FLY WITH THE EAGLES: EVALUATION OF THE "ACCeSS" SPOKEN LANGUAGE DIALOGUE SYSTEM

Gerhard Hanrieder

Daimler-Benz Aerospace AG
Defense and Civil Systems
D-89070 Ulm, Germany
hanriede@vs.dasa.de

Paul Heisterkamp

Daimler-Benz AG
Speech Understanding Systems
D-89013 Ulm, Germany
heisterkamp@dbag.ulm.
daimlerbenz.com

Thomas Brey

University of Regensburg
FG Informationswissenschaft
D-89053 Regensburg, Germany
thomas.brey@sprachlit.uni-
regensburg.de

ABSTRACT

This paper reports the experiences we had in evaluating the ACCeSS system using the EAGLES evaluation metrics both at the input/output (black box evaluation) and component levels (glass box evaluation). We deliver an example of a complete evaluation of a continuous speech/mixed initiative system using these standards. Furthermore, we discuss some useful extensions to them.

1. INTRODUCTION

Evaluation is needed during the development of spoken language dialogue systems to ensure that changes in the system design and implementation really result in improvement. For objective measurements, metrics are needed. When and if different systems are to be compared, it is necessary to have common metrics for both of them, better still, to have a metrics standard. Because of this, we chose the metrics proposed by the EAGLES group ([6], cf. section 2 below) as the reference framework for the evaluation that accompanies the design process of the ACCeSS system (cf. section 3). As Fraser ([6]:604f.) states, these metrics are provisional and far from complete. Their usefulness should be tested in applying them to different systems. When in the ACCeSS project the need arose to quickly change to a new application, we took the opportunity to do just this. Our development methodology is the 'System-in-the-loop' (cf. [1]) rather than 'Wizard-of-Oz', as we can do prototypes fast and make rapid changes in the design. In our first evaluation round (section 4), the tests with only 40 test persons already enabled us to identify some problems in the dialogue design and make the appropriate changes (section 5). For this, it was crucial that we extended the metrics to include a 'subtask' success rate. We will continue using the metrics as presented in this paper for the further development of ACCeSS as well as other systems.

2. THE EAGLES METRICS

For our development evaluation we chose the EAGLES metrics, because they are evolving into a standard and thus are useful for comparisons. The core evaluation metrics are described in [6], but previous versions (cf. e.g. [4],[8]) suggest that the still emerging field of spoken dialogue management needs some discussion and examples to make these metrics really useful. EAGLES shows two principles of evaluation: Black Box (aka performance evaluation) and Glass Box (aka diagnostic evaluation). The metrics for our Black Box evaluation are:

- Turn Duration (TD): Measures length of utterances and is an indicator of complexity.
- Dialogue Duration (DD): indicates problems with dialogues much shorter (break-off) or longer (many corrections) than average. Mostly, shorter average lengths are considered better.
- Correction Rate (CR): Indicator of number of corrections and thus misrecognitions or misinterpretations.
- Transactions Success (TS): Indicator of success of whole dialogue systems in terms of achieving the user's goals.

CR and TS can only be measured manually, because what counts as a correction and what as success, is a question of interpretation of the data. Especially TS is critical, as it presumes a very narrow band of task success. We show in the next section that the current ACCeSS application and the resulting dialogues have a very complex task structure. This is why we expected TS to be problematic from the start. There is one further metric, viz. Contextual Appropriateness, used to measure output generator performance. This we did exclude from using because the process proposed for its measurement proved too unwieldly, and lack of appropriateness should also result in corrections and other problems. Thus, the diagnostic value of this metric is rather small.

Glass Box evaluation is mostly single component performance matched against a reference or manual transcription containing the optimal or desired result. Metrics are:

- Word Accuracy (WA): Recognition result vs. transcription.
- Concept Accuracy (CA): Parser output vs. annotation of semantic (task) items.

We extended the measurement of CA to include the current semantic dialogue context. Thus, not only we can measure the performance of the semantic construction (cf. [2], [3]), but also that of the semantic interpretation.

