Are You Being Addressed? - real-time addressee detection to support remote participants in hybrid meetings

Harm op den Akker
Roessingh Research and Development
Enschede
the Netherlands
h.opdenakker@rrd.nl

Rieks op den Akker
Human Media Interaction Twente
Enschede
the Netherlands
infrieks@cs.utwente.nl

Abstract

A meeting assistant for (remote) participants in hybrid meetings has been developed. The agent has to follow the conversations in the meeting to see if his buddy is being addressed. This paper presents the experiments that have been performed to develop machine classifiers to decide if "You are being addressed" where "You" refers to a fixed (remote) participant in a meeting. The experimental results back up the choices made regarding the selection of data, features, and classification methods. We discuss variations of the addressee classification problem that have been considered in the literature and how suitable they are for addressing detection in a system that plays a role in a live meeting.

1 Introduction

In order to follow what's going on in a meeting, it is important to know who is talking, what is being said, and who is being addressed (talked to). Here, we focus on "the addressing problem". We present results obtained in developing a classifier for real-time addressee prediction to be used in an assistant for a remote participant in a *hybrid* meeting, a meeting where some people are in the same room and others take part at remote sites.

The question who is being addressed by the speaker is not only of interest for the *participants in* a meeting and in particular for remote participants in hybrid meetings, who often encounter problems in following the conversation. The question who is being addressed has long been of interest for science, for group therapists (Bales, 1950), for small group research, for *outside observers* that play the recorded meeting and analyse it after the fact. How speakers address listeners, what procedures speakers use to design their audience and to make clear whom they address has been the focus of conversational analysis, socio-linguistics and ethnomethodology for quite some time. An analysis of addressee selection is presented in (Lerner, 1996). Addressing as a special type of multi-modal interactional referring expression generation behavior is considered in (Op den Akker and Theune, 2008). The problem of *automatic addressee detection* is one of the problems that come up when technology makes the move from *two-party* man-machine natural dialogue systems to systems for *multi-party* conversations. In this context the addressing problem was raised by Traum (2004). After the presentation of Jovanović at SigDial Cambridge-USA in 2004, quite a few publications have appeared about automatic addressee prediction in meetings. Jovanović used a number of multi-modal meeting corpora developed in European Projects M4 and AMI. The public release of the multi-modal AMI meeting corpus (Carletta, 2007), a 100 hours annotated corpus of small group meetings has already shown to be an important achievement for research; not only for conversational speech recognition and tracking of visual elements but also for automatic multi-modal conversational scene analysis (Takemae and Ozawa, 2006). Addressing detection in robot human interaction is studied in (Katzenmaier et al., 2004) and in a multi-

party tutoring system in (Knott and Vlugter, 2008). Addressing in face-to-face conversations is achieved by "multi-modal behavior" and addressee detection is thus a multi-modal recognition task, that requires not only speech recognition but also gaze and gesture recognition, the recognition of deictic references in speaker's expression, and ideally the understanding of the "what's going on", the conversational floor (Edelsky, 1981). It requires the detection of who is involved in current (parallel) activities.

In AMIDA, the European follow-up project of AMI, the new targets are: (1) real-time processing (real-time speech recognition (Hain et al., 2008), focus of attention recognition (Ba and Odobez, 2009), real-time dialogue act labeling (Germesin et al., 2008) and addressee detection, etc.) and (2) technology for (remote) meeting support. Technology based on the analysis of how people behave and converse in meetings is now going to re-shape the meetings, and hopefully make them more effective and more engaging. Social interaction graphs that show who is talking to whom and how frequent in a meeting may help the group by mirroring its interpersonal relations, dominance, and group dynamics, and understand social mechanisms as possible causes of ineffectiveness. Although, feedback about the social interactions may also be usefull *during* meetings, it doesn't put a strict real-time constraint on the prediction of speaker's addressees. A participant in a meeting, however, needs to know who is being addressed by the speaker *at "the time of speaking"*. This holds for humans as well as for an artificial partner, a robot or a virtual ECA, in a multi-party conversation.

