a set of texts $\Gamma = \{t_1, t_2, t_3 \ldots, t_J\}$ with vocabulary $\{w_1, w_2, w_3 \ldots, w_I\}$, LSA calculates an $I \times J$ matrix $\mathbf{M}$, where M_{ij} is the frequency of w_i occurring in t_j. Singular value decomposition (SVD) [10] is then applied to $\mathbf{M}$ to yield

$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T. \quad (1)$$

The columns of $\mathbf{U}$ and $\mathbf{V}$ are the eigenvectors of $\mathbf{MM}^T$ and $\mathbf{M}^T\mathbf{M}$, respectively, and the diagonal values of Σ are the corresponding singular values sorted in descending order. $\mathbf{MM}^T$ is called inter-word similarity matrix where the 'meaning' of a word w_i is expressed by its dot-product with all words $\{w_1, w_2, \ldots, w_I\}$. The first K columns of $\mathbf{U}$, represented by Λ_K, are the K most significant eigen vectors (principle basis) that can be used as an approximation of $\mathbf{MM}^T$ in a K-dimensional space. The rest less salient $I-K$ columns are thus removed that are considered as 'noise'. As a result, a word w_i is represented by a 'feature vector', i.e.,

$$w_i \rightarrow \Lambda_K(i) \quad (2)$$

where $\Lambda_K(i)$ is the ith row of Λ_K.

4. MOTIVATIONS OF USE OF SUBWORDS

Latent semantic analysis aims at describe the 'meaning' of a word by its relations with other words. However, the inevitable speech recognition errors in ASR transcripts may induce word matching failures and thus destroy inter-word relations. Words are made up from subwords where subwords can be phonemes, syllables or characters in Chinese. The recognition error rates at subword levels are much lower than the word level. Subwords have the advantage of partial matching and this will partially recover the relations among words. Therefore, LSA on subword level could be more effective than word level. Here we show the robustness of Chinese subword matching in errorful ASR transcripts due to the special features of Chinese.

4.1. Robustness to Flexibility in Word Segmentation

Chinese is different from western languages such as English both in written and spoken expressions. A Chinese word is formed by one to several component characters, and there is no space between words serving as word boundaries. In fact, 'word' is not defined clearly in Chinese and word segmentation of a Chinese text is definitely not unique. As a result, the same string of characters might be segmented into different word sequences in different places of the same ASR transcript. These word sequences are probably syntactically valid and semantically meaningful. For example, word 北韩 (North Korea) is segmented to 北 (North) and 韩 (Korea) and they are both appeared in a news story from the TDT2 Mandarin corpus. In this case, it is impossible to relate them by rigid word matching. The flexibility of Chinese word segmentation may cause word matching failures and thus destroy inter-word relations. However, the above problem can be solved by character matching because different segmentations still share the same component characters.

Table 1. Samples of recognition errors from TDT2. English transliterations are in brackets.

Original word	ASR errors	Base syllables
阿尔及利亚 (*Algeria*)	鲍尔 激励 要 (*Bauer drive want*)	*a er ji li ya* *bao er ji li yao*
奥尔布莱特 (*Albright*)	二 步 莱特 (*two step Wright*)	*ao er bu lai te* *er bu lai te*
互联网 (*internet*)	互 连网 (*mutual connection*)	*hu lian wang*
赈济 (*relieve*)	震级 (*quake magnitude*)	*zhen ji*
过失 (*defect*)	国事 (*national affair*)	*guo shi*

4.2. Robustness to Speech Recognition Errors

Each Chinese character is pronounced as a tonal syllable. Syllables with different tones convey different meanings. In Mandarin, about 1200 phonologically allowed tonal syllables correspond to over 6500 commonly used simplified Chinese characters. When tones are disregarded, the number of syllables is reduced to only about 400, known as base syllables. This indicates that there are a large number of *homophones* sharing the same base syllable. Tones are often mis-recognized and tone recognition itself is a research topic in Chinese. 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. Table 1 shows some word matching failures due to speech recognition errors excerpted from the TDT2 ASR transcripts. For example, country name 阿尔及利亚 (Algeria) is substituted by a pair of three words /鲍尔 激励 要/ (Bauer drive want) with some homophone characters. Foreign person name 奥尔布赖特 (Albright) is mis-recognized as another three words /二 步 莱特/ (two step Wright). Rigid word matching cannot link these samples together. However, matching at subword scale can recover their connections. For example, we can employ syllalbe subword 'er ji li' to partially match 阿尔及利亚 and /鲍尔 激励 要/, and use syllable subword '*bu lai te*' or character subword 赖特 to match 奥尔布赖特 and /二 步 莱特/, to restore their connections.

