INVITED PAPER

Broadcast News Story Segmentation Using Conditional Random Fields and Multi-modal Features

Xiaoxuan WANG[†a)], **Lei XIE**[†b)], **Mimi LU**[†c)], **Bin MA**[††d)], **Eng SIONG CHNG**[†††e)], ***and*** **Haizhou LI**[††f)],

SUMMARY This paper proposes to integrate multi-modal features using conditional random fields (CRF) for broadcast news story segmentation. We study story boundary cues from lexical, audio and video modalities, where lexical features consist of lexical similarity, chain strength and overall cohesiveness, acoustic features involve pause duration, pitch, speaker change and audio event type, and visual features contain shot boundary, anchor face and news title caption. These features are extracted in a sequence of boundary candidate positions in the broadcast news. A linear-chain CRF is used to detect each candidate as a boundary/non-boundary tags based on the multi-modal features. Important inter-label relations and contextual feature information are effectively captured by CRF's sequential learning framework. Story segmentation experiments show that the CRF approach outperforms other popular classifiers, including decision tree (DT), Bayesian network (BN), naive Bayesian classifier (NB), multi-layer perception (MLP), support vector machines (SVM) and maximum entropy (ME) classifier.

key words: *Story Segmentation, Conditional Random Fields*

1. Introduction

With the development of multimedia and web technologies, ever-increasing multimedia collections are available, e.g., broadcast news, meetings and lectures. Given the vast amount of multimedia data, automatic approaches for multimedia processing are urgently in demand, especially for automatic indexing, summarization, retrieval, visualization, and organization technologies. Among these technologies, automatic story (or topic) segmentation is an important precursor since other tasks usually assume the presence of individual topical documents. Story segmentation is such a task that divides a stream of text, speech or video into topically homogeneous blocks, known as *stories*. Specifically for broadcast news (BN), a popular media repository, the objective is to segment continuous audio/video streams into distinct news stories, each addressing a central topic.

Story boundary cues (features) from different modalities are of great importance for automatic story segmentation. Lexical cues reveal story boundaries via semantic variations across the text, mainly including the exploration of word cohesiveness and use of cue phrases [1]. For example, the TextTiling approach [2], [3] measures pairwise sentence lexical similarities in a text and the local similarity minima are detected as story boundaries. The lexical chaining method [4] chains up the related words such as word repetitions and positions with high counts of chain starts and ends are considered as story boundaries. Recently, speech prosody has draw much attention because it provides an acoustic knowledge source with an embedded rhythm on topic shifts [5], [6]. For example, broadcast news programs often follow editorial prosodic rules, e.g., (1) news topics are switched by musical breaks or significant pauses; (2) two announcers report news stories in turn; (3) a studio anchor starts a topic and then passes it to a reporter for a detailed report. Besides the editorial prosody, speakers naturally separate their discourses into different semantic units (e.g., sentences, paragraphs and topics) through durational, intonational and intensity cues, known as *speech prosody* [5], [7]. Compared with lexical and acoustic cues, visual cues are more reliant on the editorial rules and news production patterns. The transition of stories is usually followed by the change of video shots. For example, field-to-studio shot transition is a salient story boundary cue. This is because many broadcast news programs follow a clear pattern: each news story starts from a studio shot and then moves to field shots [8]. Anchor face is another visual feature of which the presence is an indicator of topic transitions [9]. Also, in a broadcast news video, a news story is often accompanied by a caption describing the content of the news.

Story segmentation approaches can be categorized into generative topic modeling [10], [11] and story boundary detection [5], [8], [12], [13]. The former category treats the word sequence (speech transcripts) as observations of some pre-defined topics and the topic labels are assigned to the speech transcripts under an optimal criterion. In the detection-based framework, boundary candidates are first determined across the spoken document. Story segmentation is then viewed as a sequential classification/tagging problem, i.e., classifying each candidate into a boundary or non-boundary based on a set of features. In this paper, we focus on the story boundary detection approach. Some recent efforts have shown that integrating different features is able to significantly improve the detection performance [8], [13], [14]. Decision tree (DT) has been used to the integration of lexical and acoustic features [13], [15], due to their effective ability to model feature interactions, to deal

Manuscript received March 1, 2011.

[†]The author is with the School of Computer Science, Northwestern Polytechnical University, China

[††]Institute for Infocomm Research, Singapore

[†††]School of Computer Engineering, Nanyang Technological University, Singapore

a) E-mail: xwang@nwpu-aslp.org

b) E-mail: lxie@nwpu.edu.cn

c) E-mail: mlu@nwpu-aslp.org

d) E-mail: mabin@i2r.a-star.edu.sg

e) E-mail: aseschng@ntu.edu.sg

f) E-mail: hli@i2r.a-star.edu.sg

DOI: 10.1587/transinf.E0.D.1

with missing features, and to handle a large amount of training data. Tür *et al.* [13] adopted a hidden Markov model (HMM) to fuse features from different knowledge sources. The word usage and lexical cues were represented by a language model embedded in the HMM while prosodic cues, such as pause durations and pitch resets, were modeled by a decision tree based on automatically extracted acoustic features and alignments. The system developed in the *Informedia* project [12] was one of the earliest rule-based broadcast news video story segmentation system, in which, some ad-hoc rules were designed to combine visual, acoustic and lexical features. Recently, support vector machine (SVM) [16] and maximum entropy (MaxEnt) model [9] have also been used for story segmentation.