3. THE ”ACCeSS” SYSTEM

The ACCeSS (Automated Call Center through Speech understanding System) project is concerned with bringing speech understanding and dialogue technology to large-scale commercial call centers. In [5] we gave an outline of the ACCeSS system as it was in 1997. In the meantime, due to external circumstances, we had to adapt the system to a completely new and complex application domain. ACCeSS is no longer dealing with car insurance. Rather, the spoken dialogue system now has as an application domain the transactions involved in a bonus program, comparable to airline mile bonusses, where people can gather bonus points through a certain documented activity, accumulate them on an account, and get some premium for a certain number of these points. The premiums are general merchandise.

This application has a considerable complexity. Callers first have to be identified (via a customer number). They then get their points account balance. After this, they are free either to request information on any of the premiums currently available, to order premiums, to ask for premiums available for a certain number of points, and to do this all over again several times. An example dialogue to illustrate this:

System: Hello. Bonus points service. Do you have a customer number?

Caller: Yes

S: Please speak your customer number in single digits.

C: one one eight nine nine five three

S: Your customer number is one one eight nine nine five three, is that correct?

C: Yes

S: Your account currently has seven hundred points. Do you want to have information on premiums or order one?

C: What do I get for a hundred points

S: For one hundred points you can choose between a crate of Jever Pilsener, a toaster, or a pair of sunglasses. Do you want more information or to order one of these?

C: What’s the beer?

Etc.

In terms of evaluation, one of the main problems is that there are many regular exit points in the dialogue, and that the intentions of the callers may change during the interaction. In this, the application is considerably more flexible than, for example, a timetable information dialogue, where it can reasonably be assumed that there is a certain consistency in what information the caller wants. Here, the caller gets offers, and may (and does) switch from a purely informative call to an order dialogue and back again. It is thus not possible to use a measurement of overall Transaction Success, as the change of intention from the side of the caller also changes the exit point that would count as a successful completion for the original intention.

Therefore, we specified the overall application into a number of subtasks as listed here:

1. Identification
2. Offers for current points
3. Offers for X points
4. Product information
5. Orders
6. PIN identification

Note that except for the initial identification, there is no predefined sequence of these subtasks. They all can be initiated by the user, only the request for the pin code is necessarily initiated by the system when the caller makes an order for the first time. There may also be loops like product information - order - offers for X points - order etc. Any sequence of subtasks may constitute an overall successful dialogue in terms of meeting the callers intentions.

4. EVALUATION PROCEDURE

For the verification of the new ACCeSS application we adopted the system-in-the-loop methodology. This approach relies on a software engineering life-cycle model of implement, test, and-revise. Thus, instead of verifying the initial functional system design with WOZ simulations, we immediately implemented a first running prototype to get started with real user tests.

4.1. Test and Corpus Characterisation

The tests were carried out in a quiet office environment over the public telephone network (PTN). Test persons were ”naive users” which were not involved in system development. They were instructed with a scenario description including their customer number and basic information about the system´s abilities. In the first phase, dialogues from 42 test persons (42,5% females, average age 33,3 years) were collected. The total number of recorded utterances was 806.

The system was running on a Sun workstation connected with the PTN. For each dialogue turn, the system stored the following log files:

- Spoken user input in PCM format
- Recognition result
- Result of semantic interpretation
- Dialogue context used for semantic interpretation

In addition, the overall dialogue duration (in secs) and the number of turns were recorded. The resulting corpus was manually prepared for the glass box and black box evaluations. This corpus preparation phase was supported by a software tool for linking the logfiles together into a HTML-based structure. Thus, the necessary manual evaluation steps (transcription of what was spoken, determination of (sub)task success) could be carried out conveniently in an HTML browser environment.

4.2. Glass Box Evaluation

At the component level, we measured the performance of the speech recogniser and the semantic interpreter.

Speech Recogniser Evaluation

For recogniser evaluation we used the widely accepted standard metrics sentence recognition (SR) and word accuracy (WA) rate. SR is the percentage of sentences that were recognised without any error, and WA is determined by calculating the Levenshtein distance between the actually spoken sentence and the recognition result, where equal costs are assigned to substitution, insertion, and deletion errors (cf [2]). To calculate these metrics, the spoken utterances in the corpus were manually transcribed, and the transcriptions were compared with the recognition results logged during the tests.[1]

The recogniser was run in continuous, speaker-independent mode using a finite-state language model divided into 14 sublanguage models. These dialogue-step dependent sublanguage models were activated by the dialogue manager on the basis of the current dialogue context.

