Different ways of ending human-machine dialogues

Loredana Cerrato[1] & Susanne Ekeklint[2]

GSLT- Graduate School of Language Technology, Sweden
[1]TMH-CTT, Royal Institute of Technology, Sweden
[2]School of Mathematics and Systems Engineering, Växjö University, Sweden

Abstract

More and more dialogue systems are developed to provide public services to inexperienced users. This makes the competition harder among developers and increases the interest in finding methods to evaluate and improve the overall performance of dialogue systems. We believe that analysing the way in which users end their interactions with dialogue systems could provide some new metrics to evaluate users' satisfaction. To prove our hypothesis we carried out a preliminary analysis of the way in which users of two Swedish multi-modal conversational dialogue systems (namely August and Adapt) ended their interactions. The results indicate that prosodic cues, like pitch contour and intensity, can be used as indication for users' satisfaction.

Introduction

Evaluation plays a crucial role in speech and natural language processing, both for system developers and for technology users. Developers need to evaluate their product at different stages: before starting the actual development (Initial Evaluation), during the development (Evaluation in Progress) and at the end of the development (Final Evaluation).

Initial Evaluation: Developers are interested in finding the modules that are most likely to give the best dialogue model. Different modules are often separately validated for this purpose. Evaluation of what is adequate for the system to handle often considers parameters such as cost and performance according to the task into consideration (Hirschman and Thompson, 1996).

Work in Progress Evaluation: During the development of systems, different kinds of "Diagnostic Evaluations" are carried out. For example different generations of the system can be compared according to specific parameters that are particularly interesting for different modules or sometimes for the overall performance of the system.

Final Evaluation: In the very last stage of a system development it is necessary to determine the adequacy of a system for a specific purpose. This evaluation can be carried out both by developers and by users, since it can be used typically to compare similar systems (i.e. alternative implementations of the same system or successive generations of the same implementation). The developer can look at parameters for the overall performance of the system by counting for example elapsed time on a particular task or number of speech turns to complete a particular task.

During all these stages in the development there is the need to look at the performance of the separate modules. There are several evaluation methods that propose evaluation measures for individual components. For example

the standard way to evaluate speech recognition and language understanding modules is to (given a certain input) compare the output to the desired output (Lippmann, 1997). Many systems have some kind of confidence scoring of the recognized utterances (Glass 1999). Another module that might be evaluated separately is the dialogue manager. As suggested by Danieli and Gerbino (1995) the total score for the dialogue systems robustness can be measured in terms of the dialogue manager's ability to perform both implicit and explicit recovery when the speech recognition or the language understanding unit fails.

Evaluating the overall performance of a system is also necessary. One way to do this kind of evaluation is to look at the users' satisfaction. One of the more recent tools for evaluation of spoken dialogue systems is PARADISE. This method uses various parameters to calculate an estimation of the user satisfaction. (Walker et. al., 1998). The PARADISE paradigm breaks down the term user satisfaction into costs and success – the goal is to maximize the success and minimize the costs. PARADISE takes into account several parameters when calculating the user satisfaction. For example counting number of rejects, cancels, time-outs and means recognition score. Other things taken into account are number of turns, requests for help, barge-ins, elapsed time etc. Rejects, for example, are the number of times that the recogniser cannot produce a result with enough confidence.

Being a relatively new field, the development of conversational dialogue systems lacks both evaluation tools and established sets of evaluation criteria. Very few empirical tests have been carried out to learn about the benefits of embodied agents on different aspects such as: entertainment, mental load and system efficiency (Sanders & Scholtz, 2000).
The more modalities that are used in a system the more complex the evaluation of the individual components will be, since each component will need to collaborate with a larger number of components. The objectivity when looking at separate modules may also be influenced by other modules since one module's weakness may very well be saved by another module (Carlson and Granström, 1995). For example: if the speech recogniser is not so good, a well working dialogue manager may fix some of the problems of the speech recogniser.

The different methods that have been proposed so far to evaluate dialogue systems are based mainly on modular evaluation and very little focus is put on how users interact with different systems. It is of course possible to ask users for a subjective judgement of a system. This can be done by asking the users specific questions about the way in which their interaction with the system went on. However it is not always practical to use forms or interviews to evaluate dialogue systems, since there are the issues of costs, time and the users' integrity. Moreover there are particular questions which are difficult to formulate such as whether the interface has influenced the users' feelings and expectations during the interactions.

Motivation

Finding new parameters and metrics to evaluate complex dialogue systems is a challenging experience. We believe that the prosody of users' utterances could be considered as one of these parameters that we are searching for.