Despite years of study, most of the previous research focus on modeling the features independently. However, for time series data, such as speech and video, has a strong correlation among adjacent units. Especially when comes to most higher level understandings such as story segmentation, global information is believed to be much more helpful. In this study, we employ a detection-based story segmentation approach and propose to integrate multi-modal features using conditional random fields (CRF) for news story segmentation. A CRF is an undirected graphical model that defines a global log-linear distribution of the entire label sequence conditioned on the observation sequence [17]. The model has theoretical advantages in sequential classification: (1) it provides an intuitive method for integrating features from various sources because there is no independence assumption among features. This property of CRF will help us to investigate the relations among features from intra-modality to inter-modalities; (2) it models the sequential/contextual information and labels of a given candidate by considering its surrounding features and labels (i.e. global optimal labeling). In this way, CRF models the conditional distribution of the label sequence given the feature sequence by globally combining both the feature-to-label and label-to-label correlations, which is thus a better framework for segmenting time series data. Recently, CRF modeling has shown superior performances in various speech and language tasks such as POS tagging [17], shallow parsing [18], sentence boundary detection [19], pitch accent prediction [20] and speech recognition [21].

The remainder of this paper is organized as follows. In the next section, we give an overview on our story segmentation system. In Section 3, we describe the proposed CRF approach for story segmentation. Section 4 reports the extraction of multi-modal features. We present our experimental results and analysis in Section 5 and summarize the paper in Section 6.

2. System Overview

The detection-based story segmentation system consists of three steps: candidate identification, feature extraction and boundary/non-boundary classification, as shown in Fig. 1. We model the story segmentation task as a sequential

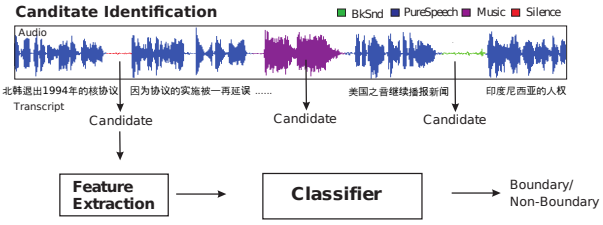

Fig. 1 Block diagram of the story segmentation approach.

boundary/non-boundary classification/tagging problem. We first identify a set of candidates (i.e. potential story boundaries), denoted as $\mathcal{B}$, in the broadcast news stream. The principle of this step is to reduce the boundary search complexity and to maintain a low miss rate (high recall rate) of story boundaries at the same time. In this study, we consider all the silence and music positions (labeled by an audio classifier) as the story boundary candidates. These positions can cover almost all the story boundaries because broadcast news use silence breaks and music intervals to keep the editorial tempo. A set of multi-modal features which include acoustic, lexical and visual features, denoted as $\mathcal{F}$, are then collected at these boundary candidates. We aim to classify the set of candidates, $\mathcal{B}$, into two classes (boundary and non-boundary) with the highest probability given the feature set $\mathcal{F}$:

$$\arg\max_{\mathcal{B}} P(\mathcal{B}|\mathcal{F}). \tag{1}$$

The CRF classifier, which is trained using multi-modal features, is designed to make the classification. For performance comparison, several state-of-the-art classifiers, including three generative classifiers based on decision tree (DT), Bayesian network (BN), naive Bayesian (NB) and three discriminative classifiers based on multilayer perception (MLP), support vector machines (SVM) and maximum entropy (ME), are evaluated. We also investigate the effectiveness of features and how different features complement with each other to improve the story segmentation performance.

3. Modeling Story Boundaries Using Conditional Random Fields

A conditional random field (CRF) is a discriminative probabilistic model recently used for labeling or segmenting sequential data [17]. It is a Markov random field in nature, where each random variable is conditioned on an observation sequence. Fig. 2 illustrates a simple linear-chain CRF frequently used in sequential data labeling, which defines a conditional probability distribution $p(\mathcal{B}|\mathcal{F})$ of a label sequence $\mathcal{B} = (\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_n)$ given input an observation sequence $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_n)$. Specifically for the story segmentation task, $\mathcal{B}$ represents a label sequence with story-boundary or non-story-boundary labels, and $\mathcal{F}$ is the feature observation sequence. We extract acoustic and visual features from the audio and video of broadcast news, and search for lexical features based on speech recognition tran-

scripts.

There are several benefits using CRF to model the features: (1) CRF is capable to accommodate statistically correlated features. It is understandable that features from the same modality usually have semantic dependencies. For two lexical features, a lower lexical similarity usually accompany with a stronger chaining strength at story boundary positions. A similar circumstance could be observed for acoustic features where a long pause usually happened with a speaker change. However, different modality features always compliment with each other, which is thus believed to lead to a more robust segmentation by integrating different sources of information. (2) modeling contextual feature information gains benefits for story segmentation. For one of the most traditional lexical similarity features, we often adopt the depthscore [2], which reflects the contextual variance tendency of lexical similarity, instead of using the lexical similarity directly indicting that the contextual information is essential for this high level structure task.

Based on a training set with the reference labels and the extracted multi-modal features, we train a linear-chain CRF classifier that can thus label an input broadcast news stream with boundary and non-boundary tags at each candidate position. The decoding problem, i.e., finding the most likely label sequence $\hat{\mathcal{B}}$ for the given observation sequence, can be calculated by

$$\hat{\mathcal{B}} = \arg\max_{\mathcal{B}} p(\mathcal{B}|\mathcal{F}), \tag{2}$$

where the posterior probability takes the exponential form:

$$p(\mathcal{B}|\mathcal{F}) = \frac{\exp \sum_k \lambda_k \cdot F_k(\mathcal{B}, \mathcal{F})}{Z_\lambda(\mathcal{F})}. \tag{3}$$

$F_k(\mathcal{B}, \mathcal{F})$ are called feature functions defined over the observation and label sequences. The index k indicates different feature functions, each of which has an associated weight λ_k. For an input sequence $\mathcal{F}$, and a label sequence $\mathcal{B}$

$$F_k(\mathcal{B}, \mathcal{F}) = \sum_i f_k(\mathcal{B}, \mathcal{F}, i). \tag{4}$$

where i ranges over all the input positions, and $f_k(\mathcal{B}, \mathcal{F}, i)$ is either a state function $s_k(\mathcal{B}, \mathcal{F}, i)$ of the entire observation sequence and the label transition at position i in the label sequence, or a transition function $t_k(\mathcal{B}, \mathcal{F}, i)$ of the label at position i and the observation sequence [22]. Z_λ is the normalization term:

$$Z_\lambda(\mathcal{F}) = \sum_{\mathcal{B}} \exp \sum_k \lambda_k \cdot F_k(\mathcal{B}, \mathcal{F}). \tag{5}$$

The CRF model is trained by globally maximizing the conditional distribution $p(\mathcal{B}|\mathcal{F})$ on a given training set. It can trade off decisions at different sequence positions to obtain a globally optimal labeling. The most likely label sequence is found using the Viterbi algorithm.

