# Initial Study on Automatic Identification of Speaker Role in Broadcast News Speech

**Yang Liu**
University of Texas at Dallas, Richardson, TX
yangl@hlt.utdallas.edu

## Abstract

Identifying a speaker's role (anchor, reporter, or guest speaker) is important for finding the structural information in broadcast news speech. We present an HMM-based approach and a maximum entropy model for speaker role labeling using Mandarin broadcast news speech. The algorithms achieve classification accuracy of about 80% (compared to the baseline of around 50%) using the human transcriptions and manually labeled speaker turns. We found that the maximum entropy model performs slightly better than the HMM, and that the combination of them outperforms any model alone. The impact of the contextual role information is also examined in this study.

## 1 Introduction

More effective information access is beneficial to deal with the increasing amount of broadcast news speech. Many attempts have been made in the past decade to build news browser, spoken document retrieval system, and summarization or question answering system to effectively handle the large volume of news broadcast speech (e.g., the recent DARPA GALE program). Structural information, such as story segmentation or speaker clustering, is critical for all of these applications. In this paper, we investigate automatic identification of the speakers' roles in broadcast news speech. A speaker's role (such as anchor, reporter or journalist, interviewee, or some soundbites) can provide useful structural information of broadcast news. For example, anchors appear through the entire program and generally introduce news stories. Reporters typically report a specific news story, in which there may be other guest speakers. The transition between anchors and reporters is usually a good indicator of story structure. Speaker role information was shown to be useful for summarizing broadcast news (Maskey and Hirschberg, 2003). Anchor information has also been used for video segmentation, such as the systems in the TRECVID evaluations.[1]

In this paper, we develop algorithms for speaker role identification in broadcast news speech. Human transcription and manual speaker turn labels are used in this initial study. The task is then to classify each speaker's turn as *anchor*, *reporter*, or *other*. We use about 170 hours of speech for training and testing. Two approaches are evaluated, an HMM and a maximum entropy classifier. Our methods achieve about 80% accuracy for the three-way classification task, compared to around 50% when every speaker is labeled with the majority class label, i.e., anchor.[2]

The rest of the paper is organized as follows. Related work is introduced in Section 2. We describe our approaches in Section 3. Experimental setup and results are presented in Section 4. Summary and future work appear in Section 5.

## 2 Related Work

The most related previous work is (Barzilay et al., 2000), in which Barzilay et al. used BoosTexter and the maximum entropy model to classify each speaker's role in an English broadcast news corpus. Three classes are used, anchor, journalist, and guest speaker, which are very similar to the role categories in our study. Lexical features (key words), context features, duration, and explicit speaker introduction are used as features. For the three-way classification task, they reported accuracy of about 80% compared to the chance of 35%. They have investigated using both the reference transcripts and speech recognition output. Our study differs from theirs in that we use one generative modeling approach (HMM), as well as the conditional maximum entropy method. We also evaluate the contextual role information for classification. In addition, our experiments are conducted using a different language, Mandarin broadcast news. There may be inherent difference across languages and news sources.

Another task related to our study is anchor segmentation. Huang et al. (Huang et al., 1999) used a recognition model for a particular anchor and a background model to identify anchor segments. They reported very promising results for the task of determining whether

[1] See http://www-nlpir.nist.gov/projects/trecvid/ for more information on video retrieval evaluations.

[2] Even though this is a baseline (or chance performance), it is not very meaningful since there is no information provided in this output.



*Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 81–84,
New York, June 2006. 

or not a particular anchor is talking. However, this method is not generalizable to multiple anchors, nor is it to reporters or other guest speakers. Speaker role detection is also related to speaker segmentation and clustering (also called speaker diarization), which was a benchmark test in the NIST Rich Transcription evaluations in the past few years (for example, NIST RT-04F http://www.nist.gov/speech/tests/rt/rt2004/fall/). Most of the speaker diarization systems only use acoustic information; however, in recent studies textual sources have also been utilized to help improve speaker clustering results, such as (Canseco et al., 2005). The goal of speaker diarization is to identify speaker change and group the same speakers together. It is different from our task since we determine the role of a speaker rather than speaker identity. In this initial study, instead of using automatic speaker segmentation and clustering results, we use the manual speaker segments but without any speaker identity information.

