Varying Input Segmentation for Story Boundary Detection in English, Arabic and Mandarin Broadcast News

Andrew Rosenberg[1], *Mehrbod Sharifi*[1], *Julia Hirschberg*[1]

[1]Department of Computer Science, Columbia University, New York City, N.Y. 10027
{amaxwell,mehrbod,julia}@cs.columbia.edu

Abstract

Story segmentation of news broadcasts has been shown to improve the accuracy of the subsequent processes such as question answering and information retrieval. In previous work, a decision tree trained on automatically extracted lexical and acoustic features was trained to predict story boundaries, using hypothesized sentence boundaries to define potential story boundaries. In this paper, we empirically evaluate several alternatives to choice of segmentation on three languages: English, Mandarin and Arabic. Our results suggest that the best performance can be achieved by using 250ms pause-based segmentation or sentence boundaries determined using a very low confidence score threshold.

Index Terms: story boundary detection, segmentation

1. Introduction

Broadcast News (BN) shows generally include a series of typically unrelated stories, with occasional commentary and commercials. The goal of story segmentation is thus similar to topic segmentation — identify where one story ends another begins.

The SRI NIGHTENGALE Y1 system searches a diverse news corpus to return answers to user queries. For BN, story segmentation is a necessary pre-processing step since information retrieval, information extraction, anaphora resolution assume the presence of single 'documents', whether from text or from audio. In this paper, we explore the ramifications of different input segmentations to the story segmentation process. In previous work, [1], we defined potential story boundary segments as a subset of hypothesized sentence boundaries provided to our system by ICSI. However, since these boundaries can be errorful, we have recently revisited this decision, testing whether story boundary detection improves if different types of segmentations are used to define our data points.

In section 2 we present a brief survey of previous approaches to story and topic boundary detection. We describe our corpus in section 3. In section 4, we identify the candidate input segmentations we evaluate. In section 6, we discuss the performance of story boundary detection on these different segmentations. In section 7 we conclude and discuss directions for future research.

2. Related Work

The majority of previous approaches to story segmentation have focussed on lexical features, such as word similarily [2], cue phrases [3], cosine similarity of lexical windows [4, 5], and adaptive language modeling [6] to identify story boundaries, generally in text. Among these lexical approaches, it is rare for any input segmentation to be used; each word boundary is considered a candidate story or topic boundary. One exception to this is [3], where the candidate boundaries were prosodic phrase boundaries. BN story segmentation has included acoustic features in detection. These approaches often apply initial segmentations to the material. The set of candidate boundaries used by Shriberg, et al. [7] were pauses with duration greater than 650ms. Others [8, 1] have used an automatic sentence unit detection technique to construct a set of potential story boundaries. Work on non-English BN has generally combined lexical and acoustic measures, as [9, 10] for Mandarin and [11] for Arabic. These approaches have not applied any input segmentation to the material prior to segmentation. Palmer [11] even goes so far as to allow story boundaries to be placed within a word, using "multi-media events" which may be lexical, acoustic or visual to define the set of potential boundary locations.

3. Corpus

The training data used for our study is the TDT4 corpus [12], which includes newswire text and broadcast news audio in English, Mandarin and Arabic. The TDT4 audio corpus includes 312.5 hours of English BN from 450 shows, 88.5 hours of Arabic BN from 109 shows and 134 hours of Mandarin BN 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 Systems and Voice of America, Mandarin Chinese – and two Arabic newscasts – Nile TV and Voice of America, Modern Standard Arabic. All shows were aired between October 1, 2000 and January 31, 2001. In addition to the raw audio signal for each BN document, our module had access to a number of automatically produced annotations, including automatic speech recognition transcripts with word boundaries [13] and inter-word durations, hypothesized sentence boundaries with confidence scores [14], and speaker segmentation (DIARIZATION) hypotheses [15].

4. Input Segmentations

To determine the most useful BN input segmentation for story boundary detection, we first created a number of alternate segmentations, which will be used to define both candidate locations for story boundaries and the unit of analysis for our feature extraction module. These candidates include word segmentation, hypothesized sentence segmentation (calculated using three different confidence score thresholds), acoustic chunking (calculated using two thresholds), and hypothesized intonational phrase boundaries.

For word segmentation inputs, we consider each ASR end-of-word boundary as a candidate story boundary. While this limits input segmentation error to ASR word segmentation error alone, it creates a very skewed distribution of story-boundary/non-story-boundary candidates, making the classification task more difficult. Moreover, since the input segmentation also determines the unit of analysis for feature extraction, using such a small unit makes the inclusion of meaningful contextual information more difficult. This limitation is, clearly, trivial to avoid, but, decoupling the unit of analysis and candidate boundaries then leaves unaddressed the question of identifying the ideal unit of analysis for story segmentation.

