Social Role annotation for AMI Corpus meetings

Ashtosh Sapru

I. Corpus Description

For the task of annotating social roles, we selected data from AMI meeting corpus [1]. AMI 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.

The formal roles in AMI meetings are scripted and participants know beforehand the overall agenda of the meeting. Each speaker assumes only one formal role that remains fixed for the entire duration of the meeting. Besides formal roles, the speakers also assume informal roles. Informal roles assumed by speakers are influenced by their individual traits, such as personality and interaction with other group members. While the personality of a speaker remains relatively stable across different scenarios, the emergent social roles develop in response to changing dynamics of group interaction. As the meeting progresses different role configurations can emerge and social role of a speaker can change from one type to another.

In order to classify speakers behavior into distinct emergent roles we follow the role coding scheme proposed in [2]. The underlying motivation behind this approach is that, while same speaker can assume different social roles, its role remains relatively stable over short time windows. Therefore, at each time instant a speaker will have a unique social role which can be defined using a set of acts and behaviors. The attributes of different roles are briefly 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 assumes a cooperative attitude, demonstrates attention and acceptance and provides technical and relational support.
- *Neutral* - a speaker that passively accepts ideas from other group members.
- *Gatekeeper* - a speaker that acts like group moderator, mediates and encourages the communication within the group.
- *Attacker* - a speaker who deflates the status of others, expresses disapproval and attacks other speakers.

For the present study a subset of 59 scenario meetings containing 128 different speakers (84 male and 44 female participants) was selected from the corpus. Subsequently each meeting was sliced into short clips (average duration less than 30 seconds). In each slice of meeting, the social role of a speaker was assumed to remain constant. Allocating social roles for short time meeting slices is supported by earlier work. In [3] manual annotations of social roles were smoothed over a one minute long sliding window for training of role recognition models. Furthermore, predicting speaker characteristics over short video clips, referred to as, “thin slices of behavior”, is very well documented in social psychology literature [4]. Considering the nature of social role annotation over meeting recordings, this is particularly advantageous since annotators can work on short video slices and need not wait for the entire meeting recording to complete.

From each meeting, a total duration of approximately 12 minutes long audio/video data was selected. Meeting slices were resampled so as to cover the entire length of recording comprising various parts of meeting such as openings, presentation, discussion and conclusions. Using this approach, we generated 1700 meeting slices, corresponding to almost 12.5 hours of meeting data.

A. Crowdsourcing

In this work, we have used an online environment for social role annotation and the human assessors were selected through the crowdsourcing platform, Amazon mechanical turk (AMT). The online platform allows raters

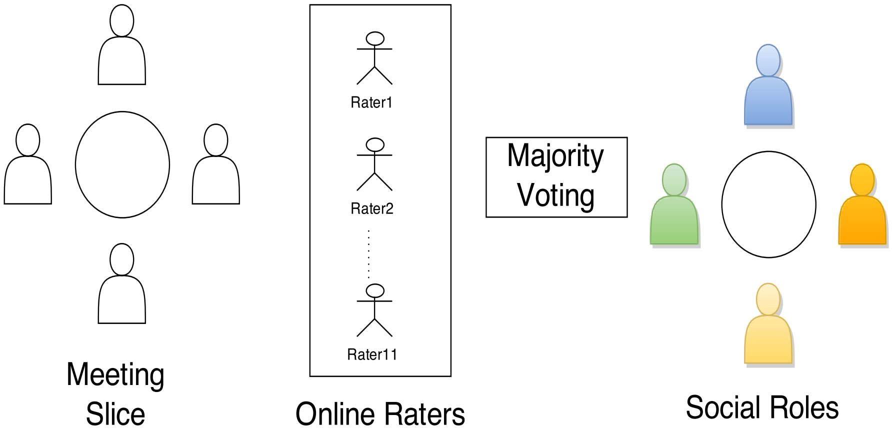

Fig. 1. Ground truth social roles using majority voting.

to work on Human Intelligence Task's (HIT's), where they have an option to accept or reject a HIT, and are paid a small amount of money in exchange for providing annotations. The HIT requester can select raters using a set of inbuilt rater qualifications, including raters location and their HIT approval rate, i.e, the fraction of completed tasks that were accepted by other HIT requesters in the past. The requester can also specify the number of unique annotations for a set of HITs as well as reward payment for each HIT. All the completed annotations can be downloaded and reviewed by the requester who also has the option to reject any HIT which does not meet the requisite quality.

For the task of social role annotation we prespecified the inbuilt rater qualifications, i.e., location of raters and their HIT approval rate. As the meetings are in English, we decided to set the location of raters to United states (US), where most people speak English as their first language. Since a large proportion of AMT raters are based in US, this requirement was not considered to adversely effect the quality of annotations. For the second qualification we decided to use raters whose HIT approval rate exceeds 95%.

