A proposal for experimental end-to-end evaluation

MADCOW

October 18, 1993

1 Introduction

End-to-end (or "logfile") evaluation provides a global evaluation for ATIS systems that goes beyond the current utterance by utterance SLS evaluation. It makes use of a complete scenario and provides more complete information about system performance.

While valuable, the current evaluation procedure has several shortcomings. For example, a system is evaluated off-line using material that may not have been generated in the context of that particular system's operation. As well, evaluation is based on only a subset of utterances (A, A+D) a procedure that destroys the continuity of an ongoing interaction. Moreover, the current techniques are not readily extended to aspects of performance thought to be important for effective human-computer interaction and cannot quantify the impact of different system strategies such as mixed-initiative or the generation of partial responses.

The proposed end-to-end evaluation procedure attempts to maintain a black-box approach that permits direct comparison between different systems while expanding the number of characteristics assessed. Note that the logfile data gathered as part of this evaluation would also be available for more detailed glass-box evaluations. In addition, the evaluation provides a mechanism for assessing the limitations of the current benchmark metrics, and may provide insight that will allow the development of more meaningful metrics.

Since end-to-end evaluation represents a new undertaking for us, a dry-run will be conducted in the Fall of 1992 with participation by four sites. The remainder of this memo defines a set of metrics and sketches out a procedure for carrying out logfile evaluation.

2 Metrics

The following metrics will be computed from information contained in files generated by each system during the evaluation. A separate log file will be generated for each scenario attempted. All necessary information for the evaluation (with the exception of the satisfaction ratings) will be derived from the logfile. The structure of the logfile will be substantially the same as that currently in use for data collection, with some minor modifications as detailed below. It is understood that the purpose of the logfile evaluation dry-run is to determine the usefulness of the proposed metrics. It is likely that some metrics may be modified or even dropped altogether.

The following metrics will be computed:

A. PERFORMANCE METRICS

(A.1) Speech Recognition

The utterances collected in this evaluation will be transcribed (by the participating sites) and the current SPREC statistics computed. Computation will exclude all utterances that are session markers and will include a break-down of word error for class A, class D, class X, and for A+D+X.

In previous evaluations, we had classified utterances into three classes: A, D, or X. Class X was used as a catch-all for those utterances that were not classifiable as either A or D. With logfile evaluation, we will be able to evaluate (with human evaluation) more of what we now classify as class X. So, for a speech recognition metric, we should measure performance for all "evaluable utterances". The term evaluable here refers to those utterances that make sense in the dialogue. There will continue to be utterances that may judged as unevaluable, for example, totally broken utterances. So, just like we had previously reported error rates for A+D utterances, i.e., evaluable utterances in the context of comparator-based evaluation, we should continue to show a single performance measure for speech recognition based on all evaluable utterances, which will include all class A and D utterances, plus a fraction of what we now classify as class X.

(A.2) Comparator Evaluation

The following section specifies an expanded set of comparator metrics suitable for End-to-End evaluation. CAS-based evaluation *will not take place* at this time; data from the Human Evaluation, described below, will be used instead.

A.2.a. PERCENT OF QUERIES ANSWERED CORRECTLY

This is based on standard comparator analysis as currently performed for the A/A+D analyses, except that it may also include in Class D utterances that depend on the answer to a previous query as part of the necessary context. (There may be other additions to Class D as well, but we need to explore exactly what they are.) Note that this analysis will require annotation of all utterances from the logfile evaluation, but it has the advantage of producing annotated training data that can be used after the evaluation.

A.2.b. PERCENT OF QUERIES ANSWERED INCORRECTLY

This is based on the standard comparator analysis, with the same remarks as above.

A.2.c. PERCENT OF QUERIES NOT ANSWERED

This is based on the standard comparator analysis, with the same remarks as above, except that it includes utterances that are responded to with an error message of any kind.

