Story Segmentation of Brodcast News in English, Mandarin and Arabic

Andrew Rosenberg
Computer Science Department
Columbia University
New York City, N.Y. 10027
amaxwell@cs.columbia.edu

Julia Hirschberg
Computer Science Department
Columbia University
New York City, N.Y. 10027
julia@cs.columbia.edu

Abstract

In this paper, we present results from a Broadcast News story segmentation system developed for the SRI NIGHTINGALE system operating on English, Arabic and Mandarin news shows to provide input to subsequent question-answering processes. Using a rule-induction algorithm with automatically extracted acoustic and lexical features, we report success rates that are competitive with state-of-the-art systems on each input language. We further demonstrate that features useful for English and Mandarin are **not** discriminative for Arabic.

1 Introduction

Broadcast News (BN) shows typically include multiple unrelated stories, interspersed with anchor presentations of headlines and commercials. Transitions between each story are frequently marked by changes in speaking style, speaker participation, and lexical choice. Despite receiving a considerable amount of attention through the Spoken Document Retrieval (SDR), Topic Detection and Tracking (TDT), and Text Retrieval Conference: Video (TRECVID) research programs, automatic detection of story boundaries remains an elusive problem. State-of-the-art story segmentation error rates on English and Mandarin BN remain fairly high and Arabic is largely unstudied. The NIGHTINGALE system searches a diverse news corpus to return answers to user queries. For audio sources, the identification of story boundaries is crucial, to segment material to be searched and to provide interpretable results to the user.

2 Related work

Previous approaches to story segmentation have largely focused lexical features, such as word similarily (Kozima, 1993), cue phrases (Passonneau and Litman, 1997), cosine similarity of lexical windows (Hearst, 1997; Galley et al., 2003), and adaptive language modeling (Beeferman et al., 1999). Segmentation of stories in BN have included some acoustic features (Shriberg et al., 2000; Tür et al., 2001). Work on non-English BN, generally use this combination of lexical and acoustic measures, such as (Wayne, 2000; Levow, 2004) on Mandarin. And (Palmer et al., 2004) report results from feature selection experiments that include Arabic sources, though they do not report on accuracy. TRECVID has also identified visual cues to story segmentation of video BN (cf. (Hsu et al., 2004; Hsieh et al., 2003; Chaisorn et al., 2003; Maybury, 1998)).

3 The NIGHTINGALE Corpus

The training data used for NIGHTINGALE includes the TDT-4 and TDT5 corpora (Strassel and Glenn, 2003; Strassel et al., 2004). TDT-4 includes newswire text and broadcast news audio in English, Arabic and Mandarin; TDT-5 contains only text data, and is therefore not used by our system. The TDT-4 audio corpus includes 312.5 hours of English Broadcast News from 450 shows, 88.5 hours of Arabic news from 109 shows, and 134 hours of Mandarin broadcasts from 205 shows. This material was drawn from six English news shows – ABC "World News Tonight", CNN "Headline News", NBC "Nightly News", Public Radio International "The World", MS-NBC "News with Brian Williams", and Voice of America, English three Mandarin newscasts — China National Radio, China Television System, and Voice of America, Mandarin Chinese — and two Arabic newscasts — Nile TV and Voice of America, Modern Standard Arabic. All of these shows aired between Oct. 1, 2000 and Jan. 31, 2001.

4 Our Features and Approach

Our story segmentation system procedure is essentially one of binary classification, trained on a variety of acoustic and lexical cues to the presence or absence of story boundaries in BN. Our classifier was trained using the JRip machine learning al-

gorithm, a Java implementation of the RIPPER algorithm of (Cohen, 1995).[1] All of the cues we use are automatically extracted. We use as input to our classifier three types of automatic annotation produced by other components of the NIGHTINGALE system, speech recognition (ASR) transcription, speaker diarization, sentence segmentation. Currently, we assume that story boundaries occur only at these hypothesized sentence boundaries. For our English corpus, this assumption is true for only 47% of story boundaries; the average reference story boundary is 9.88 words from an automatically recognized sentence boundary[2]. This errorful input immediately limits our overall performance.