When $t_k(\mathcal{B}, \mathcal{F}, i) = t_k(\mathcal{B}_{i-1}, \mathcal{B}_i, \mathcal{F}, i)$, a first-order linear-chain CRF is formed, which only includes two sequential labels in the feature set. For an N-order linear-chain CRF, the feature function is defined as $t_k(B_{i-N}, \cdots, B_i, \mathcal{F}, i)$. The probability of a transition between labels depend not only on the current observation, but also on past, future observations and previous labels. Although there are only two classes in our label set, we do believe previous labels affect current decision making. In our task, it is impossible for two adjacent candidates to be both boundaries. In contrast, if several former labels are all non-boundaries, the current candidate have more chance to be a boundary. Training is only practical for lower orders of N since the computational cost increases exponentially with N. Specifically, if we substitute $\mathcal{F}$ and $\mathcal{B}$ in Eq. (2) - (5) with F_i and B_i, the CRF model is downgraded to an ME model. The ME classifier individually classifies each data sample without using any contextual information, whereas a CRF models sequential information and performs a global optimal labeling.

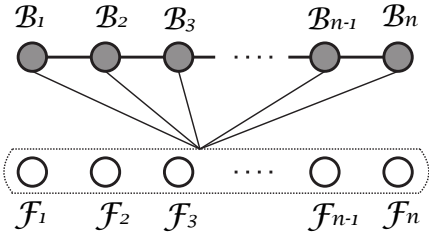

Fig. 2 A linear-chain conditional random field.

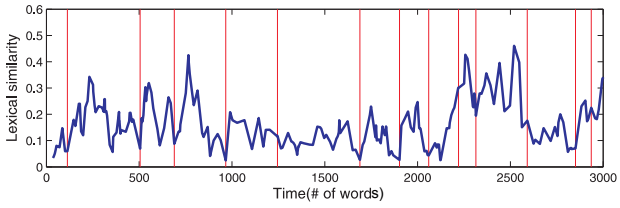

Fig. 3 Lexical similarity curve for a CCTV news episode. Vertical red lines denote the reference story boundaries.

4. Multi-modal Features Extraction

We extract story boundary features from lexical, audio and video modalities. Lexical features consist of lexical similarity, chain strength and overall cohesiveness; acoustic features involve pause duration, pitch, speaker change and audio event type; visual features contain shot boundary, anchor face and news title caption.

4.1 Lexical Features

All lexical features are extracted from Chinese Character (instead of Chinese word) unigram sequences based on the Mandarin Large Vocabulary Continue Speech Recognition (LVCSR) transcripts. The corpora we evaluated on are TDT2 (Topic Detection and Tracking) Mandarin audio corpus from Linguistic Data Consortium (LDC) and the home-grown China Central Television (CCTV) video corpus. We

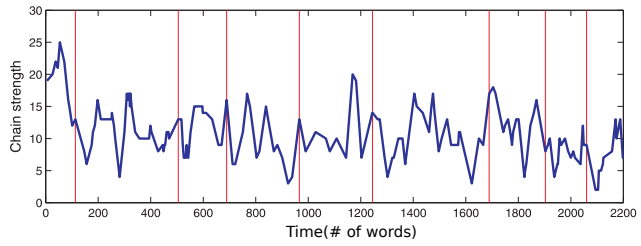

Fig. 4 Chain strength curve for a CCTV news episode. Vertical red lines denote reference story boundaries.

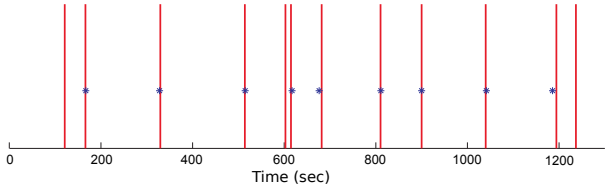

Fig. 5 Detected story boundaries (blue stars) by the overall cohesiveness-based indicator versus reference topic boundaries (vertical read lines) for a CCTV news episode.

got the transcription of TDT2 corpus also from LDC. For CCTV corpus, we construct our own broadcast news recognizer [23]. The Word Error Rate (WER) and Character Error Rate (CER) are 37% and 20% for TDT2, 25% and 18% for CCTV, respectively.

Lexical Similarity: Lexical cohesion indicates the lexical relationship of words within a story while different stories employ different sets of words. [2]. As a result, a story boundary may be detected based on the shift of word usage or the lexical similarity between sentences. We extract the lexical similarity scores as a story boundary feature from the broadcast news transcripts. To capture the variation tendency of lexical similarity, we also compute the difference among its neighbors named SimDelta. The cosine similarity is calculated at each inter-sentence position g in the transcripts:

$$\begin{aligned} lexscore(g) =& cos(\mathbf{v}_s, \mathbf{v}_{s+1}) \\ =& \frac{\sum_{i=1}^{I} v_{s,i} v_{s+1,i}}{\sqrt{\sum_{i=1}^{I} v_{s,i} v_{s,i} \sum_{i=1}^{I} v_{s+1,i} v_{s+1,i}}} \end{aligned} \quad (6)$$

where $\mathbf{v}_s$ and $\mathbf{v}_{s+1}$ are term (i.e. word) frequency vectors for the left and right sentences of g, respectively, and $v_{s,i}$ is the frequency of term w_i occurred in the sentence s with a vocabulary size of I. Since sentences boundaries are not given in the speech recognition transcripts, we apply a block of fixed-length text as a sentence. Fig. 3 shows a lexical similarity curve calculated from the speech recognition transcripts of a CCTV broadcast news episode. We can observe a good match between the story boundaries and the similarity valleys.

