Towards automatic addressee identification in multi-party dialogues

Natasa Jovanovic
Department of Computer Science
University of Twente
PO Box 217 Enschede, the Netherlands
natasa@cs.utwente.nl

Rieks op den Akker
Department of Computer Science
University of Twente
PO Box 217 Enschede, the Netherlands
infrieks@cs.utwente.nl

Abstract

The paper is about the issue of addressing in multi-party dialogues. Analysis of addressing behavior in face to face meetings results in the identification of several addressing mechanisms. From these we extract several utterance features and features of non-verbal communicative behavior of a speaker, like gaze and gesturing, that are relevant for observers to identify the participants the speaker is talking to. A method for the automatic prediction of the addressee of speech acts is discussed.

1 Introduction

Communication, between humans or between humans and conversational computer agents, involves addressing. Addressing has received attention in the tradition of conversation analysis (Clark and Carlson, 1992; Clark and Schaefer, 1992), but not that much in the community of computational dialogue systems. One exception is (Traum, 2003). An explanation for this lack of attention may be that most research in computational dialogue systems concerns systems that were designed for interaction between one human user and one conversational agent. In dialogues in which only two participants take part addressing goes without saying. Addressing becomes a real issue in multi-party conversations and that is the subject of this paper.

There are a number of application areas that could benefit from studying addressing behavior in human human interactions. It can provide valuable data for learning more about human interaction and the way humans interact with intelligent environments. The result can be used by those who develop communicative agents in interactive intelligent environments, meeting managers and presentation assistants. If we could induce from recorded meetings the *”who said what, when and to whom”* we can use this information for making summarizations of meetings, and for real-time tracking.

Research on small group discussions (Carletta et al., 2002) has shown that there is a noticeable difference in the interaction patterns between large and small groups (up to seven participants). A small group discussion looks like two-way conversations but conversations occur between all pairs of members and every member can initiate conversation. A large group discussion is more like a series of conversations between a group leader and various individuals with the rest participants present but silent. We will focus our research on *small group discussions* in meetings.

In this paper we propose research that aims at the automatic determination of the addressee of a speaker in small meetings. Analysis of the mechanisms that people use in identifying their addressees leads to a model of a conversation that describes the features that play a role in these mechanisms. These features can be of several types: verbal, non-verbal, and features of the situation. Our research is partly based on analysis of the IDIAP multi-modal meeting data corpus made available through the Media File Server [1].

2 Addressee detection - problem overview

One of the question of interest concerning a meeting is: *”Who talked to whom and about what during the meeting?”*. This question refers to three very important aspects of a conversational event: source of the message (speaker identification), topic of the message (topic detection) and addressee of the message (addressee identification).

Speaker and addressee roles are the basic conversational roles. There are different ways to categorize the audience of a speech act. We use a taxonomy of conversational roles proposed in (Clark and Carlson, 1992). People around an action are divided in those who re-

[1] http://mmm.idiap.ch

ally participate in the action (*active participants*) and those who do not (*non-participants*). The active participants in a conversation include speaker and addressee as well as other participants taking part in conversation but currently not being addressed. Clark called them *side-participants*. All other listeners who have no rights to take part in conversation are called *overhearers*. Overhearers are divided in two groups: bystanders and eavesdroppers. *Bystanders* are overhearers who are present and the speaker is aware of their presence. *Eavesdroppers* are those who are listening without the speakers awareness. In determining the conversational roles in a meeting situation we will focus on the active participants. The problem of addressee identification amounts to the problem of distinguishing the addressee from the side participants in a conversation.

According to dialogue act theory (Bunt, 2000) an utterance can consist of several segments which carry different dialogue acts. Each of these dialogue acts can have it's own addressee. The following example is an example of multi-addressee utterances.

