

TABOR - A NORWEGIAN SPOKEN DIALOGUE SYSTEM FOR BUS TRAVEL INFORMATION

Magne H. Johnsen¹ Torbjørn Svendsen¹ Tore Amble¹ Trym Holter² Erik Harborg²

¹Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway

²SINTEF Telecom and Informatics, N-7465 Trondheim, Norway

{mhj,torbjorn}@tele.ntnu.no {trym.holter,erik.harborg}@sintef.no

ABSTRACT

This paper describes the development and testing of a pilot spoken dialogue system for bus travel information in the city of Trondheim, Norway. The system driven dialogue was designed on the basis of analyzed recordings from both human-human operator dialogues, Wizard-of-Oz (WoZ) dialogues, and a text-based inquiry system for the web. The dialogue system employs a flexible speech recognizer and an utterance concatenation procedure for speech output. Even though the system is intended for research only, it has been accessible through a public phone number since October 1999. During this period all dialogues have been recorded. From these, approximately 350 dialogues were selected for annotation and comparison to 120 dialogues from the WoZ recordings.

The experiments showed that the turn error rate was more than twice as large for the real dialogues as for the WoZ calls, i.e., 13.3% versus 5.7%. Thus, the WoZ results did not give a reliable estimate for the true performance. Our experiments indicate that the current flexible speech recognizer should be further optimized.

1. INTRODUCTION

Spoken dialogue information systems over the telephone line have enjoyed growing commercial interest over the last few years, and a large number of such systems have been developed and tested. The systems vary a lot in complexity, not only due to the variation in tasks and design techniques, but also because of the difference in targeted user friendliness. Recently, standard methods for evaluation of spoken dialogue systems have been suggested [1].

The main components in a spoken dialogue system are the speech recognizer and text-to-speech modules, the natural language processing (NLP) unit and the dialogue manager. The dialogue manager keeps track of the dialogue state and history, accesses the database, and forms a textual output prompt for the text-to-speech module. The NLP unit performs the semantic interpretation of the output from the speech recognizer. This unit is interfaced with the dialogue manager, and can employ knowledge about the current dialogue state and the dialogue history.

The dialogue style has a large impact on the system complexity. The ultimate goal is to develop spoken dialogue systems with a natural and flexible dialogue. This involves that both the system and the user should be able to influence the dialogue flow, and is often termed a *mixed initiative* system [2]. It should allow the

user to take control over the dialogue, i.e., give subjectively relevant information in a possibly over-informative manner, correct, negate or verify the system hypotheses.

In a *system driven* spoken dialogue system, the user cannot take much initiative himself, but is restricted to answer the questions posed by the system. Some minimum user control is however always required, e.g., in order to ask for help or to negate the last system hypothesis. For a given task, a user will normally prefer the mixed initiative dialogue over the system driven one, but this obviously involves a large increase in system complexity. Also, even the mixed initiative system requires the option of using a system driven dialogue as a fall-back-strategy, typically in situations where the speech recognizer breaks down due to a poor signal.

This paper describes the first version of a spoken dialogue system for Norwegian. The pilot system TABOR is implemented for the domain of bus travel information for the city of Trondheim in Norway. This task was chosen for several reasons. First of all, the dialogue has a manageable complexity. We also have access to the bus company's databases, and a text-based NLP inquiry system, BusTUC [3], already exists. Our primary goal was to develop a demonstrator which could form the basis for further research, and a system driven approach was chosen. In the current phase, this work is extended towards a mixed initiative spoken dialogue system in the same domain.

2. THE PRELIMINARY DATA ANALYSIS

In order to get a coarse impression of the way naive users communicate with an operator, more than 100 real human-human dialogues were recorded and annotated. These recordings showed that:

- In order to *accurately simulate* real dialogues, a very complex mixed initiative dialogue structure is required.
- In contradiction, *most user goals* could be met by a straightforward system driven approach.

The web-based information system BusTUC [3] has been operational for several years, and a significant amount of inquiry text data is thus available. BusTUC requests the user to formulate a complete inquiry in a single, preferably grammatically correct sentence, and thus there is no real dialogue involved. The logs from BusTUC were analyzed together with the annotated real dialogues, and on basis of this, the first version of a system driven dialogue structure was designed and implemented for use in a WoZ setup.

