

Evaluating an Algorithm for the Generation of Multimodal Referring Expressions in a Virtual World: A Pilot Study

Werner Breitfuss¹, Ielka van der Sluis², Saturnino Luz²,
Helmut Prendinger³, and Mitsuru Ishizuka¹

¹ University of Tokyo, 7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-8656, Japan
werner@mi.ci.i.u-tokyo.ac.jp

² Trinity College Dublin
Dublin Ireland

{ielka.vandersluis, luzs}@cs.tcd.ie

³ National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan
helmut@nii.ac.jp

Abstract. This paper presents a quest for the most suitable setting and method to assess the naturalness of the output of an existing algorithm for the generation of multimodal referring expressions. For the evaluation of this algorithm a setting in Second Life was built. This paper reports on a pilot study that aimed to assess (1) the suitability of the setting and (2) the design of our evaluation method. Results show that subjects are able to discriminate between different types of referring expressions the algorithm produces. Lessons learnt in designing questionnaires are also reported.

Keywords: Embodied Conversational Agents, Automatic Behavior Generation, Generation of Multimodal Referring Expressions, Virtual Worlds.

1 Introduction

Research in Human Computer Interaction (HCI) shows an increased interest in developing interfaces that closely mimic human communication. The development of “embodied conversational agents” (ECAs) with appropriate verbal and non-verbal behavior with regard to a concrete spatial domain clearly fits this interest (e.g. [10]; [6]; [1]). Currently, the ability of an ECA to interact with human users is very limited. Interactions rely mostly on pre-scripted dialogue, whereby the manual generation of natural and convincing agent behavior is a cumbersome task.

An issue addressed in many HCI systems is that of identifying a certain object in a visual context accessible to both user and system. This can be done by an ECA that points to the object combined with a linguistic referring expression. The work presented in this paper uses one of these algorithms, which is arguably the most flexible in the sense that it can generate referring expressions that uniquely identify objects, which may include pointing gestures that vary in their precision.

Although many evaluations of ECAs have been performed, systematic studies on specific aspects of interaction are scarce (cf. [15] and [8]). This paper presents a carefully designed evaluation method to assess the quality of automatically generated multimodal referring expressions by ECAs in a virtual environment. The method is demonstrated through a pilot study conducted within a setting built in Second Life.

2 Generating Multimodal Referring Expressions

The generation of referring expressions (GRE) is a central task in Natural Language Generation (NLG), and various algorithms which automatically produce referring expressions have been developed ([19]; [20]; [9]; [11]). Most GRE algorithms assume that both speaker and addressee have access to the same information. This information can be represented by a knowledge base that contains the objects and their properties present in the domain of conversation. A typical algorithm takes as input a single object (**the target**) and a set of objects (**the distractors**) from which the target object needs to be distinguished (borrowing terminology from [7]). The task of a GRE algorithm is to determine which set of properties is needed to single out the target from the distractors.

The multimodal GRE algorithm that was taken as a starting point for the work presented in this paper, approaches GRE as a compositional task in which language and gestures are combined in a natural way and in which a pointing gesture does not always need to be precise. The algorithm co-relates speech and gesture dependent on the distance between the target referent and the pointing device. The decision to point is based on a notion of effort that is defined by a cost function. In practice, an ECA can identify an object located far away, by moving close to the object to distinguish it with a very precise pointing gesture and the use of only limited linguistic information. Alternatively, the algorithm could generate a less precise pointing gesture that also includes other objects in its scope. In this case more linguistic information has to be added to the referring expression to ensure that the object can be uniquely identified by the addressee. As an example, an ECA can say ‘the large blue desk in the back’ and accompany this description with an imprecise pointing gesture directed to the location of the desk. For a detailed description of the algorithm we refer to [21], for other multimodal GRE algorithms see (c.f., [12]; [13]; [2]).

3 Evaluating the Output of a Multimodal GRE Algorithm

3.1 Virtual Reality, Scripted Dialogue and Referring Expressions

For the evaluation design we use the virtual world Second Life (SL) for the design of evaluation experiments. It enables us to choose a specific domain of conversation in which all objects and their properties are known. This allows for complete semantic and pragmatic transparency, which is important for a content determination task like the generation of referring expressions.



Fig. 1. The agents and the furniture shop

The stage built for the experiment is a virtual furniture shop (Figure 1, right), in which two agents (Figure 1, left), a buyer and a seller interact with each other. The furniture shop contains over 43 objects, 13 of which are actually referred to in the dialogues. The other items in the shop are used as distractor objects.

