PROMISE – A Procedure for Multimodal Interactive System Evaluation

Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel, Uli Türk

Institute of Phonetics and Speech Communication
University of Munich
80799 Munich
+49 89 2180 5751
[beringer, ukartal, kalo, schiel, tuerk]@phonetik.uni–muenchen.de

Abstract

This paper describes a general framework for evaluating and comparing the performance of multimodal dialogue systems: PROMISE (**Pro**cedure for **M**ultimodal **I**nteractive **S**ystem **E**valuation). PROMISE is a possible extention to multimodality of the PARADISE framework ([1, 2] used for the evaluation of spoken dialogue systems), where we aimed to solve the problems of scoring multimodal inputs and outputs, weighting the different recognition modalities and of how to deal with not directed (non–directed) task definitions and the resulting, potentially uncompleted tasks by the users.

PROMISE is used in the end–to–end–evaluation of the SmartKom project – in which an intelligent computer–user interface that deals with various kinds of oral and physical input is being developed. The aim of SmartKom is to allow a natural form of communication within man–machine interaction.

Keywords

Multimodality, SmartKom, Dialogue system evaluation, Evaluation Framework

1. Introduction

The aim of this paper is to give an extended framework on dialogue system evaluation for multimodal systems in the end–to–end evaluation of SmartKom.

In the SmartKom project, an intelligent computer–user interface is being developed which deals with various kinds of oral and physical input. Potential benefits of SmartKom include the ease of use and the naturalness of the man–machine interaction which are due to multimodal input and output. However, a very critical obstacle to progress in this area is the lack of a general methodology for evaluating and comparing the performance of the three possible scenarios provided by SmartKom:

1. SmartKom Home/Office allows to communicate and to operate machines at home (e.g. TV, workstation, radio),
2. SmartKom Public provides public access to public services, and
3. SmartKom Mobile

Because of the innovative character of the project, new methods for end–to–end evaluation had to be developed partly through transferring established criteria from the evaluation of spoken dialogue systems, and partly through the definition of new multimodal measures. These criteria have to deal with a fundamental property of multimodal dialogue systems, namely the high variability of the input and output modalities with which the system has to cope.

The following section gives an overview of standard problems of dialogue system evaluation which principly can be solved by the PARADISE framework [1, 2]. Section three describes the problem of how to describe the task and define the attribute value keys out of the description – which is a problem not uniquely belonging to multimodal dialogue evaluation. The problem of what to do with incomplete tasks or tasks that get a very low task success measure, due to incooperativity of the user, is described in section four. Sections five to seven give a detailed description of the status of multiple–to–one input facilities, i.e. the possibility to express the same user intention via multiple input as well as via different input modalities. Section eight defines the approach of PROMISE as a multimodal dialogue evaluation strategy which normalizes over dialogue strategy and dialogue systems. In the last section we sum up some ideas to be implemented in our framework.

2. Standard problems of dialogue system evaluation

Of course, multimodal dialogue evaluation has to deal with the same problems spoken dialogue system evaluation has to deal with, namely

1. How can we abstract from the system itself, i.e. the different hardware and software components, in order to evaluate the dialogue?
2. How can we abstract from different dialogue strategies?

The PARADISE framework (for detailed description please refer to [1, 2]) gives a useful and promising approach how to compare different spoken dialogue strategies and different spoken dialogue systems via attribute value matrices (AVM), to compute the (normalized) task success measure (provided that a clearly defined task description is given to the user) define several (normalized) cost functions (Gaussians), and to weight their importance for the performance of the system via multiple linear regression dependent on the User Satisfaction value (cumulative function on the questionnaire completed by the subjects).

3. Task descriptions and key definition

Unfortunately, in dealing with multimodal systems – in particular with SmartKom – we find a number of components which does not fit into the PARADISE approach, which are:

1. The user is given a rather unprecise task definition, in order to enable a mostly natural interaction of user and system. Therefore exist no static definitions of the keys to compute an AVM. This means that we have to compute a multidimensional AVM to cover all possible keys. In the example of an electronic programming guide – one of the SmartKom tasks – we would get a 2000 X 2000 matrix, in which each of the values is potentially correct. The possibility of a mismatch between actual value and key is much higher, and, considering the unprecisely defined task description, is possibly not in the range of "error". One partial solution is to cluster the keys in different superordinate concepts, e.g. movie title, genre, channel, timeslots, actors etc. for the named epg domain in SmartKom, but
2. Since SmartKom has no slot oriented task description and a dynamic definition of the attribute values, we may observe dialogues which require a certain set of keys whereas in other dialogues within the same domain (e.g. EPG) some of the keys within the same set become redundant. To make this clear, imagine the following situation:
   a) Task description: "Imagine, you live in Heidelberg (a German town) and you have just obtained SmartKom. You want to have a nice evening in front of the TV. Your task is to plan your TV programme for this evening via SmartKom."

   b) Keys (clustered): movie title, genre, channel, timeslot, actors

   c) User 1: I'd like to see the news ("movie" title, implicitely genre) at eight o'clock (timeslot) on RTL (channel).

