Multimodal Context Management within Intelligent Rooms

Petra Gieselmann, Hartwig Holzapfel

Interactive System Labs,
Universität Karlsruhe, Am Fasanengarten 5
76131 Karlsruhe, Germany
petra@ira.uka.de, hartwig@ira.uka.de

Abstract

In this paper we present our current work on multimodal context management within an intelligent room. Spoken user utterances and interaction with an augmented table are used as supplementing modalities. At any time the user is free to select one of the two modalities. Both of them share the same context model which is created out of the preceding user utterances and interactions on the table. In this way, the context model can be used to restrict the search space of the speech recognizer to improve the recognition performance, as far as the dialogue part is concerned. In addition, the augmented table uses the context model to update the current situation on the table so that the user can always see the current context there regardless which modality he used.

1. Introduction

Smart room environments lead to new challenges the next generation dialogue systems have to cope with [1, 2, 3], resulting from the physical environment shared by the user and the system, the situated and context-dependent communication, the changing environment, the multimodal interaction, etc. Here, we want to concentrate on context-dependent communication and multimodal interaction in multimodal information systems including augmented reality to provide the user with additional information.

Within the European Union funded project FAME (Facilitating Agent for Multicultural Exchange), we developed an intelligent meeting room which serves as an information butler to assist the users by providing information on recorded lectures, retrieving information about these lectures, displaying some pictures, video streams or other information about the lectures or speakers of a conference, switching on or off different lamps, etc.

Therefore, the whole FAME system can be divided into two different parts: The offline part which consists of recording the conference lectures, transcribing and indexing these lectures and storing all this information in the database. On the other hand, there is the online part which consists of the actual intelligent room with the following components:

- an augmented table [4] to display the time table of the conference, the names of the speakers and topics of their talks, as you can see in figure 1. In the center, you can see the time schedule of the conference which can be rotated at the user's best convenience. Each lecture is represented by a slice that shows place, time, and title of the lecture as well as five blue circles to access detailed information about the speaker of the lecture, the topics addressed there, and the slides related to this lecture. Furthermore, the lecture circle can be selected to get video segments of the live lecture, and the testimonies circle is used to report a testimony and also to access testimonies reported by other conference attendees. The left part of the table renders all the topic names addressed in the conference and the right one shows all the speakers.
- speech recognizers with close-talk voice input for different languages (English, Spanish, German) to recognize the utterances of the user
- a multilingual dialogue manager to understand the user utterances and trigger the corresponding services
- an augmented wall to display information about speakers, lectures or also some parts of a lecture covering a specific topic.
- a topic detector to detect the topics about which the users are talking. For this, a distant speech microphone is located in the middle of the table to capture speech.
- an information retrieval component to find indexed parts of the lectures matching the user queries.
- a testimony recorder with a camera and a microphone so that the users can record testimonies about lectures they listened to.
- a database with all the information of this conference and its speakers, talks and topics. Here, not only the videos, but also the slides of the talks are stored already indexed for information retrieval.

All the different components communicate by means of an Open Agent Architecture (OAA) [5] which is an OpenSource middleware. It is used to connect the different components in the FAME project, which may be hosted on different computers, by sending messages over a central facilitator.

This paper deals with multimodal context management within the dialogue management of an intelligent room. Mechanisms for improving recognition performance by means of the current context are also explored. Section two gives an overview of related work about context management and the use of subgrammars in a given context to improve speech recognition performance. Section three deals with our dialogue manager and its features and context management issues within this framework, such as multi-user and multi-lingual context management and dynamic information from databases. Section four deals with the multimodal integration. After a short example, the extended use of subgrammars is explained in detail and finally multimodal generation and multi-lingual information retrieval is explored. Section five deals with the evaluation of the whole framework. Section six gives a conclusion and an outlook on future work.

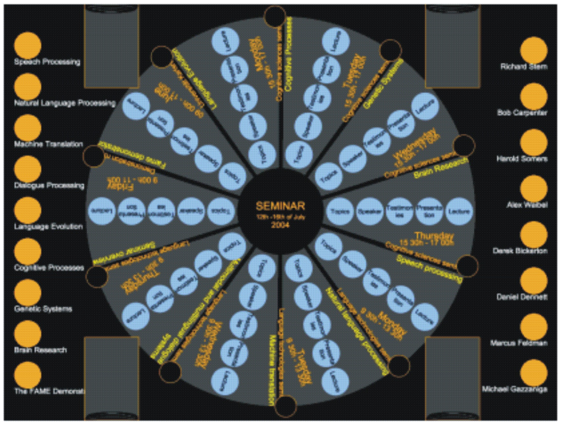

Figure 1: The Augmented Table

2. Related Work

As explained by [6], there are different possibilities how multiple modalities can be combined:

- Sequential Multimodal Input: This is the simplest case of multimodality where at every dialogue state only one modality is possible. Inputs from different modalities are always interpreted separately.
- Uncoordinated Simultaneous Multimodal Input: All modalities can be used at any time, but they are always interpreted sequentially.
- Coordinated Simultaneous Multimodal Input: Here multimodal fusion takes place.