In recent years, an alternative paradigm for computational work on agents has emerged ([1], [23]), with which entire dialogues are produced by one generator. Initially, scripted dialogues made heavy use of canned text, but recently this approach has been integrated with Natural Language Generation techniques, resulting in the Fully Generated Scripted Dialogue (FGSD) ([18]; [14]). FGSD allows us to produce dialogues, without implementing a full natural language interpretation module.

For our evaluation we manually prepared a dialogue, consisting of 19 utterances with 5 references to furniture items (3 singletons and 2 sets), featuring a conversation between an agent purchasing furniture for her office, and a shop-owner guiding her through the store while describing different items. The dialogue was used as a template in which the referring expressions that indicated particular pieces of furniture were varied. The referring expressions that were used to fill the slots in the dialogue were automatically produced with the algorithm discussed above.

Three types of output were implemented in three dialogues, with referring expressions ranging over two extremes with respect to linguistic and pointing information. One extreme, the imprecise version, used a version of the algorithm that generates very detailed linguistic descriptions of objects in which all the attributes of the target object were included. The pointing gestures generated to accompany these descriptions are, however, vague and the ECA can direct them from a considerable distance from the target object. The other extreme, the precise version, used another version of the algorithm that generates limited linguistic information (e.g. “this one”) combined with precise pointing gestures. Between these two extremes a ‘mixed version’ was implemented, in which 2 targets in the dialogue were identified with precise pointing gestures (1 singleton and 1 set) and 3 targets were identified with imprecise pointing gestures (2 singletons and 1 set).

3.2 Script Generation and Method of Evaluation

To control and animate our agents we use an existing gesture generation system that automatically adds nonverbal behavior to a dialogue and produces a play-able script.

We extended this system to add pointing gestures based on the algorithm. This gesture generation system consists of three steps. First the system analyzes the input text based on semantic and morphological information. Then the data used to suggest gestures which should appear along with the spoken utterance, like beats, metaphoric gestures and iconic gestures. In a third step the system filters the gestures, adds the most appropriate ones and produces a playable MPML3DSL script (c.f. [17]). The system is described in detail in ([3]; [4]). This system was extended to generate three levels of pointing gestures, precise, very precise and imprecise, as suggested in [21]. This involved: (1) object identification by the ECA, (2) detection of the position of these objects in relation to the ECA to select the right direction for the pointing gesture, and (3) choice of pointing gesture to be displayed.

To evaluate our setting, subjects were first introduced to the environment and asked to complete a questionnaire designed to obtain general judgments on the setting. They were then instructed to view and judge three presentations. Finally subjects were asked to compare the three presentations. Three kinds of questionnaires were used, which we will refer to as A, B and C. A aimed at obtaining a baseline and contained ten questions about the agents, the setting and the conversation plus some general questions about the subject's background. Some questions were open and some used a Likert scale that ranged from one ("strongly agree") to seven ("strongly disagree"). B was used for evaluating the three presentations and consisted of four sections addressing the interaction between the agents, the agents themselves, their role-play and the conversation. In total there were twenty-one questions. For all questions, questionnaire B used the same Likert scale as A. C compares the three presentations. Possible answers were ('Dialogue 1', 'Dialogue 2', 'Dialogue 3', 'Don't know', 'Now Difference'). All questionnaires allowed subjects to enter free comments.

4 The Pilot Study, Results and Discussion

Ten people participated in the study, all native speakers of English (4 males and 6 females). Half of them were familiar with virtual worlds, but no one visited SL regularly. After entering the experiment room individually they received a written introduction to the experiment. First, subjects were asked to familiarize themselves with the environment by moving and looking around in the shop. When ready, subjects were told to sit down at a predefined location and watch the three life versions of the dialogue presented in a random order. For the evaluation, we used the method described in 3.2, showing them three different settings and letting them fill out questionnaires A, B and C. Each Sessions lasted around 45 minutes per person. Subjects were not paid and participated voluntarily.

In general, the data obtained with A showed that subjects were content with their view of the stage, found the presentation easy to follow and enjoyable. With respect to the characters, they rated the female voice as more pleasant, and clearer than the male voice. The outcomes of B showed that the ECAs were perceived as friendly, trustworthy and talkative, and that the conversations were easy to follow. It proved difficult for the subjects to judge the naturalness of the acting and conversation.

