UNSUPERVISED BROADCAST CONVERSATION SPEAKER ROLE LABELING

Brian Hutchinson, Bin Zhang and Mari Ostendorf

Electrical Engineering Department, University of Washington, Seattle, WA 98195

ABSTRACT

We present an approach to unsupervised speaker role labeling in talk show data that makes use of two complementary sets of features: structural features that encode the participation patterns of speakers, and lexical features, which capture characteristic phrases. Techniques for using multiple clusterings are explored, leading to more robust results. Experiments on English and Mandarin talk shows yield performance similar to that reported for broadcast news using supervised learning.

***Index Terms*—** Unsupervised learning, meta-clustering, speaker role classification, broadcast conversations

1. INTRODUCTION

With a substantial and increasing amount of broadcast audio available, there is interest in automatically analyzing the broadcast content. To do so, it is useful to annotate transcripts with the structure of the show, including speaker segmentation and diarization, topic and story segmentation, and the task we address in this research: speaker role labeling.

Initial work on speaker role classification [1] was on broadcast news, categorizing speakers into three categories: anchor, journalist, and guest. In order to learn words related to speaker introductions, the authors employed a large number of features: n-grams close to the word, relative word frequency, the position of the word in a segment, and the capitalization of the word. Sentence durations and the labels of surrounding context were also used as features, and the feature weights were learned on labeled training data via Boostexter or maximum entropy model. An accuracy of 80% was achieved on the ASR derived transcripts. Liu et al. [2] studied the classification of speaker roles on TDT-4 Mandarin broadcast news (BN) audio data. Hidden Markov and maximum entropy models were used to label the sequence of speaker turns with the roles anchor, reporter, and other, based on n-gram features extracted from human transcripts. The algorithm reached 80% classification accuracy. Vinciarelli [3] proposed to use social network analysis on speaker clustered news bulletins, and achieved an accuracy of 85% on a task of classifying six speaker roles: anchor, secondary anchor, guest, interview participant, abstract (speaker who gives summary at the start of the show), and meteo (speaker who gives a weather report).

While this past work has been reasonably successful, it requires a large amount of hand-annotated data. Because of the wealth of unannotated data and the costs associated with manual annotation, we approach speaker role labeling as an unsupervised learning task. To discriminate between speakers of different roles, we designed two complementary feature sets. The first exploits the different patterns of participation employed by different roles, characterizing a speaker by a set of *structural features*. The second exploits lexical usage patterns, quantifying a speaker's use of "signature phrases" and conversational n-grams in a set of *lexical features*. We apply several clustering methods to these features and combinations thereof, followed by either a *meta-clustering* step to combine the results of multiple clusterings or a *partition selection* algorithm that chooses a typical clustering from the candidate set. Experiments on English and Mandarin broadcast conversations are presented, comparing the feature subsets and different clustering approaches across languages.

2. SPEAKER ROLE FEATURES

2.1. Structural Features

In talk shows, or broadcast conversations (BC), speakers of different roles often exhibit very different patterns of participation. For example, a journalist's sole contribution might be a two minute news update with minimal back-and-forth with the host, whereas an invited guest in a panel is likely to have a longer presence and interact with other panelists. Our "structural" feature set is designed to encode these patterns, quantifying speaker participation in a feature vector. Features include: the total duration of the speaker's utterances; the total number of words (or characters, for Mandarin), utterances and turns; the duration of the longest single turn; and the length of time the speaker is conversationally involved (end time of last turn minus start time of first turn). Each feature is normalized to account for differences in show length.

2.2. Lexical Features

According to the roles they play in the conversation, different speakers have different usage of lexical terms. For instance, hosts introduce the show and guests to the audience before the

This research is supported by IARPA. The information herein does not necessarily reflect the position or policy of the U.S. Government.

conversation, using phrases such as "Welcome back. Our next guest..." These phrases are referred to as *signature phrases* [1] — the n-grams that characterize hosts. Without labeled training data, it is not possible to learn these phrases in a supervised manner as in [1], nor is it practical to hand pick the phrases. We propose to use statistics derived from speaker and document labels to select the signature phrases. For each n-gram w_1^n, we define speaker frequency (SF) as the percent of speakers who have uttered w_1^n and document frequency (DF) as the percent of documents in which w_1^n has appeared.