Before starting each HIT, the raters were asked to follow a set of annotation guidelines. First, annotators were told that each HIT is a sequence of presentations and discussions according to a predefined meeting agenda. Second, attributes of all the five social roles were described. Third, annotators were asked to watch each clip individually and judgments should be based on behavior of participants with the clip, with focus on their interaction and what participants say and how they say it. Fourth, more than one participant can take the same role. Fifth, participants who are silent during a clip should be perceived as neutrals. Along with the annotation guidelines, the HIT also incorporates the video clips which the raters need to view before submitting their judgments. The video clip for each meeting slice 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.

To facilitate the annotation process, we grouped together the video clips from a single meeting in one HIT. Pilot studies revealed that a very large number of video clips in a HIT increases the task submission time. As a compromise about 10-11 meeting slices were grouped in a HIT. Annotators were provided with audio and video for each meeting and tasked with assigning a speaker to role mapping for each meeting participant appearing in the clip. We asked 11 annotators to rate each HIT. An analysis of completed annotations revealed that a majority of accepted HITs (70%) were completed by 10 or more than 10 raters and 95% of HITs were completed by 8 or more than 8 raters. Only HITs completed by 5 or more than 5 raters were used for further analysis.

B. Ground truth annotation

The ground truth social role label for each clip was derived by taking a majority vote over rater assignments. Figure 1 shows the schematic for deriving the ground truth social role annotation.

The ground truth annotation can be explained by considering the case of a single meeting. Let us consider the AMI meeting 'IS1004b' recorded at Idiap. A HIT is composed of video clips (thin slices) from this meeting. The raters view the videos from this this meeting, one clip at a time. For example, consider a video clip in this meeting which starts at 181 seconds and ends at 212 seconds. After the raters have seen the entire clip of 33 seconds duration, they rate the behavior of each participant that appears in the video. There are four participants in meeting 'IS1004b', indexed using labels 'A','B','C' and 'D'. A rater labels each participant with a unique social role (out of 5 possible). For example, 11 raters viewed the video of clip 'IS1004b_181_212' (clip in meeting 'IS1004b' starts at 181 seconds and ends at 212 seconds). The ratings assigned by 11 raters for this clip are shown in Table I. Out of 11 raters, 8 assign the social role of 'Gatekeeper' to participant 'A' for the duration of this clip. Taking the majority vote over raters assigns gatekeeper as the ground truth social role of 'A' for the duration of this clip. The ground truth social roles for all the participants is shown in Table II

TABLE I
NUMBER OF VOTES ASSIGNED TO EACH SOCIAL ROLE BY 11 RATERS.

Clip	Participant	Protagonist	Supporter	Gatekeeper	Neutral	Attacker
IS1004b_181_212	A	3	0	8	0	0
IS1004b_181_212	B	2	9	0	0	0
IS1004b_181_212	C	0	0	0	11	0
IS1004b_181_212	D	1	10	0	0	0

TABLE II
GROUND TRUTH SOCIAL ROLES.

Clip (IS1004b_181_212)	A	B	C	D
Social Roles	Gatekeeper	Supporter	Neutral	Supporter

The ground truth social roles were obtained for the selected video clips corresponding to meetings in AMI corpus. In the annotated database, we use six tags (see Table III) to determine the social roles. Since social roles of a participant change in the meeting

- MeetingID: Meeting identifier in the AMI Corpus (see 'http://groups.inf.ed.ac.uk/ami/corpus/meetingids.shtml' for details) .
- ClipStartingTime: Starting time of the clip in seconds (measured as time elapsed since the beginning of meeting)
- ClipEndingTime: Ending time of the clip in seconds (measured as time elapsed since the beginning of meeting)
- ClipID: A thin slice of meeting (<MeetingID_ClipStartingTime_ClipEndingTime>)
- ParticipantID: Unique label for each participant (e.g. 'A','B','C','D', see 'Agent' description in AMICorpus for details)
- SocialRole: Ground truth social role obtained by majority voting.

TABLE III
SOCIAL ROLE (AMICORPUS) ANNOTATION FORMAT.

MeetingID	ClipID	ClipStartingTime	ClipEndingTime	ParticipantID	SocialRole

The social role annotations were used to automatically train a supervised classifier. The details are described in [5].

REFERENCES

[1] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, pp. 181–190, 2007.

[2] M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic detection of group functional roles in face to face interactions.," in *ICMI*, Francis K. H. Quek, Jie Yang, Dominic W. Massaro, Abeer A. Alwan, and Timothy J. Hazen, Eds. 2006, pp. 28–34, ACM.

[3] F. Valente and A. Vinciarelli, "Language-Independent Socio-Emotional Role Recognition in the AMI Meetings Corpus," *Proceedings of Interspeech*, 2011.

[4] N. Ambady and R. Rosenthal, "Thin Slices of Expressive behavior as Predictors of Interpersonal Consequences : a Meta-Analysis ," *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.

[5] A. Sapru and H. Bourlard, "Automatic recognition of emergent social roles in small group interactions," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 746–760, May 2015.