AUTOMATIC SOCIAL ROLE RECOGNITION IN PROFESSIONAL MEETINGS

A. Sapru Hervé Bourlard

Idiap-RR-35-2012

DECEMBER 2012

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

AUTOMATIC SOCIAL ROLE RECOGNITION IN PROFESSIONAL MEETINGS

Ashtosh Sapru[1,2] and Herve Bourlard[1,2]

[1] Idiap Research Institute, 1920, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
ashtosh.sapru@idiap.ch,herve.bourlard@idiap.ch

ABSTRACT

This paper investigates the influence of social roles on the conversation style and linguistic usage of participants in professional meeting recordings. At first, we implement a generative model to capture the sequential nature of conversations in terms of participants, turn-taking behavior. In parallel, the system also employs a probabilistic discriminative classifier on a set of high level features. The final step involves combining evidence from both generative and discriminate models. Experiments suggest that both generative and discriminative models can reach a recognition accuracy of 65% in classifying four social roles. Moreover, the recognition accuracy increases to 69% when information from both models is taken into consideration.

***Index Terms*—** Social Role Labeling, Turn Taking, Linguistic, Lexical and Prosody.

1. INTRODUCTION

Analyzing spoken documents in terms of speaker role information is useful for enriching the content description of multimedia data. It can be used in applications like information retrieval, enhancing multimedia content browsing and allowing summarization of multimedia documents [1]. Speaker roles are stable behavioral patterns in an audio recording and the problem of role recognition consists in assigning a label, i.e., a role to each of the speakers. Automatic labeling of speaker roles has been widely studied in case of Broadcast News (BN) recordings. These roles are imposed from the news format and relate to the task each participant performs in the conversation like anchorman, journalists, interviewees, etc. In the last few years automatic role recognition has also been investigated for meeting recordings and broadcast conversations. Typical roles in these studies can vary with environment and applications such as Project Manager in AMI corpus [2], student, faculty member in ICSI corpus [3]. Most of the research has focused on use supervised methods (both generative or discriminative classifiers) though some works [4] have also explored unsupervised methods. Common features used in these studies extract relevant information from conversation features, lexical features, prosody and Dialog act tags [5, 6, 4].

For the studies mentioned above participants role is formal and considered to remain constant over the duration of entire audio recording. Other role coding schemes have also been proposed in literature which put roles in a more dynamic setting, such as socio-emotional roles (here after referred to as social roles) [7, 8, 9, 10]. Social roles describe relation between conversation participants and their roles "*oriented towards functioning of group as a group*". Social roles are useful to characterize the dynamics of the conversation, i.e., the interaction between the participants and can be generalized across any type of conversation. They are also related to phenomena studied in meetings like social dominance, engagement and also hot-spots [11]. Besides these social roles can also provide cues about state of meeting. Meeting segments where participants take more active roles are likely to have richer information flow compared to segments where participants only take passive roles.

Previous approaches to automatic social role recognition can be broadly classified into discriminative and generative approaches. Common format in many of these studies is to predict social role within a segment of recording, where the role of each participant is assumed to stay constant. Among the discriminative approaches, automatic social role recognition was first investigated in [7]. They used a support vector machine classifier to discriminate between social roles in meetings recorded for problem solving sessions. The feature set used in this work was extracted from audio and video activity states within a segment of meeting recording. Other studies have also investigated social role recognition in professional meetings (AMI corpus). The research in [12] revealed that automatically extracted subjectivity features from lexical and prosodic cues are correlated with social roles. They also report that combining speech activity features with subjectivity features can improve recognition over using activity features alone. More recently, in [10] a multi-class boosting classifier was used to integrate evidence from several information streams i.e. speech activity, dialog act tags, lexical and prosody for social role recognition. Investigations in this work also highlighted that some social roles are more correlated with lexical content and dialog act tags. The generative models for social role recognition were considered in [9, 8]. While the work in [8] used speech activity and video features, in [9] the generative framework was used to combine prosody and turn duration. These works also investigated the influence of other participants on the distribution of these features.

