Multi-Modal Feature Integration for Story Boundary Detection in Broadcast News

Mi-mi LU, Lei XIE, Zhong-hua FU, Dong-mei JIANG and Yan-ning ZHANG
Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an

***Abstract*—This paper investigates how to integrate multi-modal features for story boundary detection in broadcast news. The detection problem is formulated as a classification task, i.e., classifying each candidate into boundary/non-boundary based on a set of features. We use a diverse collection of features from text, audio and video modalities: lexical features capturing the semantic shifts of news topics and audio/video features reflecting the editorial rules of broadcast news. We perform a comprehensive evaluation on boundary detection performance for six popular classifiers, including decision tree (DT), Bayesian network (BN), naive Bayesian (NB) classifier, multi-layer peceptron (MLP), support vector machines (SVM) and maximum entropy (ME) classifier. Results show that BN and DT can generally achieve superior performances over other classifiers and BN offers the best F1-measure. Analysis of BN and DT reveals important inter-feature dependencies and complementarities that contribute significantly to the performance gain.**

***Index Terms*—story boundary detection, story segmentation, topic detection and tracking, multi-modal, feature integration**

I. Introduction

A multimedia document is usually composed of multiple segments in different semantic granularities. For example, a broadcast news audio/video episode generally contains various news stories, each addressing a central topic. Story boundary detection (or story segmentation) aims to identify where one story ends and another begins in a stream of text, speech or video [1]. It serves as a necessary precursor to various tasks, such as topic detection and tracking, information extraction, indexing, retrieval and summarization, etc. A typical broadcast news retrieval system is able to locate the particular positions in a repository that match the user's query, but lack the ability of determining where the user-interested stories begin and end. Story boundary detection exclusively targets to this purpose. As the overwhelming proliferation of multimedia documents, automatic story boundary detection is highly in demand.

Story boundary detection approaches can be categorized to *detection*-based [2]–[7] and *model*-based [8]–[10]. The former directly detects story boundaries through intuitive cues/features. Based on the features, a detector or classifier is learned to make boundary decisions. The latter models a multi-topical document and segments it into story units under some optimal criterion.

For detecting story boundaries in broadcast news, various boundary features have been studied in different modalities. Word similarity measures are frequently used as boundary indicators in textual documents due to the intuitive *lexical cohesion* phenomena, i.e., words in a story agglomerate together via semantic relations. TextTiling [2] and lexical chaining [3] are two typical embodiments of lexical cohesion for story boundary detection. Different from textual features that reveal topic shifts by detecting semantic variations, audio/video cues are more heuristic and rely on editorial rules [5]. News transitions are signaled by a long silence or a musical break; two announcers report news stories in turn; a studio anchor leads in a news topic and a field reporter elaborates it the detail. Originated from editorial rules, major audio/video cues include anchor-face show-up, news-title show-up, significant pause, speaker change and music [4]–[6], [11].

Despite years of research, story boundary detection is far from practical applications because of its inferior performance. Due to the complexity and generality of the problem, no particular single feature is enough to handle story boundary detection for a large volume of broadcast news. Some recent efforts have shown that integrating different features is able to significantly improve the detection performance [4]–[6], [11]. Among these previous studies, decision tree (DT) and maximum entropy (ME) model are frequently used as the boundary classification scheme. Recently, support vector machines (SVM) [5] and naive bayesian (NB) classifier [7] were adopted for story boundary detection. However, despite these tremendous efforts, we notice that a comprehensive comparison on different classifiers for multi-modal feature integration is still missing and the potential state-of-the-art story boundary detection performances remain unknown. In this paper, we present an extensive evaluation and analysis of different classifiers for story boundary detection on broadcast news. Specifically, we use a diverse set of lexical, audio and video features automatically derived from broadcast news videos, which includes frequently-used typical features and newly-released features [12]. We conduct extensive experimentation on six popular classifiers, including generative classifiers (DT, NB and BN) and discriminative classifiers (SVM, MLP and ME). We investigate feature effectiveness and how different features complement with each other to push forward the state-of-the-art performance of story boundary detection.