Sentence unit (SU) inputs as provided by ICSI are examined in three versions, including the default hypothesis defined by that system and relaxing confidence score threshold from the default .5 to .3 and .1. In previous work [1], we observed that the .5 default was a significant source of error for story boundary detection. However, as the confidence threshold is reduced, the number of hypothesized SUs increases, thereby lowering the target boundary distribution, but increasing the rate of exact coverage.

We also explore pause-based chunking as an input segmentation. Using ASR hypothesized word boundaries, we calculate the pause duration between each word pair, inserting a segmentation boundary at every pause that exceeds one of two predetermined thresholds – 250ms and 500ms. The smaller threshold was chosen to avoid potential confusion of intonationally meaningful pauses with stop gaps [16], and selected the larger based on a hypothesis that shorter pauses may appear between sentences, but longer pauses may signify story boundaries.

In order to evaluate a more linguistically meaningful input segmentation unit, the intonational phrase, we had one expert ToBI labeler manually annotate the ASR-defined words of one TDT4 show, 20010131_1830_1900_ABC_WNT, for two prosodic annotations: a binary annotation of pitch accent presence, and an annotation of intonational phrase boundaries.[1] We use the manual annotation of this show to train a single decision tree using the weka [17] implementation of of Quinlan's C4.5 algorithm [18] to predict intonational phrase boundaries. Using this model we hypothesize intonational phrase boundaries for every word in every BN document in TDT-4 – even those in Arabic and Mandarin. These hypothesized intonational phrase boundaries represent a final input segmentation for story segmentation.

We train the intonational phrase decision tree model using feature vectors containing only acoustic information: pitch, duration and intensity features. All pitch features are speaker normalized based on automatically hypothesized speaker identities. We extract pitch and intensity features that are normalized by the surrounding acoustic material in order to account for local context. Additionally, we extract some acoustic features from regions near the end of word boundary, where realizations of intonational phrase boundary events would be localized.

We are aware that these hypothesized intonational phrase boundaries are errorful. Using ten-fold cross-validation on the training document, accuracy of 89.1% was achieved. However, the f-measure of the intonational phrase boundary class was only 66.5% (precision: 68.3%, recall: 64.7%).

In Table 1 we present statistics relevant to evaluating the 'goodness' of the candidate input segmentations. First, we calculate the percentage of manually annotated story boundaries that align with input segmentation boundaries. We also calculate the average distance in words from the gold-standard story boundary to the closest input segmentation boundary, as a crude assessment of the minimum error introduced by the input segmentation. Finally, we examine the ratio of target story boundaries to input segments.

Input Segmentation	Target Boundary Distribution	Exact Coverage	Mean Alignment Error (words)
Word	0.48%	100%	0
Hyp. SUs	8.3%	68.3%	3.6
SU thresh=0.3	6.4%	74.4%	1.8
SU thresh=0.1	4.3%	82.9%	0.61
250ms pause	5.1%	83.5%	0.66
500ms pause	12.2	71.8%	12.7
Hyp. IPs	2.6%	62.0%	1.1

Table 1: Input Segmentation diagnostic statistics

5. Story Segmentation Approach

To detect story boundaries, we construct feature vectors of lexical and acoustic features for each candidate input segmentation as the unit of analysis. We use these feature vectors to train decision tree classifiers specific to each show using J48, weka's [17] implementation of C4.5 [18]. That is, for example, we build unique models for ABC's "World News Tonight" and CNN's "Headline News". This style of show-specific modeling has been shown to significantly improve story segmentation accuracy [1, 7]. For training purposes, we match each manually annotated story boundary to its closest preceding input segment boundary. These 'matched' input segment boundaries represent the set of 'boundary'-class data points for classification.

Lexical Features

To capture lexical cues to story boundaries, we extract LCSeg [5] hypothesized segments and TextTiling [4] coefficients based on window sizes of three, five and ten segments preceding and following the current boundary. TextTiling and LCSeg have been shown to be useful in topic segmentation in text and in meeting transcripts. We also compute features based on lexical consistency immediately or following story boundaries from those lexical items, for each show type, that are statistically likely to occur within a three, seven or ten word window preceding or following a story boundary.[2] For English BN these lexical items are stemmed using an implementation of the Porter Stemmer [20]. We include in our feature vector the number of words that occur in a three, seven, or ten word window preceding or following the current boundary that also occur on the corresponding keyword list. Note we do not include the identity of these words in the feature vector, only the number of matches. For English BN, we also include the number of pronouns in the segment preceding each boundary, identified by a part-of-speech tagger based on the Brill tagger [21]; our use of this feature is based on the hypothesis that a speaker may begin or end a story by identifying themselves with a pronoun – e.g. "I'm X reporting live for CNN" – , or more generally that pronoun use may change over the course of a story, e.g. persons may be more likely to be referred to by a pronoun at the

[1] Since ASR-hypothesized word boundaries may not align with true words, the annotator was asked to mark an ASR hypothesized word as ending an intonational phrase if he believed an intonational phrase was ended anywhere within the ASR-defined word.

