Phoneme Lattice based TextTiling towards Multilingual Story Segmentation

Xiaoxuan Wang[1,3], *Lei Xie*[1] , *Bin Ma*[2], *Eng Siong Chng*[3], *Haizhou Li*[2,3]

[1]School of Computer Science, Northwestern Polytechnical University, China
[2]Institute for Infocomm Research, Singapore
[3]School of Computer Engineering, Nanyang Technological University, Singapore

xwang@nwpu-aslp.org, lxie@nwpu.edu.cn, {mabin,hli}@i2r.a-star.edu.sg, aseschng@ntu.edu.sg

Abstract

This paper proposes a phoneme lattice based TextTiling approach towards multilingual story segmentation. The phoneme is the smallest segmental unit in a language and the number of phonemes in a language is usually far smaller than the number of words. Furthermore, many phonemes are shared by different languages. These properties make phonemes particularly appropriate for representing multilingual speech. As phoneme recognition is far from perfect, phoneme lattices, which carry much richer statistics than the 1-best hypotheses, are adopted in this paper as the input to the TextTiling approach. The term frequencies used in traditional TextTiling are replaced by the expected counts of phoneme *n*-gram units calculated from phoneme lattices. Experiments on TDT2 English and Mandarin corpora show that the phoneme lattice based TextTiling outperforms the phoneme 1-best based TextTiling and word based TextTiling in broadcast news story segmentation.

Index Terms: story segmentation, topic detection and tracking, spoken document retrieval, phoneme lattice, speech processing.

1. Introduction

Automatic story segmentation is to divide a stream of text, speech or video into topically homogeneous segments. It is an important step that facilitates further processing such as topic tracking, information extraction, summarization, indexing and retrieval. Taking broadcast news retrieval for example, users usually expect short clips relevant to their queries rather than entire news programs. With the current trend of globalization, the exponential proliferation of multilingual materials poses significant challenges to multimedia content management. It takes tremendous effort to transfer a spoken language system from one language to another due to problems such as lack of data, tools, linguistic resources and expertise [1]. For story segmentation of multilingual repositories, a language independent solution is highly demanded.

Research on automatic story segmentation for spoken documents has been focused on story/topic modeling and boundary detection. For example, in the topic modeling approach, hidden Markov models (HMMs) are adopted to model the relations between the hidden topics and the observable words or sentences and story boundaries are indicated by the topic transitions in HMMs [2]. For detection based approaches, different boundary cues, e.g., acoustic/prosodic cues [3], visual cues [4] and lexical cues [5, 6], have been extensively studied. Acoustic and visual cues, such as significant pauses, speaker changes and anchor face appearances, rely heavily on editorial rules. By contrast, lexical cues reveal topic shifts via semantic variation of text. Moreover, lexical cues can be extracted from both pure texts and multimedia sources such as speech and video caption. These merits make lexical cues more robust and general comparing to acoustic and visual cues. TextTiling [6] is a classical lexical cohesion based text segmentation approach. It is based on an intuitive assumption that different topics usually employ different sets of words; thus the shifts of word lexicon could probably indicate story boundaries. Due to its simplicity and efficiency, it has been recently introduced to segmenting spoken documents such as broadcast news and meetings [7].

Despite years of research, few efforts have been made to *cross-language* automatic story segmentation. Some limited studies tried to investigate and select suitable multimodal cues/features for story segmentation in different languages [8, 9], e.g., English, Chinese and Arabic. These studies reveal that most features are language or source dependent. Specifically, the lexical features are mostly extracted from language-dependent large vocabulary continuous speech recognition (LVCSR) transcripts with the assumption that the source language is known. For example, in [8], language- and source-dependent phrase *n*-gram features were derived from the word transcription based on a language specific LVCSR. Moveover, for some domain specific tasks which lack enough data for training, the word recognition results may not be reliable enough for lexical based segmentation [9].