A.2.d. PERCENT OF QUERIES ANSWERED WITH MIXED INITIATIVE

Queries for which the system responds with a directed request for further input. For example, User: `I want to go to LA`, System: `Where are you starting from?`

A.2.e. PERCENT OF QUERIES ANSWERED WITH OTHER RESPONSES

Response should imply comprehension of query. For example, User: `I want to go to Chicago`, System: `Chicago isn't in the database!`

A.2.f. PERCENT OF QUERIES CAUSING ERRORS OR NO INTERPRETATION

Queries for which the system is unable to produce a coherent response. This category implies that the system has a "bug" of some sort. For example, User: `American and is nine date`, System: `[internal error] message.` it is different from A(A.2)A.2.c which implies that the system has diagnosed itself to be in an anomalous state and responds to the user accordingly.

A.2.g. WEIGHTED ERROR

This is based on the usual weighted-error formula and is defined as $[2 * (\%False) + (\%No\_Answer)]$.

(A.3) Human Evaluation

Subject-based evaluation uses human evaluators to judge the "correctness" of system responses to user inputs. A methodology has been described in papers from MIT and in a paper from Unisys on evaluating Voyager. Human evaluation provides an alternative evaluation to comparator-based evaluations: we hypothesize that it will be more flexible in allowing evaluators to give partial credit for partially correct answers; on the other hand, it is quite possible that evaluators may not reliably be able to determine whether the answer is correct. Both MIT and Unisys achieved good evaluator agreement on sample data collected with a wizard system. The advantage of human evaluation is cost—it is quite cheap to perform, since no prior annotation is required, and the evaluation of a single scenario can be done in a few minutes. Of course, multiple evaluators are required for each scenario, and the sessions need to be transcribed.

An evaluation workbench program will present each query-response pair to the evaluator. The query text presented will consist of the (possibly cleaned up) `.sro` transcription, the response will include all information presented to the user: the result, paraphrases, context information, etc. Relevant system state information should also be presented. The evaluator will judge the query as a new input, repeat or a rephrase. The system's response will be categorized using the following categories and modifiers:

- **Answer**: Correct, Partially Correct, Incorrect or Can't Decide.
- **System-Initiated Directive**: Appropriate, Inappropriate or Can't Decide.
- **Failure-to-Understand**
- **Diagnostic Message**: Appropriate, Inappropriate or Can't Decide.

A separate document **Evaluation of ATIS (Air Travel Information System) Dialogues** describes the human evaluation procedure in greater detail.

B. TASK METRICS

(B.1) Task Completion

The subject will be asked to record the answer to the problem posed in the scenario (or indicate that none is available). Thus this indicates whether the user believes that she or he found the desired answer.

Whether the answer is actually the one that the test designers anticipated is not relevant, since the database is rich enough, and the scenarios complex enough so that more than one answer might well exist for many problems, and constraining the scenario enough to guarantee a single answer might be undesirable. Also, failure to indicate the "right" answer might be caused by something like the user failing to read accurately the answers provided by the system, rather than by any deficiency of the system itself. Failures to complete the task should therefore be differentiated as to reason. For example, they could be tabulated using the following categories: task completed correctly as far as the subject is concerned, but with an inadmissible final answer; technical failure ("crash"); insufficient time; other.

(B.2) Mean Time to Completion

This is the elapsed clock time between the beginning of the first utterance spoken into the system to the completion of the display for the last utterance in the scenario. The timestamps in question would be extracted from the session logfile.

(B.3) Query Characteristics

B.3.a. MEAN NUMBER OF QUERIES

Figure 1: A timing model

```
|<--- transaction ------------------->|
|------------|------------|------------|
|<--- t1 --->|<--- t2 --->|<--- t3---->|
```

The number of utterances spoken by the subject. This count includes any retries on the part of the subject that are due to system malfunction (for example, recognition error that results in incorrect retrieval).

B.3.b. MEAN NUMBER OF REPETITIONS
These are defined as unambiguous attempts to re-enter the current input by simple repetition. Slight wording changes are permissible so long as in the evaluator's judgement they constitute a repetition.

B.3.c. MEAN NUMBER OF REPHRASES
Slight wording changes are permissible so long as in the evaluator's judgement the subject was attempting to ask for exactly the same information as before, by using a different request.

B.3.d. MEAN NUMBER OF WORDS PER QUERY
As computed from a transcription of the utterances in a scenario (excluding marker utterances and retries). Provides a first-order indication of the complexity of utterances elicited by a particular system.