A: We could use Java as a standard?
[suggestion] addressee B,C
B: yes— but what about C++ ?
[agreement]addressee A— [suggestion]addressee A,C
C: Both is OK for me
[accept] addressee A,B

3 Observation analysis - addressee detection in meetings

Three main questions considering addressee detections are: 1. What are the relevant sources of information for the addressee detection in face-to-face meetings? 2. How does the speaker express who is the addressee of his utterance? 3. How can we combine all this information in order to determine the addressee of the speaker utterance?

In order to find answers on these questions we observed meetings recorded at the IDIAP and annotated several of them. For annotation we used the NITE Workbench for Windows (NWB3) annotation tool [2]. We defined our annotation scheme based on the initial assumptions about the information sources that can be used for the addressee identification. These assumptions are the result of our meeting observations.

3.1 Sources of information

When two or more people are engaged in interaction they communicate using verbal and/or non-verbal elements. The most natural and powerful human communication is in combined use of words, gaze, facial and gestural movements, posture, bodily contact, etc.

3.1.1 Speech

Speech is the main communication channel used in the meeting conversation. Therefore, it is the main source for addressee detection. The most common heuristics that may guide the addressee recognition process is the search for **linguistic markers** in the utterance. Table 1 contains linguistic markers that can be used as cues for addressee detection. For instance, *you* is the personal pronoun that refers to the meeting participants excluding the speaker of the utterance. Usage of quantifying determiners, numerals and indefinite pronouns may help in distinguish *you* as a particular person from *you* as a group. If an utterance contains noun phrases like *some of you, few of you, most of you, etc.,* then it is addressed to all meeting participants. The speaker doesn't know who he is actually addressing (*He saw some of you yesterday*).

Name detection is a powerful method for addressee determination. The name in vocative form is used for direct addressing the person with that name *(What about you, John?)*. Using the name of the participant the speaker can claim something about the participant addressing the utterance to the other addressee *(John was not present et the last meeting)*.

Dialogue acts. There is a relation between addressees of an utterances and the type of the dialogue act the speaker performed. Sometimes the description of a dialogue act includes the possible addressees of the act. Therefore, knowledge about the dialog act is used as a cue for addressee detection. For dialogue act annotation we use the Meeting Recorder Dialogue Acts (MRDA) tag set (Dhillon et al., 2003). The MRDA is a tag set for labeling multiparty face-to-face meetings. The tags in the MRDA set are organized into thirteen groups according to syntactic, semantic, pragmatic and functional characteristic of the utterance they mark. For addressee detection purposes we used a large subset of the MRDA tag set but we organized them at two levels: **forward looking function** (FLF) and **backward looking function** (BLF). FLF represents the effect that an utterance has on the subsequent interaction. BLF indicates how an utterance relates to the previous discourse. If an utterance has both functions the corresponding addressee is the addressee of the BLF.

When an utterance is marked with a BLF it is related to one of the previous utterances in the conversation. The addressee of the associated dialogue act in most cases is the speaker of the related utterance. However, it is possible that the speaker of the related utterance is the same as the current speaker. For instance, a speaker can repeat or correct himself. The addressees of these utterances are addressees of the related utterances. Most of the BLFs are related to the previous utterances of the other speaker (acceptance tags, answer tags, etc.). In the multiparty case there is a number of interesting interaction patterns with respect to addressee identification.

[2] http://nite.nis.sdu.dk/download/. NWB The NITE Workbench is a general-purpose natural interactivity coding tool

Word classes	Example	Example
Personal pronouns PP	I/me, you, she/her, he/him, we/us	What do you think about that?
Quantifying determiners+PP	all of you, some of you, few of you	He saw some of you yesterday.
Numerals+PP	two of you, three of you, last of you	Three of you should prepare a presentation.
Indefinite pronouns+PP	anyone of you, someone of you	Did anyone of you finish the job?
Possessive pronouns	mine, yours, hers, his, ours, theirs	Is this yours?
Personal adjectives	my, your, his, her, our, their	I like your style.
Indefinite pronouns	everybody, somebody, anyone	Does anyone have any question?