Chain Strength: Lexical chaining is another embodiment of lexical cohesion. A lexical chain links up repeating terms where a chain starts at the first appearance of a term and ends at the last appearance of the term. Due to the lexical cohesion, chains tend to start at the begin of the story and terminate at the end of the story. Therefore, a high concentration of starting and/or ending chains is an indicator of story boundary [24]. We measure the chaining strength at each inter-sentence position g by

$$chainstrength(g) = endchain(s) + startchain(s+1) \quad (7)$$

where $endchain(s)$ and $startchain(s+1)$ denote the number of chains end at the sentence s and the number of chains begin at the sentence $s+1$ of g, respectively. Similarly, fixed-length text blocks are used as 'sentences'. The variation tendency of chain strength is also adopted as one dimension of lexical features. We set up a maximal chain length and beyond which no chains are allowed. This is because some terms in a news story may re-appear in another story. For example, some chains may span across the entire text if two news reporting the same topic are situated at the beginning and end of a news program. Fig. 4 plots a chain strength curve of a CCTV broadcast news episode. We can clearly observe that story boundary positions tend to have higher boundary strength scores.

Overall Cohesiveness: When a topic has sharp variations in the lexical distribution, lexical similarity and chain strength, which focus on local cohesiveness, are quite effective. However, sometimes topic transitions among stories in broadcast news are smooth and the distributional variations are very subtle. Therefore, we adopt an overall cohesiveness that directly maximizes the total cohesiveness of all topic fragments split out from the text. This boundary indicator can effectively catch smooth story changes.

The lexical cohesiveness of a fragment f is defined by

$$Cohscore(f) = A[length(f)] \sum_{i=1}^{I} [R(w_i) S(w_i)] \quad (8)$$

where w_i is the ith term of fragment f. $R(w)$ is the repetition of the term w, indicating that each pair of identical words contained in fragment f contributes equally to the cohesiveness of f. Thus the total contribution of the word w_i is given by

$$R(w_i) = \sum_{k=1}^{Freq(w_i)-1} k = \frac{1}{2} Freq(w)[Freq(w_i) - 1] \quad (9)$$

where $Freq(w_i)$ is the term frequency of w_i in fragment f.

$S(w_i)$ is used to measure the inter-fragment discriminability for the term w_i, reflecting the fact that the term appearing the more fragments are less useful in discriminating a specific fragment:

$$S(w_i) = \frac{Freq(w_i)}{Total(w_i)} \quad (10)$$

where $Total(w_i)$ is the number of times that term w_i occurs in the whole text.

As a normalization factor, $A(length)$ should be decrease

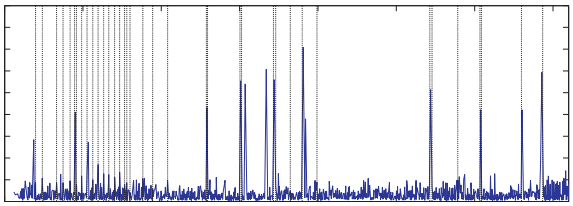

Fig. 6 Pause duration time trajectory for a VOA broadcast news episode. Dotted vertical lines denote story boundaries.

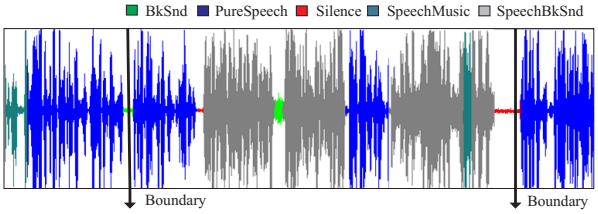

Fig. 7 A clip of annotated CCTV broadcast news audio. News stories usually starts from clean speech.

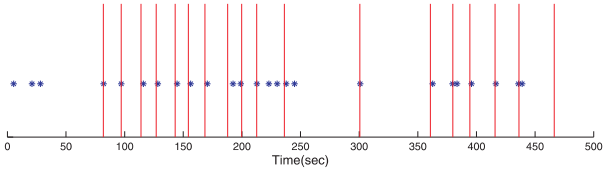

Fig. 8 Detected speaker changes (blue stars) versus reference story boundaries (vertical read lines) for a brief news audio clip in the TDT2 corpus.

slowly when $length(f)$ is not very large as it should not offset the cohesiveness gained by the increase of word repetition. If the length of a fragment is much longer than the average length of the topic, $A(length)$ should provide a considerable negative effect as a penalty factor. We found that an exponential function with a base close to 1.0 serve our needs well. Formally, the length factor is defined as

$$A(length(f)) = \alpha^{-length(f)} \tag{11}$$

where α is a constant parameter slightly larger than 1.0.