## 3 Speaker Role Identification Approaches

### 3.1 Hidden Markov Model (HMM)

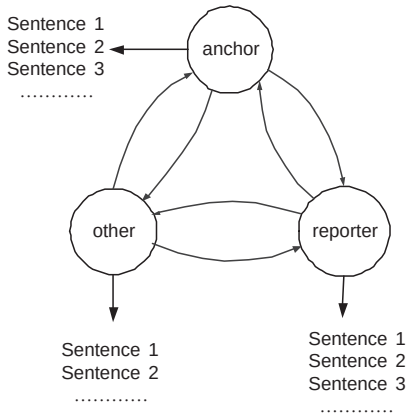


Figure 1: A graphical representation of the HMM approach for speaker role labeling. This is a simple first order HMM.

The HMM has been widely used in many tagging problems. Stolcke et al. (Stolcke et al., 2000) used it for dialog act classification, where each utterance (or dialog act) is used as the observation. In speaker role detection, the observation is composed of a much longer word sequence, i.e., the entire speech from one speaker. Figure 1 shows the graphical representation of the HMM for speaker role identification, in which the states are the speaker roles, and the observation associated with a state consists of the utterances from a speaker. The most likely role sequence $\hat{R}$ is:

$$\hat{R} = \underset{R}{\operatorname{argmax}} P(R|O) = \underset{R}{\operatorname{argmax}} P(O|R)P(R), \quad (1)$$

where $O$ is the observation sequence, in which $O_i$ corresponds to one speaker turn. If we assume what a speaker says is only dependent on his or her role, then:

$$P(O|R) = \prod_i P(O_i|R_i). \quad (2)$$

From the labeled training set, we train a language model (LM), which provides the transition probabilities in the HMM, i.e., the $P(R)$ term in Equation (1). The vocabulary in this role LM (or role grammar) consists of different role tags. All the sentences belonging to the same role are put together to train a role specific word-based N-gram LM. During testing, to obtain the observation probabilities in the HMM, $P(O_i|R_i)$, each role specific LM is used to calculate the perplexity of those sentences corresponding to a test speaker turn.

The graph in Figure 1 is a first-order HMM, in which the role state is only dependent on the previous state. In order to capture longer dependency relationship, we used a 6-gram LM for the role LM. For each role specific word-based LM, 4-gram is used with Kneser-Ney smoothing. There is a weighting factor when combining the state transitions and the observation probabilities with the best weights tuned on the development set (6 for the transition probabilities in our experiments). In addition, in stead of using Viterbi decoding, we used forward-backward decoding in order to find the most likely role tag for each segment. Finally we may use only a subset of the sentences in a speaker's turn, which are possibly more discriminative to determine the speaker's role. The LM training and testing and HMM decoding are implemented using the SRILM toolkit (Stolcke, 2002).

### 3.2 Maximum Entropy (Maxent) Classifier

A Maxent model estimates the conditional probability:

$$P(R_i|O) = \frac{1}{Z_\lambda(O)} exp(\sum_k \lambda_k g_k(R_i, O)), \quad (3)$$

where $Z_\lambda(O)$ is the normalization term, functions $g_k(R_i, O)$ are indicator functions weighted by $\lambda$, and $k$ is used to indicate different 'features'. The weights ($\lambda$) are obtained to maximize the conditional likelihood of the training data, or in other words, maximize the entropy while satisfying all the constraints. Gaussian smoothing (variance=1) is used to avoid overfitting. In our experiments we used an existing Maxent toolkit (available from http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html).