"The addressing problem" comes thus in different flavors, depending on the relations that the subject who is in need of an answer, has with the event itself. *Time* is one of the aspects that plays a role here: whether the subject needs the solution at the fact or after the fact, real-time or off-line. But not only time plays a role. The addressing problem is an *interactional problem*, meaning that it is determined by the role that the subject has in the interaction itself; if and how the speaker and others communicate with each other and with the subject. Is he himself a possible addressee of the speaker or is he an outside observer? What communications channels has he available and what channels of communication are available to the conversational partners in the meeting? It's often harder to follow a face-to-face discussion on the radio than to follow on the radio a broadcasted multi-party discussion in a point-to-point telephone connection. What speakers do to make clear whom they address depends on the status and capacities of the communication lines with his interlocutors, with those subjects that the speaker considers as taking part in the conversation. Discussion leaders in tv shows are aware of the tv viewer. They now and then address themselves explicitly at the "virtual" audience at home. They also design their questions so to make the tv viewer clear whom they address. Outside observers in the form of a video camera will however not affect the way speakers make clear whom they address as long as the camera is not considered as a participant interested in the speakers intention. It is because remote participants are often out of sight that speakers in the meeting room don't take them into account when they converse to others in the meeting room. Remote participants become a kind of remote outside observers then and share the same problems that annotators have that watch video recordings of meetings to see what's happening out there in the meeting and who is being addressed by the speaker.

In section 2 we will specify the particular variant of "the addressing problem" that we tackle here. We make clear how our problem and approach differs from other's and what this means for the applicability of previous results and available data. In section **??** we present the data we used for testing and training. We set a baseline for the performance of our classifiers as well as a hypothesized roof value based on the complexity of the task at hand. In section 4 we discuss the experiments, for selecting the optimal features, classifiers, and parameters. In section 5 we present the experimental results. In section 6 we discuss how the currently implemented addressing module works in the meeting assistant and what is required to use all the features of the addressee predictor in a hybrid meeting.

2 The Addressing Problem Considered Here

In (Jovanovic and Op den Akker, 2004; Jovanovic, 2007) Jovanović describes her ADR classifiers

trained and tested on the AMI corpus. The classification problem is to assign an addressee label to a dialogue act, a hand-labeled and hand-segmented sequence of words, which is obtained by manual transcription of a speaker's utterance. The output of the classifier, is one of a set of possible addressee labels: Group, or P0,P1,P2,P3, the four fixed positions of the four participants in the meeting. Since the AMI data contains several meetings of different groups of four people, the class value cannot be the name of a participant, because that is not an invariant of the meeting setting. Positions at the rectangular table are invariant. This implies that the classifiers can only be used for meetings with this setting and four participants. The same data is used by Galley et al. in (Gupta et al., 2007) in their study of a related problem, finding the person the speaker refers to when he uses a second person pronoun ("you", "your") as a deictic referring expression. Their class values are not positions at the table but "virtual positions" in the speaking order (next speaker, previous speaker), a solution that generalises to a broader class of conversations than four participant face-to-face meetings. Note however that in a more recent follow-up study Frampton et al. in (Frampton et al., 2009) use the same class values as Jovanovic. We will also use the AMI corpus but we will look at a different variant of the addressing problem. This is motivated by our application: to support a remote participant in a hybrid meeting. Our Addressing problem is "Are You Being Addressed" where "You" refers to an individual participant in a conversation. The possible answers we consider are "yes" or "no"[1]. The addressing classifier that solves this problem is thus dedicated to his personal buddy. Note that this makes the method useable for whatever conversational setting. Note also that the addressing prediction problem "Are You Being Addressed?" for a meeting assistant who does not himself participate in the meeting is different from the problem "Am I Being Addressed?" that a participant himself may have to solve. The meeting assistant does not have direct "internal" knowledge about the processes, and attentiveness of his buddy participant; he has to rely on outside observations. Our view on the problem implies that we have to take another look at the AMI data and that we will analyse and use it in a different way for training, testing and performance measuring. It also implies that we cannot rely for our binary classification problem on the results of Jovanović jovanovic:addressee, with (dynamic) Bayesian networks.

[1]It is certainly possible to classify in terms of a probability, or to consider a three-valued class label, including don't know. We didn't.