For each such hypothesized sentence boundary, we extract a set of features based on the previous and following hypothesized sentences. The classifier then outputs a prediction of whether or not this sentence boundary coincides with a story boundary. The features we use for story boundary prediction are divided into three types: lexical, acoustic and speaker-dependent.

The value of even errorful lexical information in identifying story boundaries has been confirmed for many previous story segmentation systems (Beeferman et al., 1999; Stokes, 2003)). We include some previously-tested types of lexical features in our own system, as well as identifying our own 'cue-word' features from our training corpus. Our lexical features are extracted from ASR transcripts produced by the NIGHTINGALE system. They include lexical similarity scores calculated from the TextTiling algorithm.(Hearst, 1997), which determines the lexical similarity of blocks of text by analyzing the cosine similarity of a sequence of sentences; this algorithm tests the likelihood of a topic boundary between blocks, preferring locations between blocks which have minimal lexical similarity. For English, we stem the input before calculating these features, using an implementation of the Porter stemmer (Porter, 1980); we have not yet attempted to identify root forms for Mandarin or Arabic. We also calculate scores from (Galley et al., 2003)'s LCseg method, a TextTiling-like approach which weights the cosine-similarity of a text window by an additional measure of its component LEXICAL CHAINS, repetitions of stemmed content words. We also identify 'cue-words' from our training data that we find to be significantly more likely (determined by χ^2) to occur at story boundaries within a window preceeding or following a story boundary. We include as features the number of such words observed within 3, 5, 7 and 10 word windows before and after the candidate sentence boundary. For English, we include the number of pronouns contained in the sentence, on the assumption that speakers would use more pronouns at the end of stories than at the beginning. We have not yet obtained reliable part-of-speech tagging for Arabic or Mandarin. Finally, for all three languages, we include features that represent the sentence length in words, and the relative sentence position in the broadcast.

Acoustic/prosodic information has been shown to be indicative of topic boundaries in both spontaneous dialogs and more structured speech, such as, broadcast news (cf. (Hirschberg and Nakatani, 1998; Shriberg et al., 2000; Levow, 2004)). The acoustic features we extract include, for the current sentence, the minimum, maximum, mean, and standard deviation of F0 and intensity, and the median and mean absolute slope of F0 calculated over the entire sentence. Additionally, we compute the first-order difference from the previous sentence of each of these. As a approximation of each sentence's speaking rate, we include the ratio of voiced 10ms frames to the total number of frames in the sentence. These acoustic values were extracted from the audio input using Praat speech analysis software(Boersma, 2001). Also, using the phone alignment information derived from the ASR process, we calculate speaking rate in terms of the number of vowels per second as an additional feature. Under the hypothesis that topic-ending sentences may exhibit some additional phrase-final lenghthening, we compare the length of the sentence-final vowel and of the sentence-final rhyme to average durations for that vowel and rhyme for the speaker, where speaker identify is available from the NIGHTINGALE diarization component; otherwise we use unnormalized values.

We also use speaker identification information from the diarization component to extract some fea-

[1] JRip is implemented in the Weka (Witten et al., 1999) machine learning environment.

[2] For Mandarin and Arabic respectively, true for 69% and 62% with the average distance between sentence and story boundary of 1.97 and 2.91 words.

tures indicative of a speaker's participation in the broadcast as a whole. We hypothesize that participants in a broadcast may have different roles, such as an anchor providing transitions between stories and reporters beginning new stories (Barzilay et al., 2000) and thus that speaker identity may serve as a story boundary indicator. To capture such information, we include binary features answering the questions: "Is the speaker preceeding this boundary the first speaker in the show?", "Is this the first time the speaker has spoken in this broadcast?", "The last time?", and "Does a speaker boundary occur at this sentence boundary?". Also, we include the percentage of sentences in the broadcast spoken by the current speaker.

We assumed in the development of this system that the source of the broadcast is known, specifically the source language and the show identity (e. g. ABC "World News Tonight", CNN "Headline News"). Given this information, we constructed different classifiers for each show. This type of source-specific modeling was shown to improve performance by Tür (2001).