The following features are used in the Maxent model:

- bigram and trigram of the words in the first and the last sentence of the current speaker turn
- bigram and trigram of the words in the last sentence of the previous turn

- bigram and trigram of the words in the first sentence of the following turn

Our hypothesis is that the first and the last sentence from a speaker's turn are more indicative of the speaker's role (e.g., self introduction and closing). Similarly the last sentence from the previous speaker segment and the first sentence of the following speaker turn also capture the speaker transition information. Even though sentences from the other speakers are included as features, the Maxent model makes a decision for each test speaker turn individually without considering the other segments. The impact of the contextual role tags will be evaluated in our experiments.

# 4 Experiments

## 4.1 Experimental Setup

We used the TDT4 Mandarin broadcast news data in this study. The data set consists of about 170 hours (336 shows) of news speech from different sources. In the original transcripts provided by LDC, stories are segmented; however, speaker information (segmentation or identity) is not provided. Using the reference transcripts and the audio files, we manually labeled the data with speaker turns and the role tag for each turn.[3] Speaker segmentation is generally very reliable; however, the role annotation is ambiguous in some cases. The interannotator agreement will be evaluated in our future work. In this initial study, we just treat the data as noisy data.

We preprocessed the transcriptions by removing some bad codes and also did text normalization. We used punctuation (period, question mark, and exclamation) available from the transcriptions (though not very accurate) to generate sentences, and a left-to-right longest word match approach to segment sentences into words. These words/sentences are then used for feature extraction in the Maxent model, and LM training and perplexity calculation in the HMM as described in Section 3. Note that the word segmentation approach we used may not be the-state-of-art, which might have some effect on our experiments.

10-fold cross validation is used in our experiments. The entire data set is split into ten subsets. Each time one subset is used as the test set, another one is used as the development set, and the rest are used for training. The average number of segments (i.e., speaker turns) in the ten subsets is 1591, among which 50.8% are anchors. Parameters (e.g., weighting factor) are tuned based on the average performance over the ten development sets, and the same weights are applied to all the splits during testing.

[3]The labeling guideline can be found from http://www.hlt.utdallas.edu/~yangl/spkr-label/. It was modified based on the annotation manual used for English at Columbia University (available from http://www1.cs.columbia.edu/~smaskey/labeling/Labeling_Manual_v_2_1.pdf).

## 4.2 Results

A **HMM and Maxent**: Table 1 shows the role identification results using the HMM and the Maxent model, including the overall classification accuracy and the precision/recall rate (%) for each role. These results are the average over the 10 test sets.

| | HMM | | Maxent | |
|---|---|---|---|---|
| | precision | recall | precision | recall |
| anchor | 78.03 | 87.33 | 80.29 | 87.23 |
| reporter | 78.54 | 66.42 | 73.34 | 77.01 |
| other | 83.05 | 68.19 | 89.52 | 41.30 |
| Accuracy (%) | 77.18 | | 77.42 | |

Table 1: Automatic role labeling results (%) using the HMM and Maxent classifiers.

From Table 1 we find that the overall classification performance is similar when using the HMM and the Maxent model; however, their error patterns are quite different. For example, the Maxent model is better than the HMM at identifying "reporter" role, but worse at identifying "other" speakers (see the recall rate shown in the table). In the HMM, we only used the first and the last sentence in a speaker's turn, which are more indicative of the speaker's role. We observed significant performance degradation, that is, 74.68% when using all the sentences for LM training and perplexity calculation, compared to 77.18% as shown in the table using a subset of a speaker's speech. Note that the sentences used in the HMM and Maxent models are the same; however, the Maxent does not use any contextual role tags (which we will examine next), although it does include some words from the previous and the following speaker segments in its feature set.

