Given that, Should I Respond? Contextual Addressee Estimation in Multi-Party Human-Robot Interactions

Dinesh Babu Jayagopi[1] and Jean-Marc Odobez[1,2]
[1]Idiap Research Institute [2]École Polytechnique de Fédérale de Lausanne (EPFL), Switzerland
{djaya, odobez}@idiap.ch

***Abstract*—In this paper, we investigate the task of addressee estimation in multi-party interactions. For every utterance from a human participant, the robot should know if it was being addressed or not, so as to respond and behave accordingly. To accomplish this various cues could be made use of: the most important being gaze cues of the speaker. Apart from this several other cues can act as contextual variables to improve the estimation accuracy of this task. For example, the gaze cue of other participants, and the long-term or short-term dialog context. In this paper we investigate the possibility to combine such information from diverse sources to improve the addressee estimation task. For this study, we use 11 interactions with a humanoid robot NAO[1] giving quiz to two human participants.**

***Keywords*—HRI; Social robots; Addressee estimation**

I. Introduction

Socially-interactive robots have wide-ranging applications such as assistive, educational, and medical [2]. Enabling multi-party conversations with robots is both a necessity and a research challenge in these settings. An important perceptual task towards this goal is to estimate who the current addressee is i.e. 'to whom a spoken utterance is addressed at'. This information is useful for the robot to decide automatically if he 'should' or 'should not' respond (refer to Fig. 1 for an illustration). Though gaze information about 'who the current speaker is looking at' carries valuable information, previous research has shown that this cue is not always sufficient. Therefore other contextual cues have been explored in the literature.

The problem of addressee estimation has not received much attention in the HRI literature (except [6]) as compared to Human Computer Interaction (HCI) / Virtual Avatar [8], [1], [3] or Human-human interaction literature [7], [5]. Though an early work, Katzenmaier et al. [6] used a pseudo-robot that could not move or speak, and therefore this scenario was rather artificial. More realistic scenarios have been explored in HCI literature, for example Bohus et al. explore the case of game-playing interactions with multiple interaction partners [1]. The other dimension along which the works on addressee estimation have varied has been whether manually [8], [5] or automatically extracted cues were used [6], [7], [1], [3]. Regarding the type of cues explored, gaze cues have been the primary ones. In some works, prosodic cues [3] and cues about spoken key-words [6] have been used. An issue with these features being that they are slightly scenario specific.

As compared to existing addressee literature in HRI, we estimate addressees in a realistic scenario, where a humanoid robot with significant nonverbal displays induces unconstrained nonverbal behaviors in human partners. We predict the addressees using manually annotated utterances and visual focus of attention (VFOA), to investigate the best case performance. As contextual cues, we investigate the gaze cues from the side-participant, prior information about the current activity (here the quiz), and the current dialog context.

[1]http://www.aldebaran-robotics.com

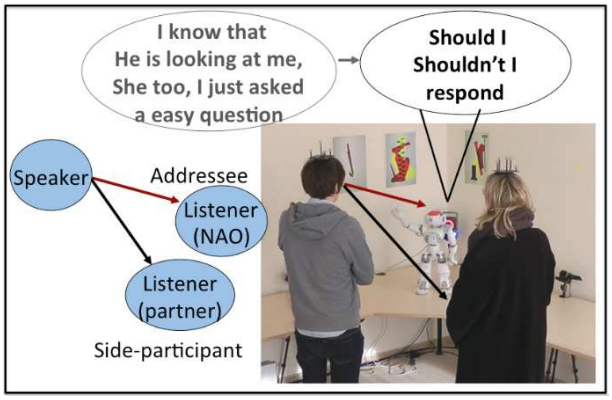

Figure 1. Overview: Addressee estimation task.

II. Addressee estimation

Setup: We use 11 interactions from the Vernissage corpus [4], where a humanoid robot NAO gives a quiz to two human participants. All participants are involved in only one interaction. The quiz consists of nine questions (or quiz episodes) in art and culture, which is same across the participant set. Some of the questions are about a set of paintings that NAO introduces to the participants before the quiz. Typically, the participants discuss among themselves before answering.

On this dataset, we have manually annotated the utterances and the addressees. An automatic method (speech detection with cross-talk suppression) was used to segment the speech and silence segments, and then an annotator revisited and adjusted the segmentation. Following the literature on addressee estimation, we define an utterance as 'a speech turn followed by silence more than 0.5 seconds'. Later, the annotator manually assigned the addressees of the utterances. Apart from this, access to NAO system data gives the start and end of all the utterances of NAO as well. There are 374 utterances of human participants in total, of which 176 were directed towards NAO, whereas 198 were directed towards a human partner (denoted Ptr henceforth). A single annotator annotated the whole dataset. In order to check the reliability, a secondary annotator carried out annotation for a subset of the dataset. The

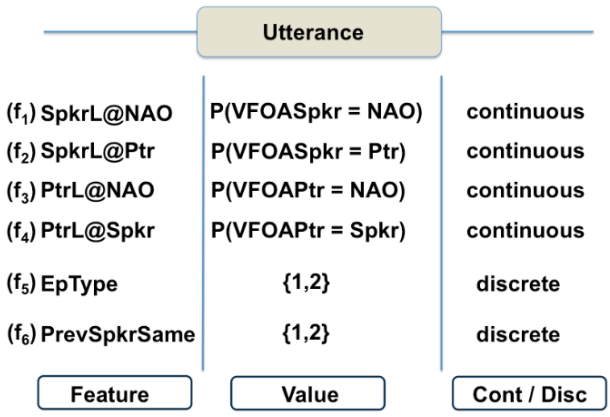

Figure 2. Addressee estimation task.

results show that Cohen's Kappa, the interannotator agreement, was 0.93, meaning they are in fact quite reliable.

