The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts

Regina Barzilay
regina@cs.columbia.edu
CS Dept., Columbia University
New York, NY 10027, USA

Michael Collins **Julia Hirschberg** **Steve Whittaker**
{mcollins,julia,stevew}@research.att.com
AT&T Labs – Research
Shannon Laboratory, 180 Park Ave., Florham Park, NJ 07932, USA

Abstract

Previous work has shown that providing information about story structure is critical for browsing audio broadcasts. We investigate the hypothesis that Speaker Role is an important cue to story structure. We implement an algorithm that classifies story segments into three Speaker Roles based on several content and duration features. The algorithm correctly classifies about 80% of segments (compared with a baseline frequency of 35.4%) when applied to ASR derived transcriptions of broadcast data.

Introduction

The amount of browsable online spoken broadcast news data is rapidly increasing. However, due to the serial nature of speech and the costs of hand-indexing, it is very difficult to navigate this data effectively. New technologies use automatic speech recognition (ASR) and information retrieval (IR) techniques to allow audio data to be searched by content. However, in developing such retrieval techniques, it becomes clear that simple term-based retrieval of such large speech "documents", does not enable users to browse audio effectively. Audio is inherently hard to skim, so that accessing a relevant newscast does not guarantee finding the crucial segment within that newscast. Therefore, a critical problem for audio data is to provide information about the internal structure of newscasts. For genres such as broadcast news corpora, we can assist local browsing by exploiting their structural regularities. Regularities include introductory headline teasers, story structuring by correspondents, and predictable program formats. Presenting this information should enable users to navigate to the relevant part of the broadcast. To present such structural information, however, we need to identify structural elements automatically.

We describe a technique for acquiring the structure of broadcast news programs by identifying *participant role*. By identifying role – anchor, journalist, or program guest – we are able to infer a STRUCTURAL SUMMARY of the broadcast. Anchors typically introduce stories and guide the program, appearing throughout. Journalists usually report a specific news story. Program guests are generally interviewed by journalists. Identifying speaker type provides clues about newscast structure. Both anchors and journalists present explicit structural information. Speaker transitions also yield structural information: when a journalist stops speaking, this is a strong cue that a story has ended. This technique is being developed in the context of an audio browsing project SCAN (Hirschberg *et al.* 1999) on the DARPA Broadcast News corpus.

We present a machine learning algorithm for speaker *role* identification from audio data. The algorithm's input is ASR transcriptions from "All Things Considered" programs, with boundaries between speakers identified, but the identity of speakers unknown. The algorithm's output is a label for each segment, identifying it as either *Anchor*, *Journalist* or *Program Guest*. We use a set of segments with known labels to train a classifier. A separate test corpus evaluates accuracy. The main classifier features relate to the text of each segment. We also include segment duration and textual context features. Our method makes the three-way classification with around 80% accuracy, compared to a baseline result of 35.4% accuracy when every segment is assigned to the most frequent class (Anchor).

We first briefly describe the audio browsing system. We then motivate our use of participant role to define program structure. We describe the Broadcast News Corpus in detail and the task of participant role identification. We then describe the algorithm used for role identification, the features it uses, and how they are computed. We present results of our learning experiments and evaluate their success. Finally, we discuss our future research directions.

The Audio Browsing System

Our system operates on the NIST TREC SDR corpus, a subset of DARPA HUB-4 Broadcast News. The system uses ASR to produce an errorful transcription of each story, after segmenting the speech into audio paragraphs. Stories relevant to a text query are retrieved by a modified version of the SMART IR system. Recognition and retrieval results are then passed to a graphical user interface (GUI). The GUI is designed to support local navigation within speech documents, as well as document retrieval. We employ well-understood text formatting conventions (e.g. headers and paragraphs) to provide useful analogues for speech browsing. The role-based structural information about broadcast programs is intended to augment and extend this interface by providing additional information about where program information is summarized, where individual stories begin and end, and where the most general summarizing portions of these stories are likely to occur.

Motivation

We hypothesized that, in news broadcasts, speaker type should be correlated with program structure. Figure 1 shows a role-based segmentation for the NPR radio program “All Things Considered”[1]. Different parts of the program exhibit different speaker change patterns. Anchor segments, which usually represent headlines or introductions to stories, occur in particular places in broadcasts, and are uninterrupted by guest segments. Anchors also tend to occur more frequently in the program, and to alternate with the journalist they introduce. Individual stories are often characterized by an anchor introduction, a journalist introduction, and then an alternation between journalist and guest segments. A typical journalistic story of this type is marked in figure 1. Given this relationship between speaker role and program structure, roles can be used to categorize program segments according to their type, e.g. headlines or interviews. We hypothesize that this will help users to browse within a broadcast.