3 The Data and How Complex Our Task Is

We use a subset of the AMI corpus, containing those 14 meetings that have not only been annotated with dialogue acts, but where dialogue acts are also attributed an addressee label, telling if the speaker addresses the Group, or the person sitting at position P0,P1,P2 or P3.[2] They have also been annotated with visual focus of attention: at any time it is known for each partner where he is looking at and during what time frame. Annotated gaze targets are persons in the meeting, white board, laptop, table, or some other target object. An other level of annotations that we use concerns the topic being discussed during a topic segment of the meeting. Persons in the AMI corpus play a role following a scenario, the group has to design a remote tv control and team members each have on of four roles in the design project: PM - project manager; UI - user interface designer; ID - industrial designer; or ME - marketing expert. Details of the scenario can be found in (Post et al., 2004). In training and testing the classifiers we select a fixed position in the meeting as the target for addressee prediction.

3.1 Base-line and Roof-value

This makes that a baseline for the binary classification task is already quite high: 89.20%, being the percentage of all dialogue acts "*not* addressed to You", which is 5962 out of a total of 6648 dialogue acts. The performance of a supervised machine learning method depends on the (1) selection of features (2) the type of classifier including the settings of the hyper-parameters of the classifiers (Daelemans et al., 2003), but also on the quality and the amount of training data (see (Reidsma, 2008) and (Reidsma and Carletta, 2008)). Since we measure the classifier's performance with a part of the

[2]Annotators could also use label *Unknown* in case they could not decide the addressee of the speaker.

annotated data it is interesting to see how human annotators/classifiers perform on this task. One of the AMI meetings[3] has been annotated with addressing information by four different annotators. We will use this to measure how ambiguous the task of addressee labeling is. Table 1 shows the confusion matrix for two annotators: *s95* and *vka*. This shows the (dis-)agreements for labelling the 412 dialogue acts as addressed to A, B, C, D or to the Group. [4] However, because we use our data differently, we will look at the confusion matrices in a different way. We split it up into 4 matrices, each from the view of one of the four meeting participants. Table 2 is an example of this, taking the view of participant A (i.e. for the binary decision task *"Is A being addressed?"*, and having annotator s95 as gold standard.

	A	B	C	D	Group	Total
A	**29**				10	39
B		**14**			8	22
C			**32**		7	39
D	1		1	**49**	18	69
Group	21	10	19	22	**171**	243
Total	51	24	52	71	214	412

Table 1: Confusion matrix for one pair of annotators ($\kappa = 0.55$).

	A	$\neg A$	Total
A	**29**	10	39
$\neg A$	22	**351**	373
Total	51	361	412

Table 2: Confusion matrix for one pair of annotators, considering addressed to A or not (derived from the matrix in Table 1).

Table 2 shows that when taking annotator s95 as gold standard, and considering annotator vka as the classifier, he achieves an accuracy of 92,23% (380 out of 412 instances classified correctly). We can argue that we can use these human annotators/classifiers scores as a measure of "maximum performance", because it indicates a level of task ambiguity. Classifiers can achieve higher scores, because they can learn through noise in the data. Thus, the inter-annotator confusion value is not an absolute limit of actual performance, but cases in which the classifier is "right" and the test-set "wrong" would not be reflected in the results. Since the inter-annotator confusion does also say something about the inherent task ambiguity, it can be used as a measure to compare a classifier score with. Table 3 contains the overal scores (taken over all 4 individual participants) for the 6 annotator pairs. The average values for Recall, Precision, F-Measure and Accuracy in Table 3 are considered as a *roof* values for the performance measures for this binary classification task. The Hypothesized Maximum Score (HMS) is the average accuracy value, 92.47.

Pair	Rec	Prec	F	Acc
s-v	73,37	62,63	67,58	92,78
m-s	59,75	70,59	64,72	91,87
m-v	69,92	74,78	72,27	93,11
m-d	37,77	81,61	51,64	91,79
v-d	42,04	80,49	55,23	92,22
s-d	43,68	77,55	55,88	93,02
Average:	**54,42**	**74,61**	**61,22**	**92,47**

Table 3: Recall, Precision, F-measure and Accuracy values for the 6 pairs of annotators.

The baseline (89.20) and the HMS (92.47) accuracy values will be used to compare the performance of our classifiers with.