B **Contextual role information**: In order to investigate how important the role sequence is, we conducted two experiments for the Maxent model. In the first experiment, for each segment, the reference role tag of the previous and the following segments and the combination of them are included as features for model training and testing (a "cheating" experiment). In the second experiment, a two-step approach is employed. Following the HMM and Maxent experiments (i.e., results as shown in Table 1), Viterbi decoding is performed using the posterior probabilities from the Maxent model and the transition probabilities from the role LM as in the HMM (with weight 0.3). The average performance over the ten test sets is shown in Table 2 for these two experiments. For comparison, we also present the decoding results of the HMM with and without using sequence information (i.e., the transition probabilities in the HMM). Additionally, the system combination

results of the HMM and Maxent are presented in the table, with more discussion on this later. We observe from Table 2 that adding contextual role information improves performance. Including the two reference role tags yields significant gain in the Maxent model, even though some sentences from the previous and the following segments are already included as features. The HMM suffers more than the Maxent classifier when role sequence information is not used during decoding, since that is the only contextual information used in the HMM, unlike the Maxent model, which uses features extracted from the neighboring speaker turns.

| | Accuracy (%) |
|---|---|
| 0: Maxent (as in Table 1) | 77.42 |
| 1: Maxent + 2 reference tags | 80.90 |
| 2: Maxent + sequence decoding | 78.59 |
| 3: HMM (as in Table 1) | 77.18 |
| 4: HMM w/o sequence | 73.30 |
| Maxent (0) + HMM (3) | 79.74 |
| Maxent (2) + HMM (3) | 81.97 |

Table 2: Impact of role sequence information on the HMM and Maxent classifiers. The combination results of the HMM and Maxent are also provided.

C **System combination**: For system combination, we used two different Maxent results: with and without the Viterbi sequence decoding, corresponding to experiments (0) and (2) as shown in Table 2 respectively. When combining the HMM and Maxent, i.e., the last two rows in Table 2, the posterior probabilities from them are linearly weighted (weight 0.6 for the Maxent in the upper one, and 0.7 for the Maxent in the bottom one). The combination of the two approaches yields better performance than any single model in the two cases. We also investigated other system combination approaches. For example, a decision tree or SVM that builds a 3-way superclassifier using the posterior probabilities from the HMM and Maxent. However, so far we have not found any gain from more complicated system combination than a simple linear interpolation. We will study this in our future work.

## 5 Summary and Future Work

In this paper we have reported an initial study of speaker role identification in Mandarin broadcast news speech using the HMM and Maxent tagging approaches. We find that the conditional Maxent generally performs slightly better than the HMM, and that their combination outperforms each model alone. The HMM and the Maxent model show differences in identifying different roles. The impact of contextual role information is also examined for the two approaches, and a significant gain is observed when contextual information is modeled. We find that the beginning and the end sentences in a speaker's turn are good cues for role identification. The overall classification performance in this study is similar to that reported in (Barzilay et al., 2000); however, the chance performance is quite different (35% in that study). It is not clear yet whether it is because of the difference across the two corpora or languages.

The Maxent model provides a convenient way to incorporate various knowledge sources. We will investigate other features to improve the classification results, such as name information, acoustic or prosodic features, and speaker clustering results (considering that the same speaker typically has the same role tag). We plan to examine the effect of using speech recognition output, as well as automatic speaker segmentation and clustering results. Analysis of difference news sources may also reveal some interesting findings. Since our working hypothesis is that speaker role information is important to find structure in broadcast news, we will investigate whether and how speaker role relates to downstream language processing applications, such as summarization or question answering.

## Acknowledgment

The author thanks Julia Hirschberg and Sameer Maskey at Columbia University and Mari Ostendorf at the University of Washington for the useful discussions, and Mei-Yuh Hwang for helping with Mandarin word segmentation and text normalization. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-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 DARPA.

## References

R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In *Proc. of AAAI*.

L. Canseco, L. Lamel, and J Gauvain. 2005. A comparative study using manual and automatic transcription for diarization. In *Proc. of ASRU*.

Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray. 1999. Automated generation of news content hierarchy by integrating audio, video, and text information. In *Proc. of ICASSP*, pages 3025–3028.

S. Maskey and J. Hirschberg. 2003. Automatic summarization of broadcast news using structural features. In *Eurospeech*.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurasky, P. Taylor, R. Martin, C.V. Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

A. Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proc. of ICSLP*, pages 901–904.