Proactive Spoken Dialogue Interaction in Multi-Party Environments

Petra-Maria Strauß • Wolfgang Minker

Proactive Spoken Dialogue Interaction in Multi-Party Environments

Springer

Petra-Maria Strauß
Ulm University
Institute of Information Technology
Albert-Einstein-Allee 43
89081 Ulm
Germany
petra-maria.strauss@uni-ulm.de

Wolfgang Minker
Ulm University
Institute of Information Technology
Albert-Einstein-Allee 43
89081 Ulm
Germany
wolfgang.minker@uni-ulm.de

ISBN 978-1-4419-5991-1 e-ISBN 978-1-4419-5992-8
DOI 10.1007/978-1-4419-5992-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009944071

Preface

This book describes the development and evaluation of a novel type of spoken language dialogue system that proactively interacts in the conversation with two users.

Spoken language dialogue systems are increasingly deployed in more and more application domains and environments. As a consequence, the demands posed on the systems are rising rapidly. In the near future, a dialogue system will be expected, for instance, to be able to perceive its environment and users and adapt accordingly. It should recognise the users' goals and desires and react in a proactive and flexible way. Flexibility is also required in the number of users that take part in the interaction. An advanced dialogue system that meets these requirements is presented in this work.

A specific focus has been placed on the dialogue management of the system on which the multi-party environment poses new challenges. In addition to the human-computer interaction, the human-human interaction has to be considered for dialogue modelling. A prevalent approach to dialogue management has been adapted accordingly. To enable proactive interaction a detailed dialogue history has been implemented. As opposed to common dialogue systems which start from scratch when the interaction begins, the system developed in the scope of this book starts modelling as soon as the conversation enters its specified domain. The knowledge acquired during this early stage of the conversation enables the system to take the initiative for meaningful proactive contributions, already from the first interaction.

In order to develop this interaction assistant comprehensive data recordings have been conducted in a multi-modal Wizard-of-Oz setup. A detailed overview and analysis of the resulting corpus of multi-party dialogues is presented. An extensive evaluation of the usability and acceptance of this novel sort of dialogue system constitutes a further significant part of this book.

Contents

List of Figures

List of Tables

1

Introduction

HAL: 'Excuse me, Frank.'
Frank: 'What is it, HAL?'
HAL: 'You got the transmission from your parents coming in.'
Frank: 'Fine. Put it on here, please. Take me in a bit.'
HAL: 'Certainly.'

Quote from '2001 – A Space Odyssey' (1968) by Stanley Kubrick.
The HAL 9000 computer is addressing Dave who is resting on his sun bed.
app. 1:00 hour into the film

As it was predicted already in 1968 by Stanley Kubrick (1928-1999) and Arthur C. Clark (1917-2008) in the science fiction movie *2001 – A Space Odyssey* [Kubrick, 1968] the future has arrived. Computers are by now playing a prominent role in our everyday lives. Over the past decades they have evolved from big, monstrous mainly industrial machines to small mobile and extremely powerful devices that are in one way or another used by presumably every human being in the developed world. The quote by the 'supercomputer' *HAL 9000* from Kubrick's movie shows that the computer is equipped with human-like qualities. It possesses natural language capabilities for both, understanding and speaking, the ability of logical reasoning and proactive behaviour, just to name a few character traits. The human characters of the movie are quoted in the movie to describe the computer as a sixth member of their space ship crew.

A 'HAL-like' computer has not been developed at present, however, HAL's characteristics, i.e. his human-like features, are starting to appear in more and more computer systems. **Natural language interaction** plays an important role due to the fact that speech is for humans still the easiest and most natural way of interaction. Big displays become superfluous which opens the way for *ubiquitous computing* which lets computers disappear more and more into the background. Security is a further supporting factor for interaction by speech.

P.-M. Strauß and W. Minker, *Proactive Spoken Dialogue Interaction in Multi-Party Environments*, DOI 10.1007/978-1-4419-5992-8_1,

This becomes apparent especially in the scope of automotive applications. While operating a vehicle, the driver can interact with the navigation, telephony and media applications by speech without taking the eyes off the road.