4 The Methods and Their Features

The following types of classifiers have been developed:

1. Lexical Context Classifiers

2. Visual Focus of Attention Classifier

3. Combined Classifiers

4. Topic and Role Extended Classifier

For each of these classifiers a large number of experiments have been performed (using Weka (Witten and Frank, 1999)) to select optimal feature sets, feature parameters, and hyper-parameters. In this section we summarize the most important findings. For a more detailed analysis refer to (op den Akker, 2009)

[3] IS1003d

[4] Note that the annotators first independently segmented the speaker's turns into dialogue act segments; then labeled them with a dialogue act type label and then labeled the dialogue acts with an addressee label. The 412 dialogues acts are those segments that both annotators identified as a dialogue act segment.

4.1 Lexical Context Classifiers

Optimal feature subsets For the offline classifier, the best results are achieved with the Multilayer Perceptron classifier, with an accuracy score of 91.89%. The corresponding best-performing feature subset is given in Table 1.

LexCont Classifier The best results for the online feature evaluation are also achieved with the Multilayer Perceptron classifier, with a score of 90.93% accuracy. The best-performing feature subset is given in Table **??**, column LexCont.

4.2 Justification of parameters

In the list of fifteen classifiers tested above, the Multilayer Perceptron has a parameter to set the number of training epochs, and the RandomForest classifier has a parameter to set the number of trees. These two parameters have both been set to an arbitrary value of 20. We could not evaluate to an optimal value for every feature set because that would explode the search space of the problem. Therefore, we now take the best-performing feature subset from the experiments with default parameters above and vary these epocs and trees variables from 1 to 50 to see how justified our choice of 20 really was.

The results can be seen in Figure 2 for the MultiLayerPerceptron classifier and in Figure 3 for the RandomForest classifier.

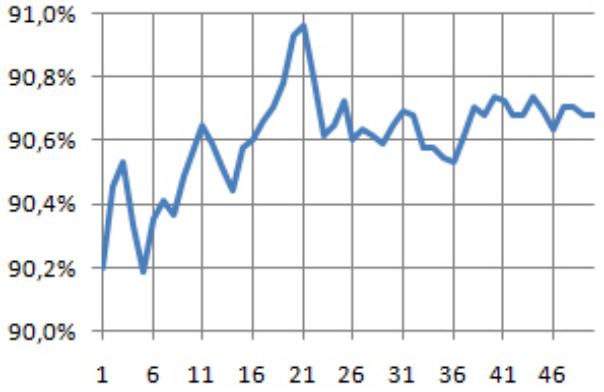

Figure 2: Results for the Online MultiLayerPerceptron classifier with best performing feature subset and varying number of traing epochs (horizontal axis).

For the MultiLayerPerceptron, the optimal result was gained with 21 training epochs with a score 90.96% accuracy versus 90.93% for 20 epochs (0.03% increase). For the RandomForest classifier, the optimal result was achieved with 16 trees, with a score of 90.39% versus 90.35% when using 20 trees (0.04% increase).

Although these findings do not definitely rule out the possibility that there is a parameter/feature-set combination that scores significantly better than any of the above results, it does show that our initial value of 20 seems reasonable and probably didn't

Type	Feature	LexCont	FoA	TopicRole
L.1	Type of the current Dialogue Act	X		
L.2	Short Dialogue Act	X		
L.3	Number of Words in the current Dialogue Act	X	X	X
L.4	Contains 1st person singular Personal Pronoun	X	X	X
L.5	Contains 1st person plural Personal Pronoun	X	X	X
L.6	Contains 2nd person singular/plural Personal Pronoun	X	X	X
L.7	Contains 3rd person singular/plural Personal Pronoun	X	X	X
C.1	Leader Role	X	X	X
C.2	Type of Previous Dialogue Act	X	X	X
C.3	Addressed History ($\eta = 6$)	X	X	X
C.4	Previous Dialogue Act addressed to me	X	X	X
C.5	Activity History ($\eta = 5$)	X	X	X
C.6	Previous Dialogue Act uttered by me	X	X	X
C.7	Speaker Diversity History ($\eta = 3$)	X	X	X
V.1	Total Time Everyone Looks at Me (Normalized)		X	
V.2	Total Time Speaker Looks at Me		X	
V.3	Total Time Speaker Looks at Me (Normalized)		X	
V.4	Number of Participants Looking At Me			
T.1	Topic			
T.2	Role			
T.3	Prior Prob of You being addressed			
T.4	Prior Prob Normalised			

Figure 1: Features used in the optimized MultilayerPerceptron Lexical and Visual FoA Classifiers, in the Combined Rule Learner and in the Topic/Role Logistic Model Tree classifier.