4.3. Robustness to OOV Words

The flexibility in Chinese word-building makes the limited Chinese characters to produce unlimited words. Hence there dose not exist a commonly accepted lexicon for Chinese. Consequently, the OOV problem is more pronounced in Chinese ASR transcripts, especially in the BN domain that focuses on timely events. Named entities (NE) like proper names are most common OOV words in spoken news, where they account for about 10% of BN words. An OOV word appeared in different places of a spoken document may share part of the characters or be substituted by several totally different character strings with the same (or partially same) syllable

Table 2. Samples of OOV words from TDT2. English transliterations are in brackets.

Characters		Base syllables
OOV word: 王有才 (*a Chinese name*)		*wang you cai*
ASR output	当 有 财 (*when have money*)	*dang you cai*
	王 油菜 (*king rape*)	*wang you cai*
	邦友 才 (*national friendship talent*)	*bang you cai*
OOV word: 莱温斯基 (*Lewinsky*)		*lai wen si ji*
ASR output	来 文 斯 基 (*come article this base*)	*lai wen si ji*
	来 问 司机 (*come ask driver*)	*lai wen si ji*
	来 的 司机 (*show-up driver*)	*lai de si ji*
OOV word: 科索沃 (*Kosovo*)		*ke suo wo*
ASR output	克 祖国 (*gram motherland*) 客座 我 (*guest me*)	*ke zu guo* *ke zuo wo*

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, probably their homophones. Table 2 shows some samples of misrecognized OOV words from TDT2. For example, the OOV foreign person name 莱温斯基 (Lewinsky) is substitued by three different character sequences, i.e., /来 问 司机/, /来 文 斯 基/ and /来的 司机/, resulting in word matching failures. Matching at syllable level can recover this highly-topic-related OOV word because the last two syllables are the same for the three ASR outputs ('si ji').

5. WORD AND SUBWORD LSA IN TEXTTILING

5.1. TextTiling-based Story Segmentation

The classical TextTiling algorithm in text segmentation includes three steps [7]: tokenization, lexical score determination and boundary identification. As a preprocessing step, the tokenization step divides the input text into lexical term units (e.g. words). Since the ASR transcripts in TDT2 are segmented words, tokenization for word level is thus bypassed. Tokenization for subword levels is described in Section 5.3.

In lexical score determination, the text stream is first divided into segments of sentences or pseudo-sentences. The lexical similarity is then determined at each inter-sentence gap *g* via *lexical score*:

$$\begin{aligned}\text{lexscore}(g) &= \cos(\mathbf{v}_s, \mathbf{v}_{s+1}) \\ &= \frac{\sum_{i=1}^{I} v_{s,i} v_{s+1,i}}{\sqrt{\sum_{i=1}^{I} v_{s,i} v_{s,i} \times \sum_{i=1}^{I} v_{s+1,i} v_{s+1,i}}}\end{aligned} \qquad (3)$$

where $\mathbf{v}_s$, $\mathbf{v}_{s+1}$ are term frequency vectors of the two adjacent sentences s and $s+1$ separated by gap g, and $v_{s,i}$ is the ith element of $\mathbf{v}_s$, i.e., the frequency of term w_i appeared in s. In this study, we measure the lexical scores between two adjacent pseudo-sentences defined as a fixed number of terms (T). This is because: 1) real sentence boundaries are not readily available in the ASR transcripts and sentence segmentation is another challengeable task that is out of the focus of this paper; 2) the number of shared terms between two long sentences and between a long and a short sentence would probably yield incomparable scores[7]. Since story boundaries are searched at each inter-sentence gap, we increase the boundary hypothesizes by sliding. Lexical scores are calculated at $\{T, T+\Delta, T+2\Delta ...\}$ term positions, where Δ is the sliding length and $\Delta \leq T$.

TextTiling adopts *depth score* to determine the story boundaries in the boundary identification step. Depth scores are calculated at valleys detected on the lexical score time trajectory:

$$\begin{aligned}\text{depthscore}(u) =& (\text{lexscore}(p_l) - \text{lexscore}(u)) \\ &+ (\text{lexscore}(p_r) - \text{lexscore}(u))\end{aligned} \qquad (4)$$

where u is a valley point, and p_l and p_r the left and right nearest peaks around u, respectively. The depth score considers that a sharp drop in lexical similarity is more probable to be a story boundary. Finally, boundary identification is carried out on the time trajectory of the depth score, in which a time point whose depth score is over a pre-defined threshold θ is determined as a story boundary.

5.2. Applying LSA to TextTiling

In original document retrieval task, LSA is performed on a set of text documents [12]. As described in Section 5.1, story boundaries are investigated across sentences in TextTiling. Thus LSA for the TextTiling task is performed on a set of sentences. As mentioned before, since real sentence boundaries are not available in the ASR transcripts, we divide all the ASR transcripts in the corpus into sentence-like segments by a silence threshold (empirically set to 0.8s in this study) and LSA is performed on this set of sentence-like text segments. Training LSA on sentence-like segments divided by silence is more reasonable than that by fixed term units (as in Section 5.1). This is because of the following two reasons. 1) Sentences are commonly separated by pauses in speech and pause is a salient speech prosody highly related to semantic units and sentence boundaries [1]. 2) Cutting sentence-like segments by fixed window across the text will probably combine each part of two adjacent real sentences into a segment. This will destroy the semantic self-containness of sentences and induce unreasonable latent semantic analysis.

