Social Role Discovery from Spoken Language using Dynamic Bayesian Networks

Sibel Yaman[1], *Dilek Hakkani-Tur*[1], *Gokhan Tur*[2]

[1]International Computer Science Institute, Berkeley, CA USA
[2] Microsoft, Microsoft Research, Mountain View, CA 94041
sibel@icsi.berkeley.edu, dilek@icsi.berkeley.edu, gokhan.tur@ieee.org

Abstract

In this paper, we focus on inferring social roles in conversations using information extracted only from the speaking styles of the speakers. We model the turn-taking behavior of the speakers with dynamic Bayesian networks (DBNs), which provide the capability of naturally formulating the dependencies between random variables. More specifically, we first explore the usefulness of a simple DBN, namely, a hidden Markov model (HMM), for this problem. As it turns out, the knowledge of the segments that belong to the same speaker can be augmented into this HMM structure, which results in a more sophisticated DBN. This information places a constraint on two subsequent speaker roles such that the current speaker role depends not only on the previous speaker's role but also on that most recent role assigned to the same speaker. We conducted an experimental study to compare these two modeling approaches using broadcast shows. In our experiments, the approach with the constraint on same speaker segments assigned 89.5% turns the correct role whereas the HMM-based approach assigned 79.2% of turns their correct role.

Index Terms: Social role discovery, speaker turn detection, spoken language understanding

1. Introduction

Spoken conversation is a growing source of intelligence that provides rich information and nuances beyond those available from written text alone. Conversation analysis [1] examines the way conversation partners manage turn-taking, the sequential relationship between utterances, communication difficulties, word selection, and the rights to knowledge and to action. How speakers do so can reflect the speaker's role in a conversation, his or her relationships with others, and characteristics of the speaker, such as his or her beliefs or group affiliation, in other words, speaker roles, social relations, and social identities.

Several researchers have extracted speaker roles from broadcast news. Some of these [2, 3, 4] explored assigning a role to a speaker by extracting features considering all the appearances of the same speaker in the same show or even in the entire data collection. For detecting roles in radio broadcasts, Vinciarelli [2] has employed duration-based and social network analysis-based features such as centrality and relative interaction. Garg *et al.* [3] extended this approach with the inclusion of the content of the interaction. These approaches have the implicit disadvantage that many turn-specific information is lost. Barzilay *et al.* [5] explored approaches that assign a role to each speaker turn. Their goal was mainly to find out those features that help best distinguish different role classes. However, our main focus in the present paper is to exploit the turn-taking patterns of speakers as well as additional information such as the information on the segments spoken by the same speaker with the help of graphical models.

Unlike these more conventional approaches that map directly from features to end representations, this paper explores learning a mapping from speaking style to social roles using a training corpus of broadcast conversations. In our future work, we will develop methods to detect the important speaking style information automatically as well but here we use manual annotations of them to present the capabilities of such a modeling scheme. DBNs provide us with the distinct benefit that a broad variety of modeling schemes can be conceptualized in a single framework with an intuitively appealing graphical representation. Such a graphical notation helps us formulate and better understand the interdependencies between linguistic phenomena and social roles.

In the next section, we briefly present the social role categories common for broadcast conversations, such as talkshows. In Section 3, we describe the proposed DBN framework for the detection of social roles. Section 4 includes our experimental results showing the effectiveness of the proposed approach compared to HMMs.

2. Speaker Roles in Broadcast Shows

When we listen to a radio program, we can usually tell whether the speaker is the program host, a reporting journalist, an audience participant, or a guest speaker in terms of their *speaking styles* even when the language is unknown. A program host is responsible for reading news, introducing reports from journalists, and announcing upcoming events. A reporting journalist is a professional speaker, generally in some remote location where a story is taking place. A guest speaker is usually a professional or non-professional speaker speaking from a subjective point of view. An audience participant is a non-professional speaker who asks questions or make personal comments.

Broadcast conversations as well as many other forms of conversations usually have a regular structure. The main functionality of the program host is to organize in what turn the participants take part in the show and to conduct the subject of discussion. Therefore, program hosts tend to occur frequently in the program and to alternate with the other participants. The majority of the conversation is among the program host and the guest speakers. Journalists typically appear only a few times throughout the show and converse with the program host while audience participants usually have questions to program guests.

