Soundbite Detection in Broadcast News Domain

Sameer Maskey, Julia Hirschberg

Department of Computer Science
Columbia University
New York, NY
{smaskey,julia}@cs.columbia.edu

Abstract

In this paper, we present results of a study designed to identify SOUNDBITES in Broadcast News. We describe a Conditional Random Field-based model for the detection of these included speech segments uttered by individuals who are interviewed or who are the subject of a news story. Our goal is to identify direct quotations in spoken corpora which can be directly attributable to particular individuals, as well as to associate these soundbites with their speakers. We frame soundbite detection as a binary classification problem in which each turn is categorized either as a soundbite or not. We use lexical, acoustic/prosodic and structural features on a turn level to train a CRF. We performed a 10-fold cross validation experiment in which we obtained an accuracy of 67.4% and an F-measure of 0.566 which is 20.9% and 38.6% higher than a chance baseline.

Index Terms: soundbite detection, speaker roles, speech summarization, information extraction.

1. Introduction

The primary speakers in Broadcast News (BN) are news ANCHORS. Anchors introduce stories which are generally presented by REPORTERS. Both anchors and reporters may in turn introduce segments of speech by others, which support a news story. These speakers may be interviewed, or clips of their speech (e.g. a speech or interview quotations) may be included in the newscast. These clips, when they can be identified by speaker, are of considerable value in news corpora, since they contain material representing views that are clearly and directly attributable to the speaker, rather than third party commentary. We term such material SOUNDBITES here, and interviews as well as other segments of a speaker's production included directly in the newscast; we term the speakers of such material SOUNDBITE-SPEAKERS. In this paper we describe experiments on the detection of soundbites in BN. This research is motivated by a larger goal, to extract answers to questions of the form ('What did X say about topic Y?') from BN, as well as the more general summarization of BN. For this purpose we have annotated a large corpus of BN with both soundbite boundaries and with the names or, where names are lacking, descriptions of soundbite-speakers provided in the transcripts.

In Section 2 we describe research related to this task. In Section 3 we describe our news corpus. We present our approach to soundbite detection in Section 4 and discuss our results in Section 5. In Section 6 we conclude and discuss our future research.

2. Related Work

To our knowledge, no research has yet been done on soundbite detection in BN. In the literature, SPEAKER ROLE detection is perhaps the most relevant to our task [1, 19]; such work attempts to classify speech segments as to the **type** of speaker (anchor, reporter or other) producing the segment. A large proportion of the category 'other' is comprised of our soundbite-speakers. In this work, Barzilay et al [1] built a maximum entropy model and a Boostexter model to perform a three-way classification of speakers in English BN. They used key words, context, duration features and explicit speaker introductions to distinguish among speaker types, obtaining classification accuracy of about 80%. Yang [19] constructed a maximum entropy model for distinguishing among the same speaker types in Mandarin BN, reporting comparable accuracy by combining language model scores trained for each speaker type.

There is also considerable research on speaker DIARIZATION, the segmentation of spoken corpora into distinct speakers and the clustering of such segments into 'same speaker' clusters. This work does not in general attempt to identify individual speakers or their roles (but cf. [7] for work on anchor identification and [16] for more general attempts to identify speakers in diarization).

However, correct segmentation of BN into speakers is critical for us, since we benefit from accurate information about where different speakers begin and end. As [1] found, accurate diarization can also provide useful distributional information about where and how often individual speakers contribute in a news show. While anchors tend to speak often in a broadcast, for example, any individual soundbite-speaker will tend to occur very infrequently in a single newscast.

3. Our Corpus

We performed our experiments on a subset of the TDT2 BN corpus [18]. We used 24 half-hour CNN Headline News shows from this corpus, which included 1045 speaker turns. We used the Dragon ASR transcripts which are distributed with TDT2 for each show for our training and test corpus. Each turn was manually segmented in the transcripts and hand-labeled also for soundbite turns. Soundbite-speakers were identified as such when their names or a description (e.g. "one unhappy farmer") appeared in the transcript. Two annotators were provided with a detailed labeling manual and a Java-based interface and labeled soundbites and soundbite-speakers in the course of a larger labeling effort on the corpus. 345 of the turns were labeled as soundbites by our annotators.

