IDIAP RESEARCH REPORT

Role Recognition for Meeting Participants: an Approach Based on Lexical Information and Social Network Analysis

Neha P. Garg [a] [b] Sarah Favre [a] [c]
Hugues Salamin [a] [c] Dilek Hakkani Tür [b]
Alessandro Vinciarelli [a] [c]

IDIAP–RR 08-57

July 2008

to appear in
Proceedings of ACM International Conference on Multimedia (2008)

[a] Ecole Polytechnique Federale de Lausanne - 1015 Lausanne, Switzerland
[b] International Computer Science Institute - 1947 Center Street, Berkeley CA 94074, USA
[c] Idiap Research Institute - CP592, 1920 Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-57

Role Recognition for Meeting Participants: an Approach Based on Lexical Information and Social Network Analysis

Neha P. Garg   Sarah Favre   Hugues Salamin   Dilek Hakkani Tür
Alessandro Vinciarelli

July 2008

to appear in
Proceedings of ACM International Conference on Multimedia (2008)

Abstract. This paper presents experiments on the automatic recognition of roles in meetings. The proposed approach combines two sources of information: the lexical choices made by people playing different roles on one hand, and the Social Networks describing the interactions between the meeting participants on the other hand. Both sources lead to role recognition results significantly higher than chance when used separately, but the best results are obtained with their combination. Preliminary experiments obtained over a corpus of 138 meeting recordings (over 45 hours of material) show that around 70% of the time is labeled correctly in terms of role.

1 Introduction

One of the main tenets of sociology is that people involved in social interactions play roles: "*People do not interact with one another as anonymous beings. They come together in the context of specific environments and with specific purposes. Their interactions involve behaviors associated with defined statuses and particular roles. These statuses and roles help to pattern our social interactions and provide predictability*" [8]. This paper proposes an approach for the automatic recognition of roles in multiparty recordings based on two behavioural cues: the first is the *lexical choice*, i.e. the use of certain words rather than others in the interventions of each individual. The second is the *interaction pattern*, i.e. the tendency of each individual to interact with certain persons rather than others.

An overall scheme of the approach is depicted in Figure 1: the first step is the application of a speaker diarization approach that identifies the time intervals where each speaker talks. The subsequent steps follow two parallel paths corresponding to the two behavioural cues mentioned above. The right path describes the modeling of the lexical choice and it includes two stages: extraction of the lexical features from the automatic speech transcriptions, and mapping of the lexical features into roles using the BoosTexter text categorization approach [7]. The left path corresponds to the interaction pattern modeling and it includes two stages as well: extraction of a Social Affiliation Network [10] representing social interactions, and assignment of roles to people using a Bernoulli distribution [3]. The main advantage of the behavioural cues is that they are, to a large extent, identity-independent. This enables one to address the general case where an individual plays different roles in different circumstances (as it actually happens in the data used in this work).

To the best of our knowledge, only few works have been dedicated to the automatic recognition of roles. Some of them recognize *functional* roles in broadcast data [2][9], i.e. the tasks that different people perform in television and radio programs (e.g. *anchorman* or *guest*), and another recognizes *functional* roles in movies [11] (e.g. *hero* or *hero's friends*). The recognition is based on lexical features like the n-gram distribution in [2], and on Social Network Analysis [10] in [9][11]. Other works recognize the *social roles* of meeting participants [12] (e.g. *attacker* or *supporter*) using features like the overall amount of movement and speech energy, or the roles corresponding to specific actions [1] (e.g. *presentation* and *briefings*) using the total speaking time of each person and turn-taking statistics.

The main novelty of this work is the combination of approaches based on both lexical features and social networks that so far have been applied only separately (see above). This is expected to make the recognition approach more robust with respect to the two major sources of *noise* in the experiments, i.e. the errors of the Automatic Speech Recognition (ASR) system used to transcribe the recordings, and the errors of the speaker diarization approach used to segment the data into single speaker intervals. The experiments of this work are performed over the AMI corpus [6], a collection of 138 meetings with a total duration of 45 hours and 38 minutes. Each meeting involves four participants playing different predefined roles (see Section 3.1).

The results show that, on average, roughly 70% of the meetings time is labeled correctly in terms of role. The accuracy is higher for the roles associated to well defined and stable behavioural patterns, while it is lower for the roles that do not exhibit predictable behaviours. However, the performance of the system is significantly higher than a random guess for all roles. The combination of the two approaches described above slightly improves the performance of the best role recognizer (based on the lexical choice). However, the improvement appears to be significant for the roles most represented in terms of time. The overall approach seems to be more robust to the errors of the speaker diarization step than to the speech recognition errors. Speakers role can enhance browsers (users can access specific data segments based on role), summarization systems (segments corresponding to certain roles can be retained in the summary as more representative of the content than others), thematic segmentation approaches (specific roles are often related to specific topics), etc.