The program participants typically also exhibit regularities in their speaking styles in accordance with their role in the conversation. A guest speaker is entitled to talk about their opinion, make a comment, or to explain a matter of truth based on their expertise. Therefore, a guest speaker is typically confident on

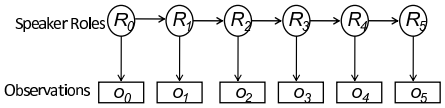

Figure 1: Basic HMM structure.

the subject being discussed. When, for instance, challenged by the program host, by another guest speaker, or by an audience participant, his/her speaking style reflects his/her reaction. For instance, when there is another opponent guest speaker in the program, a guest speaker will react to opponent views by showing disagreements, increased pitch, and frequent disfluencies.

2.1. Information in Speaking Style

Our end goal in this study is to predict the social roles of speakers by correlating their roles with automatically detected features that relate to their speaking styles. The motivation behind using this sort of feature vectors is that there is a wealth of information that is found *only* in spoken language and is not reflected in written language. We have identified the following speaking-style categories for this purpose:

Person addresses and mentions: Person addresses are the terms a speaker uses to address another participant of the discussion whereas person mentions are used to refer to a non-participant of the conversation.

Disfluencies: Disfluencies include phenomena such as filled pauses ("uh" and "um"), repetitions ("I... I ..."), repairs ("the las- the first"), and false starts ("They- it was not their choice").

Prefaces: Prefaces provide some commentary about the preceding utterance, the current utterance, or the relationship between these two, for instance "good question", "as you say", "it was like", "well". The purpose of having prefaces in speech is mainly to comment on the utterance that the speaker intends to say or on the utterance that prompted its production. For instance, in one show, one guest speaker asks to another "What did he say?". The response is "Oh come on, let's not regurgitate that, Matt." Before avoiding the question entirely, the second guest speaker produces a preface, "oh come on" that comments on the question.

2.2. Feature Extraction

The speaking style features described above reveal several important characteristics of the speaker roles. Frequent person addresses may indicate a conversation in which speakers try to come to terms on some subject while frequent person mentions may signal a reporting journalist or a guest speaker talking about a story that they have knowledge about. A speaking style with frequent hesitation may indicate that the speaker is having trouble formulating the content of a message and may suggest a lack of authority in that part of the conversation. If the speaker makes frequent prefaces, this may reveal that this speaker makes comments on what the previous speaker says.

To capture these information, we used forced-alignments of the broadcast shows and extracted two types of features using the human annotations for the linguistic phenomena described in Section 2.1.

- The first type of features is the duration of the linguistic phenomena in a given turn, i.e., the duration of disfluencies, prefaces, person addresses and person mentions.
- The second type of features is the ratio of the duration of the preface to the entire turn duration.

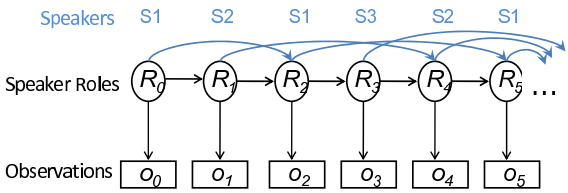

Figure 2: The segments uttered by the same speaker helps determine the role assigned to the same speaker in the subsequent turn.

3. Dynamic Bayesian Networks for Speaker Role Classification

The problem is defined as assigning a role to the speaker in each turn given the observed speaking style features. More formally, we look for the best estimates, $\hat{R}_i$, for the hidden state given the observation vectors, i.e.,

$$(\hat{R}_1, ..., \hat{R}_T) = \text{argmax } P(R_1, ..., R_T | o_1, ..., o_T) \quad (1)$$

To this end, we have segmented given broadcast shows into turns with one active speaker in each segment. We also assume the segments uttered by the same speaker are known (but the identity of the speaker is irrelevant).