We conduct LSA (described in Section 3) on the sentence-like segments using the SVDLIBC library [1]. As a result, each term w_i is represented by an LSA feature vector $\Lambda_K(i)$ as in formula (2) that embodies its relations with other terms. We thus apply $\Lambda_K(i)$ to TextTiling. The 'meaning' of each pseudo-sentence (defined in Section 5.1) is approximately represented by a K-dimensional vector

$$\hat{\mathbf{v}}_s = \sum_{i=1}^{I} v_{i,s} \times \Lambda_K(i), \qquad (5)$$

where $v_{i,s}$ is the frequency of term w_i. Therefore, the lexical score at each inter-sentence gap g is re-formulated as

$$\begin{aligned}\text{lexscore}(g) &= \cos(\hat{\mathbf{v}}_s, \hat{\mathbf{v}}_{s+1}) \\ &= \frac{\sum_{i=1}^{K} \hat{v}_{s,i} \hat{v}_{s+1,i}}{\sqrt{\sum_{i=1}^{K} \hat{v}_{s,i} \hat{v}_{s,i} \times \sum_{i=1}^{K} \hat{v}_{s+1,i} \hat{v}_{s+1,i}}},\end{aligned} \qquad (6)$$

where $\hat{v}_{s,i}$ is the ith element of vector $\hat{\mathbf{v}}_s$. Story boundaries are identified by measuring depth scores that stem from Eq. (6). Specifically, when term w_i represents a word in the ASR transcript, the above procedure is word-LSA-based TextTiling.

5.3. Subword-LSA-based TextTiling

We propose to perform LSA on different Chinese subword lexical representations, i.e., character and syllable n-gram units. For a sequence of words $\{w_1 w_2 w_3 \ldots w_Q\}$, the unigram sequence is defined as the sequence of the component characters (character unigram) or syllables (syllable unigram), i.e.,$\{c_1 c_2 c_3 \ldots c_L\}$. The subword bigram and trigram sequences are formed as

$$\text{bigram:}\{c_1c_2 \;\; c_2c_3 \;\; c_3c_4 \cdots c_{L-1}c_L\}, \qquad (7)$$

$$\text{trigram}: \{c_1c_2c_3 \;\; c_2c_3c_4 \;\; c_3c_4c_5 \cdots c_{L-2}c_{L-1}c_L\}, \qquad (8)$$

respectively. Higher order subword overlapping n-grams can be formed accordingly. To reduce the possibility of missing any useful information embedded in the subword sequence, overlapping between subwords is used.

[1] http://tedlab.mit.edu/ dr/svdlibc/

According to the procedure in Section 5.2, subword LSA is performed on the subword n-gram representation of the sentence-like segments, where term w_i is the corresponding overlapping n-gram unit. Lexical scores are then calculated by Eq. (6) and story boundaries are determined by thresholding on the depth score curve.

6. EXPERIMENTS

We carried out story segmentation experiments to compare (1) the original word-based TextTiling [6], (2) our previous work on subword TextTiling [7], (3) word-LSA-based TextTiling and (4) subword-LSA-based TextTiling. Empirical parameter tuning was first performed on the development set that selects parameters achieving the best F1-measure of story segmentation. Parameter tuning shows that a combination of $T = 50$ and $\Delta = 20$ achieves the best performance on the baseline word-based TextTiling. The selected T and Δ were then fixed in the empirical searching for the threshold θ and LSA dimensions K for each word and subword scales on the development set[2]. Experimental results on the test set in terms of F1-measure are shown in Fig. 1.

Results indicate that applying word and subword LSA to TextTiling can improve the story segmentation performance on Chinese BN except unigram LSA that degrades the performance a little bit compared with those corresponding TextTiling approaches without LSA. Bigram-LSA-based TextTiling exhibits superior performance. Character-bigram-LSA achieves the best F1-measure of 0.6598 with relative improvement of 17.4% over the conventional word-based TextTiling (on the character sequence of word) and 6.5% over our previous syllable-bigram-based TextTiling. The failure of uigram LSA is probably because of the small vocabulary of unigram singletons (2957 for characters and 395 for syllables in our corpus). The 'meanings' (LSA feature vectors) of characters/syllables might be too close due to the small vocabulary. The success of bigram LSA mainly owes to the fact that the most frequently used words in Chinese are bi-character and the probability of long sequences with correctly recognized characters is smaller than two character units. This also explains why trigram and 4-gram TextTing perform the worst. We also observe an interesting result that trigram-LSA-based and 4-gram-LSA-based TextTiling significantly outperform trigram and 4-gram TextTiling (without LSA). Their performances are even comparable to word-LSA-based TextTiling. This may be explained by the 'noise-removal' merits of LSA [9]. The inter-similarity matrices for tri-gram and 4-gram in LSA are large sparse ones. A small dimensional LSA space removes a lot of noises and tends to be an appropriate approximation of a large sparse matrix.