Because most shows have a single host and multiple guests, the signature phrases spoken by hosts should have low SF but high DF. We define the composite statistic for signature phrases,

$$\theta_1 = \frac{DF}{SF} + \alpha \log(DF). \tag{1}$$

The first term advocates high DF/SF, while the second logarithm term advocates high DF to partially suppress low-frequency n-grams. The signature phrase list is generated as the top n-grams ranked by θ_1. Trigrams are used in this paper. α is a corpus-dependent parameter which balances the dynamic range of two terms, which was tuned to 10 for English and 10^{-4} for Mandarin by inspecting the signature phrase list generated on unlabeled corpora.

Another set of phrases that discriminate speaker roles is the set of *conversational phrases*. Most soundbites lack the conversational phrases used by the host and guests who, unlike soundbites, participate in active conversation. To approach the problem of selecting conversational phrases, we introduce the cross-genre ratio $GR = f_{\mathrm{BC}}(w_1^n)/f_{\mathrm{BN}}(w_1^n)$, where $f_{\mathrm{BC}}(w_1^n)$ and $f_{\mathrm{BN}}(w_1^n)$ are the frequencies of w_1^n in the BC and BN corpora, respectively. GR is used in the composite statistic for conversational phrase selection:

$$\theta_2 = SF \times \log(GR + 1), \tag{2}$$

with the motivation that conversational phrases should be frequent n-grams and have high GR. The use of the logarithm makes GR less important, which also suppresses some BC signature phrases. We explored different order n-grams, and inspection of the resulting lists suggested using bigrams since trigrams introduced too many topic-related words.

The signature phrases learned include expressions that would typically be associated with a host, such as "be right back" and "joining us now." Conversational phrases often include "but", the pronoun "I", and filler phrases ("you know").

We assign weights $w(i) = e^{-i/N}$ to the phrases in the list according to their rank i (effectively controlling the list size with N, where $N = 100$). The speaker's feature values for signature phrases and conversational phrases are then computed by the weighted sum of the phrase counts in all of the speaker's utterances, yielding a two-dimensional lexical feature vector per speaker.

3. METHODS

Our primary task in unsupervised role labeling is to take the set of T speaker (or, talker) feature vectors, which we denote $\mathcal{X} = \{x_1, \ldots, x_T\}$, and partition it into K subsets (roles). Our approach involves performing several clustering runs in a first stage, and then leveraging these using a variety of ensemble clustering techniques.

3.1. First-Stage Clustering

In the first level of clustering, we employ three standard clustering algorithms: k-means, diagonal covariance Gaussian mixture models (GMMs), and spectral clustering, denoted `kmeans`, `gmm`, and `spectral`, respectively. We randomly select K samples as initial centroids for `kmeans`. We initialize the GMM with the same samples as initial means and with global diagonal covariance and uniform mixing weights. Hard decisions are used in the final assignment. The implementation of `spectral` follows the Shi-Malik algorithm in [4], with edge weight between nodes equal to $e^{-\|x_i - x_j\|^2/2\sigma^2}$.

3.2. Leveraging Multiple Clusterings

Intuitively, since clustering systems can fall into local optima, one might hope to improve results by combining a set of P clusterings (or, partitions) into a single clustering that outperforms the component clusterings, much like ensemble classifiers that are able to outperform individual classifiers (as in system combination). Unlike the combination of classifiers, however, the combination of clusterings is complicated by the fact that the cluster labels produced by two clustering systems are not necessarily aligned. Cluster 1 from system A may correspond to cluster 2 from system B, and they could have small differences in cluster membership. We consider three methods for addressing this problem: two meta-clustering approaches that use a set of clusterings as input to a subsequent clustering algorithm, and a partition selection algorithm that looks for multiple identical clusterings. In each case, the goal is to map T objects (speakers) to K clusters (roles).

Meta-clustering, or *clustering clusterings*, can be formulated as hyper-graph partitioning [5], meta-graph clustering [5], integer linear programming (ILP) [6] and singular value decomposition (SVD) [6]. Our work explores the last two methods. In both cases, the first-stage collection of clusterings (each of which partition the space into M clusters, not necessarily equal to the final desired K) is used to generate new (meta) feature vectors for each object (f_t for speaker t), which are then used with the k-means algorithm to determine the final K clusters. The ILP approach (`ipc`) finds a mapping of each of the P clusterings to a set of meta-clusters by iteratively maximizing the average similarity of clusters from different runs assigned to the same meta-cluster, subject to the constraint that each meta-cluster contains only one cluster