II. Corpus

We experiment with the CCTV broadcast news corpus [10], which contains 71 Mandarin broadcast news videos (over 30 hours in duration) from China Central Television. Each video is associated with a large vocabulary continuous speech recognition (LVCSR) transcript. The speech recognition error

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

rates for word, character and base syllable are 25%, 18% and 15.5%, respectively. We divide the corpus into two portions: a set of 40 episodes for training (1209 boundaries) and a set of 31 episodes (892 boundaries) for testing. The reference story boundaries (i.e. real story boundary positions) were manually annotated. We compare the detected story boundaries with the reference boundaries in terms of F1-measure, i.e., the harmonic mean of recall and precision. In accordance with the topic detection and tracking (TDT) 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.

III. Story Boundary Detection Scheme

We adopt a detection-based method that involves three stages: candidate determination, feature extraction and boundary/non-boundary classification. We first determine the boundary candidates across the broadcast news stream. The principle of this stage is to reduce the boundary search complexity and to maintain a very low miss rate of story boundaries. For the video modality, we consider all the shot boundary positions as the story boundary candidates. All the silence positions are used as story boundary candidates for the audio and text modalities. For a candidate π, a set of features is then extracted. We aim to find the boundary class $\mathcal{B}_{\pi}^{*}$ with highest probability given feature set $\mathcal{F}_{\pi}$ for each candidate π:

$$\mathcal{B}_{\pi}^{*} = \arg\max_{\mathcal{B}_{\pi}} P(\mathcal{B}_{\pi}|\mathcal{F}_{\pi}), \mathcal{B}_{\pi} \in \{\text{Bnd}, \text{Non-Bnd}\}. \quad (1)$$

To achieve this, a classifier is trained to assign $\mathcal{B}_{\pi}$ for each candidate.

IV. Multi-modal Feature Extraction

A. Lexical Features

1) Lexical Similarity: Lexical cohesion describes that a story with a central topic is created by use of words with related meanings while different topics employ different sets of words [2]. As a result, a story boundary may accord with a lexical similarity minimum. Based on this intuitive assumption, we measure lexical similarity at each candidate across the broadcast news LVCSR transcript. The cosine similarity is calculated at each candidate π across the transcript:

$$s_{\pi} = \cos(\mathbf{v}_l, \mathbf{v}_r) = \frac{\sum_{n=1}^{N} v_{n,l} v_{n,r}}{\sqrt{\sum_{n=1}^{N} v_{n,l} \sum_{n=1}^{N} v_{n,r}}} \quad (2)$$

where $\mathbf{v}_l$ and $\mathbf{v}_r$ are term (e.g. word) frequency vectors for the left and right text blocks of the candidate π, respectively, and $v_{n,\xi}$ is the frequency of the term w_n occurred in the block ξ with a vocabulary size of *N*. Fig. 1 shows a lexical similarity curve calculated on a CCTV broadcast news transcript, from which we can observe clear similarity valleys at story boundary positions.

2) Chaining Strength: Lexical chaining is another embodiment of lexical cohesion [3]. A lexical chain links up term repetitions. A chain starts at the first appearance of a term and ends at the last appearance of the term. Lexical cohesion leads to the fact that chains tend to start at the story beginning and terminate at the end of the story. Therefore, a high concentration of starting and/or ending chains is an indicator of story boundaries [3]. We construct lexical chains for the LVCSR transcript and measure the chaining strength at each candidate π by

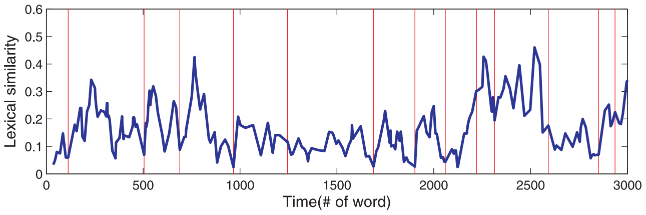

Fig. 1. Lexical similarity curve for a CCTV transcript. Vertical red lines denote reference story boundaries.