Semantic Interpreter Evaluation

The semantic interpreter of our system comprises two functional components, namely parsing and contextual interpretation (see [7] for architectural details). To measure the performance of the semantic interpreter, we used the Concept Accuracy (CA) metric that was first proposed in the SUNDIAL project ([4],[8]). The calculation of CA normally requires labour-intensive manual annotation of the dialogue corpus with semantic reference answers. In order to avoid this time-intensive work, we further developed the approach described in [2] as follows: the necessary semantic reference annotations are generated automatically from the transcriptions of what was spoken (these transcriptions have to be provided in any case for recogniser evaluation). For this purpose, the transcriptions are passed to the semantic interpreter together with the corresponding dialogue context. As mentioned in section 4.1, the system records the interpretation context of each utterance in the dialogue corpus. This information is used during evaluation to restore the dialogue context in which a semantic interpretation took place. In contrast to the work described in [2], this allows us to evaluate not only context-neutral parsing, but also contextual interpretation.

The outcome of this semantic interpretation process is then treated as the ”correct” reference answer which can be compared with the interpretation result logged by the system. In other words, we compare a reference REF with a hypothesis HYP, where REF is the result of the semantic interpretation of the transcribed string, and HYP is the result of the semantic interpretation of the recognition result.

[1] For this purpose the standard evaluation program from the German VERBMOBIL project was used.

Of course, the generation of reference annotations from transcriptions is only an approximation of ”correct” reference answers assuming that the semantic interpretation can correctly process the transcribed strings. Nevertheless, we consider this a viable approach, since we use a robust partial parsing strategy which is able to extract grammatical substrings when the whole utterance is not covered by the application grammar. Thus, only if a user utterance is completely out of domain, an empty semantic reference will be generated.

In this evaluation environment, we obtained the results shown in Table 1.

Evaluation Metric	Result
Sentence recognition rate (SR)	71.5 %
Word accuracy rate (WA)	58.0 %
Concept accuracy rate (CA)	75.8 %

Table 1: Results of Glass Box Evaluation

These figures will be contrasted with the evaluation results of future system versions. Thus, we will be able to monitor progress within the project.

4.3. Black Box Evaluation

The main purpose of a black box evaluation in the first phase of a system-in-the-loop setting is to test whether the system can successfully solve its task. Therefore, from the evaluation metrics proposed by EAGLES, we considered the transaction success (TS) rate the most relevant. However, as explained above, in our application task it is problematic to measure transaction success as a single binary measure, since the amount of information requested by the users may vary considerably. In the scenario description, users were only informed about their identification numbers and general possibilities, e.g. ”You may ask for product information or order something”. To define more specific tasks, e.g. ”Try to order the coffee machine”, would have been unrealistic, because the real-life service will be called by people with vague intentions, too.

This complex task structure led us to the conclusion that the calculation of transaction success rates at a subtask level is better suited to detect the weak points in the system design. We identified the six subtasks listed in section 3. This set of subtasks was used during the manual evaluation of the dialogue corpus to subdivide the dialogues into a sequence of subtasks. Each of these subtasks was then judged as a success or failure. The average subtask success rate can then be considered as an approximation of the overall transaction success rate.

In addition, there was also a manual determination of the correction rate as proposed by EAGLES. Together with the automatically calculated measures for dialogue and turn duration, this yielded the results shown in Table 2.

Evaluation Metric	Result
TD: Average Turn Duration	5.2 secs
DD: Average Dialogue Duration	175.9 secs

CC: Average Correction Rate	7.4 %
TS_1: Subtask 1 Success Rate	78,5 %
TS_2: Subtask 2 Success Rate	97.1 %
TS_3: Subtask 3 Success Rate	No attempt
TS_4: Subtask 4 Success Rate	85.0 %
TS_5: Subtask 5 Success Rate	96.7 %
TS_6: Subtask 6 Success Rate	96.7 %
TS: Average Subtask Success	90.8 %

Table 2: Results of Black Box Evaluation

5. EVALUATION RESULTS

The measurement of subtask success proved to be very useful, as low success rates in one of the subtasks show that there is a design or implementation flaw here, rather than an overall problem such as recognizer performance etc. However, the results of measurements like those reported here can only be indicators of problematic areas, and do not in themselves reveal the source of the problem. The evaluators and designers still have to look closely at the dialogue data of the cases that caused these low rates.