This paper reports our preliminary study on language-independent automatic story segmentation of spoken documents. Specifically, we propose a phoneme lattice based TextTiling approach for broadcast news segmentation. The number of phonemes in a language is usually far smaller than the words, so a higher order of *n*-gram units can be applied to provide discriminative statistics. Although each language has its own phoneme set, the phonemes are shared heavily across languages. A possible solution to story segmentation of multilingual spoken documents is to use the statistics from phoneme recognition results. However, the phoneme recognition usually has a higher error rate than word recognition. To address this problem, we adopt the phoneme *lattices* rather than the 1-best hypotheses as the input to the subsequent segmentation. A lattice is a directed acyclic graph in which each edge is labeled with a hypothesis and the acoustic probability of generating that hypothesis. Lattices have been widely adopted in spoken document retrieval [10] and language identification [11], while word lattices from an LVCSR were used in topic analysis of English broadcast news[12]. In this paper,we use lattices generated from a phoneme recognizer as the input to the TextTiling based story segmentation. The word counts in lexical similarity measure of traditional TextTiling are replaced by the expected counts of phoneme *n*-gram units calculated from phoneme lattices.

Experiments on TDT2 Mandarin and English broadcast news corpora show the superior performance of the phoneme

lattice based TextTiling over the phoneme 1-best and word based TextTiling in story segmentation. The phoneme-lattice approach is promising in multilingual story segmentation.

2. TextTiling for Story Segmentation

The classical TextTiling algorithm includes three steps [6]: tokenization, lexical score determination and boundary identification. Tokenization is a preprocessing step which divides the input text into lexical term units (i.e. words). For texts in a language such as English, since words are already separated by space and punctuation, tokenization may only involve a morphological analysis to reduce words to their roots. Word tokenization is a necessary step for some Asian languages such as Chinese, since a Chinese text comes as a character sequence without word delimiters. For spoken documents, regardless of the language, transcripts are segmented into sequences of words at the speech recognition stage.

In lexical score determination, the tokenized text stream is first divided into sentences (for texts) or blocks of words (for speech recognition transcripts without sentence boundaries). The lexical similarity is then calculated at each inter-sentence (or inter-block) gap g via *lexical score*:

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

where $\mathbf{v}_s$, $\mathbf{v}_{s+1}$ are the term frequency vectors of 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 v_i in s with a vocabulary size of I. The algorithm regards each inter-sentence gap as a boundary candidate. Lexical scores are measured at $\{N+\Delta, N+2\Delta \cdots\}$ boundary candidate positions, where N and Δ are the block length and the sliding length, respectively, and $\Delta \leq N$.

We adopt *relative score* for story boundary identification, which reflects the variation tendency of lexical similarity. Relative score is defined as:

$$\begin{aligned} relativescore(g_i) = &(lexscore(g_{i-1}) - lexscore(g_i)) \\ &+ (lexscore(g_{i+1}) - lexscore(g_i)) \end{aligned} \quad (2)$$

where g_i is an inter-sentence gap with its left and right neighbors g_{i-1} and g_{i+1}. Finally, boundary detection is carried out on the time trajectory of the relative score, in which a time point whose relative score over a pre-defined threshold θ is selected as a story boundary.

In this paper, we measure the lexical similarity using phoneme *n*-gram statistics instead of word statistics, and apply expected counts of phoneme *n*-grams calculated from phoneme lattices instead of term frequencies.

3. TextTiling with Phoneme Lattice Statistics

3.1. Phoneme Recognition

In traditional lexical based story segmentation for spoken documents, LVCSR is an indispensable component to generate word level transcriptions. LVCSR is usually language dependent with a closed vocabulary and involves complex acoustic and language models trained using a large collection of domain specific speech and text data. For many languages or domains, this set of language resources is not always readily available. The recognition process of LVCSR is also computationally expensive, which limits its deployment in practice. As we know, in each language, the phoneme is the smallest segmental unit of sound employed to form meaningful contrasts between utterances (e.g. words). The phoneme set of a certain language is usually far smaller than the word set and many phonemes are shared by different languages. There are at least three explicit advantages of adopting phoneme recognizers for story segmentation. Firstly, sophisticated acoustic and language model are not needed. Only a simple phoneme loop is required in the phoneme recognition system. Secondly, the out-of-vocabulary (OOV) problem doesn't exist. Finally, it is feasible to construct a language independent phoneme recognition system using a universal phonetic inventory.

