Reference Resolution in Multi-modal Interaction: Position paper

Anton Nijholt

Dept of Computer Science
University of Twente, PO Box 217
7500 AE Enschede, the Netherlands
anijholt@cs.utwente.nl

Abstract: *In this position paper we present our research on multimodal interaction in and with virtual environments. The aim of this presentation is to emphasize the necessity to spend more research on reference resolution in multimodal contexts. In multi-modal interaction the human conversational partner can apply more than one modality in conveying his or her message to the environment in which a computer detects and interprets signals from different modalities. We show some naturally arising problems and how they are treated for different contexts. No generally applicable solutions are given.*

Keywords: reference resolution, multi-modal interaction, virtual reality

1. Introduction

We present our research on multimodal interaction in and with virtual environments. In multi-modal interaction the human conversational partner can apply more than one modality in conveying his or her message to the environment in which a computer detects and interprets signals from different modalities. Clearly, a user chooses between modalities or combinations of modalities to convey his or her message. In speech or keyboard natural language input references will be made to previous inter-actions, assumed shared knowledge (common-sense knowledge of domain knowledge) and non-verbal display of information in an environment controlled by computers. These computers themselves make references to knowledge implicitly assumed because of context and history or because information is somewhere visible on the screen or the environment for the human partner. We present some environ-ments where we have been confronted with the necessity to model references to verbal and nonverbal display of information, both by computer and human conversational partner. Examples of problems are presented. Although in the literature attempts to handle references to nonverbal informa-tion display have been reported, the methods intro-duced hardly allow their use in contexts outside their specific applications. We show some naturally arising problems but do not give general solutions. Rather we have decided to perform more detailed research on reference resolution in uni-modal contexts to obtain methods generalizable to multi-modal contexts.

In language, anaphorically used words are words that are referring back to something that was earlier mentioned or that is known because of the discourse situation and/or the text as it is read or heard. The anaphorically used word is called 'the anaphor', the text to which it refers 'the antecedent'. The extra-lingual entity they corefer to is called the referent. Deictic used words are words that refer to something directly or indirectly present in the situation. The word is then used instead of a gesture, or the utterance of the word is accompanied by gesturing. Anaphora resolution is the process of determining the antecedent of an anaphor. The antecedent can be in the same sentence as the anaphor, or in another sentence. For deictically used words, used in the interaction with a computer or smart environment, it has to be determined where a user refers too.

There is certainly a lot of linguistic research that has been done on the resolution of anaphoric and deictic references in texts and dialogues [1]. None of the methods work perfect, nevertheless theory and terminology are useful when looking at multimodal contexts where we have sequences of multimodal

interactions and references to a visualized environment rather than to sentences in text or utterances in dialogue. We hope to be able to extend results from linguistics to our multimodal contexts later.

In this position paper we confine ourselves to the presentation of several resolution problems in some environments that have been introduced by ourselves and where until now the necessary reference resolution methods are either extremely simple or ad hoc. This will be done in section 2. As mentioned, more research is needed to have a more comprehensive approach to the problems that emerge here. In section 3 we mention related and future research.

2. Some of our Multimodal Contexts

Introduction

In previous papers we have discussed our work on multimodal interactions in virtual environments [5,8,9]. Here we review them with an emphasis on the necessity to be able to resolve references involving multiple modalities. We discuss our virtual theatre environment including an embodied information and transaction agent (Karin) and a navigation assistant. These are certainly not the only environments where we need to be able to handle multimodal references. In [5], for instance, we have introduced an embodied educational agent that knows how to solve the problem of the Towers of Hanoi and monitors a student who uses the mouse to manipulate the towers and natural language to communicate with the agent. Rather natural references are contained in questions like: "What should I do now?", "Is this allowed?", "Should I do the red block now?" (in a context containing different red blocks), etc. In [12] our research initiatives are mentioned that deal with the choice of output modalities when an embodied agent has to convey a message. Obviously, in order to do this in a natural way references should be generated rather than resolved. In development in our group is also a virtual environment where we have a piano teacher that guides and monitors a student learning to play the piano [3]. Also in this case, both interpretation and generation of references to different modalities have to be done.

