Addressee Identification in Face-to-Face Meetings

Natasa Jovanovic, Rieks op den Akker and **Anton Nijholt**
University of Twente
PO Box 217 Enschede
The Netherlands
{natasa,infrieks,A.Nijholt}@ewi.utwente.nl

Abstract

We present results on addressee identification in four-participants face-to-face meetings using Bayesian Network and Naive Bayes classifiers. First, we investigate how well the addressee of a dialogue act can be predicted based on gaze, utterance and conversational context features. Then, we explore whether information about meeting context can aid classifiers' performances. Both classifiers perform the best when conversational context and utterance features are combined with speaker's gaze information. The classifiers show little gain from information about meeting context.

1 Introduction

Addressing is an aspect of every form of communication. It represents a form of orientation and directionality of the act the current actor performs toward the particular other(s) who are involved in an interaction. In conversational communication involving two participants, the hearer is always the addressee of the speech act that the speaker performs. Addressing, however, becomes a real issue in multi-party conversation.

The concept of addressee as well as a variety of mechanisms that people use in addressing their speech have been extensively investigated by conversational analysts and social psychologists (Goffman, 1981a; Goodwin, 1981; Clark and Carlson, 1982).

Recently, addressing has received considerable attention in modeling multi-party interaction in various domains. Research on automatic addressee identification has been conducted in the context of mixed human-human and human-computer interaction (Bakx et al., 2003; van Turnhout et al., 2005), human-human-robot interaction (Katzenmaier et al., 2004), and mixed human-agents and multi-agents interaction (Traum, 2004). In the context of automatic analysis of multi-party face-to-face conversation, Otsuka et al. (2005) proposed a framework for automating inference of conversational structure that is defined in terms of conversational roles: speaker, addressee and unaddressed participants.

In this paper, we focus on addressee identification in a special type of communication, namely, face-to-face meetings. Moreover, we restrict our analysis to small group meetings with four participants. Automatic analysis of recorded meetings has become an emerging domain for a range of research focusing on different aspects of interactions among meeting participants. The outcomes of this research should be combined in a targeted application that would provide users with useful information about meetings. For answering questions such as "*Who was asked to prepare a presentation for the next meeting?*" or "*Were there any arguments between participants A and B?*", some sort of understanding of dialogue structure is required. In addition to identification of dialogue acts that participants perform in multi-party dialogues, identification of addressees of those acts is also important for inferring dialogue structure.

There are many applications related to meeting research that could benefit from studying addressing in human-human interactions. The results can be used by those who develop communicative agents in interactive intelligent environments and remote meeting assistants. These agents need to recognize when they are being addressed and how they should address people in the environment.

This paper presents results on addressee identi-

fication in four-participants face-to-face meetings using Bayesian Network and Naive Bayes classifiers. The goals in the current paper are (1) to find relevant features for addressee classification in meeting conversations using information obtained from multi-modal resources - gaze, speech and conversational context, (2) to explore to what extent the performances of classifiers can be improved by combining different types of features obtained from these resources, (3) to investigate whether the information about meeting context can aid the performances of classifiers, and (4) to compare performances of the Bayesian Network and Naive Bayes classifiers for the task of addressee prediction over various feature sets.

2 Addressing in face-to-face meetings

When a speaker contributes to the conversation, all those participants who happen to be in perceptual range of this event will have "some sort of participation status relative to it". The conversational roles that the participants take in a given conversational situation make up the "participation framework" (Goffman, 1981b).

Goffman (1976) distinguished three basic kinds of hearers: those who *overhear*, whether or not their unratified participation is unintentional or encouraged; those who are ratified but are not *specifically* addressed by the speaker (also called *unaddressed* recipients (Goffman, 1981a)); and those ratified participants who are *addressed*. Ratified participants are those participants who are allowed to take part in conversation. Regarding hearers' roles in meetings, we are focused only on ratified participants. Therefore, the problem of addressee identification amounts to the problem of distinguishing addressed from unaddressed participants *for each dialogue act* that speakers perform.