All of the features we extract from the corpus are extracted from these ASR transcripts, except for the turn segmentation. This

is based on the manual segmentation of the human transcriptions, automatically aligned with the ASR transcripts. Note that we use the automatic cluster ids generated for TDT2 in computing features such as the distribution of speaker turns in broadcasts; after turn segment alignment we label each turn with the automatically generated cluster id that covers most of the (true) segment.

4. Approach

Since BN shows are critically temporal in nature — news shows exhibit clear patterns as they unfold, we want to take advantage of various types of such patterns in our classification. Some of these arise from the temporal sequence of speaker turns or from the repeated occurrence of particular phrases before or after soundbites. Markov models, Hidden Markov Models (HMM) and maximum entropy models (MEMM), have been used successfully for modeling such data for the extraction of speaker role in BN. However, for many Natural Language Processing tasks, modeling a given joint distribution is difficult when rich local features with complex dependencies are used in classification. Here, we employ a Conditional Random Field (CRF) model.

CRF models have been successfully used in various Natural Language Processing tasks including named entity detection [13] and Chinese word segmentation [14]. CRFs are undirected graphical models proposed by Lafferty et. al [9] that directly model the conditional distribution $p(s|o)$ where s represents classes and o represents features. Such models have been shown to be effective in taking account of local dependencies while decoding the optimal output classes in a globally optimal framework, since dependencies do not need to be represented explicitly. For special cases of CRF when we join the output class nodes in a linear chain, the CRF corresponds to a Finite State Machine (FSM), with a first-order Markov assumption. Such CRFs represent a globally normalized extension to MEMM models without the label-bias problem.

We define our CRF with the following parameterization. Let $\mathbf{o} =< o_1, o_2, ..., o_T >$ be the observation sequence of turns in each broadcast show. Let $\mathbf{s} =< s_1, s_2, ..., s_T >$ be the sequence of states. The values on these T output nodes are limited to 0 or 1, with 0 signifying 'not a soundbite' and 1 signifying 'a soundbite'. The conditional probability of a state sequence s given the input sequence of turns is defined as

$$p_\wedge(\mathbf{s}|\mathbf{o}) = \frac{1}{Z_o} exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)\right)$$

where Z_o is a normalization factor over all state sequences and $f_k(s_{t-1}, s_t, \mathbf{o}, t)$ is an arbitrary feature function. λ_k is a weight for each feature function. The normalization factor Z_o is obtained by summing over the scores of all possible state sequences:

$$Z_o = \sum_{s \in S^T} exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)\right)$$

This can be computed efficiently in our case using dynamic programming, since our CRF is a linear chain of states.

4.1. Features

We used Prosodic/Acoustic, Structural and Lexical features to identify soundbites in our corpus.

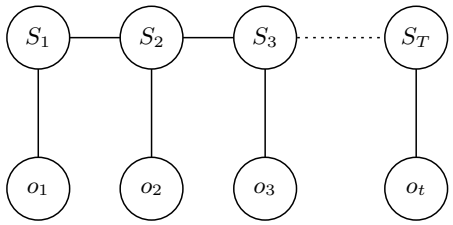

Figure 1: CRF Structure for Soundbite Detection

4.1.1. Prosodic/Acoustic Features

Prosodic/Acoustic features are useful for detecting soundbites in BN. Change in pitch, amplitude or speaking rate often differentiation between speech segments produced by various speakers. There is also considerable evidence that topic shift is marked by changes in pitch, intensity, speaking rate and duration of pause [6, 17]. We further hypothesize that a turn uttered by a soundbite-speaker may exhibit different acoustic features from a turn spoken by an anchor or reporter, due not only to changes in speaking style but also variation in signal quality and the background noise of the soundbite recording. While anchors and reporters are often recorded in the studio, soundbites are generally recorded in the field (and often spliced in for an interview) or are cut from other recorded events. So, recording conditions vary considerable for soundbites, and acoustic features may capture some of this variation and thus aid in prediction.