In natural conversations, the role of the previous speaker as well as his/her speaking style determines the role of the next speaker. For instance, when a question is asked to request information (by a program host or audience participant), a speaker with the required expertise is expected to take the next turn. Similarly, when a controversial statement is made, an opponent speaker is expected to take the next turn. Therefore, we model the dynamics of the turn-taking problem as a first-order hidden Markov model as depicted in Figure 1. Each hidden state, R_i, represents the role of the speaker in the i^{th} turn and emits an observation (or feature) vector, o_i. The probability, $P(o_i|R_i)$, denotes the probability of emitting o_i at state R_i, and the probability $P(R_i|R_{i-1})$ denotes the probability of the transition from speaker role R_i to R_{i-1}. We model the observation probabilities, $P(o_i|R_i)$, with Gaussian Mixture Models (GMMs).

Such a graphical structure assigns a role to a given speaker without any regards to the roles assigned to the same speaker at different turns. In case human annotations are available, the information on what segments are uttered by what speaker is readily available. Similarly, in case automatic speech recognition annotations are used, speaker diarization systems cluster the same speaker segments. Therefore, in either case, the segments uttered by the same speaker are already known. This knowledge can be appended to the network structure such that there is a path among every pair of the nodes that belong the same speaker. This network structure would be too complicated as every pair is connected, but as shown in Figure 2, it is possible to simplify it by leaving only the path between any two consecutive nodes belonging to the same speaker. Then, the problem is defined as a sequence classification task where the goal is finding the most probable path given the previous speaker's role as well as the most recent role assigned to the same speaker.

3.1. Constraints on Same-Speaker Segments

If the segments are known to belong to a certain speaker, S, such as "Speaker 1" (without any need for speaker identity), this information helps us constrain $\hat{R}_i$ so that R_i depends not

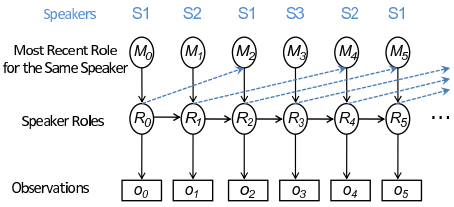

Figure 3: The most recent role assigned to the same speaker is embedded as an additional node.

only on R_{i-1} but also the previous roles assigned to the same speaker, S_i. Consider as an example that the speaker sequence is "$S_1, S_2, S_1, S_3, S_2, S_3, S_1, S_2$" with the truth speaker roles $Role(S_1) = H$ for host, $Role(S_2) = G$ for guest, and $Role(S_3) = A$ for audience participant. The corresponding speaker role sequence is "$HGHAGAHG$". In other words, what this turn-taking pattern shows is that the program host (S_1) and a guest speaker (S_2) are initially talking to each other. Then, a person from audience (S_3) makes a comment addressing the guest speaker. Upon getting a response from the guest speaker, the audience participant takes a turn once again and then the program host resumes talking to the guest speaker. Decoding this sequence based on only information from the observation vectors and previous speaker is apparently prone to error. However, the knowledge of the segments spoken by the same speaker (for instance, the segments spoken by S_2) constrains the roles that can be assigned to each of segments.

To keep generality, we propose here a Bayesian formalism and derive a corresponding graphical model. The constraint on the same speaker segments can be formally written as

$$R_i = R_j \Leftrightarrow S_i = S_j, \forall i, j \tag{2}$$

which helps make sure that a speaker will be assigned a single consistent role. A more relaxed version is to constrain the speaker roles so that the most recent estimate of the speaker role is a determining factor for the current role. This is depicted in Figure 2, where there is an additional connection from each estimated speaker role node to the next node that belongs to the same speaker.

We denote with M_i a random variable for the *estimate* of the role of the speaker in his/her most recent turn. Hence, the graphical structure of Figure 2 is equivalent to that of Figure 3. The dashed lines in Figure 3 stand for the fact that the nodes M_i just represent a copy of the most recent speaker role assigned to the same speaker, i.e.,

$$M_i = \hat{R}_j, i < j, \text{ and } \nexists\, k : j < k < i \text{ for which } S_k = S_i. \tag{3}$$

This graphical structure requires computing $P(R_i|R_{i-1}, M_i)$.

$$\frac{P(R_i, R_{i-1}, M_i)}{P(R_{i-1}, M_i)} = \frac{P(R_{i-1})P(R_i|R_{i-1})P(M_i|R_i, R_{i-1})}{P(R_{i-1})P(M_i)} \tag{4}$$