(Stolcke *et al.* 1999) observe that speaker change is a useful feature for story segmentation. However, not all speaker changes correlate with story boundaries. Anchor segments can be used to hypothesize a set of story boundaries, because anchor speech usually separates stories. Guest segments, in contrast, never introduce a story: their contributions always occur within a story.

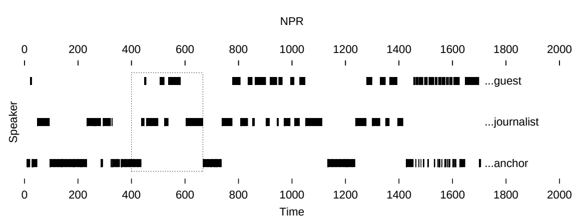

Figure 1: Speaker-based segmentation types.

Information about speaker role may also be beneficial for summarization of news stories. Some segments (for example the anchor segment just before a journalist begins a story, or the first journalist segment within a story) may be particularly important for constructing the summary of a story, and an NLP system for summarization should take this positional information into account.

[1]For simplicity, figure 1 does not show explicitly all speaker turns. For example, sequence “Anchor, Anchor” appears as one block.

The Data

The data set used for the development of our speaker type identification consisted of 35 half-hour broadcasts of the radio program “All Things Considered,” a subset of the TREC-7 SDR track corpus, which, in turn, represents a subset of the DARPA Broadcast News corpus.

We used both human transcripts provided by NIST and automatically recognized transcripts(ASR) with manually-labeled speaker boundaries from this corpus; segments were tagged for speaker identity and duration. Commercials are excluded from the corpus. We used the ASR transcription of the data provided for the SCAN systems, with an error rate of 31%. We enriched ASR transcripts with manually-labeled speaker boundaries by aligning ASR transcripts with human transcripts using the word time feature, which were provided in both transcripts.

Speaker type was not explicitly provided by the transcripts. Fortunately, it was relatively easy to acquire this information from the hand-labeled speaker names. We compiled a list of the names of anchors and journalists in the corpus[2], and used this to label the segments; any name not appearing in the list was labeled as a program guest. Thus we formed training and test sets of segments labeled with the Anchor/Journalist/Program Guest distinction.

Related Work

Researchers have noted that speaker type change is an important feature for indexing broadcast news (Mani *et al.* 1997; Huang *et al.* 1999; Reynar 1999). Huang et al. base their segmentation method on the identification of anchor segments, assuming that two adjacent stories are separated by anchor speech. Detection of anchor segments is implemented using a text-independent speaker recognition technique, based on the acoustic characteristics of speakers. A model for the target speaker, and a background model, are constructed from labeled training data. For the test case, the target speaker is detected according to likelihood ratio values from the constructed models. This method achieved impressive results — a 92% hit rate and a 1.5% false alarm rate, when tested on its ability to detect whether or not one anchor person (Tom Brokaw) was talking. Our method differs from this approach in that we assume no prior knowledge of the acoustic characteristics of the different speakers in the program.

A number of researchers (Chen & Gopalakrishnan 1998; Couvreur & Boite 1999) have considered methods which cluster speaker segments into groups of acoustically similar segments. The goal of this work has usually been the adaptation of speech recognizers to different types of speakers or different channel conditions.

[2]We used the list of anchors/journalists on the “All Things Considered” site (`http://www.npr.org/programs/atc/`).

Typically, different recognizer models are trained on different clusters. This task is similar but not identical to ours. The output of speech-based systems contains no information about the types of different speakers, although speaker identity would be very useful information in deriving speaker type, if it could be recovered with high accuracy.

Unfortunately, speaker clustering is a difficult task; (Couvreur & Boite 1999) report 70% classification accuracy on broadcast news, even when the number of speakers is given a priori. (Note that this classification accuracy may be quite sufficient for speech recognizer adaptation; in contrast, browsing tasks will usually require a lower error rate.) Furthermore, this method is sensitive to an increase in the number of speakers and changes in background conditions. This is problematic for broadcast news, where typically there may be thirty speakers, and where channel conditions often vary from microphone speech to telephone speech.

