From V. Matoušek (ed.), Proc. of 4th Int. Conf.on Text, Speech and Dialogue (TSD 2001), Czech Republic, pp. 341–348, 2001

Research Issues for the Next Generation Spoken Dialogue Systems Revisited

E. Nöth[1], M. Boros[2], J. Fischer[2], F. Gallwitz[2], J. Haas[2], R. Huber[2], H. Niemann[1], G. Stemmer[1], V. Warnke[2]

[1]Universität Erlangen–Nürnberg, Lehrstuhl für Mustererkennung,
Martensstr. 3, 91058 Erlangen, Germany

[2]Sympalog Speech Technologies AG,
Karl–Zucker–Straße 10, 91052 Erlangen, Germany

noeth@informatik.uni-erlangen.de

Abstract. In this paper we take a second look at current research issues for conversational dialogue systems addressed in [17]. We look at two systems, a movie information and a stock information system which were built based on the experiences with the train information system EVAR, described in [17].

1 Introduction

Two years ago, at TSD '99 we presented the EVAR system, a conversational spoken dialogue research prototype for train information [17]. There we discussed research issues for conversational dialogue systems. At the same time first plans were made to start a spin-off company out of our research group with the aim to market dialogue systems. Based on the philosophy presented in [17], a completely new implementation of a task– and language–independent dialogue manager was performed. Since then we implemented about 10 different conversational systems with applications ranging from movie information and movie ticket reservation to quality assurance in an industrial environment. Systems were presented at Systems '99, CeBit 2000, Systems 2000 and CeBit 2001. In October 2000 Sympalog received the IST-Prize 2001 by the European Community for its conversational dialogue technology.

We thought this conference would be a good occasion to have a second look at some of the implemented systems to see, where research issues raised in [17] were addressed and which new issues came up. Of course, we don't claim that we solved the addressed research issues. Also, even though we constantly tested the different prototypes of the systems with 'naive' users, we cannot yet provide results from an extensive and systematic field test.

We first characterize two of the implemented systems, the movie information system FRAENKI and the stock information system STOCKI (Section 2), then we look at most of the research issues addressed in [17] (Section 3 – 9) and address some additional issues (Section 10 – 11).

This research was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the VERBMOBIL Project under Grant 01 IV 701 K5. The responsibility for the contents of this article lies with the authors.

2 Two Exemplary Conversational Dialogue Systems

2.1 Fraenki

After the implementation of a completely new task– and language–independent dialogue manager in the application domain of EVAR (train information), we created a dialogue system for a typical local information retrieval task, the FRAENKI movie information system. We chose this way rather than immediately implementing a movie information system to learn more about side effects that might come up when porting a system architecture to a new application. FRAENKI knows the program of all the movie theaters in a restricted area (Middle Franconia, i.e. the greater Nuremberg area with approximately 1.5 mio. people). There are about 60 theaters in 35 locations with a total of about 350 performances shown per day. The system is hooked up to the public telephone (+49 9131 16287 and +49 9131 6166116). FRAENKI is a monolingual German system. The vocabulary is about 1500 word forms. The program is updated weekly from the Internet. New titles are added to the recognition lexicon in a semi-automatic way (see Section 6). Typical initial user utterances which can be processed successfully, range from

I want to go to the movies ⇒ (will lead to system driven dialogue)

to

I want to see 'Pearl Harbor' tonight in Erlangen at about eight ⇒
(user driven dialogue, arbitrary combination of information slots)

2.2 Stocki

STOCKI is a stock information system which knows about stocks listed in the Dax30 and EuroStoxx50 (the German and European equivalent to the Dow Jones) and the Nemax50 (the German equivalent to the Nasdaq). STOCKI can answer questions about information like stock ID, day's high, day's low and traded volume. It knows about 10 stock exchanges. The about 130 stocks are represented by about 250 variants like 'Mercedes', 'Daimler', 'Chrysler', and 'DaimlerChrysler' for the 'DaimlerChrysler' company. A typical question to the system is

What's the current price and the traded volume of BMW in Frankfurt?