Table 1: Linguistic markers

1. A: We could use Java as a standard [suggestion]
   B: I agree [accept]
   C: No [reject]
   D: For me, it is OK [accept]

2. A: I think that we should use Java [suggestion]
   B: I propose C++ [suggestion]
   C: I don't agree with you [reject]

In the first conversation all responses are related to A's proposal. Therefore, A is the addressee of the utterances expressed by B, C and D. It means the addressee doesn't have to be the previous speaker. In the second example it is not clear whether C rejects A's or B's proposal or both proposals. Additional information obtained from visual channels can help in resolving the addressee ambiguity.

Unlike BLFs, FLFs do not provide much information about the addressee of a speaker's utterance. Yet, some assumptions about the utterance's addressee are possible, especially if we take in consideration the linguistic markers mentioned above. For instance, the speaker of an utterance marked with some of the *question tags* directly addresses the addressee to provide information. In combination with the use of the personal pronoun *we* these questions are addressed to a subgroup of participants or to all participants rather than to a single person. Very often questions in meeting discussions are *open-ended questions*. An open-ended question is a question that does not seek a specific answer. Rather, it is asked in a broad sense. An open-ended question is more likely addressed to all meeting participants. If an open ended question contains *'you'* than a single person is the most probable addressee *(What about C? questions? What about you?)*.

Linguistic markers and dialogue acts described above provide us with starting assumptions about the most likely addressee. These assumptions are mostly related to a size of the target group i.e. whether the addressee is a single participant, a group of participant or all participants. Therefore, some other communication channels are used in combination with speech for addressing the utterance. In the following sections we will describe the role of non-verbal communication channels in addressee detection.

3.1.2 Gaze direction

Gaze is an important aspect of social interaction (Argyle, 1973). One of the functions of gaze is channel-control. Mutual gaze is important when people want to establish relationship. Unless the utterance is very short the speaker very soon breaks the mutual gaze. When finishing the utterance, the speaker gazes back to a listener. If the speaker decided to continue to talk at turn transition points, or even before, he usually gazes away. Need for feedback effects the speaker's gaze direction. Gaze direction shows a participant's *focus of attention*. In the meeting scenario where all participants are around the table the focus of attention of the current speaker are mostly the other meeting participants. Since it is almost impossible to record eye gazing of participants, gaze information is obtained and induced from head movements. In (Stiefelhagen and Zhu, 2002) it is shown that we can predict a participant focus of attention based on head orientation with a reliability of 88,7 %.

The contribution of gaze information to addressee detection is dependent on the current meeting action (discussion, presentation, note-taking, etc.), the participants' location and the utterance length. During a presentation a speaker most probably addresses utterances to all meeting participants. Therefore, information about gaze direction is less relevant for a presentation than for a discussion meeting action. When the utterance is short a speaker usually gazes only at one participant or at no one, addressing the utterance at a group of people or at the whole audience. Moreover, information about the visible areas of the participants and hence the relative positions they have in the meeting is relevant for interpreting the gaze behavior in terms of focus of attention and it's contribution to addressing. During a turn a speaker mostly looks at the participant who are in his visible area. On the other hand if he wants to address someone outside his visual area he will often move his body towards the addressee.

The result of automatic or manual gaze annotation is a list of gazed participants or objects, together with time stamps. For the BLFs the first participant in the list is of interest. If the participant is not in the speaker's visible area then the gazed participant is the most likely addressee of the speaker utterance. If the participant is in the speaker's visible area and he is a candidate from the speech analysis then the likelihood that he is the addressee of the speaker utterance is greater. For the FLFs utterance length and structure of the gaze list play a very important role. For BLFs the last participant in the gazed list is of interest.