For example: We had rather low subtask success and word accuracy rates in the first subtask, the identification/collection of the seven-digit account number. We first thought that this had only to do with digit string recognition errors. Closer inspection showed, however, that there were several reasons:

- Some test users gave only six digits and then insisted that this was enough, saying things like ”that’s it” or ”finished”. This type of patron error had not been foreseen in the dialogue design, and thus also not in the lexicon A new help prompt and lexicon additions quickly solved the problem.
- Some users had problems to initiate repair steps, since it was not obvious to them how to correct the system´s understanding in case of a misrecognition. As a consequence, we augmented the initial dialogue design with additional correction steps that are invoked whenever an error occurs during digit entry.

In subtask 4 (asking for further information about a product), we found various user expressions that were not covered by our initial grammar design, leading to a greater amount of misrecognitions. As a result, we enlarged both the recognition vocabulary and the grammar coverage.

We had originally intended to conduct about 100 user tests in the first phase. However, after having evaluated the first 40 tests, we decided to remedy the flaws found there immediately. At the point of writing, the second evaluation phase with the refined system has just been started. The evaluation metrics discussed in the preceeding sections are an essential means for an objective comparison of such successive system versions. Thus, besides serving diagnostic purposes, the figures reported here are also the basis for progress evaluation within the project.

6. CONCLUSIONS

All in all, the EAGLES metrics provided very valuable guidelines for our evaluation, but not all the metrics were unproblematic. Especially for diagnostics during development, the addition of subtask success rate proved very helpful.

Our primary motivation for using the EAGLES metrics was to apply a set of metrics that allow objective measurement of system performance. These measures are used both for diagnostic evaluation of the current version and for progress evaluation of successive system versions. One major problem remains the costly transcription and manual evaluation of bulk data. To minimise this effort, we decided not to use the EAGLES metric contextual appropriateness, since the diagnostic value of this metric is relatively small.

7. ACKNOWLEDGEMENTS

Part of this work was funded by the European Commission under contract no. LE1-1802 10347 (ACCeSS). The authors wish to thank all their colleagues in the ACCeSS project, especially Peter Regel and Ludwig Hitzenberger, for their support and help in carrying out this work.

8. REFERENCES

1. Baggia, P. (1996): Evaluation of Spoken Dialogue Systems, Summer School on Language and Speech Communication, Budapest.

2. Boros, M., W. Eckert, F. Gallwitz, G. Goerz, G. Hanrieder, H. Niemann (1996): Towards Understanding Spontaneous Speech: Word Accuracy vs. Concept Accuracy. In: Proc. of ICSLP 1996, Philadelphia.

3. Boros, M., U. Ehrlich, , P. Heisterkamp, H. Niemann (1998): An Evaluation Framework for Spoken Language Processing, In: Proc. of SPECOM 1998, St. Petersburg.

4. Ciaramella, A. (ed) (1993): Prototype performance evaluation report. Sundial workpackage 8000 Final Report.

5. Ehrlich, U., G. Hanrieder, L. Hitzenberger, P. Heisterkamp, K. Mecklenburg, P. Regel-Brietzmann (1997): ACCeSS - Automated Call Center through Speech Understanding System. In: Proc. of Eurospeech 1997, Rhodes, Greece.

6. Fraser, N. (1997): Assessment of interactive systems. In: D. Gibbon, R. Moore, R. Winski (eds): Handbook of Standards and Resources for Spoken Language Systems. De Gruyter: Berlin, 1997.

7. Hanrieder, G. (1998): Integration of a mixed-initiative dialogue manager into commercial IVR platforms. In: Proc. of IVTTA´98, Turin, Italy.

8. Simpson, A., Fraser, N. (1993): Black Box and Glass Box Evaluation of the SUNDIAL System. In: Proc. of Eurospeech 1993, Berlin.