Real Users Behave Weird — Experiences made collecting large Human-Machine-Dialog Corpora

Wieland Eckert Elmar Nöth Heinrich Niemann Ernst–Günter Schukat–Talamazzini

Friedrich–Alexander–Universität Erlangen-Nürnberg
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstraße 3, 91058 Erlangen, GERMANY
E-mail: wieland.eckert@informatik.uni-erlangen.de

Abstract

We have built a demonstrator for spoken human–machine dialogs. The system is capable to answer inquiries about the InterCity train timetable of the Deutsche Bahn. It evolved from the work done in the ESPRIT project SUNDIAL with four participating European countries (cf. [11]). First experiments have already been described in [1]. In this paper we describe our analysis of real user interaction with automated information systems: since performance figures of the automated system are already reported, we concentrate on description of some essential aspects of user behavior.

1. Introduction

Human–machine–interaction has been studied for several years. In the speech community we focus on spontaneous speech communication between humans and computers. At our institute we investigate speech recognition problems as well as understanding problems and interaction problems. Figure 1 shows the main components of our demonstrator. Current technology tends to use statistical methods to solve these problems. In order to have proper material representing "real" users and "real" inquiries, we connected our dialog system to the public telephone network. This was done mainly for two reasons: firstly to use the collected data to improve the statistical models (acoustic models, language models), secondly to study the behavior of humans in order to build better dialog models for human–machine–interaction.

Speech is completely different from natural language and spontaneous speech is completely different from read speech. There are lots of phenomena of spontaneous speech [14] not yet covered by typical NL systems. Having an operational system, we are in the lucky situation that we can collect data and examine the demands of human partners for dialogs. Although there already exist some integrated dialog systems for spontaneous speech, to the best of our knowledge there was no attempt to collect large corpora of spontaneous speech dialogs and to evaluate an existing system against these corpora.

2. Dialog Corpora

Presently, there already exist large speech corpora. Corpora of *spontaneous* speech are more and more collected in the speech community, but there is still some lack of human machine dialog corpora and the best approach is to collect this kind of data by a bootstrapping method. Initially, WOZ simulations are performed to extract a user model (cf. [6]). With this preliminary model a first operational system was built (cf. [9]) and exposed to real users. Adaptation and enhancement of that system is done for increased stability and user friendliness. The same approach was used in our setup and at each phase of system enhancement the dialogs are collected to build a new corpus. Figure 2 gives an outline of the collected data.

We started our collection[1] with phase 2 using a high quality microphone and made a first corpus of in house inquiries to our dialog system. Due to poor stability of the dialog system, the tests were supervised by an operator who knows the principal system behavior and logs manually every user utterance. The next step of increased difficulty incorporated the switch to telephone quality in phase 3. Again, the experiments were supervised by an operator. Additionally we switched to a more sophisticated parser resulting in a much higher accuracy of the semantic description. An intermediate phase 3+ was collected while we made substantial improvements to our system. This phase reflects the attempts of the system developers to enhance the overall stability and to adapt the feature processing component for telephone line adaptation. Phase 4 was the first corpus of spontaneous speech collected over public telephone line. Starting with this corpus, anyone could call the

[1] Phase 1 utilized read speech for recognizer evaluation and is not reported in this paper.