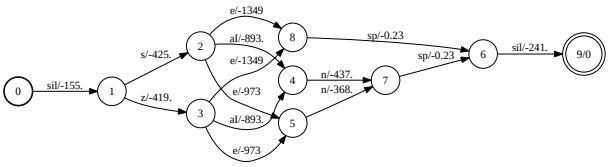

Figure 1: *Phoneme lattice for a speech utterance "say" in TDT2 English broadcast news.*

There are also two problems when we use a phoneme recognizer (instead of a word recognizer) in multilingual story segmentation. Firstly, phoneme recognition is less accurate than word recognition. Secondly, phonemes are less discriminative than words due to the small phoneme inventory in a language. In this paper, we address the first problem by multiple hypotheses from the output lattice of a phoneme recognizer. We increase the discriminative capability of phoneme units by forming phoneme *n*-grams instead of phoneme unigrams.

3.2. Lattice Generation

Due to the lack of lexical constraints, phoneme recognition typically leads to low accuracy as compared with word recognition in LVCSR. One way to overcome this problem is to use the phoneme recognition lattice which contains multiple hypotheses. As the production of Viterbi search of a phoneme recognizer, a connected loop-free directed graph [13], in which each edge is labled with a phoneme hypothesis and its acoustic probability, can be generated. For example in Figure 1, the value on each arc in the phoneme lattice is the acoustic probability of the coressponding phoneme. We first detect the non-speech segments (e.g. music and silence) from audio using an in-house audio classifier and each audio is divided into short speech utterances that are separated by the detected non-speech. A phoneme recognizer then generates a lattice for each speech utterance.

3.3. Expected Counts for Lexical Similarity

For the 1-best phoneme recognition, the calculation of *n*-gram statistics is straightforward by simply counting the number of times each phoneme *n*-gram appears in the hypothesized phoneme sequence. For the phoneme lattice, we can compute the expected counts of a *n*-gram term d, given an input speech utterance $\mathbf{o}$ by

$$E[c(d|\mathbf{o})] = \sum_{q} p(q|\mathbf{o}) c(d|q), \quad (3)$$

where q represents a lattice path and $p(q|\mathbf{o})$ represents its posterior probability computed from acoustic probabilities in the lattice using standard forward-backward algorithm. The sum is taken over all the paths in the lattice. The count $c(d|q)$ refers to

the number of times d appears in the phoneme sequence q while d can be in any order of n-gram. We employ the SRILM toolkit[1] to compute the expected counts in Eq. (3) for each speech utterance.

In the classical TextTiling approach, lexical similarity is calculated at each inter-sentence gap. Since sentence boundaries are not readily available in broadcast news audio, we measure the lexical similarity at each inter-block gap. A block $\mathbf{B}$ is composed of N adjacent speech utterances, i.e., $\mathbf{B} = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_N\}$. Thus, the expected count for a phoneme n-gram term d in the block $\mathbf{B}$ is calculated by

$$E[c(d|\mathbf{B})] = \sum_{n=1}^{N} E[c(d|\mathbf{o}_n)], \tag{4}$$

i.e., the sum of the expected counts of d of all the speech utterances in the block $\mathbf{B}$.