[2] Statistical significance is determined using χ^2 with a threshold value of 20 for inclusion in the list of keywords.

end of a story, where their identity may already be established.

Acoustic Features

Acoustic information has been shown to correlate with story boundaries [7, 1], topic shift [22] and changes in discourse structure [23], so we include such features in our detection of story boundaries. We extract the maximum, minimum, mean, median, standard deviation and mean slope of pitch, and intensity from the segment immediately preceding the current boundary. Based on speaker diarization output, we also extract these features based on speaker (z-score) normalized f0 values. We include in the feature vector the length of the segment. In addition to these, we calculate the difference of the above features extracted from the segment preceding and the segment following the current boundary. We also extract features based on speaking rate, hypothesizing that segments at the end of stories will be spoken at different rates and that vowel length may be prolonged preceding boundaries. These features include frame-based speaking rate (ratio of voiced to unvoiced frames), mean vowels per second, mean vowel length, and lengths of segment final rhyme and segment final rhyme. Each feature is also speaker normalized and, when possible, is normalized by vowel identity. We also extract differences in these values across each candidate boundary.

Structural Features

To capture structural consistencies in each news broadcast, such as the airing of commercials or the regularities in story length, we include the relative position of a candidate boundary within the show in our feature vector. We also calculate a set of features based on each identified speaker's participation in the current show. In some shows, story boundaries often co-occur with speaker boundaries. In others, one story is closed and another begun by the same (anchor) speaker. To capture such patterns we extract three binary features: Is the current segment boundary also a hypothesized speaker boundary? Is the word immediately preceding the current boundary this speaker's first spoken segment in the broadcast? last? We also include in the feature vector the percentage of segments spoken by the speaker of the segment immediately preceding the current boundary.

6. Results and Discussion

Results of story boundary detection based on our different input segmentations is shown in Table 2. We evaluate these using the Window Diff measure of [24], an extension of Beeferman's P_k [6]. The Window Diff score is incremented for each false alarm and each miss in a hypothesized segmentation such that near-errors, where a hypothesized boundary is placed close to a target boundary, incur a lesser penalty than more egregious misses or false alarms. So, lower Window Diff scores represent better segmentations. The appropriate window size for applying both WinDiff and P_k is approximately one half the length of the average segment, which in the TDT4 corpus, 215.9 words per story. We thus use a window size of 100.

The story boundary detection model produces a story-boundary/non-story-boundary prediction for each input segment. As each input segmentation defines a different data set, we need to insure that the evaluations of these data sets are comparable. To do this, we align the every set of input segment-based predictions to the word level. This allows us to apply the Window Diff evaluation technique equivalently to the results of story boundary detection based on each input segmentation, and determine which demonstrates the best segmentation performance.

Input Segmentation	English	Arabic	Mandarin
Word	0.300	0.308	0.320
Hyp. SUs	0.357	0.361	0.278
SU thresh=0.3	0.324	0.318	0.258
SU thresh=0.1	0.308	0.304	0.253
250ms pause	0.298	0.312	0.248
500ms pause	0.344	0.419	0.295
Hyp. IPs	0.340	0.333	0.266

Table 2: Story Segmentation Results - (WinDiff; k=100)

Across all languages we find that hypothesized SU boundaries using the default threshold confidence level fail to produce the best story segmentation. SU boundaries detected with lower confidence (.1) perform best for Arabic, while boundaries detected from 250ms pauses perform best for English and Mandarin. However, note that a simple word-based segmentation produces surprisingly good results; while not the best performing for any language, they are second best in English and Arabic. In general, our results show that shorter input segmentations tend to produce better results. We expected the contextual information captured in the feature vectors extracted from larger segmentations to be highly discriminative of story boundaries. However, these large segmentations introduce a significant amount of error based on their misalignment with target story boundaries. The smaller input segmentations provide very little *a priori* error. Despite using features with a narrow view of the source data, these segmentations are able to produce the best story boundary predictions, likely as a result of this small amount of baseline error.