We also annotated the visual focus of attention of the participants with labels: *NAO*, *Ptr*, the three paintings *Pai1*, *Pai2*, and *Pai3*, *Unfocussed*, and a catch-all-class *Don't know*.

Features: For every utterance, we defined the following features, summarized in Fig. 2: SpkrL@NAO (the proportion [%] of time when the speaker looked at NAO during an utterance), SpkrL@Ptr (% of time when the speaker looked at the partner), PtrL@NAO (% of time when the partner looked at NAO), PtrL@Spkr (% of time when the partner looked at the speaker), EpType (the difficulty of the quiz question: 1 being easy and 2 being difficult), and PrevSpkrSame (whether the previous speaker is the current speaker coded as 2 and 1 if not). In this work, we assigned the difficulty of the question manually. The difficulty level of quiz questions could also be learned over multiple sessions i.e. with experience. A question could be difficult because the listeners do not follow what the robot is saying or they follow the question but do not know the answer. We do not distinguish between these two cases in this work. While PtrL@NAO and PtrL@Spkr are contextual cues from the side-participant, EpType is a task-related long term context, and PrevSpkrSame is a short-term context about the dialog.

Classifer: We used Logistic Regression, a discriminative classifer, to estimate the addressee of an utterance. The log-ratio of the probability of addressing the partner vs NAO is a linear function of the features. The β parameters are estimated during training.

$$\log\left(\frac{P(Ad = Ptr|f_{1:N})}{P(Ad = NAO|f_{1:N})}\right) = \beta_0 + \beta_1 f_1 + \beta_2 f_2 + ... \beta_N f_N \quad (1)$$

Experimental results: The results of the addressee estimation task are given in Table 1. We did a leave-an interaction-and-quiz question-out evaluation. A baseline classifier, predicting the majority class i.e. the partner will have an error rate of 47.1%. From the single features, the best one is SpkrL@Ptr i.e the proportion of time the speaker looks at the partner, followed by SpkrL@NAO, 24.4% and 21.1% better than the baseline. Best feature combinations are also shown. SpkrL@Ptr and EpType have complementary properties giving a relative improvement of 5.7% on the error rate, w.r.t. the best single feature. We could obtain further improvement with the following four features: SpkrL@NAO, SpkrL@Ptr, PtrL@Spkr, and EpType (rel. improvement of 12% on the error rate). Finally, a five feature combination with all the above features along with PrevSpkrSame feature obtains the best accuracy of (rel. improvement of 17% on the error rate). This shows that the contextual information is complementary to the gaze cues from the speaker. With this classifer, there were 70 misclassifications and the confusion was assymmetric i.e. 39 times Ptr confused with NAO and 31 times NAO confused with Ptr.

Addressee Estimation							
f_1	f_2	f_3	f_4	f_5	f_6	β_0	Err(%).
-3.5						1.1	26.0
	3.7					-1.1	**22.7**
		-1.3				0.4	40.0
			1.8			-0.8	34.0
				0.9		-1.4	38.8
					-0.2	0.33	52.0
-2.1	2.8					-0.2	**22.2**
	3.7			0.8		-2.5	**21.4**
-2.6	2.2		1.0	1.1		-2.2	**19.8**
-2.5	2.2		1.0	1.1	0.12	-2.2	**18.8**

Table I. EXPERIMENTAL RESULTS: ADDRESSEE ESTIMATION. THE COLUMN ELEMENTS ARE THE β COEFFICIENTS ESTIMATED AND TASK ESTIMATION ERROR (%).

III. CONCLUSION

We have reconfirmed in our setting that gaze cues from the speaker is the most important feature for addressee estimation. We also show that additional contextual features from the fellow participant, short-term and long-term dialog-context features helps improve the estimation accuracy. In the future, we plan to use automatically estimated VFOA cues and check the loss in accuracy as compared to this gold standard results. We hope context can play a more important role in the case of degraded VFOA estimation. We also want to implement our addressee estimator on a real-time NAO platform and perform subjective user studies.

Acknowledgment: This research was funded by the EU HUMAVIPS project. The authors would like to thank Daniel Gatica-Perez for useful discussions.

REFERENCES

[1] D. Bohus and E. Horvitz. On the challenges and opportunities of physically situated dialog. In *2010 AAAI Fall Symposium Series*, 2010.

[2] T. Fong et al. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.

[3] H.H. Huang, N. Baba, and Y. Nakano. Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation using nonverbal information. In *Proc. ICMI*, 2011.

[4] D. Jayagopi et al. The vernissage corpus: A multimodal human-robot-interaction dataset. In *Idiap research report (Idiap-RR-33-2012)*, 2012.

[5] N. Jovanovic, R. op den Akker, and A. Nijholt. Addressee identification in face-to-face meetings. In *Proc. EACL*, 2006.

[6] M. Katzenmaier, R. Stiefelhagen, and T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. ICMI*, pages 144–151. ACM, 2004.

[7] Y. Takemae and S. Ozawa. Automatic addressee identification based on participants' head orientation and utterances for multiparty conversations. In *Proc. ICME*, pages 1285–1288. IEEE, 2006.

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