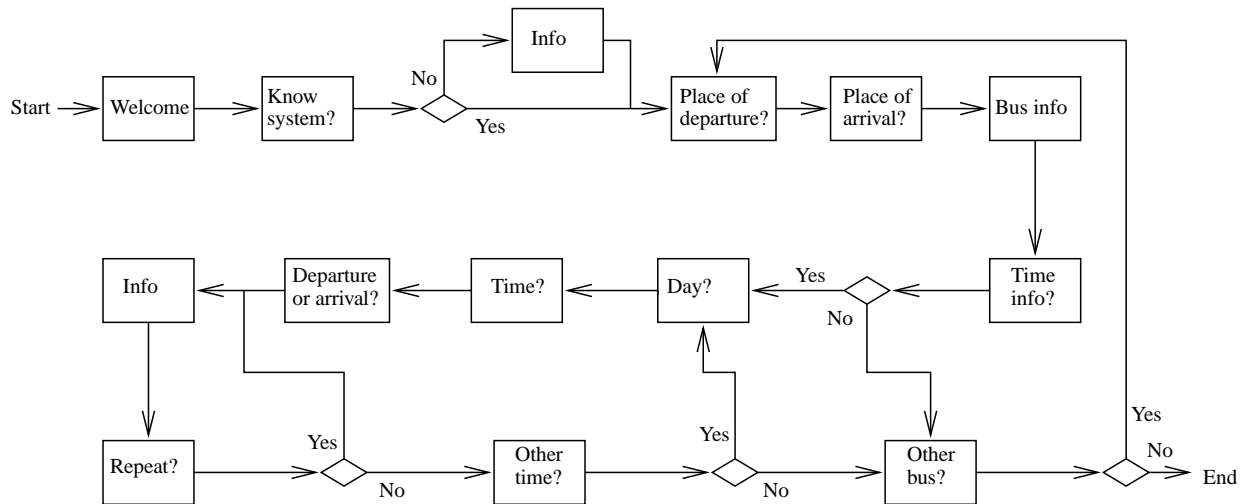


Figure 1: The system driven dialogue structure.

The WoZ dialogue system was tested for an 8 week period. During this time, the dialogue structure was continuously improved. The resulting structure was then used to record 150 WoZ dialogues. These recordings have been annotated according to the SpeechDat standard [4].

3. THE DIALOGUE STRUCTURE

A slightly simplified version of the final dialogue structure is shown in figure 1. The dialogue manager has a total of ten states, each corresponding to a distinct dialogue stage and a single user prompt. The ten states only need five different word networks for speech recognition, corresponding to five sub-tasks:

1. *Yes/no*: In five different dialogue manager states, the user is prompted to answer yes or no.
2. *Bus stops*: In two dialogue manager states ('Place of departure' and 'Place of arrival'), the user is prompted to give the name of the bus stop. The vocabulary in these states includes nearly 600 names.
3. *Day of week*: The user is prompted to say the day of travel. The vocabulary contains the seven days of the week, in addition to expressions like "today" and "tomorrow".
4. *Time information*: The user is prompted to give the desired travel time. This is actually the most complex network (though not the largest vocabulary), both due to pronunciation and syntactic variations which are experienced in Norwegian [5].
5. *Departure/arrival*: The user is prompted to answer whether the specified time relates to departure or arrival.

In addition to the state specific vocabularies, all networks included a common set of words for control, negation and help.

The dialogue system is implemented on a Linux-based PC. A software library supplied by Telenor R&D, TabuLib, has been used

for handling of the ISDN interface, the speech I/O module interface and for interface to the HAPI recognizer [6].

3.1. The Speech Recognizer

A flexible speech recognizer was designed according to the procedure developed in COST Action 249, *Continuous Speech Recognition over the Telephone* [7]. The SpeechDat database [8] was used for training of the context dependent phone set with a relatively strong degree of state-tying. The HMM state output observations were modeled by 8-component Gaussian mixture models. In addition to the triphones, models for the SpeechDat defined man-made noise were trained. Finally, a lexical based filler model was designed from a set of monophones, according to a simplified version of the procedure presented in [9].

The overall vocabulary consisted of approximately 700 words, of which nearly 600 represented different bus stop names. Less than 100 of the vocabulary words existed in the SpeechDat dictionary, and thus an initial lexicon was designed by a phonetician. This lexicon included more than 1500 entries, i.e., an average of more than 2 pronunciations per word. The lexicon was pruned to 800 entries (1.1 pronunciations per word) in a semi-automatic manner. A part of this procedure involved forced alignment recognition on a subset of 30 dialogues from the WoZ recordings. With this new lexicon, the word error rate was significantly reduced for the rest of the WoZ recordings.