Data Set	Hosts	Exp. Guests	Soundbites	Total
Man Dev	10 (14%)	29 (41%)	32 (45%)	71
Man Eval	14 (10%)	39 (28%)	87 (62%)	140
Eng Eval	9 (6%)	74 (49%)	67 (45%)	150

Table 1. Speaker role counts (and percentages) by data set.

from a particular run. The i-th entry of f_t is the percentage of clusters mapped to meta-cluster i that contain speaker t. The SVD approach (svd) constructs an $N \times MP$ matrix $\mathbf{R}$ by stacking the assignment vectors from the different clusterings together and then finds the optimal rank M approximation to $\mathbf{R}$ using SVD $\mathbf{R} \approx \mathbf{USV}^T$. Diagonal matrix $\mathbf{S} \in \mathbb{R}^{M \times M}$ contains the M largest singular values; $\mathbf{U}$ and $\mathbf{V}$ contain the corresponding left and right singular vectors. The meta feature f_t is row t of $\mathbf{US}$.

Empirically, on development data, we find that the majority of clusterings converge to the same high performing partition. Therefore, we also explore partition selection (ps) as an alternative way of using multiple clusterings, where we pick the most common clustering among the candidates. A clustering is represented by a vectorized $N \times N$ adjacency matrix, where the i, jth entry is 1 if x_i and x_j belong to the same cluster and 0 otherwise. If there are ties, we find the component-wise mode over the candidates, and the vectorized adjacency matrix closest to the mode in ℓ_1 distance is picked.

4. EXPERIMENTS AND RESULTS

4.1. Data

We use five data sets for this research, drawn from BC corpora used for the GALE project. The small amount of data labeled with speaker roles is used only for testing; this includes Mandarin development and evaluation data sets and an English evaluation data set. Table 1 summarizes the three labeled data sets; in each case the number of shows is equal to the number of hosts. Additional unlabeled data in English and Mandarin is used for intuition-driven tuning of lexical features. In testing, we compute the structural and lexical features from the "quick rich transcription" annotations, using manual sentence time alignments and transcriptions.

English and Mandarin BC shows, though both BC, have significant differences that can result in variation of our features. An English show tends to have more speakers, more utterances and turns, and more informal segments than a Mandarin show. Whereas in Mandarin there tend to be more soundbite speakers than guests, in English the reverse is true. Also, English soundbites more often involve spontaneous speech, while Mandarin soundbites are primarily news reporting. There are many highly interactive debates in English shows, which are rarely seen in Mandarin.

4.2. Experimental Details

In our experiments we classify speakers into three roles: hosts, expert guests (e.g. journalists, panelists, interviewees), and soundbites (non-interactive call-in and man-on-the-street guests are also included in this category). Because the final output of each clustering (or meta-clustering) system is a partitioning of speakers, we need to apply rules to map clusters to roles before computing accuracies. To perform this mapping we use a simple heuristic: the cluster whose members have the largest average number of turns is the host cluster, that with the smallest average number of turns is the soundbite cluster, and the remaining cluster contains the expert guests. We use similar rules to form a baseline system: per show, the single speaker with the largest number of turns is the host; any speakers with ≤ 5 utterances are soundbites, and the remaining are expert guests. This baseline yields accuracies of 81%, 72% and 87%, respectively on the Mandarin development and evaluation and English evaluation sets.

In preliminary experiments, we observed that including outputs of several different low-level clustering methods in the meta-clustering stage did not offer advantages over using the best low-level method alone. These trends persisted in additional preliminary experiments where meta-clustering was applied to the outputs of multiple meta-clustering methods. We therefore conduct our experiments as follows.

A single set of accuracies is generated by first running a low-level clustering method 50 times (varying $\sigma \in \{1, \ldots, 50\}$ in the case of spectral clustering) and storing average accuracy. All 50 clusterings are then fed into the two meta-clusterers and into the partition selection routine. The accuracies for the partition selection method and each of the meta-clustering methods is stored. This process is repeated 50 times and results averaged to produce the accuracies reported below.

The feature sets we use were selected by preliminary experiments. They include: struct, a subset of the structural features chosen on the Mandarin development set consisting of total duration of speech, number of turns, and the duration of the single longest turn; and lex, the two lexical features chosen on the unlabeled data.