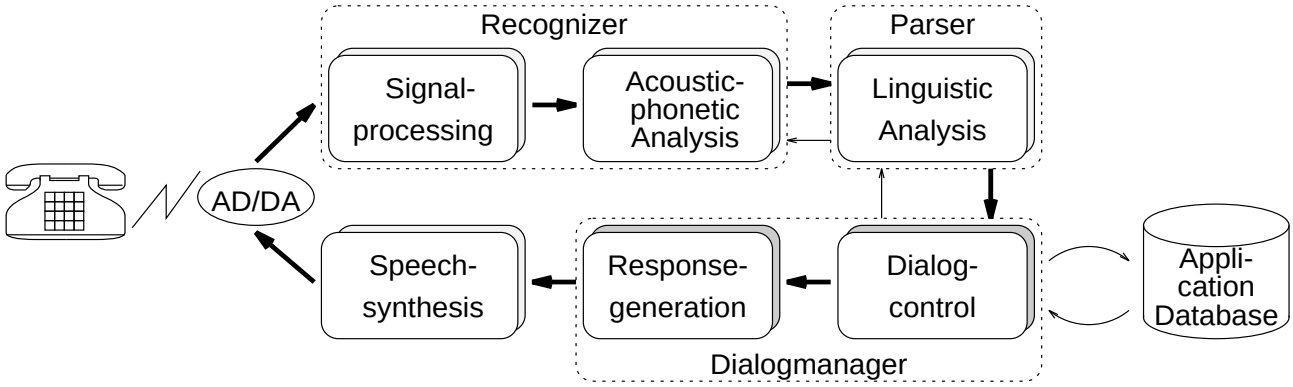

Fig. 1. Structure of the Erlangen Train Time Table Inquiry (T3I) System.

system. In phase 5 we used a completely different dialog model which resulted in much longer user utterances with a higher word frequency. For phase 6 the language models of the recognizer were enhanced and in phase 7 we incorporated a new release of the parser and of the dialog manager.

Since the performance of particular modules is already reported in other publications [8, 4], we concentrate on global experiences concerning the user model and the interaction with spoken dialog systems.

3. User and Dialog Models

Based on these dialog corpora we evaluated and enhanced our demonstration system step by step to fit the users demand. The dialog model has undergone only minor modifications whereas the semantic analysis component contextual pragmatic interpretation have been completely redesigned for the sake of robustness. We found the dialog model fairly stable and it could manage even in worst cases when fed with at least partial correct semantic and pragmatic interpretations of user utterances. Since our system is based on the principle of partial processing, the stability of the system is achieved by the interaction of robust cooperating modules (cf. [2, 5]).

No automatic system works perfectly well without any flaws. In our system we decided to use a dialog strategy which confirms every parameter given by the user. This results in a slightly longer dialog, but several of the failures due to misrecognition and misunderstanding are managed without complete failure. Users get used to confirm every useful information they provided in the utterance before.

Initially, we intended to utilize the collected material of former sessions in a stepwise improvement of the dialog models and user models. Since no dialog corpora were available, this approach would start with a minimal system and incorporate all experiences made with previous users into an improved dialog model. Training a word recognizer follows this approach. Recognition performance benefits from every utterance added to the training material.

However, we found a surprising fact: when the enhanced system was online, users did behave completely different than they did before. This implies that for the development of user models their interaction with the system is determined by the system itself. The direct feedback is an indication of the highly flexible communication strategy in human minds.

Modeling of human behavior is difficult. Since just a small modification of parameters results in major differences of behavior, the current method of dialog modeling is just a starting point. Full description of the complex interaction of human and machine is not yet possible. Eventually, we hope to reach a point of asymptotic stability where the gradual enhancement of the system leads to just a minor change in the humans behavior or no change at all.

4. Main Results

In the following paragraphs we describe some major results observed while collecting the dialog corpora. These results are just overall experiences, detailled results and figures are already published (cf. [1, 2, 4, 5, 8, 12]).

Training of the word recognizer needs spontaneous speech: Due to the lack of spontaneous speech corpora the word recognizer of the initial system was trained only with read speech. Additional spontaneous training data increased the performance substantially, even if the spontaneous training data belongs to a completely different domain. In the training phase the recognizer is able to extract the typical features of spontaneous speech and to transfer it to an already built model. In Phase 4, for example, the additional spontaneous training data increased the word accuracy from 59.5 % to 72.1 %.