The accumulated expected counts $E[c(d|\mathbf{B})]$ are adopted to replace $v_{s,i}$ in Eq. (1). The phoneme lattice based lexical score is calculated by:

$$\begin{aligned} lexscore(g) &= cos(\mathbf{B}_s, \mathbf{B}_{s+1}) \\ &= \frac{\sum_{i=1}^{I} E[c(d_i|\mathbf{B}_s)]E[c(d_i|\mathbf{B}_{s+1})]}{\sqrt{\sum_{i=1}^{I} E[c(d_i|\mathbf{B}_s)]^2 \sum_{i=1}^{I} E[c(d_i|\mathbf{B}_{s+1})]^2}} \end{aligned} \tag{5}$$

where I is the number of the phoneme n-gram units. Relative scores are calculated from the lexical score trajectory according to Eq. (2) and boundaries are detected by a preset threshold.

4. Experiments

4.1. Experimental Scheme

We experimented with the TDT2 VOA Mandarin and English broadcast news corpora[2] for performance evaluation. Table 1 shows the details of the corpora. The broadcast news audio files are accompanied with manually annotated meta-data including story boundaries, manually-annotated references (namely Man-Ref and Eng-Ref) and LVCSR transcripts (namely Man-LVCSR and Eng-LVCSR). The tuning sets were used for empirical tuning of the block size N, sliding length Δ and boundary detection threshold θ; the testing sets were for story segmentation performance evaluation. According to the TDT2 standard , 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. F1-measure, i.e., the harmonic mean of recall and precision, was used as the evaluation criterion.

We first conducted experiments with the Reference transcripts for TextTiling on both word and phoneme n-gram ($n = 1, 2, \cdots, 5$) level. For the English word transcripts (Eng-Ref), we used the Porter stemmer to reduce words to their linguistic roots before word TextTiling. We mapped the word transcripts into phoneme transcripts using an in-house Mandarin lexicon (for Mandarin) and the CMU dictionary (for English). We used real sentence boundaries as the inter-sentence gaps for similarity measurement. The purpose of these experiments is to provide the performance reference for word/phoneme based TextTiling with error-free Reference transcripts.

The TextTiling experiments with LVCSR transcripts were also conducted in a similar way, by using LVCSR transcripts, Man-LVCSR and Eng-LVCSR, as word level transcripts, and mapping the word transcripts into phoneme transcripts for phoneme level transcripts. Different from the Reference transcripts, sentence boundaries are not available for LVCSR transcripts. We adopted recognized silence as sentence boundaries.

[1] http://www-speech.sri.com/projects/srilm/

[2] http://www.ldc.upenn.edu/Projects/TDT2

Table 1: *Corpora for story segmentation experiments.*

Corpus		*TDT-2 Mandarin*	*TDT-2 English*
Source		VOA newscast, Feb. to June 1998	
No. of programs		177	111
Audio duration		53h	111h
WER		37%	35%
Data assign.	Tuning	90 (1321 boundaries)	56 (2629 boundaries)
	Testing	87 (1262 boundaries)	55 (2627 boundaries)

Table 2: *Summary of phoneme recognizers*

Recognizer	Language	Phonemes	Training Data	PER(%)
Hub4Man	Mandarin	36	28h	34.06
Hub4Eng	English	39	52h	54.55

We then applied both Mandarin and English phoneme recognizers for TextTiling experiments, with both the 1-best phoneme sequence and phoneme lattice. We not only tested the *same-language* case, i.e., using a phoneme recognizer to transcribe the broadcast news speech in the same language, but also tested the *cross-language* case, i.e., using a phoneme recognizer to transcribe the speech in a different language.

Table 2 shows the summary of the two phoneme recognizers. The Mandarin phoneme recognizer, namely Hub4Man, was trained using HUB4-NE Mandarin broadcast news speech corpus that contains audio from VOA, CCTV and KAZN-AM. The English phoneme recognizer, namely Hub4Eng, was trained using HUB4 1996 English broadcast news speech corpus, which contains speech from ABC and CNN. We used HTK to train phoneme HMMs, each of which has three states with 32 Gaussian mixtures. The acoustic feature vector consists of 12 MFCCs, the normalized energy, and their first and second order derivatives. Simple phoneme loop was used in the decoding.