What we would like to achieve with our investigation is a method to evaluate the degree of users satisfaction in machine interactions based on the analysis

of the users' last utterance. So far prediction of the users satisfaction have been based on different measures like TTS accuracy, ASR accuracy, time to solve the task, expertise of the user, the success of the dialogic transaction. The final goal with this study is to propose a prediction of the users' satisfaction based on prosodic cues of the speaker's last utterance in the interaction. We believe that by interpreting prosodic aspects of the user's contribution (in particular the pitch contour and intensity of the last utterance) it could be possible to get some cues about the users' satisfaction.

Humans are able to understand if a speaker is happy, sad, worried, disappointed; the cues for the interpretation of these emotions are in the utterances spoken by the speakers, therefore we aim at finding out if it is possible to exploit prosodic information for the evaluation.

Our method does not aim at providing an overall evaluation for dialogue systems, but it proposes to exploit prosodic information as a complement to other traditional subjective and modular evaluation methods. The acoustic measurement can be used as an indication to wether or not the users are satisfied with the way the interaction with the system went on. The reasons for the degree of satisfaction will not be given by this measurement. It will however point out when something did not go as well as expected. Even though the method does not cover all the aspects of users' satisfaction it has several positive sides:

- It can be used as an indication of whether or not the system is working as well as the user expected it to.
- It is a method that can be implemented at low-cost, since it is possible to do the tests automatically.
- It can be used as a trigger to start certain modules, for example help modules.
- It might work as a parameter to decide when to save data from dialogues for further investigation.

Material

To carry out our analysis we used material collected by means of two Swedish multimodal conversational dialogue systems: *August* and *Adapt*.

The August database counts more than 10.000 utterances produced by 2685 users, all visitors at the Stockholm Cultural Centre, where the dialogue system was displayed as part of the *Cultural Capital of Europe '98 program*. The August system (Gustafson et al.1999) was endowed with a talking head with a "distinctive personality" (resembling the author August Strindberg), which invited users from the public to try the system and somehow induced them to socialize with it, rather than just seek for straightforward information. The users received no instructions about how to interact with the system. The system was able to give information about three different domains, the Royal Institute of Technology, the city of Stockholm and the life and work of Strindberg. The material of the August database is quite heterogeneous, (users of different ages and nationalities). Most of the recorded dialogues are quite short (less then 5 turns), for this reason we selected a sub-corpus of 280 interactions among those that presented more than 5 turns. We will refer to this material as the *August corpus*.

The Adapt database was collected in 1999 by means of the Wizard of Oz technique. It includes a total of 50 dialogues produced by 33 users. Adapt is a Swedish conversational multi-modal dialogue system (Gustafson et al. 2000). The system is able to provide information about apartments on sale in Stockholm. The dialogues in the Adapt database can be described as "information seeking" or "task oriented" since the users were given the specific task of finding apartments in Stockholm that fulfilled certain criteria. From this database we selected 32 dialogues; we will refer to this material as the *Adapt corpus*.

Analysis

We carried out an analysis of user's last turns in the two corpora. We categorised the final utterances using the typology proposed by Bell and Gustafson (1999 a). They propose a grouping of the utterances in the August database in two main categories: socializing and information seeking. To these two main categories they added further subcategories as reported in table 1.

Socialising	Information seeking
Social	Domain
Insult	Meta
Tests	Facts

Table 1 Utterance typology proposed by Bell and Gustafson

The results of the distribution of the last users' turn in our August corpus are reported in figure 1. The category blank includes those final turns, which consisted of empty sounds file.
The majority of final utterances have been categorised as social. Also in the whole August database a great number of utterances were categorised as social and Bell and Gustafson explained this result by the existence of the animated agent, an explanation which we completely support.

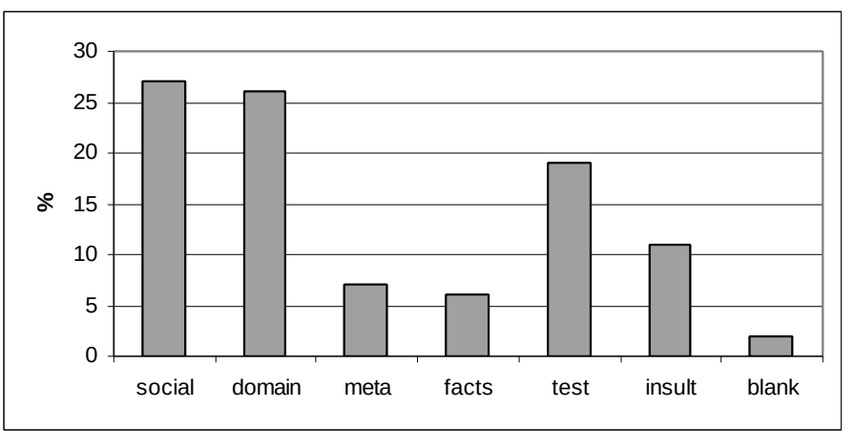

Figure 1 Distribution of the last utterance in the August corpus, according to Bell and Gustafson typology

Since the number of social utterances is so high, we propose a sub-categorisation of them based on the lexical and semantic content:

– Greetings
– Negative comments

– Thanks
– Positive comments
– Insult
– No end

The category “No end” includes all those final turns that could not be categorised as a “Conventional-closure” (i.e. not a greeting, not a thank you, not a comment, not an insult) and also those turns that consisted of empty sound files. We cannot judge if the turns categorised as no-end depend on the user’s choice of concluding the interaction without a conventional closing or whether they are due to some problems of recording. In figure 2 it is reported the distribution of the social final utterance in the August corpus according to the above mention sub-categories.