Karin: An Embodied Conversational Agent

One of our main research environments is the virtual theatre environment. It serves as a laboratory for agent research, embodied conversational agents and multimodal interactions. The environment is the virtual equivalent of an existing theatre in our hometown. The theatre has different floors, a main performance hall, a lounge, stairs, etc. A receptionist in the form of an embodied agent and called Karin is available to answer questions about performances, performers, available seats and can make reservations. Questions can be asked using the keyboard and natural language. The receptionist has a database available with the actual theatre performances for the current year. A text-to-speech synthesis system is used to mouth her answers to the visitor. The environment has been built using VRML. Visitors can walk around in the environment, visit the different locations and the receptionist. In Nijholt & Hulstijn [9] a rather comprehensive survey of the environment is given. See also [6] for a more recent paper on this continuously changing environment.

Clearly, communication situated in a visible or otherwise observable (virtual) shared environment allows the communicating partners to support their linguistic communicative acts by other means of reference to objects like gazing or pointing. Introducing this multi-modal support for language communication may help the agents to understand each other but it also introduces some new and challenging problems as well. One of these is the problem of coreferencing to shared visible objects. An agent interpreting the phrase "that door" will assume that it refers to some visible object in the environment and that it shares the visibility of this object with the agent uttering the phrase. The 'geometrical' virtual environment (described in VRML code or in some other virtual modelling language) must be accessible on an abstract, conceptual and linguistic level as well. The agent should somehow be able to know what object the user points at even in case it is not in direct view of the agent and it must be able to match this pointing-type reference with the linguistic reference ("that door").

An other multimodality and reference resolution issue is the following. Karin decides to present a table on the screen if there are too many

performances she has to read out. Clearly, when there are too many performances that satisfy the request we can not expect that the user still remembers details about the first performance after Karin has read out all the information about four or five performances. Therefore we decided to embed Karin and her information desk in a windows environment which allows us to present information in tables with clickable items and pop-up menus of frequently asked questions. The dialogue system can interpret and generate references to items in this table. A question like: "Please give me more information about the third performance", making a reference to the third item in the table of available performances, will be understood correctly. Instead one could also click on one of the frequently asked questions.

Recently we also added gaze behavior, in particular behavior that give cues for turntaking, to Karin. From experiments we know that this nonverbal behavior allows more efficient interaction between Karin and visitor [4]. Although it will not be that common that users will make explicit references to this behavior, we may nevertheless assume that implicitly a user may assume that when she makes an implicit reference to where Karin is looking at, she should understand this reference.

From the visitor's point of view the need of an other agent emerged. To whom do we address our questions about the environment itself? To whom do we address our questions about how to continue, where to find other visitors or where to find domain-related information? For that reason, in addition to Karin who knows about theatre performances, we introduced a navigation agent that knows about the geography of the building.

Navigation Assistance in a Virtual Environment

In order to investigate the problems and solutions of communicating in natural language with a navigation agent in a virtual reality environment we introduced a version of our virtual theatre environment discussed above where we have added a window to the virtual reality browser which displays a detailed floor map with positions of different objects and locations and also possible routes between them.

Associated with the map a natural language accessible navigation agent was introduced. The visitor can ask questions, give commands and provide information when prompted by this navigation agent. This is done by typing natural language utterances and by moving the mouse pointer over the map to locations and objects the user is interested in. On the map the user can find the performance halls, the lounges and bars, selling points, information desks and other interesting locations and objects. The current position of the visitor in the virtual environment is marked on the map. While moving in virtual reality the visitor can check her position on this floor map. When using the mouse to point at a position on the map references can be made by user (in natural language) and system to the object or location pointed at.

As mentioned, this navigation agent has to be accessed by natural language. We have annotated a small corpus of example user utterances that appear in navigation dialogues. On the one hand we have complete questions and commands. On the other hand we have also short phrases that are given by the user in reply to a clarifying question of the navigation agent. An example of a question is: "What is this?" while pointing at an object on the map, or "Is there an entrance for wheel chairs?" Examples of commands are "Bring me there." or "Bring me to the information desk." Examples of short phrases are "No, that one." or "Karin." From the annotated corpus a grammar was induced and a unification-type parser for Dutch can be used to parse these utterances into feature structures.