3.2. The NLP Module and the Dialogue Manager

The NLP module for the text based inquiry system, BusTUC, is based on a complex set of rules. It is versatile with respect to understanding and answering a variety of alternative formulations requesting the same information. BusTUC was chosen as the basis for the NLP module in TABOR. A dialogue based NLP module, TaBuss, was designed as a front-end to BusTUC in order to

facilitate communication with the speech I/O modules in a system driven way. TaBuss thus accepts single semantic entities and returns an XML based prompt for each turn.

Due to TaBuss, the dialogue manager could be made very simple. Its main tasks is restricted to:

- Keep control over the dialogue states.
- Exchange information between the different modules.
- Analyze and decompose the XML prompts from TaBuss.

3.3. The Speech Output

In a system driven dialogue like TABOR, the number of possible spoken responses from the system is limited. Thus, it is feasible to utilize concatenation of pre-recorded utterances for the generation of speech prompts. For this purpose, a total of approximately 2000 sentences covering all the variables and fixed parts of the prompts were read, including versions with both a flat and a falling intonation. These recordings were then manually edited prior to a simple gain normalization. The procedure resulted in speech prompts which were found both pleasant and fully understandable.

4. EVALUATION OF THE SYSTEM

A preliminary evaluation of TABOR has been performed on basis of the logged dialogue recordings. Up till now, there still does not exist any agreed upon standard for evaluation of dialogue systems. However, serious candidates, like the PARADISE scheme [1] has been proposed. In the evaluation metric for most methods, a subjective overall score from the user is normally included. At this stage, we have not incorporated this.

4.1. The Evaluation Procedure

Our main goals was to investigate:

- Which measures will give us an indication of the real demonstrator performance?
- Can the WoZ database be used for prediction of the real system performance?
- What are the main differences between the WoZ dialogues and the real system dialogues?

The turn error rate is believed to give a good indication of the quality of the recognizer performance for the TABOR system. The user success rate was harder to define and thus to estimate, as many users had more than one goal, but did not get answers to all of them. The reason for this was mainly due to the use of out-of-vocabulary (OOV) names or out-of-domain (OOD) questions. One important reason that OOV names occurred was that aliases for the correct name of bus stops were incorporated only to a limited degree. Thus, it is fair to claim that this version of TABOR in many cases requires considerable local knowledge from the user in order to be able to pose successful queries to the system. In addition, many users violated the dialogue structure, usually by supplying information in an over-informative manner. The number of illegal turns will give a good indication of the quality of the

dialogue structure. The number of turns per goal is also an indicator for the effectiveness of the dialogue structure. Both these parameters have been investigated.

For comparison, the same evaluation was performed for the WoZ recordings. The WoZ informants were given written tasks, thus the relative amount of OOD and OOV utterances were modest. The human Wizard responded both to over-informative responses (by extracting the relevant information) and to OOD queries ("I am sorry, but I cannot answer that"). Thus, the WoZ results can be regarded as a performance limit for the TABOR system. A turn error rate and a turn illegal rate close to the WoZ results, should thus indicate a system with a satisfactory recognizer performance and dialogue structure.

4.2. The TABOR Logs

The TABOR system is primarily intended for research issues, and the system lacks some important qualities which one would have to incorporate in order to offer the system as a commercial service. Despite this, it was decided to offer the current system to naive users through a public telephone number. The main intention was to get a first impression of the popularity and usefulness of such systems.

Even though the demonstrator has not been announced to a large degree, a surprisingly large number of calls have been logged. Through a period of approximately six months, more than 800 calls were received. However, more than half of these were discarded from further use due to one or more of the following reasons:

- The call was not an serious attempt to gain information. The typical scenario in this class is a call made from a party.
- The dialogue restrictions were deliberately violated.
- The dialogue restrictions were violated in an over-informative manner.
- The questions were mainly of the OOD-type.
- The user addressed alternately the system and people in the same room, but talked into the handset microphone all the time.

Obviously, our experiments and results will be strongly dependent upon the selection and discarding of illegal calls. We ended up with approximately 350 calls for further use. To a large extent, the users in these calls tried to follow the system dialogue structure. However, still most of the users can not be characterized as cooperative. Thus we believe that these calls are representative for naive but serious users of the spoken dialogue system.