Table 1. Means for questions targeting movement and talkativeness of the seller (standard deviations in brackets)

Question	Precise	Imprecise	Mixed
The male agent moved a lot	2.2 (1.398)	6.5 (0.707)	3.1 (1.663)
The male agent was talkative	2.6 (1.173)	3.0 (1.490)	2.7 (0.674)
It was clear which item the male agent was talking about	2.6 (1.080)	3.8 (1.813)	3.1 (1.100)

Table 1 shows that in the precise presentation, (i.e. precise pointing gestures + “this one”), subjects found that the furniture seller moved around a lot and that the conversation was easy to follow. A similar result was found for the mixed presentations. In contrast, in the Imprecise presentation, (imprecise pointing + detailed linguistic descriptions), subjects judged that the agent did not move around a lot and that it was not so clear which item was under discussion.

Table 2 presents the highlights from questionnaire 3, which asked subjects to compare the three presentations. For this small set of subjects it seems that the precise and the mixed version were preferred. Surprisingly, eight out of ten subjects found that there was no difference in how much the furniture seller talked while the presentations, while the imprecise dialogue contained five referring expressions of the type ‘the large blue desk in the back of the shop’ and the precise version used ‘this one’ in all these cases. Note also that none of the subjects used the answer ‘I don’t know’, all were able to remember and judge accordingly.

Table 2. Results of a comparison between the three settings

Question	Precise	Imprecise	Mixed	No Diff
The Seller acted more naturally in	5	1	4	0
If I were a buyer I would prefer Seller	5	1	4	0
The Seller moved more in	8	0	2	0
The Seller talked more in	1	1	0	8
The conversation most easy to follow in	4	1	4	1
The conversation was most naturally in	3	1	4	2

The results of the study show that subjects were able to perceive differences between the three types of GRE outputs used in the presentations, each one using a different kind of referring expressions. The study also gave us a number of pointers to improve the setup of the study. It turned out that in the questionnaires some of the questions were not very useful. In particular, the questions where subjects had to judge the naturalness of the conversation and the characters seemed problematic and need to be rephrased. This is not surprising as the setting is highly artificial (cf. [16]). Probably other types of evaluation (cf. [6]; [12], [22]) will be necessary (e.g. performance, behavior, preference etc.) to evaluate multimodal GRE algorithms.

Another issue was that some subjects found it difficult to tell which of the three furniture sellers they preferred. This is interesting because it addresses both the physical distance from which they were asked to view the presentation as well as our use of scripted dialogue. In our setting subjects were watching a play from a stand of the type that is used in theatres and arenas. Apart from the fact that there was a physical

distance between the subjects and what was happening on the stage, subjects had no actual interest in the furniture itself. As a result, it appeared that subjects had different preferences dependent on whether the goal was to comprehend the dialogue, or whether they were asked to imagine themselves in the shoes of the customer.

Some pilot results related to technical problems in SL. For instance, it was not possible to turn the characters in a particular direction other than towards each other. Also movement was still very imprecise, which makes it difficult to be sure that the agent walks precisely the predefined route in the shop. An issue that remains open is the TTS system, which sometimes rendered the prosody somewhat unnatural.

5 Conclusion and Future Directions

In this paper we presented our approach to evaluate an existing algorithm for the generation of multimodal referring expressions embedded in an automatic gesture generation system. We employed two ECAs acting as a seller and a buyer in a virtual furniture shop. The setting aimed to test three types of referring behavior by the seller in which the precision of the pointing gestures and the linguistic descriptions were varied. A pilot study was carried out to test the setting and the methods used. The results of this study gave us some useful feedback to improve the current setup.

With respect to the questionnaires, especially the questions that aimed at the naturalness of the agents' behaviour and the conversation need to be rephrased. Other changes in the experimental setup will be of a presentational nature. In future studies we will use video of our SL presentations instead of displaying the scripts live as done in the pilot study. In the video, the camera can follow the agents through the furniture store, possibly reducing the 'overhearer' effect that is inherent to FGSD. In addition, we plan to remove the non-deictic gestures from the utterances that contain point gestures. Initially, these non deictic gestures were included to increase the naturalness of the characters. However, the pilot has shown that these gestures can have a distracting effect on the viewer. In the near future a cross-cultural study is planned, that focuses on differences and similarities in the perception of multimodal referring expressions between subjects in Dublin and in Tokyo.