Our algorithm contrasts with the speaker clustering methods in two ways. First, we focus on discovering speaker type, rather than speaker identity. We believe that this is a much more tractable task than full-blown speaker identification, while still providing very useful information for indexing or browsing news programs. Second, our algorithm exploits the lexical information found in ASR transcriptions rather than acoustic information.

Identifying Speaker Type

When we listen to a radio program, we can usually tell whether the speaker is the anchor, a journalist, or a guest speaker in terms of content as well as speaking style.

- An anchor is responsible for reading news, introducing reports from journalists, and announcing upcoming events.
- A journalist is a professional speaker, generally in some remote location where a story is taking place. Journalists often interview guests in the course of presenting their stories.
- A guest speaker is usually a non-professional speaker speaking from a subjective point of view.

Our assumption is that these major functional differences will be reflected in the following features:

Lexical features Intuitively, aspects of what is said should distinguish speaker type. Previously, (Mani *et al.* 1997) and (Reynar 1999) have observed that "signature phrases", such as "This is CNN's Primenews", are frequently used by anchors and journalists in broadcast news — almost never by guests. These professional speakers also tend to exhibit more 'planned speech' vs. the spontaneous speech of guests. So, we would expect that segments of non-professional speakers would contain more self-repairs and semantically empty words, such as "well".

In previous work, lists of lexical cues have been compiled by hand (see (Mani *et al.* 1997; Reynar 1999; Teufel & Moens 1998)), with the classifier then using their occurrence as a binary feature. In our approach, we would like to learn these patterns automatically from the training corpus. To do this, we provide as input to the learning algorithm all n-grams from unigrams up to 5-grams from the segment. Thus we generate a large number of lexical features, and allow the machine learning method to find those that are useful.

Note that some lexical patterns are only predictors of speaker type if they are followed by a proper name. For example, the phrase"I'm" is common in broadcast news, but only "I'm ⟨*proper-name*⟩" is a good predictor of an anchor. Because of such observations, we decided to augment the segment text with proper name indicators. In written text, capitalization can be used for identification of proper names, but unfortunately speech transcripts do not provide capitalization. In order to acquire capitalization information, we used a parallel text corpus of written news — the AP corpus from 1996 — which contains 44,171,587 words and 209,426 word types. For each word in the corpus we counted the number of its capitalized vs. un-capitalized occurrences, excluding occurences in initial sentence positions. We consider a word in the speech transcripts to be capitalized, if the ratio of its capitalized appearances is greater than 50%. Of 209,426 word types, 123,649 (59%) were capitalized according to this definition. This method allows us to identify words which are always capitalized, e.g., Clinton, as well as words which have a tendency to be capitalized, e.g., Flowers (in Gennifer Flowers). We substituted all occurrences of words from this list in the broadcast transcripts with a special "capitalized-word" token.

Features from the surrounding context In some cases, the label and the content of adjacent segments may predict the current speaker type. An anchor usually introduces a journalist at the start of a story, and, similarly, a journalist "hands off" the report back to the anchor at the end of a story. In addition, some sequences of labels are more frequent than others (see Figure 1). For example, the sequence "Journalist, Guest, Journalist" occurs sixteen times, while "Journalist, Guest, Anchor" never appears in the graph. We experimented with two types of contextual features: the labels of the n previous segments, and all the features of n previous segments. The first feature type captures the intuition that some label sequences are more frequent than others. The second covers cases in which speakers provide cues about the type of the following speaker.

Duration features Segment duration is another feature which we observe to be correlated with speaker role. Journalist guide books (Mencher 1987) advise controlling the time length of guest speaker segments, and also give suggested lengths for anchor lead-ins and journalist's questions. These features were computed in a straightforward manner, using time labels from the transcripts.

Explicit speaker introductions One of the tasks of professional speakers in radio programs is to introduce themselves and other program participants. Speaker introductions such as "I'm Noah Adams" or "NPR's Claudio Sanchez reports" or "thanks Claudio Sanchez for that report", occur frequently in broadcast news, and can be used for distinguishing anchors and journalists from non-professional speakers. We decided to apply a learning technique to identify speaker introductions in the text (more specifically, proper names in the text where a speaker has introduced herself or a following/preceding speaker), identifying such references in the segments, and tagging, e.g., "Noah Adams" or "Claudio Sanchez" in the above examples.