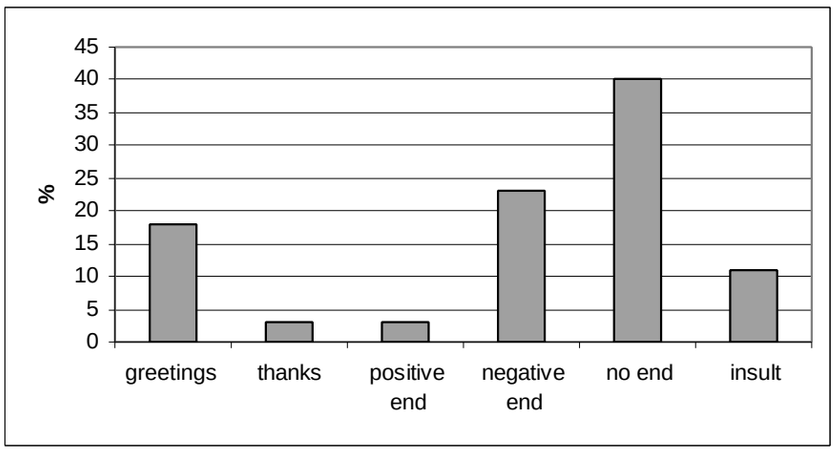

Figure. 2 Distribution of the subcategories of the social final utterances in the August corpus

.

We carried an analysis of the pitch contour of greetings (“hej” “hej då”) and thanks (“tack”) by means of the software package Wavesurfer (Sjölander & Beskow 2000) and we noticed the following trends:

- a rising pitch contour is typical when the user has had a successful interaction with the system;
- a falling pitch contour is typical when the user has not had a successful interaction with the system, due to problem of reciprocal understanding or because the user was asking questions outside the domains;
- a thank you or a greeting with a higher intensity and longer duration (respect to the rest of the words in the utterance) were produced by some users that had not had a successful interaction with the system, who, in our opinion sounded “ironic”.

The last user turn in the Adapt corpus was also categorised according to Bell and Gustafson typology. The results are reported in figure 3. In the graph we have also reported the percentage of turns consisting of an empty sound files (i.e. blank). In the Adapt corpus the types of utterances are not so many as in the August corpus, however also in this corpus the majority of the final utterances (63%) can be categorised as “social”, and they consists mostly of greetings and thanks.

In the Adapt corpus there are no final turns consisting of insults. This probably depends on the fact that the majority of the interactions with the Adapt system went on in a positive way, that is, without problems in reciprocal understanding, while in the August corpus many interactions went on in a problematic way due to problems of reciprocal understanding. 44%

of the users of the Adapt system positively completed their task (i.e. finding the apartment they were instructed to look for).

28% of the users could not find the apartment they were looking for (because there were no similar apartments in the system database), but they managed anyway to have an interaction without problems with the system. The fact that the users of the Adapt system never insulted the system at the end of the interaction can also be an indication of the fact that the interactions went on without problems of understanding.

For the remaining 28% of the users the final turns consist of an empty sound file. This can both depend on the personal choice of the users of not concluding the interaction using a conventional closing sentence or it might as well depend on some problems with the recordings.

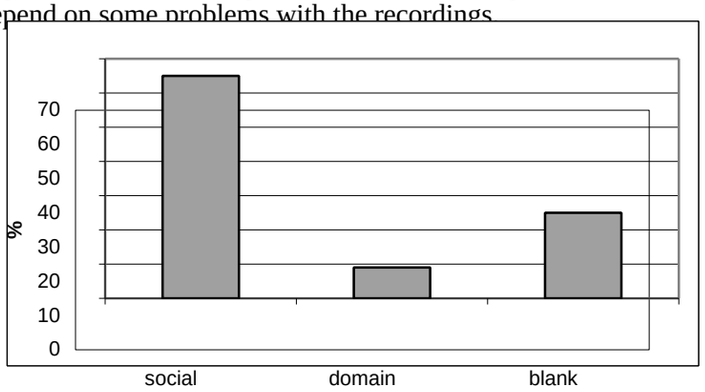

Figure 3. Distribution of the last utterance in the Adapt corpus, according Bell and Gustafson typology.