	Phase 2	P 3	P 3+	P 4	P 5	P 6	P 7	P 8
collection date, started	9306	9311	9312	9401	9404	9408	9409	9501
corpus size (MB)	127	42	53	133	50	28	204	
number of dialogs	237	49	77	161	42	35	325	
number of utterances	1742	585	533	1365	199	303	2187	
number of words	6384	1841	1668	4238	1144	1154	6773	
number of different words	239	191	168	320	196	174	1056	
number of unknown words	68	25	21	111	77			ongoing
total duration in sec	3983	1318	1677	4152	1590	902	6324	
avg length of dialog in sec	183	220	204	183		188	179	
avg length of utterance in sec	2.4	2.3	3.1	3.0	2.8	3.0	2.89	
avg words per utterance	3.67	3.16	3.13	3.11	5.75	3.80	3.10	
microphone / telephone	mic	⟵ PABX ⟶		⟵ PSTN ⟶				
supervised / unsupervised	⟵ supervised ⟶		selfsup.	⟵ unsupervised ⟶				
user skills	⟵ seminaive ⟶		expert	⟵ naive ⟶				

Fig. 2. Our corpora of human machine dialogs.

Details are reported in [12].

The linguistic complexity of user utterances is extremely low: In our domain of timetable inquiries the linguistic complexity is quite simple. Whereas the initial utterance of a dialog shows the highest complexity by far, nearly all following utterances are elliptic and consist of just a few (1–4) words. About 38% of all utterances have a length of one word, 20% have length 2 and 10% have length 3.

Typical users are not creative in their answers: When designing the dialog models we assumed a much higher degree of variations in user responses, e.g. when being asked for a departure city they might give an overinformative answer with city and specifying, say, the departure time. Apart from the initial utterance of a dialog, the typical user either corrects the previous system response or provides exactly the requested information. Astonishingly, this behavior is also observed in human–human–dialogs of the same domain (cf. [7]).

Processing of partial utterances is vital for robustness: Since we have to handle recognition errors and spontaneous phenomena, the typical utterance is not grammatically well–formed. Processing of partial descriptions in parser and dialog manager solves problems introduced by these phenomena.

Quality of synthesized speech has influence on users speech: Depending on the quality of the synthesizer (intelligible, speed, machine or human voice) the users articulative clarity as well as the speaking style vary. Clear non-human voice is prefered over human voice, since users talking to "humans" tend to be lazy in grammar, clearness and usage of words.

System responses within 10 seconds are acceptable: While immediate response is prefered, the users accept to wait for the next system utterance if this utterance directs the dialog to the goal. Waiting for a system just to respond with an isolated question for confirmation is boring and humans tend to express this in several ways (using verbal and prosodic forms).

Realizing a hung connection: In Germany there is no feedback provided by the Telekom for an on hook phone at the other side. Apart from a simple crack there is usually no audible message for a disconnected line. So the acoustic recognizer needs to have a time-out feature and the dialog strategies have to watch for subsequent timeouts in order to finish a call. This situation is not satisfying and should be characterized as a workaround.

Humans have the tendency to react in an unpredictable way and to get angry when the machine makes mistakes. Models of humans behavior are incomplete. Since the variety of possible (spoken) reactions is that large, we have to investigate in corpus analysis techniques. Large corpora can clarify the distinction between typicality and randomness in user reactions. Current handcrafted models have to be viewed in the light of these typicalities and should modified according to a better comprehension of human–machine–interaction.

The whole system performs better than just the sequence of the components. Traditional methods for sequential systems like the modules recognizer, parser and dialog manager estimate the system success by the product of each modules performance probabilities. However, we found that in a dialog system there is not a strict sequence of "filters", but a closed loop.

Robust methods in our system (cf. [2, 5, 8]) are up to a certain point resistant to minor processing mistakes like recognition errors in unimportant words, splitting sentences by means of repetitions or corrections or other phenomena of spontaneous speech or grammatical incorrectness of spoken utterances. Each module is designed specifically for robust processing of spontaneous speech and handling of difficulties in real user situations. They use prediction methods of user behavior and cooperative users usually behave according to the predictions. Large corpora of collected data help to improve the user models used in all modules.