Three agents communicate to fill in missing information in the feature structure (when the information given by the user in question, answer or command is not yet complete) and to determine the action that has to be undertaken (answering the question, prompting for clarification or missing information, displaying a route on the map or guiding the user in virtual reality to a certain position). This is done by the navigation agent in co-operation with the dialogue manager and the virtual reality display agent. The latter can 'talk' to the virtual reality browser using its EAI (External Authoring Interface) to retrieve the current position of the visitor. Not yet implemented is the possibility that not only the position but also what is in the eyesight of the visitor in virtual reality can be retrieved. This will allow more correct reference solving in the dialogue.

The natural language interaction between the navigation agent and the user allows the user to play an active role in the process of navigation. The navigation agent is reactive: the visitor can ask about existing locations in the theatre. The user can type a question like "Where do I find the coffee bar?" or a command like "Bring me to the coffee bar, please" and the system can react by answering the question in two ways: it can indicate the place on a map, or it can navigate the visitor's viewpoint through the environment along a route to this destination. In order to do so the agent needs to know:

- how objects in the inventory of the environment are referred to by means of a natural language expression ('the coffee bar')

- how the actions it can perform can be referred to by means of natural language ('bring').

- what communicative act the user is performing by his utterance (is the user asking for information, or asking the system to do something)

Because the visitors will be aware of the visual context, in natural language interaction they will probably use references to that context. Hence natural language understanding cannot be seen as an isolated activity that is carried out by some language-processing module that is independent of the virtual environment. Rather, the interpretation of natural language sentences is coupled to what is seen in the virtual world at the moment the sentence is uttered by the user. For instance, our advisor might suggest going through "the door"', in case exactly one door is visible. The use of words like 'this', 'that', 'there', 'here' (deictic references) can only be understood by a natural language capable agent if this agent is able to recognize what is in the neighborhood of the user, or what can be seen by the user. Also the agent should be able to recognize objects that have recently been referred to in the dialogue and that could have been used in the utterance at that particular position. Such objects are stored in a focus list. We illustrate this by an example dialogue:

- User action: "Where do I find the coffee bar?"

- System action: shows the coffee bar on a map

- User action: "Please, bring me there."

- System action: navigates to the coffee bar.

Since the system has been able to solve the coffee bar reference, and stored the information in the focus list, it can attach the indexical "there" to the object referred to earlier in the dialogue. If the user asked the way to the coffee bar and then tries to find his way through the environment, the navigation agent should remember what the user is looking for so he can interrupt if he notices that the user navigates in a wrong direction: "you should go left here, if you look for the coffee bar". In case the reference problem could not be solved, the system can ask for more information. When the visitor's utterance is about performances, the navigation agent may attempt to contact Karin, the information and transaction agent.

3. Related and Future Work

The aim of this paper was to present the motivation of our current work on anaphora resolution in the context of multimodal interaction as necessary for allowing natural interaction with systems that exploit visual (2D, 3D and virtual reality) and acoustic media in the interface. Speech and language are often the starting point when looking at modelling multimodal interaction. Syntax and semantics of language have been studied for a long time and formalisms have been introduced to represent syntax and semantics. A well-known example in practical natural language processing systems are feature structures.

Feature structures allow the representation of grammatical structure, actions, and, among others, spatial and temporal relations. In many systems, starting with the work of Cohen and Oviatt on systems like QuickSet, a multimodal interface for designing military simulations, integration of various input modalities is done at the level of (typed, complex, multidimensional) feature structures by unification. See e.g. [11], [2], [13] and the navigation agent that we discussed in section 2.

A more fundamental approach can be found in [7], a paper devoted to the resolution of multimodal referential expressions, 'constructive type' theory is used to represent the user's utterance and the context (domain, dialogue history and shared beliefs). The user may ask questions in natural language and may manipulate objects in the domain using the mouse. An 'assistant' can answer the user's questions or on

request perform manipulations on objects. The general outline of the resolution algorithm discussed there satisfies the approach discussed in section 2 for texts: select possible referents, apply filters (see the earlier discussion on raising and lowering salience of antecedents and/or results of unification), order the candidates by salience and evaluate the result. If necessary the system can ask a clarificatory question or report that it does not understand the user.