In this paper, we propose a novel method, which combines the strength of both discriminative and generative models which have so far been used separately. The generative framework is used to develop a model for analyzing conversations in terms of turn taking sequences while the discriminative model is trained on a novel set of high level features. The generative approach models conversation sequences as outputs of a Markov random process. It investigates effect of social roles on long pauses and overlaps in turn taking style. In comparison the study in [9] captures speaker change information in turn taking. The conversational model framework in this paper is similar to research in [13]. However, while they investigate relation of correlates such as education, gender etc. on turn taking styles in two person dialogs, the proposed work considers the influence of social roles on turn taking patterns in multiparty conversations. In addition to modeling conversation styles, the present work also investigates correlation between social roles and a novel set of high level features which includes prosodic, structural and linguistic information extracted over a segment in meeting. While structural features have been shown to be informative for social role recognition [7, 10], we hypothesize that social role also affects a participant's linguistic usage patterns. A discriminative model is trained to capture role information from the set of high level features. In summary the three main contributions of this paper are: a generative role model which captures the likelihood of participant conversational style; a discriminative model that estimates posterior role distribution from a set of high level features; and finally combination of likelihood and posterior values to predict the participant's social role.

2. GENERATIVE MODEL OF TURN TAKING

Audio from the independent headset microphones (IHM) is processed through a speech segmentation system [14] for obtaining estimated speech/non-speech boundaries for each meeting participant. The output of speech/non speech system for each speaker is a sequence of speech and silence regions in time which arise due to turn taking in conversations. However, since meeting conversations involve multiple speakers, some activity regions (speech overlaps) will have more than one participant speaking simultaneously. Also silence regions corresponding to each participant can take multiple meanings. Silence due to pause in conversation, when conversation floor changes occur or speakers pause to take breathe. On the other hand silence regions can simply be the listening silence from the perspective of some speakers when other speaker(s) is/are speaking. We hypothesize that each participant's turn taking pattern is related to its social role. For example, it is more likely that a participant with a more active role will grab the conversation floor after a pause. Similarly, the participants role is expected to affect whether it keeps control of conversation after a speech overlap or not.

Assuming that speech activity can be reliably estimated, we partition each participant's conversation sequence into four states: talkspurts (TS) i.e., a region of speech when only a single speaker speaks, pauses (PA) ,i.e., regions when all the speakers are silent, overlaps (OV), i.e., regions where multiple speakers are speaking simultaneously and listening silence (LS), i.e., regions from perspective of current participant when some other speaker is speaking. Each of these regions is smoothed using a minimum duration criterion. Furthermore we also extract prosodic and lexical features aligned with TS and OV regions. Prosodic feature vector is represented using measures like mean F0 frequency and slope, mean energy and speech rate. Lexical features are words corresponding to speaker utterances including backchannels.

More formally we consider a participant S in meeting segment k which takes a role R. The complete turn taking record for S in k can be summarized as $TT^S = \{(q_1, d_1, X_1^p, X_1^l), ..., (q_N, d_N, X_N^p, X_N^l)\}$. Here N is the number of activity states in the conversation sequence for S in k, q_t represents the state at instant t, d_t, X_t^p and X_t^l are the duration, prosodic and lexical features associated with q_t. X_t^p and X_t^l take a null value when $q_t \in \{PA, LS\}$. Conditioned on its role R a participants turn taking TT^S is modeled as a first order Markov random process. The likelihood model is given as,

$$p(TT^S|R) = \prod_t p(q_t|q_{t-1}, R)p(d_t|q_t, R)p(X_t^p|q_t, R)p(X_t^l|q_t, R) \tag{1}$$

The term $P(d_t|q_t, R)$ represents the duration distribution of each state and was modeled as Gamma distribution similar to [13]. The parameters of this distribution were trained using maximum likelihood for all states q_t where S assumes role R. Similarly to [9] the conditional distribution $p(X_t^p|q_t, R)$ was represented using a Gaussian mixture model (GMM) with number of mixture components fixed to four. The prosodic features were speaker normalized prior to their modeling. Standard EM was used for estimating the parameters of prosodic distribution. For each social role we also trained a bigram language model (LM) to represent the lexical distribution $p(X_t^l|q_t, R)$, using the word utterances time aligned to the states q_t in which participant S assumes the role R.