5 Results and Discussion

We report the results of our system on English, Mandarin and Arabic in Table 5. All results use show-specific modeling, which consistently improved our results across all metrics, reducing errors by between 10% and 30%. In these tables, we report the F-measure of identifying the precise location of a story boundary as well as three metrics designed specifically for this type of segmentation task: the pk metric (Beeferman et al., 1999), *WindowDiff* (Pevzner and Hearst, 2002) and C_{seg} (P_{seg} = 0.3) (Doddington, 1998). All three are derived from the pk metric (Beeferman et al., 1999), and for all, lower values imply better performance. For each of these three metrics we let $k = 5$, as prescribed in (Beeferman et al., 1999).

In every system, the best peforming results are achieved by including all features from the lexical, acoustic and speaker-dependent feature sets. Across all languages, our precision–and false alarm rates–are better than recall–and miss rates. We believe that inserting erroneous story boundaries will lead to more serious downstream errors in anaphora resolution and summarization than a boundary omission will. Therefore, high precision is more important than high recall for a helpful story segmentation system. In the English and Mandarin systems, the lexical and acoustic feature sets perform similarly, and combine to yield improved results. However, on the Arabic data, the acoustic feature set performs quite poorly, suggesting that the use of vocal cues to topic transitions may be fundamentally different in Arabic. Moreover, these differences are not simply differences of degree or direction. Rather, the acoustic indicators of topic shifts in English and Mandarin are, simply, not discriminative when applied to Arabic. This difference may be due to the style of Arabic newscasts or to the language itself. Across configurations, we find that the inclusion of features derived from automatic speaker identification (feature set S), errorful as it is, significantly improves performance. This improvement is particularly pronounced on the Mandarin material; in China News Radio broadcasts, story boundaries are very strongly correlated with speaker transitions.

It is difficult to determine how well our system performs against state-of-the-art story segmentation. There are no comparable results for the TDT-4 corpus. On the English TDT-2 corpus, (Shriberg et al., 2000) report a C_{seg} score of 0.1438. While our score of .0670 is half that, we hesitate to conclude that our system is significantly better than this system; since the (Shriberg et al., 2000) results are based on a word-level segmentation, the discrepancy may be influenced by the disparate datasets as well as the performance of the two systems. On CNN and Reuters stories from the TDT-1 corpus, (Stokes, 2003) report a Pk score of 0.25 and a WD score of 0.253. Our Pk score is better than this on TDT-4, while our WD score is worse. (Chaisorn et al., 2003) report an F-measure of 0.532 using only audio-based features on the TRECVID 2003 corpus , which is higher than our system, however, this allows for "correct" boundaries to fall within 5 seconds of reference boundaries. (Franz et al., 2000) present a system which achieves C_{seg} scores of 0.067 and Mandarin BN and 0.081 on English audio in TDT-3. This suggests that their system may be better than ours on Mandarin, and worse on English, although we trained and tested on different corpora. Finally, we are unaware of any reported story segmentation results on Arabic BN.

Table 1: TDT-4 segmentation results. (L=lexical feature set, A=acoustic, S=speaker-dependent)

	English				Mandarin				Arabic			
	F1(p,r)	Pk	WD	C_{seg}	F1(p,r)	Pk	WD	C_{seg}	F1(p,r)	Pk	WD	C_{seg}
L+A+S	.421(.67,.31)	.194	.318	.0670	.592(.73,.50)	.179	.245	.0679	.300(.65,.19)	.264	.353	.0850
A+S	.346(.65,.24)	.220	.349	.0721	.586(.72,.49)	.178	.252	.0680	.0487(.81,.03)	.333	.426	.0999
L+S	.342(.66,.23)	.231	.362	.074	.575(.72,.48)	.200	.278	.0742	.285(.68,.18)	.286	.372	.0884
L+A	.319(.66,.21)	.240	.376	.0787	.294(.72,.18)	.277	.354	.0886	.284(.64,.18)	.257	.344	.0851
L	.257(.68,.16)	.261	.399	.0840	.226(.74,.13)	.309	.391	.0979	.286(.68,.18)	.283	.349	.0849
A	.194(.63,.11)	.271	.412	.0850	.252(.72,.18)	.291	.377	.0904	.0526(.81,.03)	.332	.422	.0996