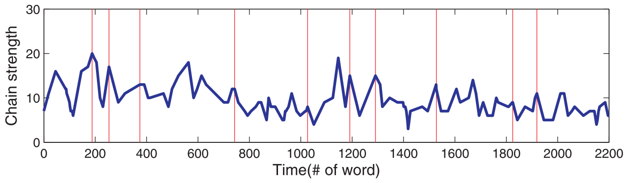

Fig. 2. Chain strength curve for a CCTV transcript. Vertical red lines denote reference topic boundaries.

$$c_{\pi} = end(l) + start(r) \quad (3)$$

where $end(l)$ and $start(r)$ denote the number of chains end at the left text block and the number of chains begin at the right text block of the candidate π, respectively. Since some chains may span across the text if two news report the related or similar topics, we set up a maximal chain length and beyond which no chains are allowed. Fig. 2 plots a chain strength curve of a CCTV broadcast news clip. We can clearly observe that story boundary positions tend to have higher chain strength.

3) Global Cohesiveness-based Boundary Indicator: Lexical similarity and chain strength depict local cohesiveness of a text, which are quite effective when story topics have salient variations in lexical distribution. However, sometimes topic transitions in broadcast news are smooth and the distributional variations are very subtle. Therefore, we use another boundary indicator that directly maximizes the total cohesiveness of all story fragments split out from the text [12]. This boundary indicator can effectively catch smooth topic shifts.

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

$$\mathcal{C}(text) = \sum_{j=1}^{J} Cohscore(f_j) \quad (4)$$

where f_j is the jth fragment. By maximizing $\mathcal{C}(text)$, we can achieve the optimal segmentation. Due to the linearity of the segmentation problem, the maximization can be solved by an efficient dynamic programming solution.

The lexical cohesiveness of a fragment f is mainly affected by three factors: the repetition of terms $- \mathcal{R}(w)$, the specificity of terms $- \mathcal{S}(w)$ and the length of the fragment $- \mathcal{A}(length)$. We combine them by

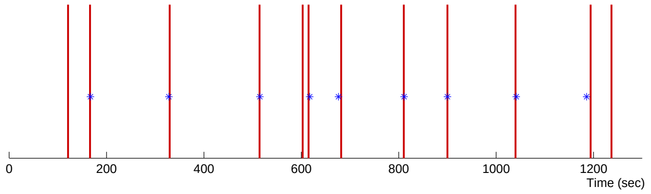

Fig. 3. The segmentation points (blue stars) versus reference story boundaries (vertical red lines) for a CCTV broadcast news transcript.

$$Cohscore(f) = \mathcal{A}[length(f)] \sum_{j=1}^{I} [\mathcal{R}(w_i)\mathcal{S}(w_i)] \quad (5)$$

where w_j is the jth different term of fragment f.

$\mathcal{R}(w_i)$ – The basic idea is that each pair of identical words contained in fragment f contributes equally to the cohesiveness of f. Thus the total contribution of a certain word w_i is given by

$$\mathcal{R}(w_i) = \sum_{k=1}^{Freq(w_i)-1} k = \frac{1}{2} Freq(w_i)\,[Freq(w_i) - 1] \quad (6)$$

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

$\mathcal{S}(w_i)$ – We introduce a specificity factor

$$\mathcal{S}(w_i) = \frac{Freq(w_i)}{Total(w_i)}, \quad (7)$$

which measures the inter-fragment discriminativity for the term w_i, reflecting the fact that the term appearing in more fragments are less useful in discriminating a specific fragment. $Total(w_i)$ is the number of times that term w_i occurs in the whole text.