The rest of the paper is organized as follows: Section 2 describes the approach proposed in this work, Section 3 presents experiments and results, and Section 4 draws some conclusions.

2 The approach

This section describes the recognition approach based on the lexical features (right path of Figure 1), the one based on Social Network Analysis (left path of Figure 1), and the combination approach. For space limitations, no details are given about speaker diarization and Automatic Speech Recognition approaches applied in this work (see [4] and [5], respectively, for a full description). The diarization accuracy (percentage of data time correctly labeled in terms of speaker) is 97.0%, while the Word Error Rate is between 35 and 40% depending on the specific recording of the corpus used for the experiments (see Section 3.1).

2.1 Lexicon Based Role Recognition

The role recognition approach based on lexical features recognizes the roles of speakers using the lexical content of their utterances. The intuition here is that the meeting structure and content are correlated with the roles of its participants, and lexical cues related to structure and topics can be useful for determining speaker roles. For example, the person leading the discussion can use phrases to return to aimed discussion, when a topic shift to an unrelated topic occurs. Also, due to his/her functional role, a speaker may only talk about certain related topics.

We model speaker role detection as a multi-class classification task, where there is one class for each speaker role, and the goal is to assign a role to a speaker in every meeting. Note that, sometimes, a speaker can play different roles in different meetings, but the role is constant in a single meeting. For classification, we use BoosTexter, a multi-class classification tool. Boosting aims to combine *weak* base classifiers to come up with a *strong* classifier [7]. This is an iterative algorithm, where at each iteration, a weak classifier is learned so as to minimize the training classification error. The algorithm begins by initializing an uniform distribution, $D_1(i, r)$, over training examples, i, and labels (i.e., speaker roles), r. After each round this distribution is updated so that the example-class combinations which are easier to classify (e.g. the examples that are classified correctly with the weak learners learned so far) get lower weights and vice versa. The intended effect is to force the algorithm to concentrate on examples and labels that will improve the most the classification rule. To represent every example i (i.e. every meeting participant in the training corpus), we use word n-grams ($n = 1, 2,$ and 3) from all the turns of a speaker in a meeting as features.

The weak classifiers check the presence or absence of word n-grams in the speaker's turns, and can therefore be used for analysis purposes. The final strong classifier is a linear combination of the individual weak classifiers. We use a held-out data set to compute the optimum number of iterations for the classifier. The classifier outputs a probability for the presence of each class for each speaker.

If $\vec{d_i}$ is the vector representing the transcription of the interventions of meeting participant i, then the BoosTexter approach estimates the probability $p(\vec{d_i} \,|\, r)$ of the participant playing role r by combining the weak classifiers described above. The participant i is assigned the role r^* that satisfies the following expression:

$$r^* = \arg\max_{r \in \mathcal{R}} p(\vec{d_i} \,|\, r), \tag{1}$$

where $\mathcal{R}$ is the set of the predefined roles.

2.2 Social Networks Based Role Recognition

This role recognition approach is based on the Affiliation Networks (see upper part of Figure 2) [10], i.e. Social Networks where there are two kinds of nodes, the *actors* and the *events*, and only links between different kinds of nodes are allowed. The rationale behind this representation is that people participating in *similar* sets of events are more likely to interact with one another. Thus, actor nodes with similar sets of connections are expected to represent individuals with high mutual interaction likelihood.

The set of the connections of an actor node a_i is represented with a binary vector $\vec{x}_i = (x_{i1}, \ldots, x_{iD})$, where D is the number of events, and $x_{ij} = 1$ if actor a_i participates in event e_j and 0 otherwise.

Role	PM	ME	UI	ID
Fraction	36.6%	22.1%	19.8%	21.5%

Table 1: Role distribution. The table reports the average fraction of time each role accounts for in a meeting.

The more two vectors $\vec{x}_i$ and $\vec{x}_l$ are similar, the more actors a_i and a_l are likely to interact because they participate together in many events. In the case of the meeting recordings, the actors are the participants, and the events are segments of uniform length that span the whole duration of a meeting (see lower part of Figure 2). If D is the total number of segments for a meeting, then the event e_n corresponds to the time interval $[(n-1)T/D, nT/D]$, where T is the total duration of the meeting. Actors are said to participate in an event when they talk during the corresponding meeting segment. Thus, the actors are supposed to have a higher probability of interaction when they talk during the same intervals of time (i.e., when they participate in the same events) than when they talk in different intervals of time.