Acknowledgements

This research was part-funded by Science Foundation Ireland under the CNGL grant.

References

1. André, E., Rist, T., Van Mulken, S., Klesen, M., Baldes, S.: The Automated Design of Believable Dialogues for Animated Presentation Teams. In: Cassell, J., Prevost, S., Sullivan, J., Churchill, E. (eds.) *Embodied Conversational Agents*. The MIT Press, Cambridge (2000)
2. Andre, E., Rist, T.: Coping with temporal constraints in multimedia presentation. In: Proc. of the 13th Conference of the AAAI, pp. 142–147 (1996)
3. Breitfuss, W., Prendinger, H., Ishizuka, M.: Automatic generation of gaze and gestures for dialogues between embodied conversational agents. *Int'l J of Semantic Computing* 2(1), 71–90 (2008)

4. Breitfuss, W., Prendinger, H., Ishizuka, M.: Automatic generation of conversational behavior for multiple embodied virtual characters: The rules and models behind our system. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 472–473. Springer, Heidelberg (2008)
5. Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., Oberlander, J.: Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In: Proc. of ENLG 2009 (2009)
6. Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vauccelle, C., Vilhjalmsson, H.: MACK: Media lab Autonomous Conversational Kiosk. In: Proc. of the IMAGINA 2002 (2002)
7. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 18, 233–263 (1995)
8. Dehn, D., Van Mulken, S.: The impact of animated interface agents: a review of empirical research. *Int. J. Human-Computer Studies* 52, 1–22 (2000)
9. Jordan, P., Walker, M.: Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research* 24, 157–194 (2005)
10. Kopp, S., Jung, B., Lessmann, N., Wachsmuth, I.: Max - A Multimodal Assistant in Virtual Reality Construction. *KI Künstliche Intelligenz* 4(3), 11–17 (2003)
11. Krahmer, E., Van Erk, S., Verleg, A.: Graph-based generation of referring Expressions. *Computational Linguistics* 29(1), 53–72 (2003)
12. Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., Wachsmuth, I.: Deictic object reference in task-oriented dialogue. In: Rickheit, G., Wachsmuth, I. (eds.) *Situated Communication*, pp. 155–207 (2006)
13. Lester, J., Voerman, J., Towns, S., Callaway, C.: Deictic believability: Coordinating gesture, locomotion and speech in lifelike pedagogical agents. *Applied Artificial Intelligence* 13(4-5), 383–414 (1997)
14. Piwek, P.: Presenting Arguments as Fictive Dialogue. In: Grasso, F., Green, N., Kibble, R., Reed, C. (eds.) *Proc. of 8th of the CMNA* (2008)
15. Ruttkay, Z., Pelachaud, C.: From Brows to Trust: Evaluating Embodied Conversational Agents. Kluwer, Dordrecht (2004)
16. Slater, M.: How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence-Teleoperators and Virtual Environments* 13(4), 484–493 (2004)
17. Ullrich, S., Prendinger, H., Ishizuka, M.: MPML3D: Agent authoring language for virtual worlds. In: Proc. of the Int'l Conf on Advances in Computer Entertainment Technology (ACE 2008), pp. 134–137. ACM Press, New York (2008)
18. Van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schroeder, M., Baumann, S.: Full Generated Scripted Dialogue for Embodied Agents. *AI Journal* (2008)
19. Van Deemter, K., Krahmer, E.: Graphs and booleans. In: Bunt, H., Muskens, R. (eds.) *Computing Meaning*, vol. 3. Kluwer Academic Publishers, Dordrecht (2006)
20. Van Deemter, K.: Generating Referring Expressions that Involve Gradable Properties. *Computational Linguistics* 32(2), 195–222 (2006)
21. Van der Sluis, I., Krahmer, E.: Generating Multimodal Referring Expressions. *Discourse Processes*. In: Piwek, P., Kuhnlein, P. (eds.) *Special Issue on Dialogue Modelling: Computational and Empirical Approaches* vol. 44(3), pp. 145–174 (2007)
22. Van der Sluis, I., Krahmer, E.: The Influence of Target Size and Distance on the Production of Speech and Gesture in Multimodal Referring Expressions. In: Proc. of the ICSLP 2004 (2004)
23. Williams, S., Piwek, P., Power, R.: Generating monologue and dialogue to present personalised medical information to patients. In: Proc. of the 11th European Workshop on Natural Language Generation, pp. 167–170 (2007)