In figure 4 is reported the sub-categorisation of the social final utterances in the Adapt corpus.

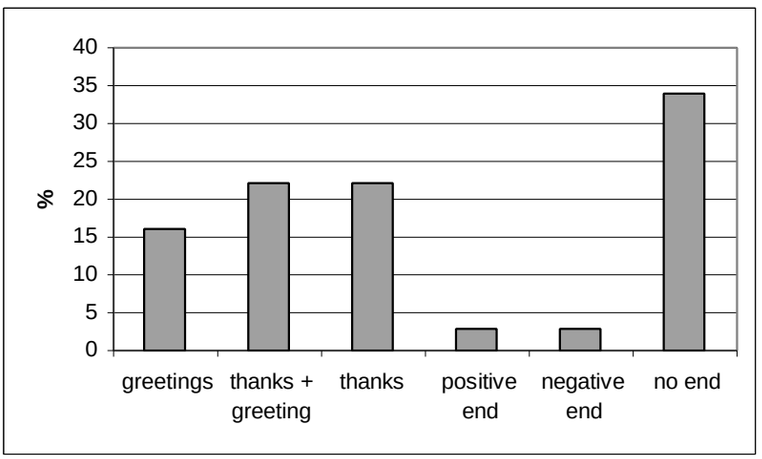

Figure. 4 Distribution of the subcategories of the social final utterances in the Adapt corpus.

An acoustic analysis of a greeting (the expression *adjö* produced by 7 different users) showed a rising pitch contour for those 3 users that successfully completed their task and a falling pitch contour for those 4 users who didn't get the information they required.

Conclusions and further investigations

The results here reported and the results reported by Bell and Gustafson (1999a,b) support the idea that the human-like appearance of the agent induces users to have a more "social" behaviour towards the system they are interacting with. This is an important observation to support the idea that special evaluation methods are needed for multimodal conversational dialogue system. The aim of our study was test our preliminary hypothesis that the way in which users end their interactions with dialogue systems could provide some new metrics to evaluate users' satisfaction.
To prove our hypothesis we carried out a preliminary analysis of the way in which users of two Swedish multi-modal conversational dialogue systems (namely August and Adapt) ended their interactions. The results show that prosodic cues, like pitch contour and intensity, can be used as indication for users' satisfaction.

Of course more thorough investigation is needed to better support our preliminary results. In particular to test our results we need bigger amount of data collected by means of conversational dialogue systems, which are very difficult to get and we need to be able to compare our results with the results of traditional evaluation method. We have planned to carry out a comparative evaluation of the user satisfaction using our method and the Paradise method on a new set of human-machine interactions that are being collected at TMH-KTH with the Adapt system.

References

Bell, L. & Gustafson, J. (1999a). Utterance types in the August System. *Proc from IDS '99*

Bell, L. & Gustafson, J. (1999b). Interacting with an animated agent: an analysis of a Swedish database of spontaneous computer directed speech, *Proc of Eurospeech '99*, 1143-1146

Carlson, R. & Granström, B. (1995), *The Waxholm spoken dialog system.* In: Palková Z, ed. Phonetica Pragensia IX. Charisteria viro doctissimo Premysl

Danieli, M. & Gerbino, E. (1995). *Metrics for evaluating dialog strategies in a spoken language system.* In Proc. of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, pages 34#39.

Glass, J. R. (1999), *Challenges for spoken dialog systems*. Spoken Language Systems Group, MIT Laboratory for computer science, Cambridge.

Gustafson, J., Lindberg N. & Lundeberg M. (1999). The August spoken dialogue system. *Proc of Eurospeech'99*, 1151-1154.

Gustafson J., Bell L., Beskow J., Boye J., Carlson R., Edlund J., Granström B., House D., Wirén M. (2000) AdApt- a Multimodal conversational dialogue system in a n apartament domain". In Proceedings of ICSLP 2000(2) 134-137. Bejiing, China.

Hirschman, L. & Thompson, H. S. (1996). Overview of Evaluation in Speech and Natural Language Processing, In: *Survey of the State of the Art in Human Language Technology.* Postscriptversion, chapter 13.1

Lippmann, R. (1997). *Speech recognition by machines and humans*. Speech communication, Elsevier Science B.V.

Sanders, G. A. & Scholtz, J. (2000). Measurments and Evaluation of Embodied conversational agents. In: *Embodied conversational agents*, Cassel, J., Sullivan, J., Prevost, S., Churchill, E., 346-373, MIT press.

Sjölander, K., Beskow J. (2000). WaveSurfer - an Open Source Speech Tool, Proc. ICSLP 2000, Bejing, China.

Walker, M. A., Kamm, C. A., Litman, D. J. (1998). *Towards Developing General Models of Usability with PARADISE*. In: Natural Language Engineering.