STOCKI is multilingual and can process inquiries and answer questions in German, English and French. The information is accessed every n seconds or on the fly after the user inquiry from a financial service provider. Currently STOCKI is only a demonstrator and cannot be accessed via the public phone.

3 WWW Database Access

In [17] we showed that EVAR accessed its information from several information sources in the WWW like the German Railways (*DB*), *Lufthansa*, and Swiss Railways (*SBB*), together with the facility for a number of local databases to be set–up for regularly–accessed data. EVAR gathered all the necessary information from the user and only accessed the remote databases once (mostly for processing speed). While FRAENKI has a local copy of its remote database which is updated weekly Stocki has to constantly access the remote database to guarantee the most recent data. This situation necessitates a closer interaction with the information provider than a simple HTML–parser for WWW–pages which suffices for a research demonstrator (not to mention the legal aspects which have to be cleared). Access speed and data security might make a different interface necessary for a commercial system but more important, the interfaces to the WWW databases have to be defined for instance via XML and respective document type definitions (DTD files) and cascading style sheets (CSS). This allows a clean separation of 'what' is represented on a WWW–page (DTD) and 'how' it is represented (CSS). Otherwise nearly any layout change in the WWW–page means a change to the HTML–parser for the database access which is an unacceptable amount of maintenance. Another important topic is the VoiceXML [2] standard. VoiceXML is becoming the voice markup standard for Interactive Voice Response (IVR) applications [14]. It allows standardized interfaces between the different modules of spoken dialogues systems such as the recognizer, the text–to–speech engine, and the dialogue manager. This will lead to an easier integration of components from different vendors.

4 Flexible and Adaptive Dialogue Strategy

One of the distinctive traits of our dialogue systems is the possibility for the users to freely formulate their queries and carry out the transaction quite flexibly. The user is allowed to take the initiative regarding the order in which task parameter specification takes place and is also usually able to change the current subgoal of the interaction; e.g. in correcting a parameter that has already been dealt with, at a time when the system is expecting information about another parameter. This contrasts to the more common approach of presenting the user with menus to which they have to comply and answer with yes or no. As a result, however, there are more possibilities regarding the content of the next user utterance, thus increasing the probability of misrecognitions and misunderstandings. To remedy this, we introduced a flexible strategy of implicit and explicit confirmation [11].

In case of misrecognitions and because of the island driven chunk–parsing approach used in the semantic analysis [16], this can lead to the insertion of semantic units. The system then tries to confirm a task parameter value which was never uttered by the user. In [5] it is reported that users react much more sensitive to the wrong insertion of task parameter values than to the deletion of uttered ones. Confidences measures as described in [9] are needed to reject an utterance rather than confirm a wrongly detected task parameter value. The tuning of when to reject an utterance is very difficult though, since it will inevitably lead to a lower understanding rate (sometimes a correctly detected task parameter value will be refused). Automatic detection of the optimal working point is very difficult and an important research issue. Besides confidences measures the explicit handling of out–of–vocabulary and out–of–domain situations, as described in the next chapter, can help very much to increase the understanding rate.