$\mathcal{A}(length)$ – A factor is introduced to reflect the reasonable fragment length. The length factor should be decreasing and decrease slowly when $length(f)$ is not very large as it should not offset cohesiveness gained by the increase of word repetition. That is to say, a term occurs intermittently within a short distance can be held in a same fragment. Besides, if the fragment is much beyond the typical topic length, $\mathcal{A}(length)$ has to provide a considerable negative effect to the cohesiveness value as a penalty factor. We find that an exponential function with a base close to 1.0 suits our needs well. Formally, the length factor is defined as

$$\mathcal{A}(length(f)) = \alpha^{-length(f)} \quad (8)$$

where α is a constant parameter slightly larger than 1.0.

The dynamic programming algorithm finally results in an optimal segmentation on the text transcript. Fig. 3 illustrates the segmentation points (blue stars) and the reference story boundaries (vertical red lines) for a CCTV broadcast news transcript. We align each segmentation point to its nearest pause (i.e. candidate) as the boundary indicator.

B. Audio Features

1) Pause Duration: Pause duration is a salient speech prosodic factor relevant to discourse structures. Speakers tend to use a long silence pause at large semantic boundaries. Broadcast news producers usually insert a silence or a music clip between consecutive news stories. Previous work has shown that pause duration (i.e. silence and music duration) is quite effective for story boundary detection in broadcast news [5], [11]. Fig. 4 shows the relative frequency histograms of pause duration for story boundaries and non-story boundaries for the CCTV corpus. We can see that story boundaries have longer pause durations. Therefore, we use the pause duration at each candidate point as an audio feature.

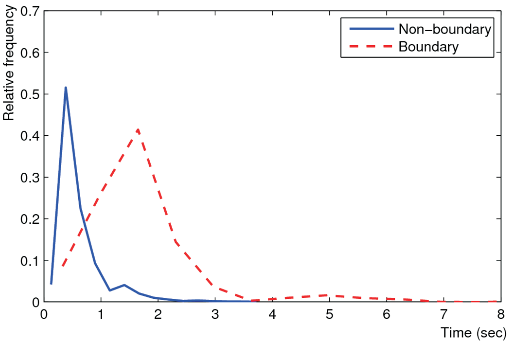

Fig. 4. Relative frequency histograms of pause duration for boundary and non-boundary pauses for the CCTV corpus.

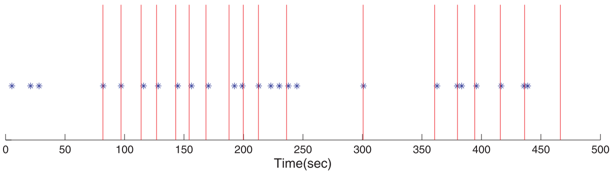

Fig. 5. Detected speaker changes (blue stars) versus reference story boundaries (vertical read lines) for a brief news session in the CCTV corpus.

2) Speaker Change: Broadcast news programs usually involve various speakers, such as anchors, reporters and interviewees, etc. Some news sessions are hosted by two anchors and they report news in turn. For example, in the CCTV brief news session, a male anchor and a female anchor usually alternate with each other to announce news. Fig. 5 shows the detected speaker changes for a brief news clip. We can clearly see that almost every story boundary is associated with a speaker change. Some news programs follow a clear syntax: a news story is led in by an anchor in the studio, and then followed by a detailed 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 integration approach [13] to automatically detect speaker changes from broadcast news audio. Speaker change is used as a binary feature (change/non-change for each candidate).

3) Speech Type: According to the editorial rules of broadcast news, studio-to-field transitions may coincide with 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). 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, speech type may indicate potential story boundaries. We use the speech type of the right side of a candidate point as a discrete audio feature (pure speech, speech with noise, speech with music).