7. SUMMARY

This paper has extended our previous work on subword TextTiling for automatic story segmentation of Chinese broadcast news. This study proposes to use subword latent semantic analysis (LSA) in TextTiling that integrates the concept matching merit of LSA and the robustness of character and syllalbe n-gram units in lexical matching on errorful Chinese ASR transcripts. After subword LSA, the 'meaning' of a subword n-gram unit is represented by its relations with other subword n-gram units. The inter-sentence lexical score is measured by the cosine of the 'meaning' of the sentences that is composed of subword n-gram units. We conclude from the experiments that applying subword LSA to TextTiling can effectively improve the story segmentation performance on Chinese BN. Character-bigram-LSA-based TextTiling achieves the best F1-measure of 0.6598 with relative improvement of 17.4% over the conventional word-based TextTiling and 6.5% over our previous syllable-bigram-based TextTiling. In the future, we plan to perform a statistical study on the relations between the story segmentation performance and the speech recognition error rate. we also plan to study the combination of word and subword LSA to incorporate the complementarity from different lexical representation scales.

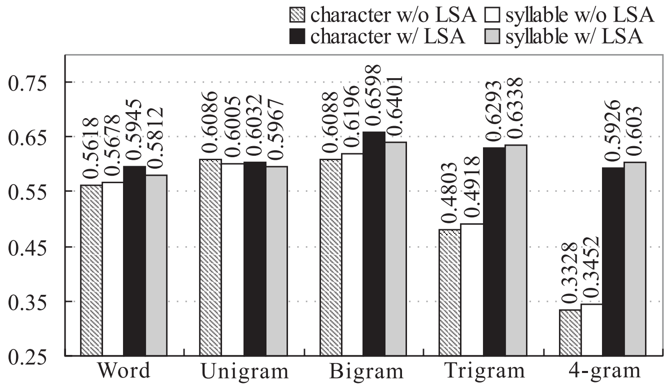

Fig. 1. Experimental results (F1-measure) on word/subword-based TextTiling and word/subowrd-LSA-based TextTiling. Word level involves character sequence and syllable sequence of a word.

8. REFERENCES

[1] L. Xie and H. Meng, "Combined use of speaker and tone-normalized pitch reset with pause duration for automatic story segmentation in mandarin broadcast news," in *Proc. HLT-NAACL*, 2007, pp. 193–169.

[2] W. Hsu, S. F. Chang, C. W. Huang, L. Kennedy, C. Y. Lin, and G. Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *Proc. IS&T/SPIE Symposium on Electronic Imaging*, 2004.

[3] N. Stokes, J. Carthy, and A. Smeaton, "Select: A lexical cohesion based news story segmentation system," *Journal of AI Communication*, vol. 17, no. 1, pp. 3–12, 2004.

[4] S. Dharanipragada, M. Franz, J. Mccarley, S. Roukos, and T. Ward, "Story segmentation and topic detection in the broadcast news domain," in *Proc. the DARPA Broadcast News Workshop*, 1999.

[5] J. Yamron, I. carp, L. Gillick, and P. Mulbregt, "A hidden markov model approach to text segmentation and event tracking," in *Proc. ICASSP*, 1999, pp. 333–336.

[6] M. A. Hearst, "Textiling: Segmentation text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[7] L. Xie, J. Zeng, and W. Feng, "Multi-scale TextTiling for automatic sroty segmentation in chinese broascast news," in *Proc. Asia Information Retrieval Symposium*, 2008, pp. 345–355.

[8] S. Banerjee and I. A. Rudnicky, "A TextTiling based approach to topic boundary detection in meetings," in *Proc. Interspeech*, 2006.

[9] F. Y. Y. Choi, P. Wiemer-hastings, and J. Moore, "Latent semantic analysis for story segmentation," in *Proc. the 2001 Conf. on Empirical methods in Natural Language Processing*, 2001.

[10] R. Jerome and Bellegarda, "Latent semantic mapping," *IEEE signal processing magazine*, vol. 5, no. 1053-5888, pp. 70–80, 2005.

[11] K. Ng and V. W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.

[12] S. Deerwester, S. T. Dumains, and et al. G. W. Furnas, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.

[2] The actual T and Δ for a subword scale are the numbers of n-gram units in the corresponding 50-word and 20-word sequences, respectively.