5 Robustness towards Out–of–Vocabulary Words

One of the most important causes of failure in spoken dialogue systems is usually neglected: the problem of words that are not covered by the system's vocabulary (Out–Of–Vocabulary or OOV words). In such a case, the word recognizer usually recognizes one or more different words with an acoustic profile similar to the unknown. These misrecognitions often result in possibly irreparable misunderstandings between the user and the system. In [13] we presented an approach to directly detect OOV words during the speech recognition process and, at the same time, to classify the word with respect to a set of semantic classes. This information was used to handle OOV words in the dialogue [8]. In Fraenki and Stocki we extended this notion towards 'out–of–domain' (OOD) questions which are likely to occur even from a cooperative user: The user might ask for an information that is not in the database, like the content of a movie. These questions are modeled and the system informs the user that it cannot answer the question. Also, we included some predictable OOV words (like big German cities outside the region for which Fraenki has information) explicitly in the recognition lexicon. Thus Fraenki can recognize that the user wants to know the movie schedule of a city outside of its region either via the 'OOV City' word ([13]) or because the user asks for the schedule in a city that is close to the region like *Regensburg* or one of the major cities of Germany like *Munich* and *Hamburg*. Similarly Stocki recognizes 6 stock exchanges like *New York* and *Tokyo*, 6 indices like *Nasdaq* and about 50 stocks listed in other indices than the ones in the database like *Microsoft* and *Exxon*. Figure 1 shows excerpts from example dialogues for OOV and OOD situations. Of course we do not handle arbitrary OOD inquiries like *'Can you bring me a Pizza?'*, since we assume a cooperative user. The automatic detection of 'cooperative' OOD and OOV situations during field tests and in routine use of the system without huge amount of human inspection is an important research area.

6 Multilinguality

Within the EC Copernicus–project SQEL, a multilingual (Czech, German, Slovak, and Slovenian) and multifunctional (airline and railroad connections) version of Evar was implemented [4]. Stocki can handle questions in German, English, and French using the same technology. Speech recognition and language identification is performed in one integrated process as described in [18]. The applications of Stocki and Fraenki have another multilingual problem, which is especially important for the update of the lexicon: Many stock names and movie titles contain foreign, especially English words; often a movie has an English title and a German subtitle and users ask for the film using all 3 possibilities (English title, German title, and both titles). We thus have the problem (typical for many dialogue applications) that the recognition vocabulary changes regularly and contains proper names and acronyms from many languages which are expected to be pronounced by non–native speakers. The pronunciation e.g. of 'IBM' or 'Carrefour' depends to a large extent on the skill level of the user in the language of the names' origin. This problem has barely been touched in research concerning multilingual speech recognition, since practically all research concerns

System: *Welcome at the stock information system from Sympalog -*
which information do you want?
User: *What is the current price of Coca-Cola?*
System: *Information about Coca-Cola is currently not available.*
Do you need any other information?

(a) Example of Stocki handling an OOV stock name

System: *Hi, this is Fraenki - which information do you want?*
User: *Hi, I want to see an action movie.*
System: *Sorry, but currently I am unable to provide you with information*
concerning the quality, content or genre of the movie.
Do you need any other information?

(b) Example of Fraenki handling an OOD question

Fig. 1. Example dialogues involving OOV and OOD situations

native speakers or non–natives from bilingual regions [3]. A lot of research needs to be done both on the level of grapheme–to–phoneme conversion and in speech synthesis to handle this difficult situation. For Fraenki new titles are retrieved from the internet source once a week, phonetized semi–manually by an expert, and spoken by the speaker whose voice is used for a concatenation of prerecorded words and phrases as system output. This is acceptable since the movie schedules only change once a week and since the manual work normally only concerns a few new titles. However, only switching to TV program rather than movie theaters makes this approach unacceptable because of the huge amount of manual work and a different approach needs to be taken (see [19] for our first experiments concerning the pronunciation of foreign words in the Fraenki scenario). Even though there is significantly more manual work involved, current 'off–the–shelve' synthesis performs far too poorly to be an alternative to concatenation of prerecorded words and phrases in the Fraenki/Stocki scenario.

7 Stochastic Methods for Semantic Analysis

In [17] we argue that statistical methods need to be explored for semantic analysis. While we are still absolutely convinced that this is the long term way to go, current statistical methods for semantic analysis require too much training data and don't generalize enough across applications to be used for fast prototyping. One can imagine though a hybrid approach from stochastic methods as described in [15] and linguistic methods as described in [7] for semantic analysis: a stochastic module can score competing hypotheses from a knowledge based linguistic module and thus decide on the order in which the semantic hypotheses are processed. Another possibility, that we currently look at is whether a stochastic module can be used across applications to detect uncooperative user utterances. Furthermore, it is an open and fascinating research topic whether it is possible to detect changes in the users' behavior over time, using stochastic semantic analysis and unsupervised learning techniques on log–files of running systems. Since the client of a dialogue system is always interested in the effectiveness of his system and since evaluation requires a lot of expensive handwork by IT–experts, we believe that this is an important question.