4.2. Results on Reference and LVCSR Transcripts

Table 3 summarizes the results on Reference and LVCSR transcripts. We observe that, in general, phoneme *n*-gram based TextTiling can achieve comparable and even better segmentation performance as compared with word based TextTiling (word unigram) on both Reference and LVCSR transcripts. Phoneme 4-gram obtains the highest F1-measures for Mandarin broadcast news (on both Reference and LVCSR transcripts) while Phoneme 2-gram and Phoneme 3-gram achieve the best F1-measures for English Reference and LVCSR transcripts. These results indicate that although words are more discriminative than phonemes, phoneme *n*-grams convey enough discriminative capability with the sequential contextual information. Furthermore, using phoneme subwords has the advantage over words for overcoming OOV problem. It is shown that the F1-measure on English corpus is much lower than that on Mandarin corpus. It is also noticed that as the value of *n* increases from 1 to 5 for the phoneme *n*-grams, F1-measure peaks at a moderate *n*. The inferior performance of phoneme 1-gram is due to its low discriminative capability, while the sparseness of higher order 5-gram units leads to a lower F1-measure too.

4.3. Results on Phoneme 1-Best Hypothesis and Lattices

Table 4 shows the story segmentation results with the phoneme 1-best and phoneme lattice based on the two phoneme recognizers. It clearly shows that the phoneme lattice consistently outperforms the phoneme 1-best. For instance, 3-gram phoneme lattice brings 3.13% absolute gain in F1-measure as compared with 3-gram phoneme 1-best on TDT2 Mandarin corpus us-

Table 4: *Story segmentation results (F1-measure) on phoneme 1-best sequence and lattices*

Recognizer	Corpus	1-gram		2-gram		3-gram		4-gram	
		1-best	Lat.	1-best	Lat.	1-best	Lat.	1-best	Lat.
Hub4Man	TDT-MAN	0.5012	0.5042	0.5089	0.5186	0.5122	0.5435	0.5116	0.5254
	TDT-ENG	0.3988	0.4	0.4033	0.4053	0.3903	0.4095	0.3997	0.3914
Hub4Eng	TDT-MAN	0.5038	0.5040	0.5047	0.5224	0.4949	0.5279	0.4916	0.5007
	TDT-ENG	0.4008	0.4047	0.4008	0.4139	0.4078	0.4266	0.4111	0.4021

Table 3: *Story segmentation results (F1-measure) on Reference and LVCSR transcripts at word and phoneme n-gram levels*

Source	Word	Phoneme n-gram				
		1-gram	2-gram	3-gram	4-gram	5-gram
Man-Ref	0.694	0.6515	0.6871	0.6947	0.7068	0.6571
Man-LVCSR	0.5298	0.4968	0.5051	0.541	0.5549	0.5541
Eng-Ref	0.552	0.5333	0.5567	0.5355	0.5184	0.4995
Eng-LVCSR	0.3669	0.3205	0.3491	0.3844	0.3713	0.3577

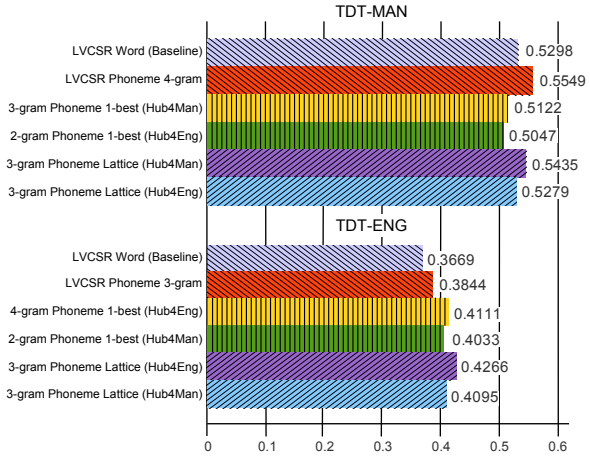

Figure 2: *F1-measure comparison between word baseline and various phoneme based approaches*