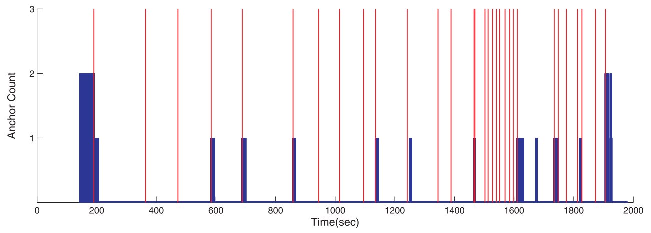

Fig. 6. Anchor face counts (blue bars) versus reference story boundaries (vertical red lines) for a CCTV news episode.

C. Video Features

1) Shot Boundary: A shot is a continuous strip of motion pictures, consisting of a series of frames. In broadcast news video, story transitions usually happen at a shot boundary. 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 $M \times N$ blocks and gray-scale histogram $\mathbf{h}(m,n,k)$ is calculated for each block (m,n). The histogram difference between frames k and $k+1$ is calculated by

$$D(k,k+1) = \sum_{m=1}^{M} \sum_{n=1}^{N} |\mathbf{h}(m,n,k) - \mathbf{h}(m,n,k+1)|. \quad (9)$$

A shot boundary is detected if the calculated distance $D(k,k+1)$ is larger than an empirically set threshold.

2) Anchor Face: With regards to the structural rules of broadcast news, quite a number of 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 [4], [5], [6]. 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 counts (0,1,2). Fig. 6 shows the anchor face counts for a CCTV news episode. We can clearly see the anchor count changes at some story boundary positions. Therefore, we use inter-frame anchor count difference at a candidate position as a visual feature for story boundary detection.

3) Title Caption: In broadcast news video, a news story is often accompanied by a caption describing the title of the news. Hence, the appearance of a title caption is a clear story boundary indicator. We detect title captions from broadcast news video based on caption region's color and structural information. Since the title caption usually comes out later than the news and lasts for a short period, we employ a pair of numeric values to characterize this feature: distance from a candidate to its right nearest title caption appearance and the time duration of the title caption. In Fig. 7, 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 a title caption.

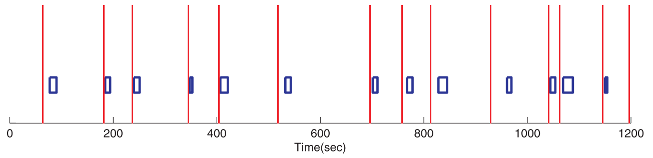

Fig. 7. Appearance of detected title captions (blue boxes) versus reference story boundaries (vertical read lines) for a CCTV news episode.

V. Experiments

A. Experiment Setup

We experimented with feature sets from a single modality (L,A,V) and integrated feature sets from multiple modalities (L+A, L+A+V), summarized in Table I. 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 II reports the accuracies of our shot boundary detection, anchor counts, caption detection, speech type detection and speaker change detection, which are tested on an extra validation set from the same broadcast news source (CCTV).

We chose the silence and music positions (labeled by the LVCSR and an audio classifier) as story boundary candidates for all the experiments except video feature experiments (V). Instead, shot boundaries were selected as story boundary candidates for video feature experiments. This ensures the video-only experiments do not rely on an audio/lexical candidate (pause). After extraction, features such as GlbCoh, SpkChg and ShotBnd need to be aligned to an appropriate candidate since they are not likely to show up exactly at a corresponding pause. We aligned boundaries reported by GlbCoh to their nearest candidates. A shot boundary was also aligned to its nearest pause. Speaker change points were matched with their left nearest candidates due to a detection delay [13]. Note that all the lexical features were calculated on character unigram sequences for the Mandarin LVCSR transcripts due to the robustness of sub-word to speech recognition errors [10].