Goffman (1981a) defined *addressees* as those "ratified participants () oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants". According to this, it is the speaker who selects his addressee; the addressee is the one who is expected by the speaker to react on what the speaker says and to whom, therefore, the speaker is giving primary attention in the present act.

In meeting conversations, a speaker may address his utterance to the whole group of participants present in the meeting, or to a particular subgroup of them, or to a single participant in particular. A speaker can also just think aloud or mumble to himself without really addressing anybody (e.g.*"What else do I want to say?"* (while trying to evoke more details about the issue that he is presenting)). We excluded self-addressed speech from our study.

Addressing behavior is behavior that speakers show to express to whom they are addressing their speech. It depends on the course of the conversation, the status of attention of participants, their current involvement in the discussion as well as on what the participants know about each others' roles and knowledge, whether explicit addressing behavior is called for. Using a vocative is the explicit verbal way to address someone. In some cases the speaker identifies the addressee of his speech by looking at the addressee, sometimes accompanying this by deictic hand gestures. Addressees can also be designated by the manner of speaking. For example, by whispering, a speaker can select a single individual or a group of people as addressees. Addressees are often designated by the content of what is being said. For example, when making the suggestion *"We all have to decide together about the design"*, the speaker is addressing the whole group.

In meetings, people may perform various group actions (termed as *meeting actions*) such as presentations, discussions or monologues (McCowan et al., 2003). A type of group action that meeting participants perform may influence the speaker's addressing behavior. For example, speakers may show different behavior during a presentation than during a discussion when addressing an individual: regardless of the fact that a speaker has turned his back to a participant in the audience during a presentation, he most probably addresses his speech to the group including that participant, whereas the same behavior during a discussion, in many situations, indicates that that participant is unaddressed.

In this paper, we focus on speech and gaze aspects of addressing behavior as well as on contextual aspects such as conversational history and meeting actions.

3 Cues for addressee identification

In this section, we present our motivation for feature selection, referring also to some existing work

on the examination of cues that are relevant for addressee identification.

Adjacency pairs and addressing - Adjacency pairs (AP) are minimal dialogic units that consist of pairs of utterances called "first pair-part" (or a-part) and the "second pair-part" (or b-part) that are produced by different speakers. Examples include question-answers or statement-agreement. In the exploration of the conversational organization, special attention has been given to the a-parts that are used as one of the basic techniques for selecting a next speaker (Sacks et al., 1974). For addressee identification, the main focus is on b-parts and their addressees. It is to be expected that the a-part provides a useful cue for identification of addressee of the b-part (Galley et al., 2004). However, it does not imply that the speaker of the a-part is always the addressee of the b-part. For example, A can address a question to B, whereas B's reply to A's question is addressed to the whole group. In this case, the addressee of the b-part includes the speaker of the a-part.

Dialogue acts and addressing When designing an utterance, a speaker intends not only to perform a certain communicative act that contributes to a coherent dialogue (in the literature referred to as dialogue act), but also to perform that act toward the particular others. Within a turn, a speaker may perform several dialogue acts, each of those having its own addressee (e.g. *I agree with you* [agreement; addressed to a previous speaker] *but is this what we want* [information request; addressed to the group]). Dialogue act types can provide useful information about addressing types since some types of dialogue acts -such as agreements or disagreements- tend to be addressed to an individual rather than to a group. More information about the addressee of a dialogue can be induced by combining the dialogue act information with some lexical markers that are used as addressee "indicators" (e.g. you, we, everybody, all of you) (Jovanovic and op den Akker, 2004).

Gaze behavior and addressing Analyzing dyadic conversations, researchers into social interaction observed that gaze in social interaction is used for several purposes: to control communication, to provide a visual feedback, to communicate emotions and to communicate the nature of relationships (Kendon, 1967; Argyle, 1969).