The automotive environment is also a pioneer domain for **proactiveness**. State of the art head units inform the driver about traffic hazards coming up on the road. According to the priority of the message, i.e. for instance in terms of the distance to the obstacle which denotes if the driver could be affected immediately, even ongoing phone calls should be interrupted for the driver to receive the message as soon as the system learns about the hazard. As an independent crew member, HAL is further able to communicate with **multiple users** at the same time while most of today's computer systems are restricted to one user, i.e. human-computer interaction. If dialogue systems can interact with several users simultaneously many applications would benefit, for instance in the process of achieving a common task.

The research presented in this book addresses the presented challenges: A spoken language dialogue system that interacts with two users as an independent dialogue partner. It engages proactively in the interaction when required by the conversational situation and also takes itself back when it is not needed anymore. We thereby focus on the dialogue management functionality of the system (Chapter 4) for which we perform an extensive data collection (Chapter 3) to support the system development. Further, evaluation of the novel sort of dialogue system builds another prominent part of this book (Chapter 5). The envisaged system is introduced in more detail in Section 1.3. First, a short introduction is given on spoken language dialogue systems in general followed by a description of current trends and related work conducted in the area of advanced dialogue systems.

1.1 Introduction on Spoken Language Dialogue Systems

1.1.1 System Architecture

The task of a spoken language dialogue system (SLDS) is to enable and support spoken interaction between human users and the service offered by the application. The SLDS deals with two types of information - the one understood by the user (natural language in speech or text) and the one understood by the system (e.g. semantic frames). The system carries out a number of tasks before it can give a response to the user. The tasks are performed by different modules which are usually connected in a sort of pipeline architecture. Figure 1.1 shows a basic architecture of a SLDS. The different modules are described in the following:

Automatic Speech Recognition (ASR). The task of the ASR module is the transcription of the speech input (i.e. acoustic signals) of the user into words (e.g. [Jelinek, 1997, Rabiner and Juang, 1993, Jurafsky and Martin,

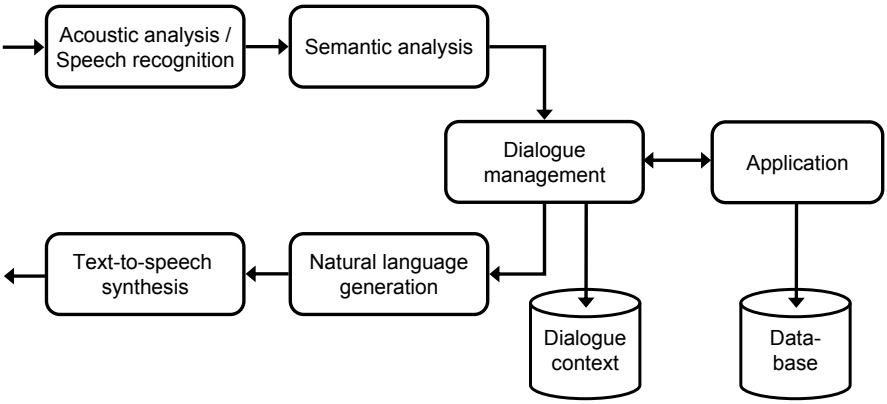

Fig. 1.1. SLDS architecture.

2000, Huang et al., 2001]). Using an acoustic model which describes potential signals, a lexicon containing all potential vocabulary and a language model, i.e. a grammar, the acoustic signals are usually mapped to the resulting words or sentences with statistical methodology. Different factors determine the complexity of speech recognition. A system that is to be used by an unknown number of different users possibly speaking in different accents and dialects is said to be speaker-independent. The opposite is a speaker-dependent system which is trained especially for the individual future user. A third intermediate option is a speaker-adaptive system which is developed as a speaker-independent system but can adapt to the actual user through training and usage. The vocabulary of the system further influences the complexity and performance: Small vocabulary is easier to be recognised than large vocabulary. Finally, continuous speech poses a greater challenge than isolated keywords.