3.1.3 Gesture

Pointing at a person is a way to address a speech act to a person. It is usually accompanied with gazing at the person. Still the addressee of a speaker's utterance is not necessarily the same as a person that the speaker points at. When X talks to Y and points at Z, at the same time X usually verbally refers to Z using a proper noun (name of people, group name, etc.), a pronoun (he/she/they, him/her/them, his/her/their, etc.) or using the role of participant (boss, chairman, etc.). This means that X talks to Y about Z. *(Yesterday I met **him** on the street.)*

3.1.4 Context

The categories of the context that contribute to addressee detection are: interaction history, meeting action history, user context and spatial context. Interaction history is related to the conversation history and to the nonverbal interaction history. Conversation history contains the temporal sequence of speakers, performed dialogue acts and their addressees. Meeting action history is a sequence of previous meeting actions including the current meeting action. For instance, if a presentation is followed by a discussion, the presenter is the more probable addressee of the other participants' utterances, especially those that are marked as questions. Spatial context includes participants' location, locations of the environmental objects, distance between participants, participants' visible area. User context includes participants names, gender, social roles (status roles and closeness), institutional roles etc.

3.2 Towards an automatic addressee detection

Although participants or outsiders are most of the time quite sure about the intended addressee of a speaker this knowledge is essentially error-prone. Using observational features obtained from different available sources they can only predict the most probable addressee of an utterance. Methods for addressee detection will either be rule based or follow a statistical approach.

A rule-based algorithm used for computing addressee in the MRE (Mission Rehearsal Exercise) project is shown in (Traum, 2003). The rule-based method we intend to apply for addressee identification first processes information obtained from the utterance. This returns a list of possible addressees with corresponding probabilities. The probabilities are estimations from annotated meeting data. The idea is first to eliminate cases where the addressee is completely determined (names in vocative forms, quantifiers and numerals in combination with 'you', etc.). According to analysis of the relation between dialogue acts and addressee, different sets of rules are applied for FLFs and BLFs. For instance, if an utterance is marked with a BLF that is related to an utterance of a previous speaker, the addressee is the speaker of the related utterance with probability P. The following steps are related to the processing of information from additional sources (gaze and gesture) adding the additional probability values to the possible addressee. Contextual features are used at each level of processing.

Given all available multi-modal information E about a conversational situation a statistical addressee identification method should classify the addressee for each dialogue act in the conversation. As a computational model we will use Bayesian networks (BN). The nodes in the Bayesian network will include all observable features as input variable and one unobservable output variable that represent the addressee. From some preliminary models, we concluded that Bayesian network used for addressee classification of FLFs is more complicated than for BLFs. We therefore consider using separate models for BLFs and FLFs.

4 Conclusions and future directions

Addressing is an interesting aspect of communication and the automatic identification of conversational roles in multi-party dialogues is an open research problem. We expect that statistical approaches can be applied at this domain. Our future work will be based primarily on obtaining a huge set of data for training and testing the models. We will also define new scenario's for new types of meetings that will show more interesting phenomena related to addressing behavior.

References

Michael Argyle. 1973. *Social Interaction*. Tavistock Publications.

H. Bunt. 2000. Dialogue pragmatics and context specification. In *Abduction, Belief and Context in Dialogue; studies in computational pragmatics*. John Benjamins, Amsterdam.

Jean Carletta, Anne H. Anderson, and S. Garrod. 2002. Seeing eye to eye: an account of grounding and understanding in work groups. *Bulletin of the Japanese cognitive sciences*, 9(1):1–20.

Herbert H. Clark and Thomas B. Carlson. 1992. Hearers and speech acts. In *Arenas of Language Use (H.H.Clark ed.)*. University of Chicago Press and CSLI.

Herbert H. Clark and Edward F. Schaefer. 1992. Dealing with overhearers. In *Arenas of Language Use (H.H.Clark ed.)*. University of Chicago Press and CSLI.

R. Dhillon, S Bhagat, H Carvey, and E. Shriberg. 2003. Meeting recorder project:dialogue act labeling guide, version 3. Technical report, ICSI.

Rainer Stiefelhagen and Jie Zhu. 2002. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems (CHI2002)*.

David Traum. 2003. Issues in multi-party dialogues. In *Advances in Agent Communication (F. Dignum, ed.)*. Springer-Verlag LNCS.