Recent studies into multi-party interaction emphasized the relevance of gaze as a means of addressing. Vertegaal (1998) investigated to what extent the focus of visual attention might function as an indicator for the focus of "dialogic attention" in four-participants face-to-face conversations. "Dialogic attention" refers to attention while listening to a person as well as attention while talking to one or more persons. Empirical findings show that when a speaker is addressing an individual, there is 77% chance that the gazed person is addressed. When addressing a triad, speaker gaze seems to be evenly distributed over the listeners in the situation where participants are seated around the table. It is also shown that on average a speaker spends significantly more time gazing at an individual when addressing the whole group, than at others when addressing a single individual. When addressing an individual, people gaze 1.6 times more while listening (62%) than while speaking (40%). When addressing a triad the amount of speaker gaze increases significantly to 59%. According to all these estimates, we can expect that gaze directional cues are good indicators for addressee prediction.

However, these findings cannot be generalized in the situations where some objects of interest are present in the conversational environment, since it is expected that the amount of time spent looking at the persons will decrease significantly. As shown in (Bakx et al., 2003), in a situation where a user interacts with a multimodal information system and in the meantime talks to another person, the user looks most of the time at the system, both when talking to the system (94%) and when talking to the user (57%). Also, another person looks at the system in 60% of cases when talking to the user. Bakx et al. (2003) also showed that some improvement in addressee detection can be achieved by combining utterance duration with gaze.

In meeting conversations, the contribution of the gaze direction to addressee prediction is also affected by the current meeting activity and seating arrangement (Jovanovic and op den Akker, 2004). For example, when giving a presentation, a speaker most probably addresses his speech to the whole audience, although he may only look at a single participant in the audience. A seating arrangement determines a visible area for each meeting participant. During a turn, a speaker mostly looks at the participants who are in his visible area.

Moreover, the speaker frequently looks at a single participant in his visual area when addressing a group. However, when he wants to address a single participant outside his visual area, he will often turn his body and head toward that participant.

In this paper, we explored not only the effectiveness of the speaker's gaze direction, but also the effectiveness of the listeners' gaze directions as cues for addressee prediction.

Meeting context and addressing As Goffman (1981a) has noted, "the notion of a conversational encounter does not suffice in dealing with the context in which words are spoken; a social occasion involving a podium event or no speech event at all may be involved, and in any case, the whole social situation, the whole surround, must always be considered". A set of various meeting actions that participants perform in meetings is one aspect of the social situation that differentiates meetings from other contexts of talk such as ordinary conversations, interviews or trials. As noted above, it influences addressing behavior as well as the contribution of gaze to addressee identification. Furthermore, distributions of addressing types vary for different meeting actions. Clearly, the percentage of the utterances addressed to the whole group during a presentation is expected to be much higher than during a discussion.

4 Data collection

To train and test our classifiers, we used a small multimodal corpus developed for studying addressing behavior in meetings (Jovanovic et al., 2005). The corpus contains 12 meetings recorded at the IDIAP smart meeting room in the research program of the M4[1] and AMI projects[2]. The room has been equipped with fully synchronized multi-channel audio and video recording devices, a whiteboard and a projector screen. The seating arrangement includes two participants at each of two opposite sides of the rectangular table. The total amount of the recorded data is approximately 75 minutes. For experiments presented in this paper, we have selected meetings from the M4 data collection. These meetings are scripted in terms of type and schedule of group actions, but content is natural and unconstrained.

The meetings are manually annotated with dialogue acts, addressees, adjacency pairs and gaze direction. Each type of annotation is described in detail in (Jovanovic et al., 2005). Additionally, the available annotations of meeting actions for the M4 meetings[3] were converted into the corpus format and included in the collection.