Our Prosodic/Acoustic feature-set includes features similar to those described [8, 4, 11] as well as some additional features. It includes **speaking rate** (the ratio of voiced/total frames); **F0 minimum**, **maximum**, and **mean**; **F0 range** and **slope**; **minimum, maximum**, and **mean RMS energy** (minDB, maxDB, meanDB); **RMS slope** (slopeDB); **turn duration** (timeLen = endtime - starttime). We extracted these features by automatically aligning the turn boundaries from the manual transcripts with the ASR transcripts and extracting the timestamps. We used Praat [15] to extract these features from the speech signal.

Our 'speaking rate' feature is estimated by dividing the number of voiced frames by the total number of frames. We hypothesize that our generally non-professional soundbite-speakers may have a different speaking rates from the trained speech of anchors and reporters, based on Bolinger's description of newscaster speech [3]. Similarly, our pitch and energy features are motivated by the possibility that these may vary differently for untrained soundbite-speakers. Our 'turn duration' feature captures the length in seconds of the turn. On average, 'turn duration' for soundbites was 26 seconds shorter than non-soundbite turns.

4.1.2. Structural and Lexical Features

BN programs exhibit similar structure — particularly broadcasts of the same news show. Each usually begins with one or more anchors reporting the headlines, followed by the actual presentation of those stories by the anchor and reporters. These stories may sometimes include interviews as well. Programs are usually concluded in the same conventional manner. We call the features which rely upon this typical broadcast structure *Structural* features [11], comparable to [4]'s *style features*. Maskey and Hirschberg [11] have previously shown that structural features are useful predictors of sentences to include in extractive summaries of BN.

The structural features we investigated for our current study include **normalized position of turn** in the broadcast; **speaker change**; **turn position** in the show; **speaker distribution**; **previous and next speakers**; and **top-ranking speakers** in the broadcast. The positional feature encodes where the current turn is in the broadcast; soundbites, for example, rarely occur at the beginning or end of a broadcast. 'Speaker change' is a binary feature, indicating whether the current speaker is different from the previous one. The 'speaker distribution' feature captures the percentage of turns belonging to a given speaker in the broadcast, as calculated from the automatic speaker clustering information (i.e., identification of individual speakers by unique identifier for each segment) provided in TDT2. We hypothesize that the overall percentage of turns for any soundbite-speaker should be very low compared to the anchors and reporters; note however that, with automatic speaker clustering, we are adding a degree of noise here. The information on identify of previous and next speakers should also help in identifying soundbites, as there is a very low probability of two soundbites occurring together. Soundbites are usually followed by anchor or reporter comments. 'Top-ranking speakers' indicates whether the current speaker is among the top 3 speakers in the broadcast in the number of turns produced; this feature is intended to rule out anchors in particular as speakers of a soundbite.

We extract all of our lexical features from the ASR transcript, with no capitalization or punctuation. Our lexical features include **number of words in the turn**, **cue phrases**, **distribution of cue words**.

Cue phrases are currently identified by inspection of the soundbite turns as well as the turns that precede and follow them. These can be important cues for a turn transition. Since we train and test on shows that are primarily CNN Headline News, our cues are dependent on the type of cue phrases used in these broadcasts. Cue words and phrases such as anchor names, "headline news", "reporting from" are all useful in indicating an upcoming soundbite.

5. Experiments, Results and Discussion

To classify segments as soundbite segments or not, we built CRF models using the Mallet tool [12]. This tool allows us to build CRF models with varying degrees of Markov order. In order to test the effect of previous context, we built CRF models with a Markov order of 0, 1 and 2 and compared them to MEMM models. Figure 2 compares the performance of the different models.

The first model on the left in Figure 2 is a CRF model with a markov order of 1, with the observation conditioned both on the parent state and the previous parent. The second model is maximum entropy model where the observation is conditioned on the parent state only and the current state is dependent on the previous state. The best model, shown on the right of the figure, is a 1-order CRF model with the current state depending only on the previous state. Intuitively, we would assume that for a task such as soundbite detection, a higher order model would do better. However, our experiments showed that a 2-order model overfits the data and degrades overall performance. In a 10-fold cross validation experiment our 1-order model performed 10.75% better in accuracy and 9.01% better on F-measure than a 2-order model.