ing Hub4Man phoneme recognizer (same-language case). The superior performance of lattices is due to their statistical interpretation of multiple recognition hypotheses and robustness to recognition errors. It is also found that, at cross-language case, e.g., using the Mandarin phoneme recognizer, instead of using the English phoneme recognizer, to transcribe English broadcast news audio, only suffer a slight performance degradation. It is more interesting to notice that, the Mandarin phoneme lattice brings an even better story segmentation performance on English corpus as compared with the English LVCSR word transcript and its phoneme mapping. For example, the 3-gram phoneme lattice generated by the Mandarin phoneme recognizer achieves an F1-measure of 0.4095 on English corpus (shown in Table 4), while the phoneme 3-gram mapped from English LVCSR gets an F1-measure of 0.3844 (shown in Table 3). The selected experimental results have also been illustrated in Figure 2, to show the performance comparison. These results show that it is promising to provide a multilingual story segmentation solution by adopting the phoneme lattice based TextTiling approach.

5. Conclusions

This paper has proposed a phoneme lattice based TextTiling towards story segmentation of cross-language spoken documents. Different from other linguistic units such as words, phonemes are compact and cross-langauge. Phoneme lattices convey a statistical interpretation of multiple recognition hypotheses and show robustness to recognition errors. We use the lattice output of a simple phoneme recognizer, other than a complicated LVCSR, as the input to the classical TextTiling. Specifically, expected counts of phoneme *n*-grams calculated from the lattice are used in the similarity measurement. Experiments show that the phoneme based TextTiling achieves comparable and even better results than conventional word TextTiling; phoneme lattice based TextTiling consistently outperforms phoneme 1-best TextTiling. Cross-langauge (English and Mandarin) experiments indicate that phoneme representations are promising in multilingual story segmentation. We plan to use the international phonetic alphabet (IPA) to construct an universal phoneme recognizer for multilingual story segmentation. We are currently studying more effective segmentation approaches using phoneme lattices, e.g., Normalized Cuts, to achieve a better story segmentation performance.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (60802085) and the NPU Foundation for Fundamental Research (W018103).

7. References

[1] Fung, P. and Shultz, T., "Multilingual spoken language processing," *IEEE Signal Processing Magazine*, pp. 89–97, May 2008.

[2] Yamron, J.P., Carp, I., Gillick, L., Lowe, S., and P. van Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," in *Proc. ICASSP*, vol. 1, May 1998, pp. 333 –336 vol.1.

[3] Shriberg, E., Stolcke, A., Hakkani-Tür, D. Z., and Tür, G., "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[4] Hsu, W., Chang, S.-F., Huang, C., Kennedy, L., and Lin, C., "Discovery and fusion of salient multi-modal features towards news story segmentation," in *SPIE Electronic Imaging*, 2004.

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

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

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

[8] Palmer, D. D., Reichman, M., and Yaich, E., "Feature selection for trainable multilingual broadcast news segmentation," in *Proc. HLT-NAACL*, 2004, pp. 89–92.

[9] Rosenberg, A. and Hirschberg, J., "Story segmentation of broadcast news in English, Mandarin and Arabic," in *Proc. HLT-NAACL*, 2006, pp. 125–128.

[10] Murat, S., "Lattice-based search for spoken utterance retrieval," in *Proc. HLT-NAACL*, 2004, pp. 129–136.

[11] Gauvain, J.L., Messaoudi, A., and Schwenk, H., "Language Recognition Using Phone Latices," in *Proc. ICSLP*, 2004, pp. 1283–1286.

[12] Mehryar, M., Pedro, M., and Eugene, W., "A new quality measure for topic segmentation of text and speech," in *Proc. Interspeech*, 2009.

[13] James, D. A., Young, S. J., and Pz, C., "A fast lattice-based approach to vocabulary independent wordspotting," in *Proc. ICASSP*, 1994, pp. 377–380.