The dialogue act tag set employed for the corpus creation is based on the MRDA (Meeting Recorder Dialogue Act) tag set (Dhillon et al., 2004). The MRDA tag set represents a modification of the SWDB-DAMSL tag set (Jurafsky et al., 1997) for an application to multi-party meeting dialogues. The tag set used for the corpus creation is made by grouping the MRDA tags into 17 categories that are divided into seven groups: acknowledgments/backchannels, statements, questions, responses, action motivators, checks and politeness mechanisms. A mapping between this tag set and the MRDA tag set is given in (Jovanovic et al., 2005). Unlike MRDA where each utterance is marked with a label made up of one or more tags from the set, each utterance in the corpus is marked as `Unlabeled` or with exactly one tag from the set. Adjacency pairs are labeled by marking dialogue acts that occur as their a-part and b-part.

Since all meetings in the corpus consist of four participants, the addressee of a dialogue act is labeled as `Unknown` or with one of the following addressee tags: individual $\mathtt{P_x}$, a subgroup of participants $\mathtt{P_x}$,$\mathtt{P_y}$ or the whole audience $\mathtt{P_x}$,$\mathtt{P_y}$,$\mathtt{P_z}$.

Labeling gaze direction denotes labeling gazed targets for each meeting participants. As the only targets of interest for addressee identification are meeting participants, the meetings were annotated with the tag set that contains tags that are linked to each participant $\mathtt{P}_x$ and the `NoTarget` tag that is used when the speaker does not look at any of the participants.

Meetings are annotated with a set of six meeting actions described in (McCowan et al., 2003): monologue, presentation, white-board, discussion, consensus, disagreement and note-taking.

Reliability of the annotation schema As reported in (Jovanovic et al., 2005), gaze annotation has been reproduced reliably (segmentation 80.40% (N=939); classification $\kappa = 0.95$). Table 1 shows reliability of dialogue act segmentation as well as Kappa values for dialogue act and addressee classification for two different annotation

[1] http://www.m4project.org
[2] http://www.amiproject.org
[3] http://mmm.idiap.ch/

groups that annotated two different sets of meeting data.

Group	Seg(%)	N	DA(κ)	ADD(κ)
B&E	91.73	377	0.77	0.81
M&R	86.14	367	0.70	0.70

Table 1: Inter-annotator agreement on DA and addressee annotation: N- number of agreed segments

5 Addressee classification

In this section we present the results on addressee classification in four-persons face-to-face meetings using Bayesian Network and Naive Bayes classifiers.

5.1 Classification task

In a dialogue situation, which is an event which lasts as long as the dialogue act performed by the speaker in that situation, the class variable is the addressee of the dialogue act (ADD). Since there are only a few instances of subgroup addressing in the data, we removed them from the data set and excluded all possible subgroups of meeting participants from the set of class values. Therefore, we define addressee classifiers to identify one of the following class values: individual `Px` where $x \in \{0,1,2,3\}$ and `ALLP` which denotes the whole group.

5.2 Feature set

To identify the addressee of a dialogue act we initially used three sorts of features: conversational context features (later referred to as contextual features), utterance features and gaze features. Additionally, we conducted experiments with an extended feature set including a feature that conveys information about meeting context.

Contextual features provide information about the preceding utterances. We experimented with using information about the speaker, the addressee and the dialogue act of the immediately preceding utterance on the same or a different channel (SP-1, ADD-1, DA-1) as well as information about the related utterance (SP-R, ADD-R, DA-R). A related utterance is the utterance that is the a-part of an adjacency pair with the current utterance as the b-part. Information about the speaker of the current utterance (SP) has also been included in the contextual feature set.

As **utterance features**, we used a subset of lexical features presented in (Jovanovic and op den Akker, 2004) as useful cues for determining whether the utterance is single or group addressed. The subset includes the following features:

- does the utterance contain personal pronouns "we" or "you", both of them, or neither of them?
- does the utterance contain possessive pronouns or possessive adjectives ("your/yours" or "our/ours"), their combination or neither of them?
- does the utterance contain indefinite pronouns such as "somebody", "someone", "anybody", "anyone", "everybody" or "everyone"?
- does the utterance contain the name of participant P_x?

Utterance features also include information about the utterance's conversational function (DA tag) and information about utterance duration i.e. whether the utterance is short or long. In our experiments, an utterance is considered as a short utterance, if its duration is less than or equal to 1 sec.