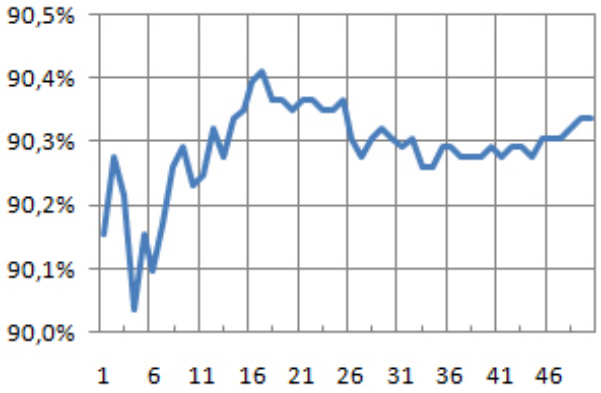

Figure 3: Results for the Online RandomForest classifier with best performing feature subset and varying number of trees in the forest (horizontal axis).

lead to a severe underestimation of the classifier results.

Table **??** gives a summary of the results of the offline- and online experiments. The last row indicates how good the results are on the scale of the Base Score (89,24%) and the Hypothesized Maximum Score (92,47%) from Section **??**.

5 Results Table

Table 4 summarizes the results for the various classifiers.

Method	Acc	Rec	Prec	F	PoM
Baseline	89,20	-	-	-	-
HMS	92,47	54,42	74,61	61,22	100
LexCont(Online)	90,93	33,10	66,02	44,09	53
LexCont(Offline)	91,85	43,35	70,18	53,59	82
VFoA	90,80	-	-	-	48
CombinedFeat	91,56	36,62	70,82	48,28	72
ClassOfResults	43,68	77,55	55,88	93,02	102
LogComb(AND)	90,24	9,86	94,23	17,85	31
LogComb(OR)	91,19	47,08	61,90	53,48	60
TopicRoleExt	92,99	-	-	-	120

Table 4: Performance values of the Methods (classifiers) discussed in this paper: Accuracy, Recall, Precision, F-measure and Percentage of Hypothezised Maximum Score (PoM).

6 How Does The Assistant Work?

At the time of writing the assistant that has been implemented is based on the simple visual focus of attention classifier. The focus of attention is inferred from the head pose and head movements of a participant in the meeting room who is being observed by a close-up camera. The real-time focus of attention module (Ba and Odobez, 2009) sends 15 times a second the coordinates of the head pose to a central dbase. The coordinates are translated into targets, objects and persons in the meeting room. For the addressing module most important are the persons and in particular the screen in the MR where the remote participant is visible. The addressing module is notified of updates who is speaking and decides if the RP is being looked at by the speaker. If the RP is not attentive (which can be detected automatically based on his recent activity) the RP is being called when he is being addressed or when the real-time keyword spotter has detected a word or phrase that occurs on the list of topics of interest of the RP. For a detailed description of the remote meeting assistant demonstrator developed in the AMIDA project refer to (Op den Akker et al., 2009). The meeting assistant allows the RP to distribute time over various tasks. The system can give a transcript of the fragment of the meeting that is of interest for the RP, so he can catch-up with the meeting if he wasn't following. The simple focus of attention based addressing module works fine. The question is now if an addressing module that uses the output of the real-time dialogue act recognizer, which uses in turn the output of the real-time speech recognizer will outperform the visual focus of attention based addressee detector.

The most explicit way of addressing is by using a vocative, the proper name of the addressed person. In small group face-to-face meetings, where people constantly pay attention and keep track of others' attentiveness to what is being said and done, this way of addressing hardly ever occurs. In remote meetings where it is often not clear to the speaker if other's pay attention people call other's name in addressing them. Other properties of the participant relevant for addressee detection include his role and his topics of interest. They can either be obtained directly from the participant when he subscribes for the meeting, or they can be recognized during an introduction round that most business meetings start with. For automatic topic detection further analysis of the meeting will be needed (see (Purver et al., 2007)). Probability tables for the conditional proba-

bilities of the chance that someone with a given role is being addressed when the talk is about a given topic, can be obtained from previous data, and could be updated on the fly during the meeting. Only if that is achieved our extended topic/role addressee classifier can be exploited fully by a live meeting assistant.