We tested several popular classifiers, i.e., C4.5 decision tree (DT), Bayesian network (BN), naive Bayesian classifier (NB), RBF-kernel support vector machines (SVM), multi-layer perceptron (MLP) and maximum entropy classifier (ME). The Weka toolkit (downloaded from http://www.cs.waikato.ac.nz/ml/weka/) was used to train DT, BN, NB, SVM and MLP classifiers and ME classifier was trained using the opennlp.maxent package (downloaded from http://maxent.sourceforge.net/).

Since features may be redundant and some features may have low discriminativity, we performed a feature selection procedure to achieve 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 do not decrease performance. All the parameter tuning, classifier training and feature selection were fulfilled on the training set. Experimental results are reported on the testing set.

TABLE III

EXPERIMENTAL RESULTS IN TERMS OF F1-MEASURE FOR DIFFERENT FEATURE SETS AND CLASSIFIERS. L+A+V*: FEATURE SELECTION PERFORMED.

Classifier	Lexical	Acoustic	Visual	L+A	L+A+V	L+A+V*	Removed in feature selection
DT	0.5938	0.7133	**0.7349**	0.7673	0.8011	0.8103	ShotBnd AchrCnt ChStr
BN	**0.6498**	**0.7185**	0.6761	0.7644	**0.8305**	**0.8335**	SpTyp ShotBnd
NB	0.6491	0.7149	0.6188	0.7451	0.8097	0.8190	LexSim AchrCnt
MLP	0.6176	0.6926	0.6116	**0.7679**	0.7865	0.8029	SpTyp
SVM	0.6077	0.5540	0.4765	0.6976	0.7027	0.7105	CapDur AchrCnt LexSim SpTyp SpkChg
ME	0.5617	0.5691	0.6278	0.6685	0.7201	0.7330	CapDur ShotBnd SpTyp

TABLE I

LEXICAL, AUDIO AND VIDEO FEATURE SETS USED IN THE EXPERIMENTS.

Set	Feature	Abbreviation	Value
Lexical	Lexical Similarity	LexSim	Continuous
	Chain Strength	ChStr	Continuous
	Global Cohesiveness	GlbCoh	Binary
Audio	Pause Duration	PseDur	Continuous
	Speaker Change	SpkChg	Binary
	Speech Type	SpTyp	Triple
Video	Shot Boundary	ShotBnd	Binary
	Anchor Count	AchrCnt	Triple
	Caption Distance	CapDist	Continuous
	Caption Duration	CapDur	Continuous

TABLE II

THE ACCURACY RATES OF FEATURE EXTRACTION METHODS.

Feature	ShotBnd	AchrCnt	Caption	SpTyp	SpkChg
Accuracy	0.935	0.967	0.948	0.96	0.813

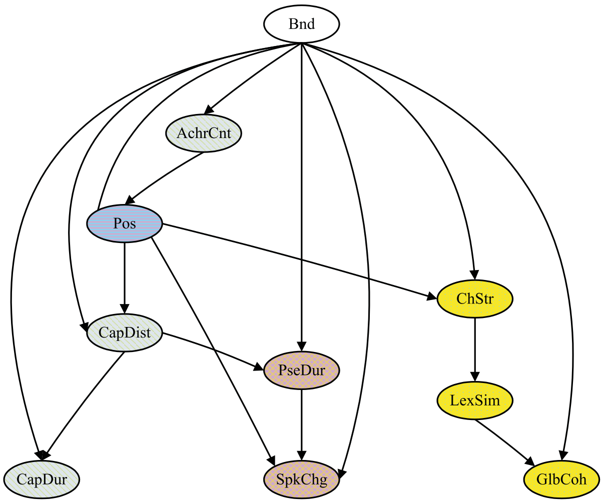

Fig. 8. The Bayesian network for L+A+V* multi-feature set.