The system described here implements an uncoordinated simultaneous multimodal input facility. Sequential multimodal input would restrict the user to one modality at a time, and multimodal fusion is not adequate for the task. Due to the implementation of uncoordinated simultaneous multimodal input, the user can decide at any time whether he prefers to select something on the intelligent table or utter spoken commands to the system. Coordinated simultaneous multimodal input, such as fusion of deictic pointing gestures and speech, has already been described in [7, 8].

2.1. Context Modeling within Multimodal Systems

Landragin and his colleagues use a multimodal dialogue history modeling for context management [9]. The main purpose of this dialogue history is to keep track of the referring actions and to resolve elliptical constructions thereafter. They use a multimodal interface language which is a uniform representation format for the histories of the different modalities. In this way, the information is synchronized between the different modalities so that multimodal fusion on a semantic level is possible.

Already in 1996, Seneff and her colleagues created a first discourse model for multimodal multi-user environments [10]. A history of previously mentioned objects is used to resolve underspecification there, too.

2.2. Using Subgrammars to Improve the Recognition Performance

As already explained by different researchers [11, 12, 13, 14], subgrammars can be used in speech recognizers to improve the recognition performance. The context determines which grammar rules are penalized in a given context and which ones are preferred. In this way, the recognition can be improved substantially.

3. Multimodal Context Management within an Intelligent Room

3.1. Dialogue Management

For dialogue management we use the TAPAS dialogue tools collection which is based on the approaches of the language and domain independent dialogue manager ARIADNE [15]. The dialogue manager uses typed feature structures [16] to represent semantic input and discourse information. A context-free grammar is used to parse the user utterance. The grammar is enhanced by information from the ontology defining all the objects, tasks and properties about which the user can talk. After parsing, the parse tree is converted into a semantic representation and added to the current discourse. If all the necessary information to accomplish a goal is available in discourse, the dialogue system calls the corresponding service. Otherwise, the dialogue manager generates clarification questions to request this information.

3.2. Context Management

Whenever a user selects something on the table respectively by means of some spoken sentences, a context is created which forms the system's expectations for the following user interactions. The context is shared by the different modalities to assure that the user can alternately select one of the possible modalities. The context representation is very simple: The system keeps a list of all active menus and a timestamp of the most recent access.

Therefore, a new context affects the representation on the table so that every active menu corresponds to a graphical representation on the table, and has been activated either by speech or by actions on the table. Speech recognition is affected by activating subgrammars that model utterances the user can say within the current context. Since the system can use multiple languages, the corresponding subgrammars are activated for each language. Semantic interpretation is affected by the context to be able to interpret elliptical expressions. Each subgrammar that is activated by the speech recognizer is associated with a conversion rule for contextual semantic interpretation by the system. The discourse is affected in a way that it gets informed which menus are activated, so that new incoming information can be unified with the one of the existing threads, supporting multiple users who might even speak different languages.

3.2.1. Multi-user Context Management

The system is designed to be a collaborative multi-user interaction scenario. The multimodal context provides various interaction tracks to support the interaction of different users. A single track is not necessarily restricted to one user. Referring to the context model, as described briefly above, one can easily see that sharing the context between multiple users is straightforward due to its simplicity.

Different users can pick up and continue a started track with graphical interaction or speech, since the table and the dialogue manager share the knowledge of opened, active menus on the table. The expectation model of the dialogue manager allows new speech input to be semantically interpreted in the given multimodal context, to select elements from a menu or open new sub-menus.

3.2.2. Multi-lingual Context Management

While the context is shared between multiple users, these users can also use different languages to communicate with the system. The context itself is based on language-independent semantic representations which are created by language-dependent grammars within the dialogue manager [17]. Each language is covered by a speech recognizer that uses language-dependent grammars as its language model. The semantic parser of the dialogue manager then creates language-independent semantic representations from the parse tree that is sent by the speech recognizer. The system can then interpret the language independent representation within its context.