We experimented with a variety of **gaze features**. In the first experiment, for each participant P_x we defined a set of features in the form P_x-looks-P_y and P_x-looks-NT where $x, y \in \{0,1,2,3\}$ and $x \neq y$; P_x-looks-NT represents that participant P_x does not look at any of the participants. The value set represents the number of times that speaker P_x looks at P_y or looks away during the time span of the utterance: `zero` for 0, `one` for 1, `two` for 2 and `more` for 3 or more times. In the second experiment, we defined a feature set that incorporates only information about gaze direction of the current speaker (SP-looks-P_x and SP-looks-NT) with the same value set as in the first experiment.

As to **meeting context**, we experimented with different values of the feature that represents the meeting actions (MA-TYPE). First, we used a full set of speech based meeting actions that was applied for the manual annotation of the meetings in the corpus: monologue, discussion, presentation, white-board, consensus and disagreement. As the results on modeling group actions in meetings presented in (McCowan et al., 2003) indicate that consensus and disagreements were mostly misclassified as discussion, we have also conducted experiments with a set of four values for MA-TYPE, where consensus, disagreement and discussion meeting actions were grouped in one category marked as discussion.

5.3 Results and Discussions

To train and test the addressee classifiers, we used the hand-annotated M4 data from the corpus. After we had discarded the instances labeled with `Unknown` or subgroup addressee tags, there were 781 instances left available for the experiments. The distribution of the class values in the selected data is presented in Table 2.

ALLP	P_0	P_1	P_2	P_3
40.20%	13.83%	17.03%	15.88%	13.06%

Table 2: Distribution of addressee values

For learning the Bayesian Network structure, we applied the K2 algorithm (Cooper and Herskovits, 1992). The algorithm requires an ordering on the observable features; different ordering leads to different network structures. We conducted experiments with several orderings regarding feature types as well as with different orderings regarding features of the same type. The obtained classification results for different orderings were nearly identical. For learning conditional probability distributions, we used the algorithm implemented in the WEKA toolbox[4] that produces direct estimates of the conditional probabilities.

5.3.1 Initial experiments without meeting context

The performances of the classifiers are measured using different feature sets. First, we measured the performances of classifiers using utterance features, gaze features and contextual features separately. Then, we conducted experiments with all possible combinations of different types of features. For each classifier, we performed 10-fold cross-validation. Table 3 summarizes the accuracies of the classifiers (with 95% confidence interval) for different feature sets (1) using gaze information of all meeting participants and (2) using only information about speaker gaze direction.

The results show that the Bayesian Network classifier outperforms the Naive Bayes classifier for all feature sets, although the difference is significant only for the feature sets that include contextual features.

For the feature set that contains only information about gaze behavior combined with information about the speaker (Gaze+SP), both classifiers perform significantly better when exploiting gaze information of all meeting participants. In other words, when using solely focus of visual attention to identify the addressee of a dialogue act, listeners' focus of attention provides valuable information for addressee prediction. The same conclusion can be drawn when adding information about utterance duration to the gaze feature set (Gaze+SP+Short), although for the Bayesian Network classifier the difference is not significant. For all other feature sets, the classifiers do not perform significantly different when including or excluding the listeners gaze information. Even more, both classifiers perform better using only speaker gaze information in all cases except when combined utterance and gaze features are exploited (Utterance+Gaze+SP).

The Bayesian network and Naive Bayes classifiers show the same changes in the performances over different feature sets. The results indicate that the selected utterance features are less informative for addressee prediction (BN:52.62%, NB:52.50%) compared to contextual features (BN:73.11%; NB:68.12%) or features of gaze behavior (BN:66.45%, NB:64.53%). The results also show that adding the information about the utterance duration to the gaze features, slightly increases the accuracies of the classifiers (BN:67.73%, NB:65.94%), which confirms findings presented in (Bakx et al., 2003). Combining the information from the gaze and speech channels significantly improves the performances of the classifiers (BN:70.68%; NB:69.78%) in comparison to performances obtained from each channel separately. Furthermore, higher accuracies are gained when adding contextual features to the utterance features (BN:76.82%; NB:72.21%) and even more to the features of gaze behavior (BN:80.03%, NB:77.59%). As it is expected, the best performances are achieved by combining all three types of features (BN:82.59%, NB:78.49%), although not significantly better compared to combined contextual and gaze features.