In the remainder of the section we describe our method for speaker introduction computation. The broadcast news human transcripts include the identity of the speaker of each segment. From this information, we created a training corpus where speaker name were labeled. Out of 133,391 words, 522 (0.4%) fit this definition of speaker name. The identification of speaker introduction was reduced to a binary word-by-word classification problem, which was addressed using the BoosTexter algorithm (described in the following section). The following features were used to represent each word:

- Lexical features aim to discover templates for speaker introduction. For a word in position n, we extracted all trigrams, bigrams and unigrams surrounding the word, including those beginning at position $n-3$ to those ending at position $n+3$. To distinguish all these, we prepend each n-gram with its length and starting position.
- The frequency of the word in the broadcast. Typically, professional speakers are introduced no more than twice during the program, therefore high frequency of the word in the broadcast is an indicator that the word is not a speaker introduction.
- Relative distance from the start and the end of the segment. Self-introduction usually occurs in the start of the segment, while the introduction of other speakers usually happens at the end of the segment.

We approximate capitalization, using the techniques described above. Figure 2 shows an input example for BoosTexter. We evaluated our method on 21,905 words of unseen data. 87 of these words were speaker introductions, our method recovered 70 of them with no false positives (80% recall, 100% precision on this test set). Therefore, it can be used as a reliable feature for our task. This method can also be used for extracting the identity of professional speakers from broadcast transcripts.

Learning Methods

We applied two algorithms to the classification task. The first, BoosTexter, is a boosting algorithm which was originally applied to text classification (Singer & Shapire 1998). The second technique, maximum entropy modeling, has been previously applied to a variety of natural language tasks, the closest application to ours being part-of-speech tagging as described in (Ratnaparkhi 1996). Both of these methods learn simple weighted rules, each rule using a feature to predict one of the labels with some weight: an example rule would be, *if the segment contains the n-gram "this is NPR news" vote for label Anchor with weight 0.3.* On test data examples, the label with the highest weighted vote is taken as the output of the algorithm. The boosting approach greedily searches for a subset of the features which predict the label with high accuracy; in the maximum entropy method all features occurring above a certain number of times (in our case 12) were used by the model ((Ratnaparkhi 1996) also used a count cut-off to select features).

```
30___npr's 31__npr's_@ 32_npr's_@_@ 33_@_@_has
34_@_has_this 35_has_this_report 20__ 21__npr's
22_npr's_@ 23_@_@ 24_@_has 25_has_this 10_ 11_
12_npr's 13_@ 14_@ 15_has, 2, 2, 16.67, 83.33, yes.
```

Figure 2: BoosTexter input for the word "Phillip Davis" in the segment "npr's Phillip Davis has this report". "*xy_www*" stands for a sequence of "_"-delimited words *www* for a window of size x in the y-th position, "@" stands for capitalized words.

Results and Evaluation

In this section we first discuss the accuracy of the method on human transcripts, focusing on the contribution of different feature types to the method's performance. We then discuss results on ASR output. We divided our data into a training set containing 27 broadcasts (2336 segments), a development set of 5 broadcasts (339 segments), and a held-out test set containing 5 broadcasts (347 segments). Table 1 shows the numbers of anchors, journalists and guests segments for the training, development and testing sets. On this particular breakdown of the data, a baseline classifier would achieve 35.4% accuracy on the test set by labeling each segment with the most frequent category in the training set — anchor.

	Training	Development	Testing
Anchor	878(37.6%)	123(36.3%)	123(35.4%)
Journalist	630(27%)	83 (24.5%)	119(34.3%)
Guest	828(35.4%)	133(39.2%)	105(30.3%)

Table 1: Number of segments per Speaker Type

Using this training/development partition, for each segment we calculated features described in the previous section. Figure 3 shows the classification error on the development set with different types of features included in the model. The following feature types were all found to be useful:

Lexical features We used four textual features: the text of the current segment, the two previous segments and the next segment. Word n-grams of up to length 5 were included. Table 2 shows the textual features with the highest weight found by BoosTexter. The majority of n-grams in the table corresponds to "signature phrases" — these phrases discriminate professional participants from guests, and also help to make the distinction between anchors and journalists. Another group of phrases picked up by BoosTexter, as a predictor of anchors and journalists, corresponds to questions, e.g., "do you, what about". The highest weight predictors of guests, such as "uh, well, you know", are words which are frequent in "everyday" spontaneous speech.

Segment duration The relative segment duration, namely the ratio of current segment duration to previous segment duration, is one of the high-weighted features. When this value is higher than a certain threshold (2.035), it is considered to be journalist predictor. This empirical result can be explained by the fact that a short summary from an anchor often precedes full coverage of the story by a journalist. Absolute segment duration also serves as a predictor of the journalist category: a duration higher than 5.26 minutes corresponds to journalist's segments. On the other hand, very short segments(duration < 0.6 minutes) are indicators that an anchor is speaking.