The decomposition of the numerator is due to the independence of R_{i-1} and M_i as implied by Figure 3. The last term can be simplified further by noticing that M_i stands for the role *already-estimated* for the current speaker, S_i, in his/her most recent turn. Therefore, we can rewrite the last term as

$$P(R_i|R_{i-1}, M_i) = \frac{P(R_i|R_{i-1})P(M_i|R_i)}{P(M_i)}, \tag{5}$$

and using Bayes' rule,

$$P(R_i|R_{i-1}, M_i) = \frac{P(R_i|R_{i-1})P(R_i|M_i)}{P(R_i)}, \tag{6}$$

	Turn-Level	Speaker-Level
Total Number	4,349	226
Host (H)	50%	16%
Guest (G)	47%	67%
Audience Participant (A)	1%	9%
Reporting Journalist (J)	2%	8%

Table 1: Speaker role characteristics used in the experiments.

i.e., a multiplication of two conditional probability distributions (CDFs), $P(R_i|R_{i-1})$ and $P(R_i|M_i)$, with proper normalization. These two CDFs are estimated from the training data as explained next.

3.2. Learning Conditional Probability Distributions

The probability distribution $P(R_i|R_{i-1})$ tells us how often there is a transition from role R_{i-1} to R_i. This corresponds to a bigram language model estimated from available training data. The probability distribution $P(R_i|M_i)$ tells us how often the role R_i should be assigned to the speaker of the i^{th} turn, S_i, given that the most recent role assigned to the same speaker is M_i. It is essentially a *correction mechanism* in which the decoder is biased based on the previous role assigned to the same speaker.

To train these probabilities, we first decode the training data. For this we tried two justifiable approaches. In the first approach, we decoded the training data with the language model of the speaker role transitions only, $P(R_i|R_{i-1})$, which corresponds to decoding with no speaker information constraint. In the second approach, we forced the constraint that R_i has to be equal to M_i. This corresponds to a probability distribution such that $P(R_i|M_i) = 1$ if $R_i = M_i$ and 0 otherwise. Moreover, if it is the first appearance of the speaker in the conversation, $P(R_i|M_i)$ was assumed uniform. Using the decoded output of the training data, we estimated the probabilities $P(R_i|M_i)$ and the current speaker role R_i using the estimated distribution $P(R_i|R_{i-1}, M_i)$.

4. Experiments

We conducted leave-one-show-out cross-validation experiments to assign a speaker role to each turn of a broadcast show using DBNs. The specific implementation of DBNs we used is GMTK[1]. In this section, we report our experimental setup and report our empirical results.

4.1. Data

We focus on data of conversational genre from broadcast conversation (BC) corpora collected under the DARPA GALE program for English. The BC corpora include interactive, spontaneous interactions in news-style TV and radio programs, including talk shows, interviews, call-in programs, live reports, and round tables. The corpora cover a wide range of shows about political, economic, civil, social, and cultural issues. The shows may include anchors, reporters, correspondents, political figures and analysts, writers, and others. The participants may discuss issues in a conversational manner, or the shows may have segments playing speeches or pre-taped recordings, or interviewing participants.

We had 36 shows annotated by human annotators and per-

[1] http://ssli.ee.washington.edu/~bilmes/gmtk

	Turn Acc.	Speaker Acc.
HMM	79.2%	76.6%
Approach1-Iteration 1	89.1%	81.7%
Approach1-Iteration 2	89.5%	81.7%
Approach 2-Iteration 1	86.9%	79.9%

Table 2: Performance of the proposed dynamic Bayesian network structures.

formed forced-alignments for these shows. We removed the sound-bites and transitional segments as our goal is analyzing the live conversational interactions between the speakers. This resulted in a total of 4,349 turns from 266 speakers where the number of speakers per show in the corpus ranged from 3 to 18. We used the LDC-supplied speaker information in constraining the speaker roles as described in the previous section. The statistics of this data is presented in Table 1.