B. *Results and Discussions*

Experimental results are summarized in Table III. We notice that the best F1-measure for the lexical feature set (L), the audio feature set (A) and the video feature set (V) are 0.6498, 0.7185 and 0.7349, respectively. This shows that the features from three modalities can achieve comparable story boundary detection performance. When three feature sets are combined, F1-measure is increased to 0.8305. After feature selection, the performance is further increased to the highest F1-measure of 0.8335. Specifically, the integration of lexical and audio features (L+A) also achieve a pretty good F1-measure with 0.7679. We believe the complementarity between features from different modalities contributes to the considerable performance gain.

When comparing different classifiers, we observe that DT and BN achieve superior performances over other classifiers in general. We notice that BN exhibits the best performance for story boundary detection, and SVM and ME show inferior performance as compared with other classifiers. Besides their outstanding performances, BN and DT offer the advantage of *visual interpretability* of features. We can interpret feature behaviors, their dependencies and interactions by observing the automatically learnt graph and tree.

Bayesian network, a directed acyclic graph (DAG), represents a set of random variables (features) and their conditional dependencies. BNs are potentially valuable to take into account the inter-dependencies among the variables that facilitates effective decision-making process. Fig. 8 demonstrates the BN graph achieving the best story boundary detection performance. We can easily understand the semantic dependencies among features by observing whether arcs exist between different features. We notice that features from the same modality are naturally grouped together through inter-feature dependencies. Specifically, node ChStr is connected with node LexSim through a direct arc. This makes sense because a lower lexical similarity with a stronger chaining strength reinforces the probability that a topic shift takes place. A similar structural relation happens between node SpkChg and PseDur. Speaker change occurs more likely with a longer pause duration accompanied.

Decision trees have been extensively used in various event detection tasks, including prosody-based story boundary detection [11]. The mechanism of a decision tree, i.e., split criterion, acts closely to human thinking mode. As compared with BN, DT offers more direct and clear interpretation of feature interactions. Figure 9 (left) shows the top levels of the DT that achieves the best performance. We notice that CapDist is the top node of the tree, reflecting its importance to the boundary decision. This is easy to interpret because almost every news story is accompanied by a news title. We also can observe the complementarity between features from different modalities. This may be summarized in terms of some major heuristic rules. For example, *Path_A* in Fig. 9 comes out naturally: the occurrence of a speaker change near the end of the episode (Pos>1076) enhances the boundary

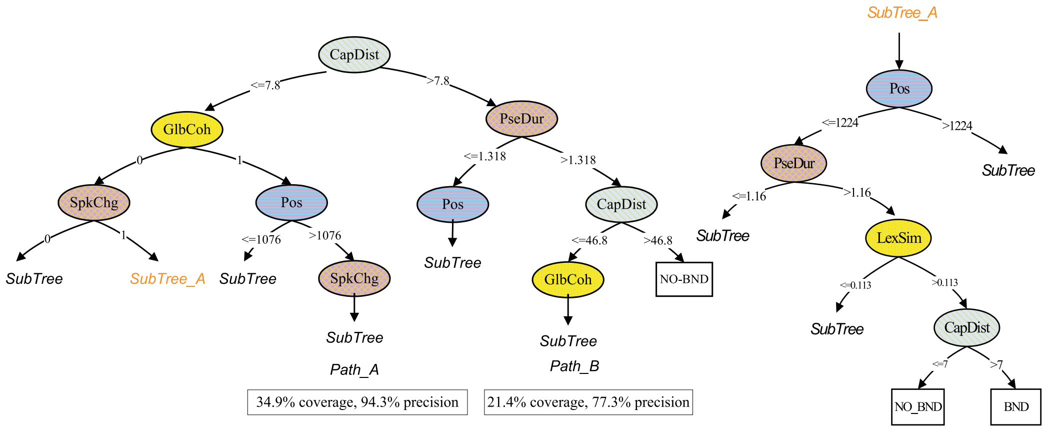

Fig. 9. The decision tree for L+A+V* multi-feature set.