In addition to experiments on single feature sets, we report results for feature-level and cluster-level feature set combinations. In the results, these combinations are marked with & and +, respectively. For the feature-level combination, all 50 clusterings are run on the single vector-concatenated feature set struct&lex. For the cluster-level combination, the 50 clusterings are split between different feature sets. That is, for the struct+lex result, 25 clusterings are created on the struct features and 25 clusterings are generated on the lex features to produce the 50 clusterings fed into the meta-clustering and partition selection routines.

	Mandarin Development				**Mandarin Evaluation**				**English Evaluation**			
Feature Set	spectral	ipc	svd	ps	spectral	ipc	svd	ps	spectral	ipc	svd	ps
struct	78.8	76.6	64.2	**78.9**	**81.4**	77.4	69.0	**81.4**	79.5	76.5	70.7	**84.0**
lex	85.7	81.1	61.1	**85.9**	80.6	82.4	64.6	**84.3**	71.5	66.2	66.5	**78.0**
struct+lex	**82.3**	71.1	71.5	78.9	81.2	70.6	70.5	**81.4**	75.3	64.4	50.9	**80.3**
struct&lex	83.8	78.9	72.5	**85.9**	**82.1**	79.3	77.2	**82.1**	85.1	78.6	67.0	**86.0**

Table 2. Accuracies for all data sets.

4.3. Results

Table 3 compares performance of the low-level clustering methods versus the cluster combination method on the struct&lex features in the Mandarin development set. The spectral low-level clustering method achieves the highest score. For this reason, the remainder of our reported results are on experiments using spectral clustering as the low-level method.

Table 2 shows clustering strategy versus feature set on Mandarin development and English and Mandarin evaluation sets. Not only are the highest accuracies obtained by partition selection, but also the lowest variances; for example, the average standard deviation of spectral clustering accuracies is 1.0 versus 0.2 for partition selection. In fact, for all data sets and feature sets except for struct+lex, the variance of partition selection accuracies is 0. Feature-level concatenation (struct&lex) also yields consistent improvements over the cluster-level combination (struct+lex). While the meta-clustering methods do not always improve accuracy, they may be useful in determining role label confidence in future work. As illustrated by both the baseline and system results, structural features are more powerful on the English data than on the Mandarin data.

Potential sources of mismatch between the data sets include the degree to which signature phrases and conversationality are linked to role and the relative distributions of speaker role. This may help explain differences in the success of feature versus cluster-level combination alternatives across language. In English shows, for example, the spontaneous speech in the soundbites reduces the discriminative power of the conversational dimension of the lexical features. Interestingly, despite having tuned the structural feature set to the Mandarin development set, we observe higher overall performance on the English evaluation set than the Mandarin evaluation set.

	low	ipc	svd	ps	avg
gmm	77.5	77.3	67.2	80.3	75.6
kmeans	78.9	70.6	69.7	78.9	74.5
spectral	83.8	78.9	72.5	**85.9**	**80.3**

Table 3. Comparison of partition selection accuracies by low-level method on Mandarin development data set, using struct&lex feature set.

5. CONCLUSION

In this paper we presented an approach to unsupervised speaker role labeling using two complementary feature sets: structural and lexical features. We applied several standard clustering algorithms to the feature sets, and clustered the clusterings using two meta-clustering strategies as well as a simple partition selection algorithm. We find the best results using low-level spectral clustering and high-level partition selection. Structural features outperform lexical features alone for English, but the reverse is true for Mandarin. In general, the best results are obtained by feature combination methods.

There are several ways one could extend this work. First, additional kinds of features could be explored, such as prosodic features. Other meta-clustering methods could also be applied, including those in [5]. Finally, incorporating redundancy compensation into the phrase list generation is likely to improve the quality of the lexical features.

6. REFERENCES

[1] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, “The rules behind roles: Identifying speaker role in radio broadcasts,” in *Proc. AAAI*, 2000, pp. 679–684.

[2] Y. Liu, “Initial study on automatic identification of speaker role in broadcast news speech,” in *Proc. HLT*, 2006, pp. 81–84.

[3] A. Vinciarelli, “Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling,” *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1215–1226, 2007.

[4] U. von Luxburg, “A tutorial on spectral clustering,” *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, December 2007.

[5] A. Strehl and J. Ghosh, “Cluster ensembles — a knowledge reuse framework for combining multiple partitions,” *Jour. Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[6] C. Boulis, *Topic Learning in Text and Conversational Speech*, Ph.D. thesis, University of Washington, 2005.