3. DISCRIMINATIVE MODEL FOR ROLE RECOGNITION

While the previous section described feature modeling at the level of talkspurts, role information is also correlated with features which represent the aggregate statistics (X_d) for the participant over the entire length of meeting segment. For each meeting participant a set of features were extracted.

Structural/Conversational features : The total speech time, difference between end of last talk spurt and start of first talk spurt, the total number of turns in the segment per speaker as well as number of times a speaker overlaps and total duration of overlap and number of speakers with which overlap occurs. Also included were features like maximum, minimum and mean and standard deviation for all talk spurts and pauses.

Linguistic features : To extract linguistic information from word usage we used word categories that have well defined meaning in social psychology studies [15]. While speaker personality and linguistic cues share a strong correlation [15], to the best of our knowledge, this is first investigation that explores the effect of social roles on the linguistic style of each participant. For this purpose a handcrafted dictionary is employed which studies word usage along emotional, cognitive, functional, and process dimensions.

Prosodic and Spectral features : Low level descriptors like fundamental frequency (F0) and intensity and MFCC were computed from the headset microphones. Three families of functionals representing the extremum, higher order statistics (standard deviation, kurtosis, skewness) and linear regression coefficients (slope and prediction error) were computed from these low level features.

If we represent X_d^S as the above feature set for participant S, then using a probabilistic discriminative classifier we can estimate probability $P(R|X_d^S)$ when S assumes role R. The discriminative model implemented in this work is based on Maximum entropy classification (Maxent). Maxent estimates the conditional probability of role given data as

$$P(R|X_d^S) = \frac{1}{Z_\alpha(X_d^S)} exp(\sum_i \alpha_i X_d^S(i)) \tag{2}$$

where $Z_\alpha(X_d^S)$ is the normalization term. The model weights α_i are obtained by maximizing the conditional entropy consistent with information in training data.

4. AUTOMATIC ROLE RECOGNITION

The problem of automatic role recognition is defined as assigning a role $\hat{R}$ to a conversation participant S. We assume $\hat{R}$ comes from a finite set of roles $\mathbf{R}$ and is the best possible assignment which explains the participants behavioral patterns. The sequential turn taking patterns are modeled as Markov random process while discriminative classifier captures information from high level structural, prosodic and linguistic usage. Both of these models compute an estimate of participant S taking a role $R \in \mathbf{R}$. Information fusion from the two models can be approached by defining an objective function based on the convex combination of their log values.

$$\mathcal{L}_\lambda(R, X_d^S, TT^S) = \lambda log(p(TT^S|R)) + (1-\lambda)log(P(R|X_d^S)) \tag{3}$$

where $\lambda \in [0,1]$. Using Equation 3 the predicted role is give as,

$$\hat{R} = \underset{R \in \mathbf{R}}{\arg\max} \mathcal{L}_\lambda(R, X_d^S, TT^S) \tag{4}$$

The minimum values of $\lambda = 0$ results in predicted role estimated using only discriminative approach while $\lambda = 1$ uses only generative model. An appropriate value of λ can be selected using cross-validation.

5. DATA AND ANNOTATION

The AMI Meeting Corpus is a collection of meetings captured in specially instrumented meeting rooms, which record the audio and video for each meeting participant. The corpus contains both scenario and non-scenario meetings. In the scenario meetings, four participants play the role of a design team composed of *Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI), and Industrial Designer (ID)* tasked with designing a new remote control. The meeting is supervised by the PM who follows an agenda with a number of items to be discussed with other speakers. A subset of 59 meetings containing 128 different speakers (84 male and 44 female participants) is selected from the entire corpus. Subsequently each meeting was segmented into short clips (with a minimum duration of 20 seconds) based on presence of long pauses i.e. pauses longer than 1 second. Within each such meeting segment social role of the participant is assumed to remain constant. From each meeting a total duration of approximately 12 minutes long audio/video data was selected. Meeting segments are resampled so as to cover the entire length of recording comprising various parts of meeting such as openings, presentation, discussion and conclusions.