8 Integrated Recognition of Words and Boundaries

In [17], we propose the direct integration of the classification of phrase boundaries into the word recognition process. HMMs are used to model phrase boundaries, which are also integrated into the stochastic language model. The word recognizer then determines the optimal sequence of words and boundaries. The approach is described in detail in [12] and is used in the semantic analysis of our systems: a word sequence containing a boundary is less likely to represent a task parameter value than one without.

9 User Emotion

In [17] we argued that it is important to identify a situation where the user gets angry in order to initiate an appropriate reaction, such as referring the customer to a human operator or starting a clarification sub–dialogue. This subject remains to be a topic of fundamental research and of growing interest [10]. In [6] it is shown in a WOZ–scenario that not only acoustic/prosodic parameters have to be exploited but dialogue structure/history as well; for instance, plain repetition of the last utterance vs. rephrasing can be an important cue to the detection of anger/frustration.

10 Multimodality

Stocki is a demonstrator system. When installed at a direct brokerage bank, questions like access control become important, especially if the functionality is extended to stock trading. Current voice based verification technology is far too error prone to be accepted by any bank and PIN or password input via voice might be unacceptable in a public environment. Touch–tone (DTMF)–based PIN input is an example where multimodal input is a necessity. The fact that a PIN might be spoken from a client driving a car but not be typed in via DTMF in that situation also demonstrates the necessity to flexibly offer several input modalities. If one thinks of unified message systems, then multimodal output is just as important. For instance, a user might ask for tasks like

give me the current price of all the car manufacturers
and fax them to my secretary.

Similarly, a Fraenki user in the near future might want to see a preview of a movie, if s/he has a UMTS–phone. We believe that speech will only be an integral part of future human–machine–interaction, but that spoken dialogue will be 'central control' of such an interaction (see [1] and [20] for a project on dialogue based multimodal human–technology interaction).

11 Various topics

In this section we want to address some topics that came up during the implementation of the various systems that are important for future systems and were not mentioned in [17].

11.1 Barge in

For a conversational system it is absolutely inevitable to have a sophisticated barge in capability and robust noise cancellation. Especially in the Fraenki scenario we experience many calls from non–home/office environments, i.e. bars and public places with significant background noise.

11.2 Rapid prototyping

Fraenki was built from scratch in two months using the existing recognition engine and dialogue manager. Due to the meanwhile available Sympalog toolkit, other prototypes were implemented within a couple of days. Fast porting to new domains is very important for 'real life' systems and has enormous consequences especially for the methods in semantic interpretation and dialogue modeling; for instance methods that require a lot of hand encoding of linguistic knowledge for each lexical item are not feasible.

11.3 Upscaling

The ability to handle n calls in parallel is no research issue but has large consequences for the system architecture and the use of resources. Questions which have to be addressed include timing issues (real time speech input, speech output, speech recognition, and database access) and redundancy towards hardware failures.

12 Conclusion

In this paper we presented two state–of–the–art conversational dialogue systems. We took a second look at the research issues raised in [17]. It turns out that most of the research issues are still more than valid and that automatic learning methods lack robustness towards insufficient data to be used in rapid prototyping for new systems. Speech synthesis is an Achilles' heel for telephony based dialogue systems, since it is THE interface to the user and since proper names and foreign words are not pronounced well by current synthesis systems. VoiceXML will be an important part of future conversational systems and multimodality will become increasingly important with strong technological changes in the mobile communication ahead of us.

References

1. SmartKom Project. http://www.smartkom.org.
2. VoiceXML Forum. http://www.voicexml.org.
3. *Proc. of the Workshop on Multi-Lingual Speech Communication*, Kyoto, Japan, 2000.