We also explored how well the addressee can be predicted excluding information about the related utterance (i.e. AP information). The best performances are achieved combining speaker gaze information with contextual and utterance features (BN:79.39%; NB:76.06%). A small decrease in the classification accuracies when excluding AP information (about 3%) indicates that remaining contextual, utterance and gaze features capture most of the useful information provided by AP.

[4] http://www.cs.waikato.ac.nz/ ml/weka/

	Baysian Networks		Naive Bayes	
Feature sets	**Gaze All**	**Gaze SP**	**Gaze All**	**Gaze SP**
All Features	81.05% (±2.75)	82.59% (±2.66)	78.10% (±2.90)	78.49% (±2.88)
Context	73.11% (±3.11)		68.12% (±3.27)	
Utterance+SP	52.62% (±3.50)		52.50% (±3.50)	
Gaze+SP	66.45% (±3.31)	62.36% (±3.40)	64.53% (±3.36)	59.02% (±3.45)
Gaze+SP+Short	67.73% (±3.28)	66.45% (±3.31)	65.94% (±3.32)	61.46% (±3.41)
Context+Utterance	76.82% (±2.96)		72.21% (±3.14)	
Context+Gaze	79.00% (±2.86)	80.03% (±2.80)	74.90% (±3.04)	77.59% (±2.92)
Utterance+Gaze+SP	70.68% (±3.19)	70.04% (±3.21)	69.78% (±3.22)	68.63% (±3.25)

Table 3: Classification results for Bayesian Network and Naive Bayes classifiers using gaze information of all meeting participants (Gaze All) and using speaker gaze information (Gaze SP)

Error analysis Further analysis of confusion matrixes for the best performed BN and NB classifiers, show that most misclassifications were between addressing types (individual vs. group): each P_x was more confused with ALLP than with P_y. A similar type of confusion is observed between human annotators regarding addressee annotation (Jovanovic et al., 2005). Out of all misclassified cases for each classifier, individual types of addressing (P_x) were, in average, misclassified with addressing the group (ALLP) in 73% cases for NB, and 68% cases for BN.

5.3.2 Experiments with meeting context

We examined whether meeting context information can aid the classifiers' performances. First, we conducted experiments using the six values set for the MA-TYPE feature. Then, we experimented with employing the reduced set of four types of meeting actions (see Section 5.2). The accuracies obtained by combining the MA-TYPE feature with contextual, utterance and gaze features are presented in Table 4.

	Bayesian Networks		Naive Bayes	
Features	Gaze All	Gaze SP	Gaze All	Gaze SP
MA-6+All	81.82%	82.84%	78.74%	79.90%
MA-4+All	81.69%	83.74%	78.23%	79.13%

Table 4: Classification results combining MA-TYPE with the initial feature set

The results indicate that adding meeting context information to the initial feature set improves slightly, but not significantly, the classifiers' performances. The highest accuracy (83.74%) is achieved using the Bayesian Network classifier by combining the four-values MA-TYPE feature with contextual, utterance and the speaker's gaze features.

6 Conclusion and Future work

We presented results on addressee classification in four-participants face-to-face meetings using Bayesian Network and Naive Bayes classifiers. The experiments presented should be seen as preliminary explorations of appropriate features and models for addressee identification in meetings.