Across languages and input segmentations, we find 62% of errors to be missed story boundaries (M) with 38% false alarms (FA). The rate of misses is slightly lower on Mandarin, and Arabic shows where they represent 60.8% and 60.9% of errors. The ratio of misses to false alarms varies significantly across input segmentations, with the greatest skew toward misses being produced by the word segmenation (70.7% M, 29.3% FA) and the greatest rate of false alarms being produced by the 500ms pause-based segmentation (56.7% M, 43.3% FA). Across languages the rate of false alarms increases with the average input segment length. Despite this relationship, the best input segmentations (250ms pause, low conf. SU) produce fewer false alarms *and* misses than the other input segmentations.

We, clearly, hesitate to make any claims about the success of the hypothesized IP segmentation in identifying intonational phrase boundaries on Arabic and Mandarin shows – the model's performance is modest even on the training document. However, we note that hypothesized IP boundaries predict story boundaries with greater success than hypothesized SUs in all languages. Whether or not story segmentation performance would improve with more accurate intonational phrase predictions remains an open question.

7. Conclusions

In this paper we evaluate the use of different input segmentations to define candidate boundaries for story boundary detection in English, Arabic, and Mandarin. These input segmentations include hypothesized sentences taken at a number of confidence thresholds, pause-based segmentations, and hypothesized intonational phrases. Our experiments indicate that, in general, shorter input segmentations produce better story segmentations, with the best results being produced by low (0.1) thresholding of

SU hypotheses and short (250ms) pause-based segmentations.

In future we will examine the interaction and potential decoupling of the definition of the unit of analysis for feature extraction and the set of potential candidate boundaries. In this paper, we have used the input segmentations to define both; however the two need not be tied. It may be that the optimal unit of analysis is independent of the location of candidate boundaries. We will explore the relationship between the intonational phrase boundary prediction and story segmentation; if more accurate IP predictions are generated, will story segmentation improve? Finally, we intend to explore the use of ensemble learners on story segmentation. In this work, we have identified six potential weak learners, which may be able to segment BN better in combination than in isolation.

8. Acknowledgements

This work was funded by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

9. References

[1] A. Rosenberg and J. Hirschberg, "Story segmentation of broadcast news in english, mandarin and arabic," in *Proc. HLT/NAACL*, 2006.

[2] H. Kozima, "Text segmentation based on similarity between words," in *31st Annual Meeting of the ACL*, 1993, pp. 286–288.

[3] R. J. Passonneau and D. J. Litman, "Discourse segmentation by human and automated means," *Computational Liunguistics*, vol. 23, no. 1, pp. 103–109, 1997.

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

[5] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *41st Annual Meeting of ACL*, July 2003, pp. 562–569.

[6] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 31, no. 1-3, pp. 177–210, 1999.

[7] 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.

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

[9] C. L. Wayne, "Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation," in *LREC*, 2000, pp. 1487–1494.

[10] G. A. Levow, "Assessing prosodic and text features for segmentation of mandarin broadcast news," in *HLT-NAACL 2004*, 2004.

[11] D. D. Palmer, M. Reichman, and E. Yaich, "Feature selection for trainable multilingual broadcast news segmentation," in *HLT/NAACL 2004*, 2004.

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

[13] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, "Recent innovations in speech-to-text transcription at sri-icsi-uw." *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1729–1744, 2006.

[14] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. P. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies." *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[15] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system," in *RT-04F Workshop*, November 2004.

[16] P. A. Luce and J. Charles-Luce, "Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production," *Journal of the Acoustical Society of America*, vol. 78, no. 1949–1957, 1985.

[17] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with java implementation," in *ICONIP/ANZIIS/ANNES International Workshop: Emerging Knowledge Engineering and Connectionist-Based Information Systems*, 1999, pp. 192–196.

[18] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5(9-10), pp. 341–345, 2001.

[20] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[21] E. Brill, "A simple rule-based part-of-speech tagger," in *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, Trento, IT, 1992, pp. 152–155. [Online]. Available: citeseer.ist.psu.edu/brill92simple.html

[22] J. Hirschberg and C. Nakatani, "Acoustic indicators of topic segmentation," in *Proc. of ICSLP*, vol. 4, 1998, pp. 1255–1258.

[23] J. Hirschberg and J. Pierrehumbert, "The intonational structure of discourse," in *Proc. of 24th Annual Meetinc og the Assoc. for Computational Linguistics*, 1986, pp. 136–144.

[24] L. Pevzner and M. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Computational Linguistics*, vol. 28, no. 1, pp. 19–36, 2002.