   It can clearly be seen, that the user's request contains ALL necessary keys to solve her task, although "actors" has not been mentioned and in this case is redundant.

   d) User 2: Is there anything special on TV this evening (partly: timeslot)? I'd like to see something with "Jeff Goldblum" (actor).

   Here, only one key is really mentioned (actor), whereas the timeslot is given in the range "this evening".

   e) User 2 (after the system has given "Jurassic Park 1" and "Jurassic Park 2", part one in RTL and ARD (2 Channels) and part 2 in SAT1 (Channel)): I'd like to see "Jurassic Park 1" (movie title) then.

   After this turn all important keys are mentioned: actor, movie title (implicitely genre), timeslot and, redundantly, channel (RTL or ARD).

   As can be seen there are different obligatory keys in different dialogues belonging to the same task.
3. Constructing AVMs within the multimodal dialogue situation of SmartKom could only be possible, if
   a) the task description is switched to a very static one. To stay in our example: "Imagine, you live in Heidelberg (a German town) and you have just obtained SmartKom. You want to have a nice evening in front of the TV. Your task is to plan your TV program for this evening via SmartKom, namely to see "Jurassic Park 1" with "Jeff Goldblum" on RTL at 20:15h". The problem is that such a kind of task description gives no possibility to really test the advances of SmartKom.

   b) all "seen" keys are defined before starting the evaluation. This, of course, defines all further dialogue task success measures, since they would be inadequate if another possible key was found in later dialogues.

   c) "seen" keys are defined per dialogue: some kind of normalization has to be found.

4. How to deal with a bad performance due to users' incooperativity?

One of the main problems of dialogue systems is an incooperative user. This, of course, is not unique to multimodal system evaluation but can occur in other situations as well. On a first cue, it is impossible to value incooperative users without lowering task success and the system performance. To avoid a bad system performance due to incooperative users there exist the following approaches:

1. Only dialogues with cooperative users are evaluated
2. Only dialogues which terminate with finished tasks are evaluated.

The first approach guarantees that there cannot be any bad system performance due to incooperative users; for the second, AVMs can be used to some extent . Using AVMs, however, could have some undesirable effects on task success in such a way that there is a value for task success although none of the evaluated dialogues shows any finished task. To make this clear, just imagine the following:

- An AVM requires four keys.
- 12 dialogues are being evaluated
- none of the 12 dialogues ends in a task success but
- all of the 12 dialogues have some matchings on the keys, which are randomly distributed.
- for each key there is at least one match with the actual value.

Intuitively, one would say, that there is no task success at all, because none of the 12 dialogues will terminate the task. But regarding the AVM, one can find at least a low value for task success.

5. How to score multimodal inputs or outputs?

In contrast to interactive unimodal spoken dialogue systems, which are based on many component technologies like speech recognition, text–to–speech, natural language understanding, natural language generation and database query languages, multimodal dialogue systems consist of several such technologies which are functionally similar to each other and therefore could interfere with each other. To make this clear, just imagine the similar functions of ASR and Gesture Recognition: while interacting with a multimodal man–machine interactive system like SmartKom users have the posibility to say what information they want to have and to simultaneously give the same, an additional, or a more specific input via "interactional gesture" [3]. There are several possible problem solving strategies for the system namely:

1. First match: the information which was recognized first is taken for further system processing, regardless of the recognition method. This would of course not help in multimodal processing.
2. "Mean" match: the system takes the information which is common to both of the recognition modules. This could be called multimodal verification.
3. Additional match: take all the information given by several recognizers for further system processing. This would be the best solution, if we assume all recognizers to be highly accurate, which leads us to the next problem:

6. How to weight the several multimodal components of recognition systems?

How can we estimate the accuracy of different recognizers? I.e., in talking about speech recognition, we have to deal with a very complicated pattern match, whereas gesture recognition has a limited set of recognizible gestures which can be found in a given coordinate plane.