6 Conclusion

In this paper we have presented results of our story boundary detection procedures on English, Mandarin, and Arabic Broadcast News from the TDT-4 corpus. All features are obtained automatically, except for the identity of the news show and the source language, information which is, however, available from the data itself, and could be automatically obtained. Our performance on TDT-4 BN appears to be better than previous work on earlier corpora of BN for English, and slightly worse than previous efforts on Mandarin, again for a different corpus. We believe our Arabic results to be the first reported evaluation for BN in that language. One important observation from our study is that acoustic/prosodic features that correlate with story boundaries in English and in Mandarin, do not correlate with Arabic boundaries. Our further research will adress the study of vocal cues to segmentation in Arabic BN.

Acknowledgments

This research was partially supported by the Defese Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023.

References

R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In *AAAI/IAAI*, 679–684.

D. Beeferman, A. Berger, and J. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 31:177–210.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9-10):341–345.

L. Chaisorn, T. Chua, C. Koh, Y. Zhao, H. Xu, H. Feng, and Q. Tian. 2003. A two-level multi-modeal approach for story segmentation of large news video corpus. In *TRECVID*.

W. Cohen. 1995. Fast effective rule induction. In *Machine Learning: Proc. of the Twelfth International Conference*, 115–123.

G. Doddington. 1998. The topic detection and tracking phase 2 (tdt2) evaluation plan. In *Proccedings DARPA Broadcast News Transcription and Understanding Workshop*, 223–229.

M. Franz, J. S. McCarley, T. Ward, and W. J. Zhu. 2000. Segmentation and detection at ibm: Hybrid statstical models and two-tiered clustering. In *Proc. of TDT-3 Workshop*.

M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multi-party conversation. In *41st Annual Meeting of ACL*, 562–569.

M. A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

J. Hirschberg and C. Nakatani. 1998. Acoustic indicators of topic segmentation. In *Proc. of ICSLP*, 1255–1258.

J. H. Hsieh, C. H. Wu, and K. A. Fung. 2003. Two-stage story segmentation and detection on broadcast news using genetic algorithm. In *Proc. of the 2003 ISCA Workshop on Multilingual Spoken Document Retrieval (MSDR2003)*, 55–60.

W. Hsu, L. Kennedy, C. W. Huang, S. F. Chang, C. Y. Lin, and G. Iyengar. 2004. News video story segmentation using fusion of multi-level multi-modal features in trecvid 2003. In *ICASSP*.

H. Kozima. 1993. Text segmentation based on similarity between words. In *31st Annual Meeting of the ACL*, 286–288.

G. A. Levow. 2004. Assessing prosodic and text features for segmentation of mandarin broadcast news. In *HLT-NAACL*.

M. T. Maybury. 1998. Discourse cues for broadcast news segmentation. In *COLING-ACL*, 819–822.

D. D. Palmer, M. Reichman, and E. Yaich. 2004. Feature selection for trainable multilingual broadcast news segmentation. In *HLT/NAACL*.

R. J. Passonneau and D. J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Liunguistics*, 23(1):103–109.

L. Pevzner and M. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

M. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. 2000. Prosody based automatic segmentation of speech into sentences and topics. *Speech Comm.*, 32(1-2):127–154.

N. Stokes. 2003. Spoken and written news story segmentation using lexical chains. In *Proc. of the Student Workshop at HLT-NAACL2003*, 49–53.

S. Strassel and M. Glenn. 2003. Creating the annotated tdt-4 y2003 evaluation corpus. http://www.nist.gov/speech/tests/tdt/tdt2003/papers/ldc.ppt.

S. Strassel, M. Glenn, and J. Kong. 2004. Creating the tdt5 corpus and 2004 evalutation topics at ldc. http://www.nist.gov/speech/tests/tdt/tdt2004/papers/LDC-TDT5.ppt.

G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27:31–57.

C. L. Wayne. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *LREC*, 1487–1494.

I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham. 1999. Weka: Practical machine learning tools and techniques with java implementation. In *ICONIP/ANZIIS/ANNES*, 192–196.