The most natural way of modeling binary vectors is to use Bernoulli discrete distributions:

$$p(\vec{x}_i \,|\, \vec{\mu}_r) = \prod_{j=1}^{D} \mu_{rj}^{x_{ij}} (1-\mu_{rj})^{1-x_{ij}}, \tag{2}$$

where $\vec{\mu}_r = (\mu_{r1}, \ldots, \mu_{rD})$ is the parameter vector of the distributuion related to role r. The maximum likelihood estimates of the μ_{ri} parameters are as follows [3]:

$$\mu_{ri} = \frac{1}{N_r} \sum_{n=1}^{N_r} x_{ni}, \tag{3}$$

where N_r is the number of people playing the role r in the training set, and x_{nj} is the j^{th} component of the vector representing the n^{th} person playing the role r. A different Bernoulli distribution can be trained for each role, and an actor represented with a vector $\vec{x}$ will be assigned the role r^* satisfying the following equation:

$$r^* = \arg\max_{r \in \mathcal{R}} p(\vec{x} \,|\, \vec{\mu}_r), \tag{4}$$

where $\mathcal{R}$ is the set of the predefined roles.

2.3 Combination Approach

Both role recognition approaches described above estimate the probability of a meeting participant playing a role r. The combination is performed by multiplying the two estimates as follows:

$$\begin{aligned} r^* &= \arg\max_{r \in \mathcal{R}} p(\vec{x}, \vec{d} \,|\, r, \vec{\mu}_r) \\ &= \arg\max_{r \in \mathcal{R}} \beta \log p(\vec{d} \,|\, r) + (1-\beta) \log p(\vec{x} \,|\, \vec{\mu}_r), \end{aligned} \tag{5}$$

where the factor β ensures that both terms are of the same order of magnitude and contribute to the final decision. The β value is selected through cross validation (see next section). The techniques to estimate $p(\vec{d} \,|\, r)$ and $p(\vec{x} \,|\, \vec{\mu}_r)$ are explained in the previous subsections.

3 Experiments and Results

This section presents the data, the experiments and the results obtained in this work.

approach	all	PM	ME	UI	ID
SNA (aut.)	43.1	75.7	16.4	41.2	13.4
lex. (aut.)	67.1	78.3	71.9	38.1	53.0
SNA+lex. (aut.)	67.9	84.0	69.8	38.1	50.1
SNA (man.)	49.5	79.0	20.3	44.9	24.6
lexical (man.)	76.7	92.0	70.3	60.1	60.9
SNA+lex. (man.)	78.0	95.7	68.8	60.1	61.6

Table 2: Role recognition results. The upper part of the table shows the accuracies obtained over automatic (aut.) speaker diarization and speech recognition. The lower part reports the accuracies obtained over manual (man.) speaker segmentation and speech transcriptions.

3.1 Data and Roles

The experiments of this work are performed over the AMI corpus [6], a collection of 138 meeting recordings for a total of 45 hours and 38 minutes of material. The meetings are simulated and are based on a scenario where the participants are the members of a team working on the development of a new remote control. Each meeting involves four participants playing one of the following roles: the *Project Manager* (PM), the *Marketing Expert* (ME), the *User Interface Expert* (UI), and the *Industrial Designer* (ID). Each participant plays a different role, and all roles are represented in each meeting. The same person can play different roles in different meetings, and the fraction of meeting time that each role accounts for, on average, is reported in Table 1.

3.2 Experiments

The training of the role recognition system is performed using a *leave-one-out* approach: all the meetings of the corpus are used for training the models with the exception of one that is used as test set. Training and test are repeated as many times as there are meetings in the corpus (138 in the case of the AMI corpus), and each time a different meeting is *left out* as test set. In this way, the whole corpus can be used as test set while still keeping rigorously separated training and test set, as required to assess correctly the system performance. The hyperparameters of the system (number of AdaBoost iterations for the lexicon based approach, and β factor for the combination) are tuned over a subset of 20 meetings randomly selected in the training set.

The performance is measured with the *accuracy* α, i.e. with the percentage of data time correctly labeled in terms of role. Table 2 reports the accuracies obtained by using only Social Network Analysis, only lexical choices, and the combination of the two. The lower part of the table shows the results obtained using groundtruth speaker segmentation and speech transcripts, while the upper part of the table shows the results obtained using the output of automatic speaker diarization and speech recognition systems. The results are reported for the overall meetings, as well as for the single roles separately.