Natural Language Understanding (NLU). The NLU module tries to extract the semantic information from the word string produced by the speech recogniser (refer to e.g. [Allen, 1995, Jurafsky and Martin, 2000]). It produces a computer readable representation of the information (e.g. as semantic frames) which is then further processed by the dialogue management module. A common approach is to perform rule-based semantic parsing to extract the semantic meaning, e.g. attribute-value pairs, out of the utterances. Another approach include statistical methods for semantic analysis (e.g. [Minker et al., 1999]).

Dialogue Management (DM). The dialogue manager is responsible for smooth interaction. It handles the input (in form of a semantic representation)

which is to be integrated into the dialogue context. It organises turn taking and initiative, and performs task or problem solving by interacting with the application. Finally, it induces the output generation to return an appropriate response (e.g. the requested information) or to ask for any information missing in order to be able to fulfil the task. The DM makes use of various knowledge sources which constitute the dialogue context. The main parts are the task model and the dialogue model [McTear, 2002]. The task model contains all task-related parts of the system, such as the task record which holds all user constraints mentioned in the ongoing dialogue so far whereas the dialogue model contains information regarding the dialogue, such as a conversation model which consists of the current speaker, addressee, speech act, etc. The dialogue history can be said to belong to this part of the context as it holds information about the previous utterances of the ongoing dialogue. Further knowledge sources are a domain and world knowledge model which holds the logical structure of the domain and world the dialogue system functions in. A user model can be deployed which holds the information about the users, either to recognise specific users or more general information to be able to make recommendations. All these components are implemented more or less explicitly depending on the type of dialogue management used. Approaches to dialogue management can be classified into three main categories (following the categorisation presented by McTear (2002)):

- **Finite-state-based approach.** The dialogue is always in a certain predefined dialogue state, certain conditions trigger state changes. In this approach, knowledge base and dialogue management strategy are not separated but are represented together in the dialogue states. The approach is rigid and inflexible but very suitable for small, clearly defined applications.
- **Frame-based approach.** The systems implementing this approach deploy a specific task model which determines a set of slots to be filled with values supplied by the user in the course of the dialogue in order for the system to fulfil the task. Conversational aspects of the dialogue are considered only in the scope of task solving. The system is not expected to hold a conversation or know details of the conversation such as regarding the order of the constraints mentioned etc. Thus, no complex models have to be deployed. The approach is suitable for dialogue systems used for information retrieval, such as train departure times etc.
- **Agent-based approach.** This approach is able to model dialogues in a more complex way. With sophisticated models of the conversation and dialogue participant it overcomes the limitations of the aforementioned approaches. Dialogue is no longer limited to certain dialogue stages but rather works towards understanding the dialogue as a whole. It models from the viewpoint of the dialogue system which is modelled as an agent who has goals (e.g. to fulfil the task), intentions, and plans to achieve its goals.

The prominent *Information State Update* approach (e.g. [Ginzburg, 1996, Larsson, 2002]) belongs to the third category. The dialogue which is seen as a state of information that is updated with every utterance is modelled from the viewpoint of the system enabling it to 'understand' the dialogue as it occurs. This approach is thus very suitable to be adopted for our dialogue system which is to constitute an independent dialogue partner. The approach is introduced in Section 2.4 and later adopted and extended to suit our setup as presented in Chapter 4.

A further categorisation differentiates between rule-based and statistical processing of dialogue management. All of the above mentioned categories of dialogue management can be implemented using either approach. The rule-based approach has been state of the art for a long time. Rules, defined by the developer, have to be supplied for all cases that can possibly occur in the dialogue. Accurate processing is thus assured, however, the development of the rule-base is very time-consuming and an increase in the complexity of the application brings about an analogical increase in the complexity of the rule set which can easily reach an unmanageable dimension. Recently, statistical approaches popular in ASR and also in NLU (e.g. [Minker et al., 1999]) are starting to be also deployed to dialogue management e.g. [Levin and Pieraccini, 1997, Singh et al., 2002, Lemon et al., 2006, Williams and Young, 2007]. Statistical techniques are based on statistical modelling of the different processes and learning the parameters of a model from appropriate data. The drawback of this approach is that for development a large amount of training data is needed which is difficult to obtain.