decision making when a candidate is close to the beginning of a title caption (CapDist<=7.8) and a global-cohesion-based boundary (GlbCoh=1) is also reported. This accords with the editorial rule of CCTV broadcast news: the last several minutes of a news episode are brief news alternately reported by two speakers. This path covers 34.9% boundary instances in the testing set and 94.3% of them are correctly classified. *Path_B* shows the employment of PseDur and GlbCoh helps to make boundary decision when the candidate has less distinct CapDist value. This rule is also understandable because the title captions sometimes show up quite late for some news stories. Coverage of boundary cases for this path is 21.4%, in which 77.3% are correctly classified. $SubTree\_A$ in Fig 9 (right) shows another advantage of multi-modal feature integration: when some cues conflict, other cues can help to make the right decision. Specifically, the employment of position information and title caption plays a rescuing role for the boundary decision when lexical and acoustic features conflict (i.e., lexical similarity is high but pause duration is long). This disagreement occurs when two neighboring news stories have related/similar topics.

Some conclusions can be drawn from the feature selection results. Not all features contribute to story boundary detection. As listed in Table III, some features are removed probably due to their lower discriminative ability or correlated with other more effective features. For example, shot boundary and speech type are excluded by most classifiers. The optimal feature set varies for different classifiers; this may suggest that features are supposed to be used according to the capability of the chosen classifier.

VI. Conclusions

This paper has investigated integrating lexical, audio and video features for story boundary detection in broadcast news. Lexical features capture the semantic shifts of news topics and audio/video features reflect the editorial structure of broadcast news. We have experimented with individual feature sets and multi-feature integrated sets. The story boundary detection abilities of six popular classifiers are also examined. Experimental results show that: (1) multi-modal feature integration can significantly boost story boundary detection performance; (2) BN and DT generally outperform other classifiers. We have discovered important inter-feature dependencies and complementarities from DT and BN, which contribute significantly to the performance improvement.

Acknowledgment

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

References

[1] Tat-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives: techniques, experience and trends," in *ACM Multimedia*, 2004, pp. 656 – 659.

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

[3] S. K. Chan, L. Xie, and H. Meng, "Modeling the Statistical Behavior of Lexical Chains to Capture Word Cohesiveness for Automatic Story Segmentation," in *Proc. of Interspeech*, Anterwerp, Belgium, 2007.

[4] C. Ma, B. Byun, I. Kim, and C.-H. Lee, "A detection-based approach to broadcast news video story segmentation," in *Proc. ICASSP*, 2009, pp. 1957–1960.

[5] W. H. Hsu, L. S. Kennedy, S. f. Chang, M. Franz, and J. Smith, "Columbia-ibm news video story segmentation in trecvid 2004," in *Proc. CIVR*, 2005.

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

[7] W. Xiong W.-K. Lo and H. MENG, "Automatic story segmentation using a bayesian decision framework for statistical models of lexical chain features," in *Proc. ACL*, 2009.

[8] 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.

[9] C.-H. Wu and C.-H. Hsieh, "Story segmentation and topic classification of broadcast news via a topic-based segmental model and a genetic algorithm," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1612 1623, 2009.

[10] J. Zhang, L. Xie, W. Feng, and Y. Zhang, "A Subword Normalized Cut Approach to Automatic Story Segmentation of Chinese Broadcast News," in *LNCS 5839*, 2009, pp. 136–148.

[11] G. Tür and D. Hakkani-Tür, "Integrating prosodic and lexical cues for automatic topic segmentation," *Computational Linguistics*, vol. 27, no. 1, pp. 31–57, 2001.

[12] L. Xie, Y. Yang, Z. Liu, W. Feng, and Z. Liu, "Integrating acoustic and lexical features in topic segmentation of chinese broadcast news using maximum entropy approach," in *Proc. ICALIP*, 2010.

[13] L. Xie and G. Wang, "A two-stage multi-feature integration approach to unsupervised speaker change detection in real-time news broadcasting," in *Proc. ISCSLP*, 2008, pp. 350–353.