Since social roles are subjective labels and require human annotators, the annotation scheme was implemented as follows. The video for each meeting segment was obtained by merging the four speaker specific closeup cameras and an overview camera with the audio from individual headset microphones that each speaker wears. Each annotator is asked to view and listen the entire video segment and tasked with assigning a speaker to role mapping based on a list of specified guidelines. These guidelines define a set of acts and behaviors that characterize each social role and is summarized in the following: *Protagonist* - a speaker that takes the floor, drives the conversation, asserts its authority and assume a personal perspective; *Supporter* - a speaker that shows a cooperative attitude demonstrating attention and acceptance providing technical and relational support; *Neutral* - a speaker that passively accepts others ideas; *Gatekeeper* - a speaker that acts like group moderator, mediates and encourage the communication; *Attacker* - a speaker who deflates the status of others, express disapproval and attacks other speakers. Figure 1 shows the distribution of roles over all the meeting segments present in the data set. As can be seen the Neutral role has been labeled most often by annotators. This is followed by Supporter, Gatekeeper and Protagonist. Comparatively the Attacker role has received the fewest labels as observed by multiple annotators. A reason for this distribution may be due to collaborative nature of AMI meetings. The reliability of labeling scheme as measured through Fliess's kappa shows a value 0.5 which is considered to have moderate agreement according to Landis and Koch's criterion [7].

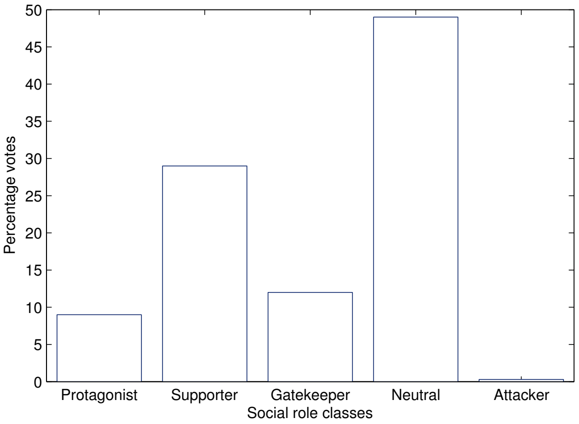

Fig. 1. Social role distribution in the annotated corpus. The vertical axis represents percentage votes for each class as labeled by multiple annotators.

6. EXPERIMENTS

For evaluation of proposed method experiments were conducted using repeated cross-validation wherein one set of meetings (all but two) was kept for training/tuning the model parameters while a distinct set (remaining two meetings) was used for evaluation. The partition of meetings was done keeping in view that participant with same speaker identity does not appear in both training and test set. The ground truth for participant role labels was derived by majority voting. An initial filtering was done to consider only those meeting segments were a participant is active, also a few meeting segments, were majority voting resulted in participant having an attacker role label were not considered(see Figure 1).

During training the parameters of model were estimated using maximum likelihood estimation. The transition probabilities in the model were computed as normalized counts. Standard EM was used for estimating the parameters of prosodic distribution. All the lexical information was extracted using output of AMI-ASR system [16]. The LM training for lexical features was implemented using SRILM toolkit [17]. The linguistic features used in discriminative model are extracted from word categories defined in Linguistic Inquiry and Word Count software (LIWC) [18]. The dictionary is composed of 4500 words and groups them into 64 different overlapping categories such as pronouns, activity and functional words and words for positive and negative emotions etc. Gaussian smoothing was used to avoid overfitting in Maxent model. The experiments for the discriminative classifier implemented in this work are based on existing toolkit [19]. The tuning parameters were selected by evaluations on a randomly sampled portion of training data. All the models were evaluated on a separated test set and performance measured in terms of recognition accuracy and F-measure/Precision/Recall.

Figure 2 reports the feature wise performance of both the generative and discriminative models. It can be seen that all the individual feature streams perform better than chance level 0.25. Modeling conversational turntaking pattern achieves the highest macro F-measure 0.55. Interestingly the next best numbers are achieved by linguistic features used in discriminative model 0.54. This confirms our two main hypotheses that social roles influence both the conversation styles of meeting participants as well as their high level linguistic usage. For turntaking patterns model statistics reveal that whenever a participant acts as protagonist it is most likely to speak

Table 1. Per role F-measure, Precision and Recalls obtained in recognizing social roles for the three considered models.