Another important task of the dialogue management is problem solving. The dialogue management communicates with the application in order to fulfil the task. The simplest form of an application is a database. The dialogue management would in this case interact by performing database queries based on the current user constraints (contained in the task model) (e.g. [Qu and Beale, 1999]). Problem solving further looks at the outcome of the query and, if necessary, tries to optimise it. For instance, in the case that the query does not yield any results, the constraint set can be modified (for instance by relaxing less important user constraints) until a more satisfying result is achieved (e.g. [Walker et al., 2004, Carberry et al., 1999]).

Natural Language Generation (NLG). The response commissioned by the dialogue management module is in this step turned into a natural language utterance. A common practise for NLG is the template based approach. Previously defined templates are filled with the current values. The NLG module is further responsible of structuring the output, i.e. choosing the best or combining the output if various are available or breaking it down into appropriate chunks if the answer is too large. The dialogue history can be consulted to assure responses that are consistent and coherent with the preceding interaction. For a multi-modal system, if e.g. visual output is deployed besides the speech output, the different modalities have to be integrated. The

respective output has to be assigned the appropriate modality always assuring conformity. In general, NLG is from concerned with three tasks [Reiter, 1994, Reiter and Dale, 2000]:

- **Content determination and text planning** to decide on what information, and in what kind of rhetorical structure it should be communicated.
- **Sentence planning** determines the structure of the utterance for instance adapting it in order to fit in well with the current flow of the dialogue. Examples are splitting or conjunction of sentences as well as adding references or discourse markers.
- **Realisation** is responsible for linguistic correctness and adaptation of the content to the actual output modality.

A common practise for NLG is the template based approach. Previously defined templates are filled with the current values. The NLG module is further responsible of structuring the output, i.e. chosing the best or combining the output if various are available or breaking it down into appropriate chunks if the answer is too large. The dialogue history can be consulted to assure responses that are consistent and coherent with the preceding interaction. For a multi-modal system, if e.g. visual output is deployed besides the speech output, the different modalities have to be integrated. The respective output has to be assigned the appropriate modality always assuring conformity.

Text-to-Speech Synthesis (TSS). Utterances generated in the previous module are converted from textual form into acoustic signals using text-to-speech (TTS) conversion [Dutoit, 2001, Huang et al., 2001]. The text is in a first step converted into a phoneme sequence and prosodic information on a symbolic level. Acoustic synthesis then performs a concatenation of speech units (e.g. for German diphones are common, while syllables are used for Chinese) contained in a database. The generated audio is then played back to the user. A different option yields the most natural sounding speech output that uses pre-recorded audio files. The duty of the NLG module is to simply select the adequate audio file to be played back to the user. A combination of these approaches, popular for commercial dialogue systems, is especially convenient for template-based NLG. The fixed template texts are pre-recorded, all variable parts are generated on the fly (preferably using the same speaker for both recordings). This way, the prompts sound as naturally as possible, however, not losing the flexibility of synthetically produced speech prompts.

1.1.2 Current Trends in Spoken Language Dialogue Systems

Today's commercial dialogue systems are usually deployed for simple tasks. They are predominantly slot-filling small-vocabulary finite-state machines, i.e. systems that match specific incoming keywords to a fixed output, a task that does not demand for elaborate dialogue systems. They are mainly found in