4.2. Experimental Results

The performances of using DBNs with speaking style features as observation vectors are reported in Table 2. We first trained the probability models involved in Figure 1, namely the LM $P(R_i|R_{i-1})$ and 2-component GMMs for the probability of observation vectors given speaker roles, $P(o_i|R_i)$. We then trained the probability models involved in Figure 3. We experimented with both approaches of Section 3.2 to observe their behavior.

As the results in Table 2 show, the simple structure of Figure 1 assigned 79.2% the correct role. To find out what percentage of the speakers were mostly assigned their correct role, we used majority rule to assign a single role to one speaker. It was found that 76.6% of the speakers received the correct role if this approach is taken.

When the network structure of Figure 2 is used, as explained in Section 3.2, there are multiple approaches that can be taken. First, no initial constraining is performed in which the training data is decoded and its decoded speaker role sequence is used to determine the probability of assigning R_i given that the most recent role assigned to the same speaker was M_i, i.e., $P(R_i|M_i)$. The second approach is based on constraining this initial decoding of the training data such that $R_i = M_i$. As the results show, the first approach was able to assign 89.5% of the turns the correct role, whereas the latter approach did 86.9% of the role assignments correctly. When majority rule is used to assign a single role to any given speaker, the first approach as correct 81.7% of the time, while the second approach was correct 79.9% of the time. We explored what happens if we decode the training data once again with this data structure and re-train $P(R_i|M_i)$, but as Table 2 shows, this did not help much. As these results suggest, the first approach is more capable of correcting errors made in assigning roles to the subsequent turns of a speaker. The second approach places a too rigid constraint regarding what is allowed while the first approach is more liberal and this explains the difference in the performance.

We analyzed the output speaker roles assigned by the network structures of Figure 2 and Figure 3 to verify that the latter is capable of correcting role assignments for the same speaker. We observed that the structure of Figure 3 was able to correct many assignments of guest speaker roles. Assigning guest speaker role to a host many times had a major degradation in assigning roles for the following turns. This resulted in improvements not only in guest speaker role assignments but in all role assignments. Moreover, the structure of Figure 3 was particularly successful at assigning the roles J (reporting journalist) and A (audience participants) less frequently.

5. Conclusions and Future Work

In this paper, we explored assigning social roles to speakers in conversations. We used information extracted only from the speaking styles of the speakers and employed DBNs to model how the speakers manage the turn-taking among themselves. We first modeled the problem with an HMM and then embedded a constraint on this structure. The constraint was that the current speaker role depends not only on the previous speaker's role but also on that most recent role assigned to the same speaker. Our experimental results proved that the approach with the constraint on same speaker segments significantly outperformed the HMM-based approach.

There are multiple research directions that we will take as our future work. First, we will explore other types of information that will help constrain the speaker role assignment problem. One such constraint might be derived from the dialog act tags, which signal whether a given utterance is a statement, a question, a back-channel an so on. A second direction that we will take is exploring other types of social role assignments, for instance, how to determine the dominant speakers in different portions of a conversation. As the present paper proves, DBNs help formulate the inter-dependencies between involved random variables in a natural way. As a third direction, we will detect the speaking style features (for instance, disfluencies) automatically and use these. We will first explore the degradation due to automatic detection of speaking-style features and then propose methods to solve related problems. Moreover, we will investigate what happens when automatic speech recognition output is used.

Acknowledgements: The authors thank Dr. Geoff Raymond and Dr. Kristin Precoda for sharing their knowledge on conversation analysis and linguistics. This work has been supported by the Intelligence Advanced Research Projects Activity (IARPA) via Army Research Laboratory (ARL) contract number W911NF-09-C-0089. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, ARL, or the U.S. Government.

6. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, 1974.

[2] A. Vinciarelli, "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," *IEEE Trans. Multimedia*, vol. 9, 2007.

[3] N. Garg, S. Favre, H. Salamin, D. Hakkani-Tur, and A. Vinciarelli, "Role recognition for meeting participants: an approach based on lexical information and social network analysis," in *in Proc. of ACM Multimedia Conference*, 2008.

[4] B. Hutchinson, B. Zhang, and M. Ostendorf, "Unsupervised broadcast conversation speaker role labeling," in *in Proc. of ICASSP*, 2010.

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