	Per-role F-measure (Precision/Recall)				Accuracy
Model	Protagonist	Supporter	Gatekeeper	Neutral	
Generative	0.58 (0.57/0.58)	0.74 (0.81/0.68)	0.56 (0.56/0.56)	0.58 (0.46/0.79)	0.65
Discriminative	0.48 (0.46/0.52)	0.74 (0.77/0.70)	0.51 (0.50/0.52)	0.71 (0.68/0.75)	0.64
Combination	0.58 (0.55/0.61)	0.77 (0.83/0.72)	0.59 (0.57/0.61)	0.71 (0.64/0.80)	0.69

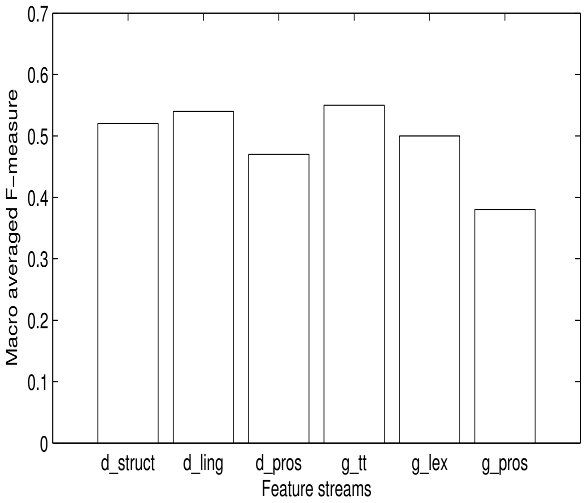

Fig. 2. Performance of various feature streams used in generative and discriminative models. Discriminative model: d_struct (structural), d_ling (linguistic), d_pros (prosody/spectral). Generative model: g_tt (turntaking and duration), g_lex (lexical) , g_pros (prosody).

immediately after a pause or overlap region, while negative is true in case of neutral and supporter roles. Protagonist and gatekeepers are also more likely to start or end a conversational segment and produce on average longer talkspurts compared to supporters and neutrals. Performance of prosody related features is different in generative and discriminative models. This can be due to the fact that features representation in discriminative model captures extreme values and higher order statistics of these features while only average statistics were considered in the generative model. Figure 3 compares the performance of best performing features against the case when all features were combined in each model. It can be seen that for both models there is an improvement in performance over all social roles when all the features are used. This reveals that various features capture different aspects of role related information.

Table 1 compares the performance of combination model against individual models, both generative and discriminative. Also reported are performance figures for different roles. It can be seen that combining evidence from both generative and discriminative approaches achieves a superior performance 69% compared to their standalone performance. Also table numbers reveal that the two models vary in their performance amongst individual roles. Generative model performs better in recognizing protagonist (0.58) and gatekeepers (0.56) compared to discriminative model. In comparison participants with neutral role have a better chance of being recognized by discriminative model. Both models perform equally for supporter role as measured in terms of F-measure (0.74). Interestingly combining the two models improves precision over all roles.

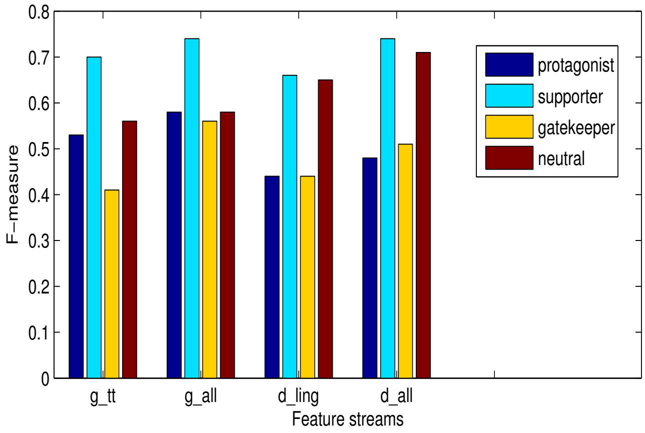

Fig. 3. Comparison of best performing features against multistream features. Discriminative model: d_ling (linguistic), d_all (all features). Generative model: g_tt (turntaking and duration), g_all (all features).