With the 1-order model, we obtained 67.43% accuracy in soundbite prediction with precision of 0.522, recall of 0.624 and an F-measure of 0.566. For this model, the maximum accuracy we obtained in a single iteration was 85.9%. For the same iteration

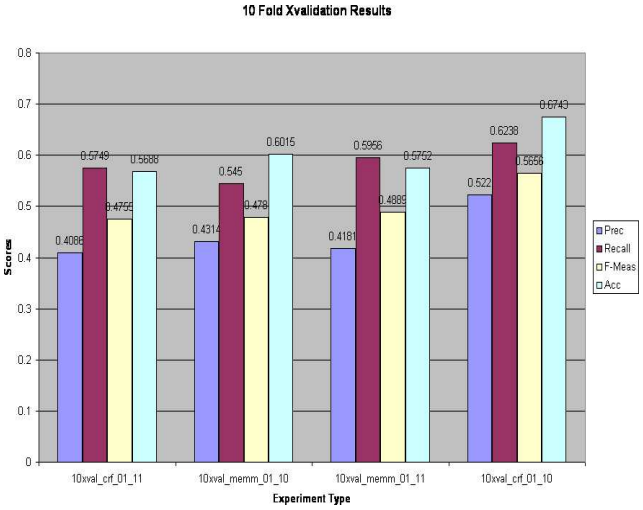

Figure 2: F-measure with 10 fold cross-validation

we obtained a precision of 0.816, recall of 0.838 and F-measure of 0.827, which were also the highest among all the iterations. The lowest accuracy and F-measure for this model were 55.1% and 0.394. The difference in results between best and worst iterations were thus quite high, with a difference of 30.8% in accuracy and 43.3% in F-measure.

The significant difference in the best and the worst iteration of the cross-validation experiment for the soundbite detection suggests either that the variability in the soundbites is quite high or that our annotation is somewhat noisy. We suspect, anecdotally, that both these possibilities are true. Our annotators reported some difficulty in labeling the data.

We next compare our results with a baseline based on a chance. We do not use a "majority class baseline" because it would result in a baseline with a recall and an F-measure of 0. The CRF 10-fold cross-validation results are significantly higher than the baseline. This F-measure is 38.56% higher than baseline and the best performing iteration has an F-measure 64.7% higher than the baseline. Similarly recall and precision for the CRF model is 45.38% and 34.2% higher than the baseline respectively.

In order to determine whether conditioning the observations on more context would improve performance, we built a CRF model in which the observations were conditioned on previous states. The model generated with such conditioning did worst than CRF models that conditioned only on the parent state. The F-measure on 10-fold cross validation F-measure was lower by 9.01%, recall was lower by 4.89% and precision was lower by 11.34%. Such a difference in performance shows that either the model is overfitting the data or that our features are not highly dependent across turns. We think it likely that some of our features — in particular, the acoustic features — may not be dependent on prior context, since soundbites are often recorded in completely different contexts from the rest of the broadcast, even for interviews, and later spliced in to the show.

We also built MEMM models for soundbite detection to compare to our CRF models. CRFs are similar to MEMMs except MEMMs suffer from a label bias problem due to normalization over local features rather than over the entire sequence. The results presented in Table 1 show that the MEMM model does slightly worst than the CRF models. For the same Markov order and simi-

ModelType	Precision	Recall	F-Meas	Acc
CRF	0.522	0.624	0.566	0.674
MEMM	0.431	0.545	0.478	0.602
Baseline	0.18	0.17	0.18	0.465

Table 1: Soundbite Detection Results

lar conditioning of features over states, the CRF model does better than MEMM by 7.28% on accuracy, 8.76% on F-measure, 7.88% on recall and 9.06% on precision.

Our experiments on soundbite detection suggest to us that this task is more difficult than the related but more general task of speaker role labeling. To explore this hypothesis on our corpus, we built a reporter detection model with the same set of features used for soundbite detection, but we used a Bayesian Network model for training purposes. We also built a Bayesian Network model for soundbite detection and compared the results. For reporter detection, the Bayesian Net model could classify reporter vs. non-reporter segments with an accuracy of 72%, an F-measure of 0.665, precision of 0.719 and recall of 0.618. However, a similar Bayesian Net model built on the same set of features classified soundbites more poorly, with an accuracy of 67.6%, an F-measure of 0.522, precision of 0.477 and recall of 0.577. These results are considerably lower than results for reporter detection.