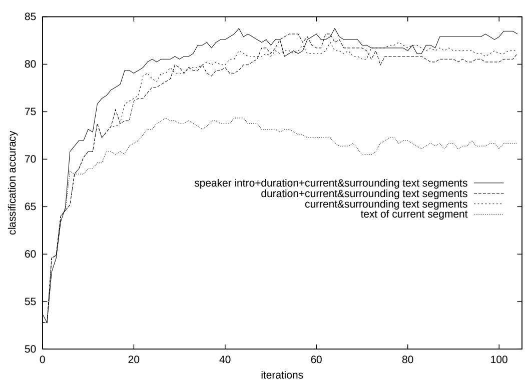

Figure 3: Classification accuracy of BoosTexter on different combinations of features (on development set, human transcripts)

	Human transcripts	ASR transcription
`Anchor`	npr's, npr, from national, all things considered i'm, and i'm @, us from, good afternoon i'm, reports, do you, what about	nbrs, nbi, things considered an, reports, this is all, commentator @, you, news in
`Journ.`	but, says, to all things, for national, is @ @ in, his, do you, we've been	reports, @ said, you, explain, @ @ says
`Guest`	i, we, yeah, well, i think, uh, our	i, i think, that we, it, you know

Table 2: Examples of n-grams with highest weight for human and ASR transcripts found by BoosTexter

Speaker introduction The presence of a speaker introduction in the current or previous segment were high-weighted features. A speaker introduction in the previous segment predicts journalist as speaker of the current segment, while a speaker introduction in the current segment predicts anchor as the speaker of the segment.

Given this set of features, BoosTexter outperforms Maximum Entropy by 4%. BoosTexter has an accuracy of 83.2% on the development set (after 100 rounds), while Maximum Entropy has 79% accuracy. However this picture changes when we add the labels of previously tagged paragraphs to the feature set. In the BoosTexter approach, the labels of the two previous paragraphs as computed by BoosTexter were given as input when tagging the current paragraph. Surprisingly, classification rate decreased significantly — 4.5%. This drop in accuracy occurs because in many cases the categories of previous speakers fully determine the category of current speaker. Therefore, when training BoosTexter weighs these features very highly, "neglecting" other features. In testing, one incorrectly predicted label often causes a "chain reaction" of incorrect labels. This greedy approach could be improved by using the confidence values computed by BoosTexter, and searching for the global sequence with the highest combined confidence. We leave this for future work.

The Maximum Entropy approach provides the conditional probability of label given the segment features and previous labels. Given a broadcast with segments $\{s_1, s_2, \ldots, s_n\}$, a label sequence candidate $\{l_1, l_2, \ldots, l_n\}$ has conditional probability:

$$P(l_1, l_2, \ldots, l_n | s_1, s_2, \ldots, s_n) = \prod_{i=1}^{n} p(l_i | l_{i-1}, l_{i-2}, l_{i-3}, s_i)$$

Beam search aims to find the labeling of the broadcast segments sequence with highest probability. History features included the label of the previous paragraph, the two previous paragraphs and the three previous paragraphs. With a beam-size of $N = 15$, Maximum Entropy outperforms BoosTexter by 1.5% (see Figure 4). Taking into account labels of previous segments improved the accuracy of the Maximum Entropy approach by 5%.

We ran BoosTexter and Maximum Entropy on an unseen test set. The classification accuracy on the test set is 79% for BoosTexter and 80.5% for Maximum Entropy. Table 3 shows prediction accuracy for each of the three speaker types.

	BoosTexter		MaxEnt	
	Recall	Precision	Recall	Precision
`Anchor`	81.3%	74.6%	91.7%	74.8%
`Journalist`	70.6%	83.2%	74.0%	90.4%
`Guest`	82.9%	76.6%	75.2%	78.2%

Table 3: Precision/recall by category on the test set(human transcripts)

After developing the two algorithms on the human transcripts, we examined their performance on ASR

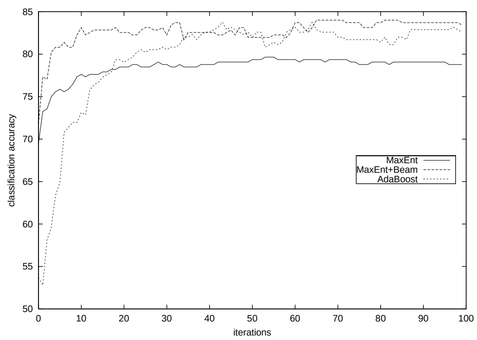

Figure 4: Classification accuracy of different learning algorithms (on development set, human transcripts)

transcripts. We formed training, development and test sets of ASR data, and trained BoosTexter and Maximum Entropy classifiers using the same feature types. BoosTexter performance is 75% for development set, and 72.8% for testing set. The accuracy of Maximum Entropy reached 79.9% on the development set and 77% on the test set. It is encouraging that the results were not substantially lower on ASR output, in spite of relatively high speech recognition error rates.