The lexical choice appears to be, at least for the AMI corpus, a more reliable cue for the recognition of the role. The overall accuracy of the lexicon based system is significantly higher for both groundtruth (76.7% against 49.5%) and automatic data (67.1% against 43.1%). A possible explanation is that the AMI corpus is particularly suitable for lexical analysis, while it is rather unfavorable to the application of SNA. On one hand, the content of the interventions is constrained by the role and this helps the former approach, on the other hand, the small number of participants limits significanly the latter approach because the social networks tend to be more meaningful when the number of people increases [10].

The SNA based system appears to be more robust when passing from the groundtruth data to the output of the automatic systems for speaker segmentation and speech recognition. A possible explanation is that the SNA based approach uses only the speaker segmentation that is performed with high accuracy (around 97%), while the lexical based approach uses the speech transcriptions that are affected by a much higher error rate (around 40%). As a result, while the overall performance remains significantly different, the accuracy for PM and UI is comparable for both systems (see upper part of Table 2). Thus, the systems have similar performance over more than 50% of the data time because PM and UI account together for roughly 57% of the total AMI corpus time (see Table 1).

The combination of the two systems improves only slightly the performance of the best system (see table 2). The main reason is probably that the performance of the SNA approach is too close to the chance (around 25%) for at least two roles (ME and ID). Thus, the SNA does not bring useful information in the combination, but simply some random noise. This seems to be confirmed by the case of the PM role, where the combination improves by almost 6% the performance of the best classifier. Not surprisingly, the performance of the SNA system over the PM is significantly better than the chance.

4 Conclusions

This work has presented a role recognition approach based on the combination of two systems relying on lexical choices and interaction patterns, respectively. The results show that roughly 70% of the data time is labeled correctly in terms of role, and that the combination improves the best classifier, in particular for the PM role.

The main limits of the approach are, on one hand, that the Affiliation Networks are not sufficiently effective because the participants are too few to give rise to a meaningful interaction structure [10] and, on the other hand, that both combined sources of information are extracted from the audio channel while the integration of different modalities seems to be the most effective technique to analyze social interactions [12]. This suggests two potential directions for future work: the recognition of roles in data where the number of participants is sufficiently high for the social networks (like in [9]), and the extraction of information from the video channel. Moreover, the existing approach can be applied for different kinds of roles in other data.

References

[1] S. Banerjee and A. Rudnicky. Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of International Conference on Spoken Language Processing*, pages 2189–2192, 2004.

[2] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind the roles: identifying speaker roles in radio broadcasts. In *Proceedings of American Association of Artificial Intelligence Symposium*, pages 679–684, 2000.

[3] C. Bishop. *Pattern Recognition and Machine Learning.* Springer Verlag, 2006.

[4] J. Dines, J. Vepa, and T. Hain. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proceedings of the Interspeech*, pages 1213–1216, 2006.

[5] T. Hain, L. Burget, J. Dines, I. McCowan, M. Karafiat, M. Lincoln, D. Moore, G. Garau, V. Wan, R. Ordelman, and S. Renals. The development of the AMI system for the transcription of speech in meetings. *Proceedings of the Conference on Machine Learning for Multimodal Interaction*, pages 344–356, 2005.

[6] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[7] R. E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[8] H. Tischler. *Introduction to Sociology.* Harcourt Brace College Publishers, 1990.

[9] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(9):1215–1226, 2007.

[10] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

[11] C. Weng, W. Chu, and J. Wu. Movie analysis based on roles social network. In *proceedings of IEEE International Conference on Multimedia and Expo*, pages 1403–1406, 2007.

[12] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 28–34, 2006.

Acknowledgments

This work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2). The authors wish to thank John Dines for his help.