It should be clear, that

1. the gesture recognizer will be more accurate than the ASR system but
2. the improvement of accuracy of the ASR system must get a higher value than the improvement of gesture recognition within the system!
3. Apart from the problems of how to weight the different multimodal system components in an end–to–end evaluation of a multimodal system there, is also the problem of synchrony:

7. Are multimodal inputs synchronous or linear within the evaluation?

Are inputs from different modalities synchronous, i.e. are they describing the same user intention, although they may not be synchronous in time? Or does the system have to cope with different inputs?

8. PROMISE – A Procedure for Multimodal Interactive System Evaluation

In the last sections we have defined the most characteristic problems which show the need for an extended framework for multimodal dialogue system evaluation. We already gave some examples of possible problem solving strategies. Within this section we will precise these ideas and present the current version of PROMISE. Given the normalized performance function of PARADISE

$$\text{Performance} = \alpha * \mathrm{N}(\kappa) - \sum_{i=1}^{n} \omega_i * \mathrm{N}(c_i)$$

with α the weight for κ (task success), the Gaussian cost functions c_i weighted by ω_i, and N z−score normalisation function, PROMISE splits this function in two parts in the way that the formula is reduced to normalized cost functions first. Instead of a multiple linear regression between the free cost variables and the dependent user satisfaction score, PROMISE searches correlations via t−test and User−Satisfaction − Cost pairs. This means objective mesurable costs will be refered to in the questionnaire.

The following costs are defined in SmartKom, some of them equivalent to the PARADISE costs, some of them extended to deal with multimodality or user incooperativity.

Cooperativity of the user, Number of Senses of multiple input, Helps (Help%), Recognition (Speech Recognition, Recognition of facial expression, Gesture Recognition), Barge−In, Cancels, Reliability, Transaction Success, Percentage of diagnostic error messages, Rejections (rejections%Error frequency of input), Timeout (timeout% Error rate of output, Error rate of input), Words/Turn (No. of spoken words or produced gestures), Possibility of misunderstanding of input/output, Task Complexity, Input Complexity, Dialogue Manager Complexity, ways of interaction (using gestures, using speech), Recordings, N−Way communication, Semantical Correctness of Input/Output, Off−Talk [4], Elapsed time (duration of input of the facial expression, duration of gestural input, duration of speech input, duration of ASR, duration of gesture recognition), Task Completion, Dialogue elapsed time, synchrony of graphical and speech output, Object manipulations, Mean System response time, Mean User response time, User/System Turns(Number of Turns, Mixed initiative dialogue management, Incremental compatibility, Percentage of appropriate/inappropriate system directive diagnostic utterances, Percentage of explicit recovery answers).

The second step is to define another way to calculate the task success. Our first approach is to split task success in

1. task complete and
2. task incomplete

each of which is being clustered in

1. user cooperative
2. user incooperative

This splitting allows us to compute different AVMs with different weights resulting in the following formula for system performance:

$$\text{Performance} = \sum_{j=1}^{4} \alpha_j * \mathrm{N}(\kappa_j) - \sum_{i=1}^{n} \omega_i * \mathrm{N}(c_i)$$

with α the weight for κ (task success), the Gaussian cost functions c_i weighted by ω_i, and N z−score normalisation function. The index “j” stands for the four possibilities of the task completion status, namely

1. task complete and user cooperative
2. task complete and user incooperative
3. task incomplete due to incooperative user
4. task incomplete although user is cooperative.

9. Conclusion and future work

Our aim was to define an extended evaluation framework for a multimodal dialogue system evaluation which can deal with multimodal dialogue processing. We defined a new way to define system performance and gave a list of multimodal costs to compute. However, there are still some unresolved or solely unsatisfactorily solved problems dealing with the linearity of multimodality. We are currently specifying different approaches in order to satisfactorily solve the remaining problems which we hope to present at the LREC pre−conference workshop on “Multimodal Resources and Multimodal Systems Evaluation” in June.

References

[1] M.A. Walker et al.: Evaluating Spoken Dialogue Agents with PARADISE:]Two Case Studies, In Computer Speech and Language, Vol. 12, No. 3, 1998.

[2] M.A. Walker, D.J. Litman, C.A. Kamm, A. Abella: PARADISE: A Framework for Evaluation Spoken Dialogue Agents. Annnual Meeting of the Association of Computational Linguistics. ACL 1997.

[3] S. Steininger, B. Lindemann, T. Paetzold: Labeling of Gestures in SmartKom − The Coding System, Gesture Workshop 2001, London, UK. to appear in: Springer "Gesture Workshop 2001, London" (LNAI 2298)

[4] Daniela Oppermann (2000): OFF−TALK − Ein Problem für die Mensch−Maschine−Kommunikation? SmartKom Memo 04−00.