telephony applications replacing human operators in call centres. Their main aim is to save cost. A nice side-effect has been achieved by some companies by personifying their dialogue systems to use them as marketing instruments. The systems are given a name and appearance and thus star in commercials and on websites to improve a company's image and level of awareness. A prominent example for such a system is the award-winning *Julie*[1] (deployed in May 2002) who answers the phone if someone calls for train schedule information to travel within the United States. Insufficient technical performance, however, has been hindering speech based interfaces to obtain large-scale acceptance and application. Broad usage requires good recognition performance of speaker-independent large-vocabulary continuous natural speech which has been posing a great challenge to speech recognition. The last years have been coined by technical advancement and further, user acceptance has been growing due to the fact that people gradually get accustomed to the SLDS. The usefulness and convenience of spoken language interaction has been recognised and thus the range of applications is starting to grow and change. With progressing technology and the quest for smart and apt computer systems the foundation for accelerated progress has been provided. Possibly, scenarios that have for a long time only been found in science fiction might become ordinary scenes of everyday life in the future.

A current trend addresses the nature of computer systems. Computers are blending more and more into the background, as described by the term *ubiquitous computing*. Computers are becoming smaller, almost disappearing, and are deployed more and more in mobile form. Everyday appliances are enriched with computational intelligence trying to ease human life building the basis for *intelligent environments*. Popular examples are intelligent heating and lighting adjustments and the intelligent refrigerator that keeps track of the contents, recipes, shopping lists and even ingested calories of the users. The overall trend is that computers adapt to the human way of interaction instead of requiring the humans to move towards the system for interaction. All of these facts pose further demands on applications and technology and at the same time show the importance of speech based interaction. It is an intuitive means of communication and does not require any space (e.g. big screens as is the case for haptic interaction) nor visual attention to be deployed and is thus also a suitable way for human-computer interaction in critical situations, such as the car where the driver's gaze should not be drawn from the road if possible[2].

Novel demands are posed on future systems in order to realise the adoption to new applications and environments. The objective of future systems is to actually understand the dialogue they are involved in and to adapt to

[1] http://www.amtrak.com

[2] In practise, as an intermediate step towards speech interaction, current systems adopt speech interaction mostly as an alternative on top of the common forms of interaction and this way trying to gain in user acceptance.

the surroundings and users, to autonomously perceive the user's needs and desires and to react flexibly and proactively. Future dialogue systems are thus endowed with perceptive skills from different sensory channels (vision, hearing, haptic, etc.) to capture the spacial, temporal, and user specific context of an interaction process. Elaborate conversational skills are required to be able to capture and analyse spoken, gestural, as well as emotional utterings. Integration of perception, emotion processing, and multimodal dialogue skills in interactive systems is expected to not only improve the human-computer communication but also the human-human communication over networked systems. There is further an increasing demand for flexibility in terms of the number of users that are able to take part in the interaction. A system could this way for instance assist a group of users already during the decision process by providing information, immediately reporting problems and thus accelerating the task-solving process. Thus, interaction between various humans and possibly also various computers will be possible that integrates the dialogue system as an independent dialogue partner in the conversation.

The research presented in this book focuses on a dialogue system of this kind: The system resembles an independent dialogue partner. It interacts with two users and engages proactively in the conversation when it finds it necessary in respect to advancing the task solving process in the best possible way. A description of the system and objective of this book is presented in detail below after taking a look at related work conducted on advanced dialogue systems.

1.2 Related Work on Advanced Dialogue Systems

Various research projects investigate possibilities that open up by enriching multi-party interaction and advanced dialogue systems with the perception of the users' state, context and needs. Most of the research on multi-party interaction at present is concerned with the meeting scenario as it can benefit greatly from the use of intelligent computer systems which enhance and assist the human communication during (and also after) the meetings. Great effort is put in the design and development of adequate software tools for meeting support and to investigate multi-party interaction phenomena. Meeting assistants could be deployed as meeting browsers or summarisers, i.e. they obtain information about the course and content of a meeting. They can be used for example during the meeting to assist participants who have come late to the meeting, summing up what has been said and who has committed to what. In the same way, easy and fast access of the meeting content is enabled at a later point in time. An example of a tool of this kind is the meeting browser developed in the framework of the *Augmented Multi-Party Interaction* (AMI) project [Renals, 2005] (and its successor AMIDA). The aim is to develop new multimodal technologies in the context of instrumented meeting rooms and