An interesting and more comprehensive approach can be found in Thorisson's Gandalf system. Gandalf, an embodied conversational agent based on the Ymir architecture [13], is an example of a system where multimodal integration of information is done at several levels. Here, speech content, attentional prosody, pointing gestures and gaze and head direction during a dialogue are integrated at four levels and different actions (e.g. a decision on turntaking, reference resolution, topic shift, asking a clarifying question) can be decided upon at these different levels.

We conclude by mentioning that in the near future more than now users will communicate with devices (hand-helds, wearables, smart environments, etc.) where communication and so, also references, are context-dependent and multimodal. Without multi-modal reference resolution methods no natural interaction will be possible.

References

[1] R. op den Akker, M. Hospers, E. Kroezen, A. Nijholt and D. Lie. A rule-based reference resolution method for Dutch discourse analysis. In: Proc. *2002 Symp. on Reference Resolution in Natural Language Processing*, University of Alicante, Spain, June, 2002, to appear.

[2] Bin, Xiao, Pu Jiantao & Dong Shihai. Multimodal integration using complex feature sets. In: Proc. *Advances in multimodal interfaces – ICMI 2000,* Beijing, 2000.

[3] A. Broersen & A. Nijholt. Developing a Virtual Piano Playing environment. In: Proc. *Intern. Conf. On Advanced Learning Technologies (ICALT '02)*, Kazan, 2002, to appear.

[4] I. van Es, D. Heylen, B. van Dijk & A. Nijholt. Gaze Behavior of Talking Faces Makes a Difference. Proc. *CHI 2002: Changing the World, Changing Ourselves*, ISBN 1-58113-454-1, L. Terveen & D. Wixon (eds.), Minneapolis, April 2002, 734-735.

[5] M. Evers & A. Nijholt. Jacob - an animated instruction agent for virtual reality. In: *Advances in Multimodal Interfaces - ICMI 2000*. 3rd Intern. Conf. on Multimodal Interfaces, Springer, Berlin, 2000, 526-533.

[6] D. Heylen, A. Nijholt & M. Poel. Embodied agents in virtual environments: The Aveiro project. In: Electronic Proc. *European Symp. on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, Tenerife, Spain, December 2001.

[7] L. Kievit and P. Piwek. Multimodal cooperative resolution of referential expressions in the DenK system. In: Proc. *CMC '98*, H. Bunt & R.J. Beun (eds.), LNAI, Springer, 1999.

[8] J. van Luin, A. Nijholt & R. op den Akker. Natural Language Navigation Support in Virtual Reality. In: Proc. *Intern. Conf. on Augmented, Virtual Environments and Three-dimensional Imaging (ICAV3D)*, V. Giagourta & M.G. Strintzis (eds.), Mykonos, Greece, 2001, 263-266.

[9] A. Nijholt & J. Hulstijn. Multimodal interactions with agents in virtual worlds. In *Future directions for Intelligent Information Systems and Information Science*, Physica-Verlag: Studies in Fuzziness and Soft Computing, 2000, 148-173.

[10] S. Oviatt & P. Cohen. Multimodal Interfaces That Process What Comes Naturally. *Communications of the ACM*, Vol. 43, No 3, 45-53, 2000.

[11] P. Paggio & B. Jongejan. Representing multimodal input in a unification-system: the Staging project. In: Proc. *Integrating Information from Different Channels in Multi-Media Contexts.* ESSLLI, 2000.

[12] M. Theune. ANGELICA: Choice of output modality in an embodied agent. Proceedings of the International Workshop on *Information Presentation and Natural Multimodal Dialogue (IPNMD-2001).* N.O. Bernsen & O. Stock (eds.), Verona, Italy, 14-15 December 2001, 89-93.

[13] K.R. Thórisson. Real-Time Decision Making in Multimodal Face to Face Communication. *Second ACM Intern. Conf. on Autonomous Agents*, Minneapolis, Minnesota, May 1998, 16-23.