We define the *overall cohesiveness* of a text segment as the sum of cohesiveness value of all fragments split out from it, i.e.,

$$C(text) = \sum_{i=1}^{I} Cohescore(f_i). \tag{12}$$

To obtain the optimal text segments, we adopt the segmentation scheme to $C(text)$ by using dynamic programming algorithm. Assume that the whole text consists of n words, represented as $w_1 w_2 ... w_n$. Let $F(n)$ denote the objective function, i.e.

$$F(n) = max[C(w_1 w_2 ... w_n)]. \tag{13}$$

The dynamic programming is conducted as follows:

$$F(i) = max_{0<=j<=i}[F(j) + Cohscore(w_{j+1} ... w_i)] \tag{14}$$

with $F(0) = 0$. Fig. 5 shows the segmentation results (blue stars) by the overall cohesiveness-based indicator versus reference story boundaries (vertical read lines) for a CCTV news episode. We align the boundary of each segmentation to its nearest pause (i.e. story boundary candidate) as the boundary indicator.

4.2 Audio Features

Pause Duration: Pause duration is one of the most important speech prosodic factors relevant to discourse structures. Speakers tend to use a long pause at semantic boundaries. The pause duration between different stories usually lasts longer than between sentences. On the other hand, broadcast news producers usually insert a clear silence or a music clip between news stories. Previous works have shown that pause duration is effective for story segmentation of broadcast news [5], [6]. Fig. 6 shows the pause duration time trajectory of a VOA (Voice of American) broadcast news episode. We can clearly see the pause melody, where pause duration is much salient at story boundaries. We used an home-grown audio classifier [25] to label a broadcast news audio stream into clips of six types: music, pure speech, speech with background sound, speech with music, background sound and silence. Here, all the detected silence are regarded as pause duration as a prosodic feature, namely PauD.

Audio Event Type: According to the editorial rules of broadcast news, studio-to-field transitions often coincide with inter-news boundaries; a news story usually starts from clean speech (e.g. anchor speech in studio) and rarely start from noisy speech (e.g. field speech), as shown in Fig. 7. Studio speech is clean in general while field speech is often contaminated with diverse background noises from news scenes such as streets, factories and buildings, etc. Therefore, audio event type may suggest potential topic boundaries. We use an SVM binary tree (SVM-BT) approach [25] to hierarchically classify an audio clip into six classes: pure speech, speech with noise, speech with music, music, silence and noise. The SVM-BT architecture can realize coarse-to-fine multi-class classification with high accuracy and efficiency.

Speaker Change: Broadcast news programs usually contain various speakers, such as anchors, reporters, interviewees, etc. Many news sessions are hosted by two anchors and they report news in turn. For example, a male anchor and a female anchor usually alternate with each other to announce news in a news session. Fig. 8 shows an example, where most detected speaker changes are at story boundaries. Some news programs follow a clear syntax: a news story is led in by an anchor in the studio, and then followed by a detail report from a field reporter or an interview. Therefore, in broadcast news, speaker changes may coincide with story transitions. We use a two-stage multi-feature in-

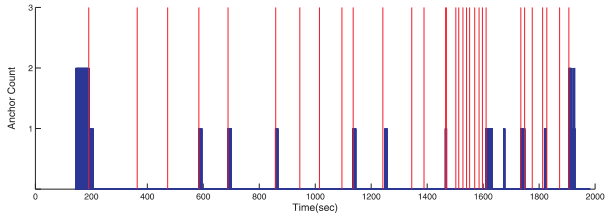

Fig. 9 Anchor face counts (blue bars) versus reference story boundaries (vertical red lines) for a CCTV news episode.

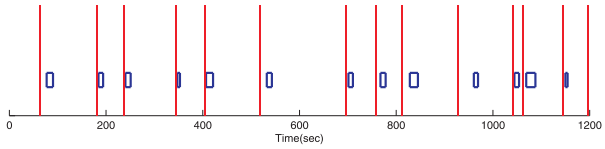

Fig. 10 Appearance of detected title captions (blue boxes) versus reference story boundaries (vertical read lines) for a CCTV news episode.

Table 1 Corpora for story segmentation experiments

Corpus		TDT-2 Mandarin	CCTV
Source		VOA newscast	China Central TV
Media		audio, text	audio, video, text
No. of programs		177	71
Audio duration		53h	27h
LVCSR WER		37%	25%
Data assignment	Training	90 programs (1321 bnds)	40 (1209 bnds)
	Testing	87 programs (1262 bnds)	31 (892 bnds)

tegration approach to automatically detect speaker changes from broadcast news audio [26]. Speaker change is used as a binary feature (Change/Not-change for each candidates).

Pitch: Pitch declination and reset phenomena are characterized by the tendency of a speaker to raise his/her pitch to the topline at the beginning of a major speech unit, and lower it towards the pitch baseline at the end of the major speech unit [7]. Therefore, pitch undergoes a declination within the major speech unit and a reset between two major speech unit. Pitch declination and reset behaviors have been observed more often at topic level than other smaller speech levels like utterances [5], [6], [19].

In this study, we extract pitch trajectory from broadcast news audio by the YIN pitch tracker [27]. The left and right nearest successive pitch contours of each boundary candidate (i.e. pause segment) are determined as our regions of interest. A set of three pitch features are extracted from each boundary candidate, including mean pitch before and after a candidate (PLmn and PRmn) and pitch reset (PReset, i.e. PRmn-PLmn). Since pitch is a speaker-dependent characteristic, we normalize the pitch contour by the speaker before the pitch feature calculation. The speaker boundaries are automatically determined by the detected speaker change [26].

4.3 Video Features

Shot Boundary: We notice that, in broadcast news video, news story transitions usually accompany with a shot change. Therefore, it is reasonable to detect whether there is a shot change at a story boundary candidate. We measure block histogram difference between two adjacent video frames to decide whether a shot boundary exists. First, a frame k is divided into $M * N$ blocks and a gray-scale histogram $h(m, n, k)$ is calculated for each block (m,n). The histogram difference between frames k and $k + 1$ is calculated by

$$D(k, k+1) = \sum_{m=1}^{M} \sum_{n=1}^{N} |h(m, n, k) - h(m, n, k+1)|. \quad (15)$$

A shot boundary is detected if the calculated distance $D(k, k + 1)$ is larger than a pre-set threshold. Shot boundary is used as a binary feature (yes, no) for each story boundary candidate.