Furthermore, all the Spanish titles of the talks and their English translations can be found in the database and in this way also translated into English. This is very useful especially for Spanish native speakers on a conference who can understand English quite well, but only speak a little bit. Therefore, the generated output is always in English so that every user in the intelligent room can understand it.

3.2.3. Dynamic Information from Databases

All elements represented on the table about which the user can talk are stored in a database and are dynamically loaded at runtime. This is necessary to be able to add lectures to the system in the offline part (e.g. during a conference) without changing the dialogue manager code or its grammar files.

Furthermore, this database contains multi-lingual information for all stored objects. When adding new objects, such as titles of the talks, weekdays, information on lecture places, etc., this information is added for all existing languages. The language specific information is then used by the language specific grammar parts and integrated as terminal symbols.

4. Multimodal Integration

4.1. Example

The user selects on the augmented table the video flower of a specific lecture because he wants to see some part of this lecture. Then a message is sent to the dialogue manager to set up the corresponding context which is then used to interpret further incoming messages and events. The flower opens and the different topics of this lecture can be seen as presented in figure 2. The user selects a topic by saying ”show me the video covering machine translation”, a message is sent again and the flower of this topic opens, too so that the user can see the different video segments which are available. He can select such a segment by speech or by means of the augmented table and then watch the corresponding video segment on the augmented wall.

4.2. Extended Use of Subgrammars

The use of subgrammars is extended so that the context completely determines what the user can say. For example, new grammar rules are generated based on the number of alternatives found in the database. This means that at first the database is accessed to get the number of different topics for the requested lecture. Based on this number, the subgrammars are generated. In our example possible cases are:

- If only one topic is found, this flower is opened immediately without any clarification questions.

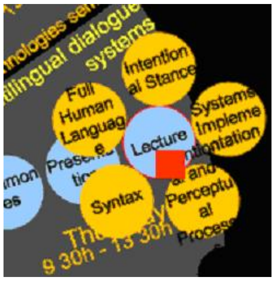

Figure 2: Example of an opened flower where the user can see the different topics mentioned in a lecture

- If there are four different topics, the system generates a clarification question so that the user has to specify which topic he is interested in. In this case, the user can simply say the names of these topics, and also utterances such as ”show me video segments on topic X”. Other formulations can be imagined, such as ”show me video segments on the left topic” or ”i am interested in video segments of the above topic”.

  These context-dependent formulations could as well be covered by the grammar, but therefore the given menu structure of round flowers has to be taken into account. This means that some more user tests are necessary to find out how users really name a specific topic and whether they take the rotation of the table into account or just the orientation of the letters to describe the place of such a circle.

- If more than four different topics are found, the system also generates a clarification question and the user has to specify the topic he is interested in by saying the number.

4.3. Multimodal Generation and Multilingual Information Retrieval

Spoken output is only generated when the input is also spoken. Otherwise, the output can only be seen at the table or the augmented walls. In this way, it is avoided that the user gets distracted by the speech synthesis when he does not expect a speech output because he was just working on the augmented table.

In addition, spoken output is always in English to assure that all the people in the intelligent room can understand it. Here again, we assume that the user understands at least enough English, to understand the program and contents of the system. To retrieve information he is given the opportunity to use his native language, making it more easy for him to formulate queries. The user can ask direct search queries to the system that cannot be covered by the table, such as for example, ”I would like to see something about Statistical Machine Translation”. Depending what kind of content can be found for the desired search keys, the system might display video segments, powerpoint slides, recorded testimonies, or reply to the user if no content can be found.

5. Evaluation

Over 50 persons used the FAME augmented meeting environment. Among them, we randomly selected 15 users (5 groups of 3 users) for a small user study in order to evaluate the system as a whole. After a brief introduction to the system, users were asked to perform some predefined tasks, such as ”can you retrieve Mr. Alex Waibel’s lecture?” or ”What topics are being addressed in the lecture about language evolution?”, etc. We observed the user behavior wile they interacted within the intelligent room. Finally, the users answered some questions about their general impression of the system.

Question	Yes	No	No Answer
Is the system reliable?	12	2	1
Would you be interested in using it?	12	0	3
Did you enjoy the system?	13	0	2
Is it useful with other people?	9	0	6
Is it fun with other people?	11	0	4

Table 1: Results from the User Survey

The results showed that multi-surface and multi-user interaction were considered useful and fun, as you can see in figure 1. Most of the people answered that they enjoyed the system and also found it useful while interacting with other people. Some people reported minor technical problems, such as that the system is sometimes too slow.