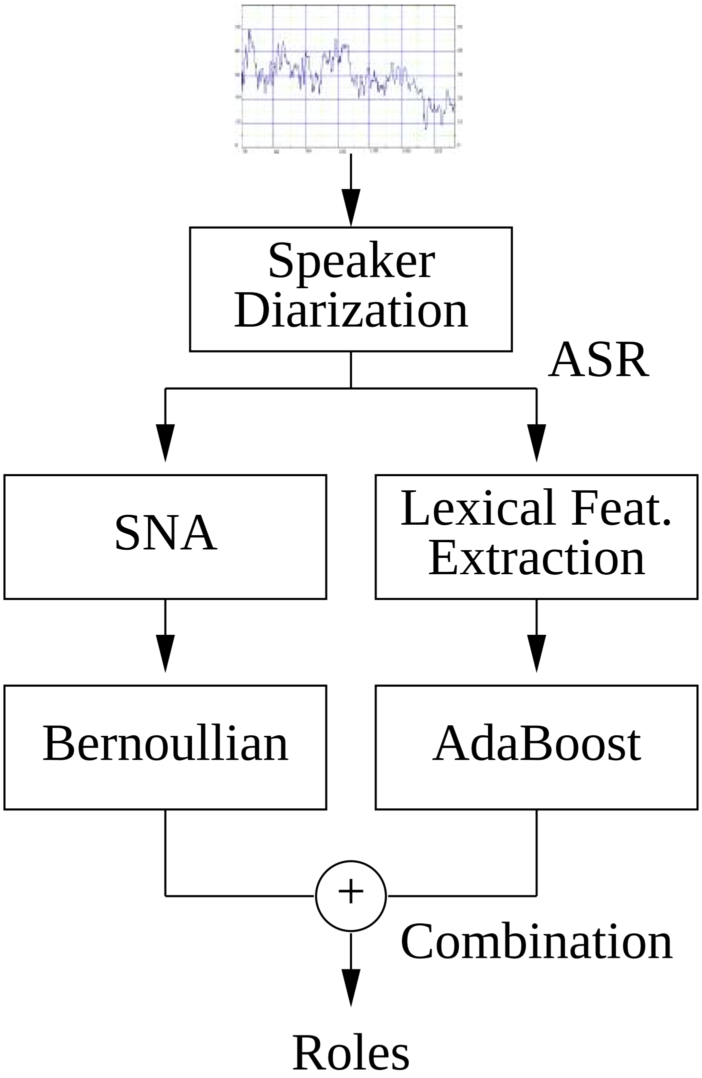

Figure 1: Overview of the approach. The two parallel paths produce separate decisions that are combined at the end of the process.

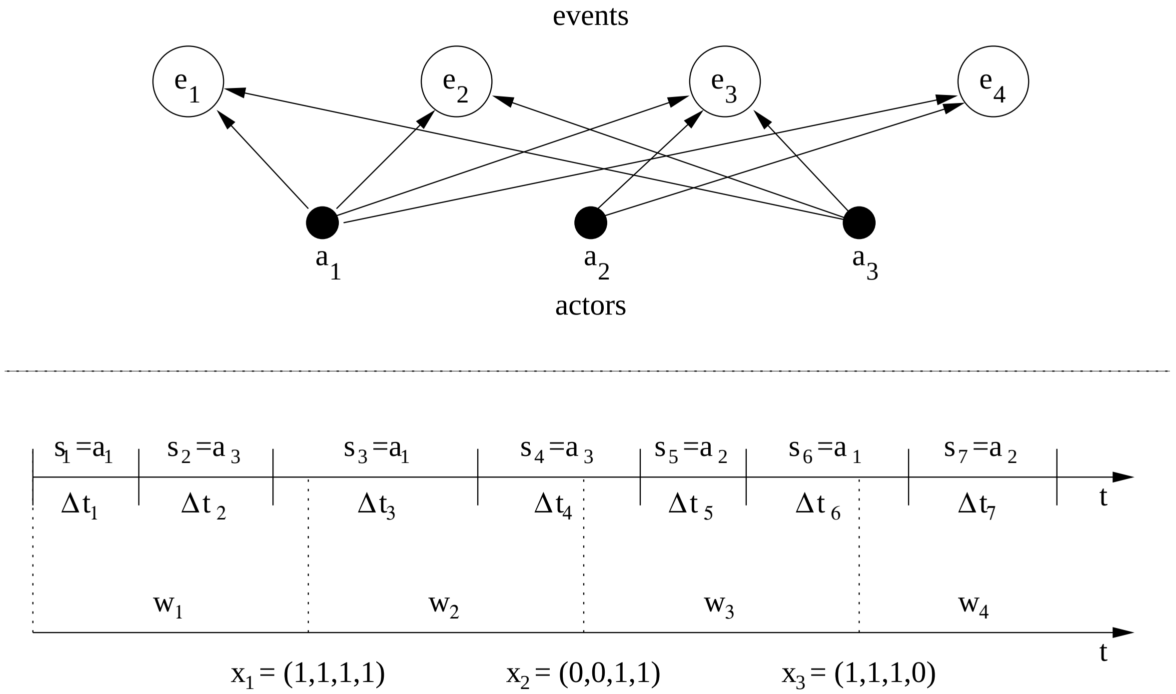

Figure 2: Social Affiliation Network. The figure shows how the Affiliation Network (upper part) is built starting from a speaker segmentation (lower part), and how the vectors $\vec{x}_i$ are obtained.