6. Conclusions

In this paper we present results of experiments in classifying soundbites in Broadcast News, segments of direct recorded speech included in a broadcast from interviewees and figures in the news. Our goal in this research is to be able to identify these segments as well as their speakers, to answer questions about what particular speakers say about particular topics, automatically. We use Conditional Random Fields to model the binary classification problem and obtain an accuracy of 67% and an F-measure of 0.566, which are 20.9% and 38.6% higher, respectively, than a chance baseline. We compare this model to MEMMs and Bayesian Networks for soundbite classification, and we compare soundbite classification to speaker role classification using the same feature-set, to show that soundbite classification is a more difficult task. In our future work, we will study the identification of soundbite-speakers also, from mentions in the transcript, and address the task of associating these speakers with the soundbites they produced.

7. Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. We would like to thank Michel Galley for letting us use the modified tools of Andrew McCallum. We would also like to thank our annotators, Tasha Moore, Paola Garcia and Shuvojit Gosh.

8. References

[1] Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S., "The Rules Behind Roles: Identifying speaker role in radio broadcasts" Proceeding of AAAI

[2] Bikel, M.D., Miller, S., Schwartz, R and Weisschedel, R. "An algorithm that learns what's in a name", Machine Learning, 34(1/2/3):211-231, 1999.

[3] Bolinger, D., "The Network Tone of Voice", Journal of Broadcasting, Vol. 26, pages 725-728

[4] Christensen, H., Kolluru, B., Gotoh, Y., Renals, S. "From text summarisation to style-specific summarisation for broadcast news", ECIR, 2004.

[5] Hirschberg J. "Communication and Prosody: Functional Aspects of Prosody", Speech Communication, Vol 36, pp 31-43, 2002.

[6] Hirschberg, J., Nakatani, C. "A Prosodic Analysis of Discourse Segments Direction-Giving Monologues", ACL 1996.

[7] Huang, Q., Liu, Z., Rosenberg, A., Gibbon, D., Shahraray, B., "Automated Generation of News Content Hierarchy by Integrating Audio Video and Text Information" Proceedings of ICASSP, 1999, pages 3025-3028

[8] Inoue, A., Mikami, T., Yamashita, Y. "Improvement of Speech Summarization Using Prosodic Information", Proc. Speech Prosody, 2004, Japan.

[9] Lafferty, J., McCallum, A., and Pereira, F., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" Proceedings of ICML, 2001

[10] Barras, C., Zhu, X., Meignier, S., Gauvain, J.L. Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain. "Improving Speaker Diarization" Proc. DARPA RT04, 2004.

[11] Maskey, S.R., Hirshberg, J., "Automatic Summarization of Broadcast News Using Structural Features" Proceedings of Eurospeech 2003

[12] McCallum, A. K., "MALLET: A Machine Learning for Language ToolKit", 2002

[13] McCallum, A., Li, W., "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web Enhanced Lexicons" Proceedings of CoNLL, 2003

[14] Peng, F., Feng, F., McCallum, A., "Chinese Segmentation and New Word Detection Using Conditional Random Fields", Proceedings of COLING, pages 562-568, 2004

[15] Boersma, P. "Praat, a system for doing phonetics by computer", Glot International 5:9/10, 341-345. 2001.

[16] Rosset, S. and Lamel, L., "Automatic Detection of Dialog Acts Based on Multi Level Information" Proceedings of ICSLP 2004

[17] Shriberg, E., Stolcke, A., Tur, D.H., Tur, G. "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics", Speech Communication, Vo. 32. pp 127-154 2000.

[18] Language Data Consortium "TDT-2 Corpus", Univ. of Pennsylvania.

[19] Liu, Y. "Initial Study on Automatic Detection of Speaker Roles in Broadcast News Speech", not published yet

[20] Witten, I.H., E. Frank, L. Trigg, M. Hall, G. Holmes and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with Java implementations," in H. Kasabov and K. Ko, eds., ICONIP/ANZIIS/ANNES'99 International Workshop, Dunedin, 1999.