Anchor Face: According to the structural rules of broadcast news, many news stories begin with a studio anchor shot and then move to field shots. Previous research shows that anchor face presence is an important visual cue for story boundary detection [16], [28]. We first use an AdaBoost detector to detect human faces in video frames, and then use a regression classifier to discriminate anchor faces from other detected non-anchor faces. Based on the characteristics of anchor appearances, e.g., face coordinates and size, the classifier labels video frames with anchor face counts (0,1,2). Fig. 9 shows the anchor face counts for a CCTV news episode. We can clearly see the anchor face count changes at some story boundaries. Therefore, we use inter-frame anchor face count difference at a candidate position as a visual feature for story boundary detection.

News Title Caption: In broadcast news video, a news story is often accompanied by a caption indicating the title of the news. Hence, the appearance of a title caption is a clear story boundary indicator. We detect news title captions from broadcast news video based on the color and structural information of the caption region. Since the title caption usually comes out later than the news, we measure the time distance from a boundary candidate to the appearance of its right nearest title caption as the feature. In Fig. 10, the blue boxes indicate the appearance of title captions and their durations. We can clearly see that almost every story boundary is associated with a right-side appearance of title caption.

5. Experiments

5.1 Experimental Setup

We carried out story segmentation experiments on two Mandarin broadcast news corpora, the LDC TDT2 Mandarin audio corpus and the home-grown CCTV video corpus, to evaluate the proposed approach. Table 1 shows the details of the two corpora and the data organization for experiments. We extracted audio, video and lexical features for the CCTV video corpus, and audio and lexical features for the TDT2

Table 2 Lexical, audio and video feature sets used in the experiments

set	Features	Abbreviation	Value
Lexical	Lexical Similarity	LexSim	Continuous
	Similarity Variation	SimDelta	Continuous
	Chain Strength	ChStr	Continuous
	Chain Variation	ChainDelta	Continuous
	Global Cohesiveness	GlbCoh	Binary
Audio	Pause Duration	PseDur	Continuous
	Speaker Change	SpkChg	Binary
	Audio Event Type	AETyp	Discrete
	Pitch Left Mean	PLmn	Continuous
	Pitch Right Mean	PRmn	Continuous
	Pitch Reset	PReset	Continuous
Video	Shot Boundary	ShotBnd	Binary
	Anchor Count	AchrCnt	Triple
	Caption Distance	CapDist	Continuous

Table 3 The accuracy rates of feature extraction methods

Features	ShotBnd	AchrCnt	Caption	AETyp	SpkChg
Accuracy	0.935	0.967	0.948	0.960	0.813

audio corpus. We conducted the experiments with feature sets from a single modality (L, A, V) and integrated feature sets from multiple modalities (L+A, L+A+V). The full list of feature sets are shown in Table 2. Note that the position of the candidate (to the beginning of the broadcast news episode), namely Pos, was inserted into all the feature sets in the experiments. The Pos feature was used as a time-dependent heuristics. Table 3 reports the accuracies of our shot boundary detection, anchor counts, caption detection, audio event type detection and speaker change detection, which are tested on an extra validation set. We compared the detected story boundaries with the manually annotated boundaries in terms of *recall, precision* and their harmonic mean-*F1-measure*. According to the TDT evaluation standard, a detected story boundary is consider correct if it lies within a 15-second tolerant window on each side of a manually- annotated reference boundary.

Since the recorded broadcast news audio may have channel noises, we consider background sound positions together with silence and music positions as the story boundary candidates in the experiments. After feature extraction, some features, e.g., GlbCoh, SpkChg and ShotBnd, need to be aligned to an appropriate candidate because they are not always show up exactly at a candidate position. We aligned GlbCoh and ShotBnd to its nearest pause. Speaker change (SpkChg) points were matched with their left nearest candidates due to a detection delay. To keep a reasonable dynamic range of feature values, we normalize all the continuous features into [0,1] by

$$\mathcal{F}_v = \frac{F_v - F_{min}}{F_{max} - F_{min}}. \tag{16}$$

5.2 Story Segmentation with CRF

We trained a CRF boundary/non-boundary classifier using the labeled candidates in the training set. We adopted the GRMM toolkit [†] to perform CRF training and testing while the GRMM is modified to support real-value feature input. Different CRF orders N ($\mathcal{B} = \mathcal{B}_{i-N}, \cdots, \mathcal{B}_i$) and feature contexts M ($\mathcal{F} = \mathcal{F}_{i-M}, \cdots, \mathcal{F}_i, \cdots, \mathcal{F}_{i+M}$) have been tested in order to achieve the best story segmentation performance. Order N is limited to 2 due to the exponential computation cost for high orders and the data sparseness problem. Feature context M indicates the amount of the preceding and following features are used besides the current $\mathcal{F}_i$.