7. CONCLUSION

Results are consistent with the initial hypothesis that participants social roles influences their conversation style. Turn taking patterns in conversation where found to be most informative features. Furthermore, by integrating lexical and prosodic cues in conversational model an overall improvement in performance over all social roles was reported, reaching an accuracy of 65%. A parallel system where a set of high level features are extracted from the entire segment and modeled using a discriminative approach reaches a similar accuracy 64%. Combination model based on information fusion from the two systems reaches an accuracy of 69% significantly higher than standalone generative and discriminative approaches. In summary proposed approach leads us to conclude that recognizing social roles requires extracting meaningful information at different layers of data. In future we plan to extend our model to integrate dependencies between social roles of multiple participants. Furthermore we also plan to extend our study on other meeting environments.

8. ACKNOWLEDGEMENT

This work was funded by the Hasler Stiftung under SESAME grant, the EU NoE SSPNet, and the Swiss National Science Foundation NCCR IM2.

9. REFERENCES

[1] Gabriel Murray, Pei-Yun Hsueh, Simon Tucker, Jonathan Kilgour, Jean Carletta, Johanna Moore, and Steve Renals, “Automatic Segmentation and Summarization of Meeting Speech,” *Proceedings of Human Language Technologies: The Annual*

Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2007.

[2] Salamin H. et al., "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," *IEEE Transactions on Multimedia*, vol. 11, November 2009.

[3] Laskowski K., Ostendorf M., and Schultz T., "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008.

[4] Damnati G. and Charlet D., " Robust speaker turn role labeling of TV Broadcast News shows," *proceedings of ICASSP*, 2011.

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

[6] Wang W., Yaman S., Precoda P., and Richey C., "Automatic identification of speaker role and agreement/disagreement in b roadcast conversation.," in *Proceedings of ICASSP*, 2011.

[7] Zancaro M. et al., "Automatic detection of group functional roles in face to face interactions," *Proceedings of ICMI*, 2006.

[8] Dong W. et al., "Using the influence model to recognize functional roles in meetings," *Proceedings of ICMI*, 2007.

[9] Valente F. and Vinciarelli A., "Language-independent socio-emotional role recognition in the ami corpus," in *Proceedings of Interspeech*, 2011.

[10] Sapru A. and Valente F., "Automatic speaker role labeling in AMI meetings: recognition of formal and social roles," *Proceedings of Icassp*, 2012.

[11] Wrede D. and Shriberg E., "Spotting "hotspots" in meetings: Human judgments and prosodic cues," *Proc. Eurospeech*, 2003.

[12] Wilson T. et. al., "Using linguistic and vocal expressiveness in social role recognition," *Proceedings of the Conference on Intelligent User Interfaces(IUI)*, 2011.

[13] Grothendieck J et al., "Social correlates of turn-taking behavior.," *Proceedings of ICASSP*, 2010.

[14] Hain, Vepa J., and J. T.Dines, "The segmentation of multi-channel meeting recordings for automatic speech recognition," *Proceedings of Interspeech*, 2006.

[15] M. R. Mehl J. W. Pennebaker and K. G. Niederhoffer., "Psychological aspects of natural language use: Our words, our selves," *Annual Review of Psychology*, 2003.

[16] Hain T., Wan V., Burget L., Karafiat M., J. Dines, Vepa J., Garau G., and Lincoln M., "The AMI System for the Transcription of Speech in Meetings.," *Proceedings of Icassp*, 2007.

[17] A. Stolcke., "Srilm an extensible language modeling toolkit.," *Proc. of ICSLP*, 2002.

[18] Liwc inc. http://www.liwc.net/index.php, ," .

[19] Andrew Kachites. McCallum, "Mallet: A machine learning for language toolkit.," *http://mallet.cs.umass.edu*, 2002.