Furthermore, some users felt confused by the fact that different people can use the system at the same time concurrently because they did not know whether the replies of the system corresponded to their own request or not. The quasi-simultaneity of text-to-speech, video, and graphical system outputs favour collaboration so that some users formed subgroups within which they coordinated their actions in order to avoid such confusions.

Some users would like the system to be able to perform more functions, such as more search facilities, a haptic interface for browsing through the slide shows, navigation in video clips, etc. To conclude, the overall design and technical integration of the FAME augmented meeting room were very well perceived, fun to use, and provided a very special experience for the users.

6. Conclusion and Outlook

We have presented our efforts in building a multimodal dialogue system supporting multi-user and multilingual context management. The different modalities share the same context so that the user can alternately select which one he wants to use. Furthermore, we use this context to generate context-dependent grammar rules so that the search space of the speech recognizer can be restricted according to the current situation.

In the future, we want to extend this context modeling and use it within a more complex intelligent room which observes and explains human-human activities and intentions. Here, we have more input modalities which can be used to create an overall context.

7. Acknowledgments

This work was supported in part within the FAME project by the European Commission as IST-2000-28323, and within the CHIL project as contract #506909.

8. References

[1] J. F. Quesada, F. Garcia, E. Sena, J. A. Bernal, and G. Amores, ”Dialogue Management in a Home Machine Environment: Linguistic Components over an Agent Architecture”, Procesamiento del Lenguaje Natural, 2001.

[2] D. Bühler, W. Minker, J. Häussler, and S. Krüger, ”The Smartkom Mobile Multi-Modal Dialogue System”, ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments, 2002.

[3] M. H. Coen, ”The Future of Human-Computer-Interaction or How I learned to stop worrying and love my Intelligent Room”, IEEE Intelligent Systems, 1999.

[4] F. Bérard, ”The Magic Table: Computer-Vision Based Augmentation of a Whiteboard for Creative Meetings”, IEEE Workshop on PROCAM of the International Conference on Computer Vision, 2003.

[5] D.L. Martin, A. J.Cheyer, and D. B. Moran, ”The Open Agent Architecture: A Framework for building Distributed Software Systems”, Applied Artificial Intelligence, vol. 13, 1999.

[6] K. Kvale, N. D. Warakagoda, and J. E. Knudsen, ”Speech-centric Multimodal Interaction with Small Mobile Terminals”, Proceedings of NORSIG 2001, 2001.

[7] P. Gieselmann and M. Denecke, ”Towards Multimodal Interaction with an Intelligent Room”, Proceedings of the Eurospeech, 2003.

[8] H. Holzapfel, K. Nickel and R. Stiefelhagen, ”Implementation and Evaluation of a Constraint-Based Multimodal Fusion System for Speech and 3D Pointing Gestures”, Proceedings of the ICMI, 2004.

[9] F. Landragin and L. Romary, ”Dialogue History Modelling for Multimodal Human-Computer Interaction”, Proceedings of the Catalog Workshop, 2004.

[10] S. Seneff, D. Goddeau, C. Pao, and J. Polifroni, ”Multimodal Discourse Modelling in a Multi-User Multi-Domain Environment”, Proceedings of the ICSLP, 1996.

[11] C. Fügen, H. Holzapfel and A. Waibel, ”Tight Coupling of Speech Recognition and Dialog Management - Dialog-Context Grammar Weighting for Speech Recognition”, Proceedings of the ICSLP, 2004.

[12] A. Stent, J. Dowding, J. Gawron, E. Bratt and R. Moore, ”The CommandTalk Spoken Dialogue System”, Proceedings of the 37th Annual Meeting of ACL, 1999.

[13] O. Lemon, ”Context-sensitive Speech Recognition in ISU-Dialogue Systems: Results for the Grammar-switching Approach”, Proceedings of the Catalog Workshop, 2004.

[14] W. Xu and A. Rudnicky, ”Language Modeling for Dialog System”, Proceedings of the ICSLP, 2000.

[15] M. Denecke, ”Rapid Prototyping for Spoken Dialogue Systems”, Proceedings of the 19th International Conference on Computational Linguistics, 2002.

[16] B. Carpenter, ”The Logic of Typed Feature Structures”, Cambridge University Press, 1992.

[17] H. Holzapfel, ”Towards Development of Multilingual Spoken Dialogue Systems”, Proceedings of the 2nd Language & Technology Conference (L&T’05), 2005.