Table 4 and 5 show the story segmentation results using CRF on the CCTV and TDT2 corpora, individually. We also report the performance on training data to comparing evaluations of closed and open. The results show that: (1) with the increase of the sequental/contextual information (M), the story segmentation performance is generally improved on both corpora; (2) multi-modal feature integration significantly outperform single-modal feature set in story segmentation. We notice that the best F1-measures

Table 4 Experimental results for CRF with different N and M on the CCTV corpus

Orders(N) \ Context(M)			0	1	2	3
Training	Lexical	1	0.7044	0.6981	0.7244	0.7354
		2	0.7243	0.7478	0.7451	**0.7581**
	Acoustic	1	0.7254	0.7334	0.7640	0.7604
		2	0.7438	0.7600	**0.7844**	0.7534
	Visual	1	0.5209	0.5769	0.5207	0.6458
		2	0.5207	**0.7178**	0.6888	0.7038
	L+A	1	0.7995	0.8226	0.8329	0.8263
		2	0.8140	0.8371	**0.8432**	0.8364
	L+A+V	1	0.8379	0.8528	0.8625	0.8729
		2	0.8374	0.8686	0.8766	**0.8832**
Testing	Lexical	1	0.6901	0.6830	0.7141	0.7119
		2	0.7006	0.7198	0.7310	**0.7361**
	Acoustic	1	0.7204	0.7334	0.7416	0.7420
		2	0.7404	**0.7518**	0.7365	0.7367
	Visual	1	0.5524	0.5299	0.5601	0.5607
		2	**0.7046**	0.6760	0.6992	0.6887
	L+A	1	0.7955	0.8086	0.8180	**0.8204**
		2	0.7965	0.8078	0.8095	0.8139
	L+A+V	1	0.8366	0.8516	**0.8576**	0.8559
		2	0.8397	0.8334	0.8565	0.8443

Table 5 Experimental results for CRF with different N and M on the TDT2 corpus

Orders(N) \ Context(M)			0	1	2	3
Training	Lexical	1	0.6393	0.6350	0.6407	0.6296
		2	0.6702	**0.6769**	0.6738	0.6734
	Acoustic	1	0.7786	0.7743	0.7978	0.7989
		2	0.7826	0.7808	0.7691	**0.7897**
	L+A	1	0.7694	0.7883	0.8007	0.7988
		2	0.7768	0.7793	0.8004	**0.8229**
Testing	Lexical	1	0.6708	0.6652	0.6690	0.6629
		2	0.6964	0.7039	0.7034	**0.7175**
	Acoustic	1	0.7051	0.7056	0.7140	0.7241
		2	0.7094	0.6908	0.7123	**0.7269**
	L+A	1	0.7492	0.7665	0.7768	0.7802
		2	0.7717	0.7729	0.7847	**0.7981**

[†] http://mallet.cs.umass.edu/grmm/

for the lexical feature set (L), the acoustic feature set (A) and the visual feature set (V) on testing set are 0.7361, 0.7518, 0.7046 on the CCTV corpus, respectively. For the TDT2 corpus, the lexical feature set (L) and the acoustic feature set (A) achieve F1-measure scores of 0.7175 and 0.7269, respectively. These results show that the features from three modalities can achieve comparable story segmentation performance. When features from different modalities are combined, F1-measure is increased to 0.8204 (L+A, $N = 1, M = 3$) and 0.8576 (L+A+V, $N = 1$, $M = 2$) on the CCTV corpus and 0.7981 (L+A, $N = 2$, $M = 3$) on the TDT2 corpus.

5.3 Comparison with different classifiers

For performance comparison, we also tested several popular classifiers, i.e., C4.5 decision tree (DT), naive Bayesian classifier (NB), RBF-kernel support vector machines (SVM), multi-layer perceptron (MLP), Bayesian network (BN) and the maximum entropy classifier (ME). The Weka tookit [†] was used to train DT, NB, SVM, MLP, BN and SVM classifiers and the ME classifier was trained using the opennlp.maxent package [††].

Since some features may have low discriminative ability, we performed a feature selection procedure to find the optimal feature subset with highest F1-measure. We adopted the backward elimination algorithm to seek the optimal subset by iteratively eliminating features whose absence did not decrease performance on different classifiers and corpora. Parameter tuning, classifier training and feature selection were performed on the training set and experimental results were reported on the testing set.

Experimental results for different classifiers on the CCTV and TDT2 corpora are listed in Table 6 and Table 7. We can clearly observe that the CRF classifier outperforms other classifiers for both individual feature sets and the integrated feature sets on the two tested corpora. From feature selection, we find that not all features contribute to story boundary detection for a particular classifier. Some features are removed due to their low discriminative ability or its correlation with other more effective features. After feature selection, the highest F1-measure for the two corpora are 0.8607 (for CCTV) and 0.7981 (for TDT2), respectively. We also notice that feature selection approach selects different optimal subsets for different corpora and different classifiers.

6. Conclusion

In this paper, we propose to integrate multi-modal features using conditional random fields (CRF) for automatic story segmentation of broadcast news. Features from different modalities, i.e., audio, visual, and lexical modalities, are extracted for sequential boundary/non-boundary tagging of a story boundary candidate set. Sequential inter-label relations and contextual information are effectively captured by a linear-chain CRF. Experimental results in story segmentation have shown that: (1) the CRF approach outperforms other competitive classifiers, i.e., DT, BN, NB, SVM and MLP; (2) multi-modal feature integration shows significant performance gain over features from single modalities.

[†]http://www.cs.waikato.ac.nz/ml/weka/

[††]http://opennlp.sourceforge.net/

Table 7 Experimental results (F1-measure) for different feature sets and classifiers on TDT2. L+A*: feature selection performed.

Classifier	Lexical	Acoustic	L+A	L+A*	Removed in feature selection
DT	0.6829	0.7144	0.7381	0.7566	PRmn
BN	0.6744	0.6936	0.7652	0.7712	LexSim
NB	0.5934	0.6543	0.6960	0.7313	AETyp ChStr SimDelta
MLP	0.6652	0.7249	0.7654	0.7600	—-
SVM	0.7076	0.7038	0.7572	0.7583	PLmn
ME	0.6687	0.6848	0.7441	0.7441	—-
CRF	**0.7175**	**0.7269**	**0.7981**	**0.7981**	—-

References

[1] M. Franz, J. McCarley, T. Ward, and W. Zhu, "Segmentation and Detection at IBM: Hybrid Statistical Models and Two-tiered Clustering," TDT-3 Workshop, 2000.

[2] M.A. Hearst, ""TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages"," Computational Linguistics, vol.23, no.1, pp.33–64, 1997.