4. M. Aretoulaki, S. Harbeck, F. Gallwitz, E. Nöth, H. Niemann, J. Ivanecky, I. Ipsic, N. Pavesic, and V. Matousek. SQEL: A Multilingual and Multifunctional Dialogue System. In *Proc. Int. Conf. on Spoken Language Processing*, Sydney, 1998.
5. H. Aust and O. Schröer. Application Development with the Philips Dialog System. In *Proceedings of 1998 International Symposium on Spoken Dialogue (ISSD 98)*, pages 27–34, Sydney, Australia, 1998.
6. A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In Cowie et al. [10], pages 195–200.
7. M. Boros. *Partielles robustes Parsing spontansprachlicher Dialoge am Beispiel von Zugauskunftdialogen.* Studien zur Mustererkennung. Logos Verlag, Berlin, 2001.
8. M. Boros, M. Aretoulaki, F. Gallwitz, E. Nöth, and H. Niemann. Semantic Processing of Out-Of-Vocabulary Words in a Spoken Dialogue System. In *Proc. European Conf. on Speech Communication and Technology*, volume 4, pages 1887–1890, Rhodes, 1997.
9. L. Chase. Error–Responsive Feedback Mechanism for Speech Recognizers. Technical report, Dissertation, Carnegie Mellon University, 1997.
10. R. Cowie, E. Douglas-Cowie, and M. Schröder, editors. *Proc. of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, Newcastle, Northern Ireland, 2000.
11. W. Eckert. *Gesprochener Mensch–Maschine–Dialog.* Berichte aus der Informatik. Shaker Verlag, Aachen, 1996.
12. F. Gallwitz. *Integrated Stochastic Models for Spontaneous Speech Recognition.* PhD thesis, University of Erlangen-Nuremberg, Germany, 2001. (to appear).
13. F. Gallwitz, E. Nöth, and H. Niemann. A Category Based Approach for Recognition of Out-of-Vocabulary Words. In *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 228–231, Philadelphia, 1996.
14. C. Günther, M. Klehr, J. Kunzmann, and T. Roß. Building Voice Enabled Internet Portals based on VoiceXML. In K. Fellbaum, editor, *Elektronische Sprachsignalverarbeitung*, volume 20 of *Studientexte zur Sprachkommunikation*, pages 102–109. TU Cottbus, Cottbus, 2000.
15. J. Haas. *Probabilistic Methods in Linguistic Analysis.* Studien zur Mustererkennung. Logos Verlag, Berlin, 2001.
16. E. Nöth, M. Boros, J. Haas, V. Warnke, and F. Gallwitz. A Hybrid Approach to Spoken Dialogue Understanding: Prosody, Statistics and Partial Parsing. In *Proc. European Conf. on Speech Communication and Technology*, volume 5, pages 2019–2022, Budapest, Hungary, 1999.
17. E. Nöth, F. Gallwitz, M. Aretoulaki, M. Boros, J. Haas, S. Harbeck, R. Huber, and H. Niemann. Research Issues for the Next Generation Spoken Dialogue Systems. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Proc. Workshop on TEXT, SPEECH and DIALOG (TSD'99)*, volume 1692 of *Lecture Notes for Artificial Intelligence*, pages 1–8, Berlin, 1999. Springer–Verlag.
18. E. Nöth, S. Harbeck, and H. Niemann. Multilingual Speech Recognition. In K. Ponting, editor, *Computational Models of Speech Pattern Processing*, volume 169 of *NATO ASI Series F*, pages 363–375. Springer–Verlag, Berlin, 1999.
19. G. Stemmer, E. Nöth, and H. Niemann. Acoustic Modeling of Foreign Words in a German Speech Recognition System. In *Proc. European Conf. on Speech Communication and Technology*, Aalborg, Denmark, 2001. (to appear).
20. W. Wahlster, N. Reithinger, and A. Blocher. SmartKom: Multimodal Communication with a Life-like Character. In *Proc. European Conf. on Speech Communication and Technology*, Aalborg, Denmark, 2001. (to appear).