5. Further Work

Further enhancements are considered to integrate prosodic features into the system. Preliminary studies of the prosody of time expressions are quite promising. Prosody is useful to guide a dialog as shown in the following example: the second user utterance could either be a confirmation (yes — at nine o'clock) or a correction (no — at nine AM). This kind of ambiguity might be resolved by considering prosody.

User(1):	at nine o'clock
System(1):	You want to leave at 9 PM ?
User(2):	at nine o'clock
System(2):	?

Since our system performs fairly well, porting it to a similar task was considered. In cooperation with the Centre de Recherche Informatique de Montréal (CRIM) we integrated their recognizer and semantic analysis component with our dialog manager. The resulting system for a subset of the ATIS domain was integrated and will be used for collecting a corpus of spontaneous spoken dialogs in English using the already described bootstrapping method.

Further evaluation of the collected data will include the cumbersome manual examination of dialog performance rates according to methods presented in [13]. Finally, we plan to bundle all phases with the transliterations and semantic annotations of the dialogs in a common method on CDROM.

6. Acknowledgements

We wish to thank all colleagues and students involved in the collection of the sample dialogs and their postprocessing. The parser for the dialog system was provided by FORWISS (Erlangen) and is sponsored by Daimler–Benz Research (Ulm) within the SYSLID project. Special thanks to Prof. de Mori and Dr. Normandin at CRIM for the fruitful cooperation.

REFERENCES

[1] W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, and E. Schukat-Talamazzini. A Spoken Dialogue System for German Intercity Train Timetable Inquiries. In EUROSPPECH 93 [3], pages 1871–1874.

[2] W. Eckert and H. Niemann. Semantic Analysis in a Robust Spoken Dialog System. In *Proc. Int. Conf. on Spoken Language Processing*, pages 107–110, Yokohama, Japan, Sept. 1994.

[3] *Proc. European Conf. on Speech Communication and Technology*, Berlin, Germany, Sept. 1993.

[4] G. Görz and G. Hanrieder. Robust Parsing of Spoken Dialogue using Contextual Knowledge and Recognition Probabilities. In *Proc. Workshop on Spoken Dialog Systems*, Aalborg, 1995.

[5] G. Hanrieder and P. Heisterkamp. Robust Analysis and Interpretation in Speech Dialog. In Niemann et al. [10], pages 204–211.

[6] L. Hitzenberger and H. Kritzenberger. Simulating Experiments and Prototyping of User Interfaces in a Multimedial Environment of an Information System. In *Proc. European Conf. on Speech Communication and Technology*, pages 2:597–600, Paris, France, Sept. 1989.

[7] M. Keil. Analyse von Partikeln in einem sprachverstehenden System. Master's thesis, Philosophical Faculty (II), University Erlangen, 1990.

[8] T. Kuhn. *Die Erkennungsphase in einem Dialogsystem.* Number 50 in Dissertationen zur Künstlichen Intelligenz. Infix, Sankt Augustin, 1994.

[9] S. McGlashan, N. Fraser, N. Gilbert, E. Bilange, P. Heisterkamp, and N. Youd. Dialogue Management for Ttelephone Information Services. In *Proceedings of the International Conference on Applied Language Processing*, Trento, Italy, 1992.

[10] H. Niemann, R. DeMori, and G. Hanrieder, editors. *Progress and Prospects of Speech Research and Technology: Proceedings of the CRIM / FORWISS Workshop*, PAI 1. Infix, Sept. 1994.

[11] J. Peckham. A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Projekt. In EUROSPPECH 93 [3], pages 33–40.

[12] E. Schukat-Talamazzini. Speech Recognition for Spoken Dialog Systems. In Niemann et al. [10], pages 110–120.

[13] A. Simpson and N. Fraser. Black Box and Glass Box Evaluation of the SUNDIAL System. In EUROSPPECH 93 [3], pages 1423–1426.

[14] W. Ward. Understanding spantaneous speech. In *Speech and Natural Language Workshop*, pages 137–141. Morgan Kaufmann, Philadelphia, 1989.