Acknowledgement

The research of the first author was performed when he was a master student at the Human Media Interaction group. The work of the second author is supported by the EU FP7 framework; the AMIDA project. We thank our collegues of the HMI research group and our partners in AMI and AMIDA for their contributions to the AMI corpus and to the remote meeting assistant.

References

Sileye Ba and Jean-Marc Odobez. 2009. Recognizing human visual focus of attention from head pose in meetings. In *IEEE Transaction on Systems, Man, and Cybernetics, Part B (Trans. SMC-B)*, volume 39, pages 16–33.

Robert Freed Bales. 1950. *Interaction Process Analysis; A Method for the Study of Small Groups*. Addison Wesley, Reading, Mass.

Jean C. Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, May.

Walter Daelemans, Véronique Hoste, Fien De Meulder, and Bart Naudts. 2003. Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, Lecture Notes in Computer Science 2837, pages 84–95, Cavtat-Dubrovnik, Croatia. Springer-Verlag.

C. Edelsky. 1981. Who's got the floor? *Language and Society*, 10(3):383–421.

Matthew Frampton, Raquel Fernndez, Patrick Ehlen, Mario Christoudias, Trevor Darrell, and Stanley Peters. 2009. Who is you? combining linguistic and gaze features to resolve second-person references in dialogue. In *Proceedings of the 12th Conference of the EACL*.

Sebastian Germesin, Tilman Becker, and Peter Poller. 2008. Determining latency for on-line dialog act classification. In *Poster Session for the 5th International Workshop on Machine Learning for Multimodal Interaction*, volume 5237.

Surabhi Gupta, John Niekrasz, Matthew Purver, and Daniel Jurafsky. 2007. Resolving "you" in multiparty dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, September.

Thomas Hain, Asmaa El Hannani, Stuart N. Wrigley, and Vincent Wan. 2008. Automatic speech recognition for scientific purposes - webasr. In *Proceedings of the international conference on spoken language processing (Interspeech 2008)*.

Natasa Jovanovic and Rieks Op den Akker. 2004. Towards automatic addressee identification in multiparty dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 89–92, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Natasa Jovanovic. 2007. *To whom it may concern : adressee identification in face-to-face meetings*. Ph.D. thesis, University of Twente.

M. Katzenmaier, R. Stiefelhagen, and T. Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of International Conference on Multimodal Interfaces (ICMI)*, pages 144–151, State College, PA.

A. Knott and P. Vlugter. 2008. Multi-agent human-machine dialogue: issues in dialogue management and referring expression semantics. *Artificial Intelligence*, 172:69–102.

Gene H. Lerner. 1996. On the place of linguistic resources in the organization of talk-in interaction: "Second person" reference in multi-party conversation. *Pragmatics*, 6(3):281–294.

Rieks Op den Akker and Mariet Theune. 2008. How do i address you? - modelling addressing behavior based on an analysis of a multi-modal corpus of conversational discourse. In *Proceedings of the AISB 2008 Symposium on Multimodal Output Generation (MOG 2008), Aberdeen, UK*, pages 10–17.

Rieks Op den Akker, Dennis Hofs, Hendri Hondorp, Harm Op den Akker, Job Zwiers, and Anton Nijholt. 2009. Engagement and floor control in hybrid meetings. In *Proceedings COST Action Prague 2008 (to appear)*, LNCS. Springer Verlag.

Harm op den Akker. 2009. On addressee detection for remote hybrid meeting settings. Technical report, University of Twente.

W.M. Post, A.H. Cremers, and O.B. Henkemans. 2004. A research environment for meeting behavior. In A. Nijholt, T. Nishida, R. Fruchter, and D. Rosenberg, editors, *Social Intelligence Design*, Enschede, The Netherlands.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, September.

Dennis Reidsma and Jean C. Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326, September.

Dennis Reidsma. 2008. *Annotations and Subjective Machines*. Ph.D. thesis, University of Twente.

Y. Takemae and S. Ozawa. 2006. Automatic addressee identification based on participants' head orientation and utterances for multiparty conversations. In *Proceedings IEEE International Conference on Multimedia*, pages 1285–1288.

David Traum. 2004. Issues in multiparty dialogues. In *Advances in Agent Communication*, pages 201–211.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1st edition, October.