[3] X. Wang, L. Xie, B. Ma, E. Chng, and H. Li, "Phoneme Lattice based TextTiling towards Multilingual Story Segmentation," Proc. of Interspeech, pp.1305–1308, 2010.

[4] S. Chan, L. Xie, and H. Meng, "Modeling the Statistical Behavior of Lexical Chains to Capture Word Cohesiveness for Automatic Story Segmentation," Interspeech, 2007.

[5] L. Xie, "Discovering Salient Prosodic Cues and Their Interactions for Automatic Story Segmentation in Mandarin Broadcast News," Multimedia Systems, vol.14, pp.237–253, 2008.

[6] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based Automatic Segmentation of Speech into Sentences and Topics," Speech communication, vol.32, no.1-2, pp.127–154, 2000.

[7] C.Y. Tseng, S.H. Pin, Y. Lee, H.M. Wang, and Y.C. Chen, "Fluent speech prosody: Framework and modeling," Speech Communication, vol.46, pp.284–309, 2005.

[8] W. Hsu, S. Chang, C. Huang, L. Kennedy, C. Lin, and G. Iyengar, "Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation," SPIE Electronic Imaging, 2004.

[9] H. Winston, H. Hsu, and S. Chang, "A Statistical Framework for Fusing Mid-level Perceptual Features in News Story Segmentation," International Conference on Multimedia and Expo, 2003.

[10] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. Van Mulbregt, "A Hidden Markov Model Approach to Text Segmentation and Event Tracking," ICASSP, pp.333–336, 1998.

[11] J. Zhang, L. Xie, W. Feng, and Y. Zhang, "A Subword Normalized Cut Approach to Automatic Story Segmentation of Chinese Broadcast News," Information Retrieval Technology, pp.136–148, 2009.

[12] A. Hauptmann and M. Witbrock, "Story Segmentation and Detection of Commercials in Broadcast News Video," IEEE International Forum on Research and Technology Advances in Digital Libraries, pp.168–179, 2002.

[13] G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, "Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation," Computational linguistics, vol.27, no.1, pp.31–57, 2001.

[14] W. Qi, L. Gu, H. Jiang, X. Chen, and H. Zhang, "Integrating Visual, Audio and Text Analysis for News Video," Proc. of International Conference on Image Processing, pp.520–523, 2002.

[15] A. Rosenberg and J. Hirschberg, "Story Segmentation of Broadcast News in English, Mandarin and Arabic," Proc. of HLT-NAACL,

Table 6 Experimental results (F1-measure) for different feature sets and classifiers on CCTV. L+A+V*: feature selection performed.

Classifier	Lexical	Acoustic	Visual	L+A	L+A+V	L+A+V*	Removed in feature selection
DT	0.6688	0.7224	0.6974	0.7744	0.8174	0.8187	AchrCnt
BN	0.6974	0.7361	0.6753	0.7907	0.8431	0.8432	PReset
NB	0.6696	0.5624	0.4087	0.7184	0.7453	0.7571	AETyp SimDelta ChStr
MLP	0.6934	0.7244	0.6934	0.7842	0.8054	0.8231	AchrCnt PRmn ChainDelta
SVM	0.6769	0.7125	0.4244	0.7845	0.8077	0.8077	—-
ME	0.6767	0.6745	0.5881	0.7606	0.7973	0.8006	PLmn PRmn PReset ChainDelta
CRF	**0.7361**	**0.7518**	**0.7046**	**0.8204**	**0.8576**	**0.8607**	PReset

pp.125–128, 2006.

[16] W. Hsu, L. Kennedy, S. Chang, M. Franz, and J. Smith, "Columbia-IBM News Video Story Segmentation in TRECvid 2004," CIVR, 2005.

[17] J. Lafferty, A. McCallum, and F. Pereira, ""Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data"," Proc. 18th Int. Conf. on Machine Learning, pp.282–289, 2001.

[18] F. Sha and F. Pereira, "Shallow Parsing with Conditional Random Fields," Proc. of HLT-NAACL, pp.213–220, 2003.

[19] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies," IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.5, pp.1526–1540, 2006.

[20] G. Levow, "Automatic Prosodic Labeling With Conditional Random Fields and Rich Acoustic Features," Proc. of IJCNLP, 2008.

[21] J. Morris and E. Fosler-Lussier, "Combining Phonetic Attributes using Conditional Random Fields," Proc. of Interspeech, 2006.

[22] H.M. Wallach, "Conditional random fields: An introduction," tech. rep., 2004.

[23] L. Xie, Y. Yang, and Z.Q. Liu, "On the Effectiveness of Subwords for Lexical Cohesion Based Story Segmentation of Chinese Broadcast News," Information Sciences, vol.181, no.13, pp.287–2891, 2011.

[24] N. Stokes, J. Carthy, and A.F. Smeaton, "SeleCT: A lexical cohesion based news story segmentation system," Jan. 2004.

[25] L. Xie, Z. Fu, W. Feng, and Y. Luo, "Pitch-Density-based Features and an SVM Binary Tree Approach for Multi-Class Audio Classification in Broadcast News," Multimedia Systems, vol.17, no.2, pp.101–112, 2011.

[26] L. Xie and G. Wang, "A Two-Stage Multi-Feature Integration Approach to Unsupervised Speaker Change Detection in Real-Time News Broadcasting," ISCSLP, pp.1–4, 2008.

[27] A. de Cheveigné and H. Kawahara, "YIN, A Fundamental Frequency Estimator for Speech and Music," The Journal of the Acoustical Society of America, vol.111, p.1917, 2002.

[28] C. Ma, B. Byun, I. Kim, and C. Lee, "A Detection-based Approach to Broadcast News Video Story Segmentation," ICASSP, pp.1957–1960, 2009.