5. EXPERIMENTS AND RESULTS

In this section, we report the experiments performed with the two data sets we have available, i.e., the recordings from the WoZ callers and the real users. In both cases, the dialogues were annotated according to the SpeechDat standard and compared to the output from the speech recognizer. Turns with truncated and/or unintelligible speech were removed from the test set. Turns including OOV bus stops and/or OOD requests were defined as illegal turns. Table 1 summarizes the results:

<i>Number of:</i>	<i>WoZ</i>	<i>Real</i>
dialogues	120	350
turns	1412	3019
turns per dialogue	11.8	8.6
illegal turns	47 (3.3 %)	321 (10.7 %)
turn errors	80 (5.7 %)	402 (13.3 %)

Table 1: Evaluation of WoZ and real dialogues.

The turn error rate is more than twice as large for the real dialogues compared to the WoZ dialogues. The performance difference of the speech recognizer for the two data sets is most evident with the largest word network, i.e., for recognition of the bus stops. This is probably due to a higher level of motivation among the WoZ callers. The rather poor performance achieved in these experiments confirms that the speech recognizer needs further optimization.

The results also show a large difference between the WoZ and real dialogues with respect to the number of illegal turns. This is not surprising, as the WoZ callers were given written instructions which described the task in some detail. It also seems like a large proportion of the real users do not have the proper knowledge about the system limitations. Despite this, only 10% asked for directions of use. We hope this can be improved by a change in the welcome prompt. Although naive users dominated also among the WoZ callers, they probably benefited from the implemented WoZ strategy in learning the system limitations. In particular, giving an intelligent answer like “Sorry, I can not answer this kind of questions” to the OOD phrases is likely to improve the callers understanding of how the system works.

6. CONCLUDING REMARKS AND FURTHER WORK

In this paper we have described the Norwegian bus travel information system TABOR, and described a preliminary evaluation of it. We have compared the performance achieved for real users with a corresponding experiment using the recordings from a WoZ session. As expected, the results for the real users showed a significant deterioration with respect to both the relative number of illegal turns and the turn error rate, compared to the WoZ callers. The results indicate that the dialogue structure as well as the speech recognition engine should be further optimized. Currently, we are working on task adaptation of the acoustic models.

A mixed initiative spoken dialogue system in the same domain as described in this paper is currently being developed. The experiences from the TABOR project will be valuable in this work. Future work will include public testing and evaluation of this new system.

7. ACKNOWLEDGMENTS

This work was financed by the Norwegian Research Council and Telenor R&D. In addition, Telenor R&D has developed the software library, TabuLib, which forms the basis for the TABOR demonstrator. The flexible speech recognizer design procedure is due to COST Action 249 *Continuous Speech Recognition over the Telephone*.

8. REFERENCES

- [1] M. Walker, C. Kamm, and J. Boland, “Developing and testing general models of spoken dialogue system performance,” in *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, (Athens, Greece), pp. 189–192, May 2000.
- [2] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide, “The thoughtful elephant: Strategies for spoken dialog systems,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8, pp. 51–62, Jan. 2000.
- [3] T. Amble, “BusTUC—a natural language bus oracle,” in *Applied Natural Language Processing Conference*, (Seattle, USA), Apr. 2000.
- [4] <http://www.phonetik.uni-muenchen.de/SpeechDat.html>.
- [5] K. Kvale and I. Amdal, “Improved automatic recognition of natural Norwegian numbers by incorporating phonetic knowledge,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, (München, Germany), pp. 1763–1766, IEEE, Apr. 1997.
- [6] J. Odell, D. Kershaw, D. Ollason, V. Valtchev, and D. Whitehouse, *The HAPI Book VI.4*. Entropic Ltd., Jan. 1999.
- [7] F. T. Johansen, N. Warakagoda, B. Lindeberg, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, “The COST 249 SpeechDat multilingual reference recogniser,” in *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, (Athens, Greece), pp. 1351–1354, May 2000.
- [8] H. Höge, C. Draxler, H. van den Heuvel, F. T. Johansen, E. Sanders, and H. S. Tropic, “SpeechDat multilingual speech databases for teleservices: Across the finish line,” in *Proc. European Conf. on Speech Commun. and Techn. (EUROSPEECH)*, (Budapest, Hungary), pp. 2699–2702, Sept. 1999.
- [9] R. E. Meliani and D. O’Shaughnessy, “New efficient fillers for unlimited word recognition and keyword spotting,” in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, (Philadelphia, USA), pp. 590–593, Oct. 1996.