We investigated how well the addressee of a dialogue act can be predicted (1) using utterance, gaze and conversational context features alone as well as (2) using various combinations of these features. Regarding gaze features, classifiers' performances are measured using gaze directional cues of the speaker only as well as of all meeting participants. We found that contextual information aids classifiers' performances over gaze information as well as over utterance information. Furthermore, the results indicate that selected utterance features are the most unreliable cues for addressee prediction. The listeners' gaze direction provides useful information only in the situation where gaze features are used alone. Combinations of features from various resources increases classifiers' performances in comparison to performances obtained from each resource separately. However, the highest accuracies for both classifiers are reached by combining contextual and utterance features with speaker's gaze (BN:82.59%, NB:78.49%). We have also explored the effect of meeting context on the classification task. Surprisingly, addressee classifiers showed little gain from the information about meeting actions (BN:83.74%, NB:79.90%). For all feature sets, the Bayesian Network classifier outperforms the Naive Bayes classifier.

In contrast to Vertegaal (1998) and Otsuka et al. (2005) findings, where it is shown that gaze can be a good predictor for addressee in four-participants face-to-face *conversations*, our results

show that in four-participants face-to-face *meetings*, gaze is less effective as an addressee indicator. This can be due to several reasons. First, they used different seating arrangements which is implicated in the organization of gaze. Second, our meeting environment contains attentional 'distracters' such as whiteboard, projector screen and notes. Finally, during a meeting, in contrast to an ordinary conversation, participants perform various meeting actions which may influence gaze as an aspect of addressing behavior.

We will continue our work on addressee identification on the large AMI data collection that is currently in production. The AMI corpus contains more natural, scenario-based, meetings that involve groups focused on the design of a TV remote control. Some initial experiments on the AMI pilot data show that additional challenges for addressee identification on the AMI data are: roles that participants play in the meetings (e.g. project manager or marketing expert) and additional attentional 'distracters' present in the meeting room such as, the task object at first place and laptops. This means that a richer feature set should be explored to improve classifiers' performances on the AMI data including, for example, the background knowledge about participants' roles. We will also focus on the development of new models that better handle conditional and contextual dependencies among different types of features.

Acknowledgments

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-153).

References

M. Argyle. 1969. *Social Interaction*. London: Tavistock Press.

I. Bakx, K. van Turnhout, and J. Terken. 2003. Facial orientation during multi-party interaction with information kiosks. In *Proc. of INTERACT*.

H. H. Clark and B. T. Carlson. 1982. Hearers and speech acts. *Language*, 58:332–373.

G. Cooper and E. Herskovits. 1992. Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. 2004. Meeting recorder project: Dialogue act labeling guide. Technical report, ICSI, Berkeley, USA.

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proc. of 42nd Meeting of the ACL*.

E. Goffman. 1976. Replies and responses. *Language in Society*, 5:257–313.

E. Goffman. 1981a. Footing. In *Forms of Talk*, pages 124–159. University of Pennsylvania Press.

E. Goffman. 1981b. *Forms of Talk*. University of Pennsylvania Press, Philadelphia.

C. Goodwin. 1981. *Conversational Organization: Interaction Between Speakers and Hearers*. NY:Academic Press.

N. Jovanovic and R. op den Akker. 2004. Towards automatic addressee identification in multi-party dialogues. In *Proc of the 5th SIGDial*.

N. Jovanovic, R. op den Akker, and A. Nijholt. 2005. A corpus for studying addressing behavior in face-to-face meetings. In *Proc. of the 6th SIGDial*.

D. Jurafsky, L. Shriberg, and D. Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13. Technical report, University of Colorado, Institute of Cognitive Science.

M. Katzenmaier, R. Stiefelhagen, and T. Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. of ICMI*.

A. Kendon. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica*, 32:1–25.

I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. 2003. Modeling human interactions in meetings. In *Proc. IEEE ICASSP*.

K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. 2005. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proc. of ICMI*.

H. Sacks, E. A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

D. Traum. 2004. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*, pages 201–211. Springer-Verlag.

K. van Turnhout, J. Terken, I. Bakx, and B. Eggen. 2005. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proc. of ICMI*.

R. Vertegaal. 1998. *Look who is talking to whom*. Ph.D. thesis, University of Twente, September.