Summary and Future Work

In this paper, we have described how to compute the speaker role in news broadcasts from an automatic transcription of those broadcasts, assuming hand-labeled segmentation of (roughly speaking) speaker turns. We distinguish among three speaker types, Anchor, Journalist (non-Anchor), and Guest Speaker. The main contributions of this paper include identification of features which characterize each category, and an implementation of an algorithm based on those features which identifies speaker roles with high accuracy. A key finding is that content-based features are robust clues to speaker identity, and can be used as a complement to traditional audio-based methods.

Our working hypothesis was that speaker type information is an important cue to story structure. Our immediate future plans therefore involve testing the utility of the speaker type information we can currently identify when it is added to the speech browsing system. We will test the effectiveness of speaker type as an aid for audio browsing of Broadcast News. In addition, the system we have developed gives rise to a number of important issues. First is the question of how to combine our method with methods based on audio features in order to increase the accuracy of our procedure. For example, if the classifier is uncertain about a segment's type based on textual information alone, the acoustic similarity of the segment to other segments classified with high confidence may provide useful information. A related question is whether a combined acoustic-textual method might be extended to a full speaker identification system.

A more ambitious goal is to use speaker roles for parsing broadcast transcripts into structural units, such as headlines, interviews and news summaries. We have observed that each of these broadcast structural units appears to have its own patterns of speaker role changes, so this form of structural identification might indeed be possible. With such additional information, we should be closer to our long-term goal of providing summaries of broadcast news programs.

Acknowledgments

We thank Michiel Bacchiani and Amit Singhal for providing help with the TREC data and ASR transcriptions. Thanks to Aaron Rosenberg, Shimei Pan, Barry Schiffman, Noemie Elhadad, and Carl Sable for useful discussions and comments on earlier versions of this paper.

References

Chen, S., and Gopalakrishnan, P. 1998. Speaker, environment and change change detection and clustering via the bayesian information criterion. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

Couvreur, L., and Boite, J.-M. 1999. Speaker tracking in broadcast audio material in the framework of the thisl project. In *Proceedings of the ESCA ETRW workshop Accessing Information in Spoken Audio*, 84–89.

Hirschberg, J.; Whittaker, S.; Hindle, D.; Pereira, F.; and Singhal, A. 1999. Finding information in audio: A new paradigm for audio browsing and retrieval. In *Proceeding of the ESCA ETRW Workshop*.

Huang, Q.; Liu, Z.; Rosenberg, A.; Gibbon, D.; and Shahraray, B. 1999. Automated generation of news content hierarchy by integrating audio, video, and text information. In *Proceeding of the IEEE International Conference On acoustics, speech, and signal processing*, volume 6, 3025–3028.

Mani, I.; House, M.; Maybury, M.; and Green, M. 1997. Towards content-based browsing of broadcast news video. In Maybury, M., ed., *Intelligent Multimedia Information Retrieval*. AAAI/MIT Press. 241 – 258.

Mencher, M. 1987. *News Reporting and Writing*. Dubuque, Iowa: William C. Brown, 4 edition.

Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging. In *Proceeding of the Conference on Empirical Methods in Natural Language Processing*.

Reynar, J. 1999. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the ACL*.

Singer, Y., and Shapire, R. 1998. Improved boosting algorithms using confidence-rated predictions. In *Proceeding of 11th Annual Conference on Computational Learning Theory*, 80–91.

Stolcke, A.; Shriberg, E.; Hakkani-Tur, D.; Tur, G.; Rivlin, Z.; and Sonmez, K. 1999. Combining words and speech prosody for automatic topic segmentation. In *Proceedings of the DARPA Broadcast News Workshop*, 61 – 64.

Teufel, S., and Moens, M. 1998. Sentence extraction and rhetorical classification for flexible abstracts. In *Spring AAAI Symposium on Intelligent Text summarization*.