(B.4) Transaction Duration

The following measures are based on the simple model shown in Figure 1; more complex models could be adopted at small initial cost, though their utility may not be immediately apparent. The logfile should contain the time in seconds (to three decimal digits) as computed from the data structure returned by the Unix `gettimeofday()` function. Summary times will be reported rounded to tenths of seconds.

These will be measured for all utterances except scenario markers. The information produced by this analysis can be correlated with other measures to better characterize system performance, such as relating the length of utterances to recognition accuracy and response delay.

B.4.a. MEAN UTTERANCE DURATION
This is `t1` in the diagram above and is the utterance duration, measured from the moment the user begins to speak to the end of speech input. All systems will collect speech data in compliance with the guidelines established by the Data Quality Committee. This is to say all utterances will be endpointed to

eliminate excessive silence and duration will be computed on the end-pointed utterances.

B.4.b. MEAN RESPONSE DELAY

This is `t2` in the diagram above. Response delay includes recognition time in excess of real-time, parse time, database retrieval time and display time. This interval is measured from the end of `t1` to the completion of display rendering.

B.4.c. MEAN ASSIMILATION TIME

Query assimilation time is `t3`, measured from the end of `t2` to the beginning of the next input utterance (`t1`).

(B.5) SATISFACTION QUESTIONNAIRE

Satisfaction is assessed by a debriefing questionnaire administered at the end of a session. The questionnaire contains a small number of scale-based questions derived from existing questionnaires. Six-point scales are used (that is, no neutral point is provided). The data will be reported averaged by question over all subjects for each site. The questionnaire assesses whether the subject felt the system performed acceptably given the task at hand and probes specific dimensions of performance, such as response appropriateness, perceived understanding, effectiveness of error recovery. The questionnaire is included as an appendix to this document.

3 CONDUCTING THE EVALUATION

For practical reasons, it may not be possible to control for all extraneous factors, such as differences between sites and scenarios, or differences between individual subjects. Nevertheless, the evaluation procedure described here attempts to minimize the influence of such factors.

We foresee the eventual adoption of an evaluation procedure that brings systems together in a single location and allows the same panel of subjects to work with each system. In the near term, our task is to determine whether the proposed metrics are sufficiently diagnostic and whether logfile evaluation has merit. We can do this through a *distributed* evaluation that can be carried out separately by each participating site. While this runs the risk of making meaningful comparisons of performance across sites difficult, it does have the virtue of providing us with useful data on the metrics described in Section 2.

3.1 General Logistics

The system configuration used for this evaluation will be entirely up to the participating site except, of course, for the requirement that it use live on-line speech input, automatically processed. This is understood to exclude any participation by a wizard. The system must be capable of generating suitably-formatted logfiles, waveform files, as well as output compatible with the Comparator.

A set of 8 problem scenarios will be selected by NIST from among the scenarios used by all the sites for collecting training data. Each site will be responsible for submitting 4 debugged scenarios to NIST. NIST will choose 2 of these. The selection should be representative of types of scenarios and sites. Specific cities, dates, and so on will be inserted into the scenarios. Minor changes may be made for clarity. Subjects will not be allowed to modify those variables. The person conducting the evaluation on site (the "experimenter") is the ONLY person other than the subjects who is allowed to examine the problem scenarios prior to the test.

Data gathering should be completed within a reasonable time, say two weeks from receipt of the scenarios. NIST will tabulate and distribute the results of the evaluation.

3.2 Design

While many factors are likely to influence system performance and it may not be possible to control for all of them, the largest source of variation will likely come from the behavior of the subjects testing the systems and from the scenarios used for evaluation. To minimize the influence of this factor, each site will run 8 different subjects. The 8 scenarios selected by NIST will be arranged in two 4x4 Latin squares and four subjects will be assigned to each set. All sites will use the same design.

The Latin square design efficiently samples the space of possible combinations and allows us to concentrate on the effects of primary interest. Figure 2 shows an example of such a design. The letters (`a`–`h`) indicate the eight scenarios in the evaluation. A subject is assigned to each row of each Latin square and completes the scenarios on that line. For example, the third subject assigned to the first square would start by doing scenario `b` and end by doing scenario `a`. The Latin square balances order of presentation for each scenario (since there is only one of each in each column).

This design will yield 32 completed scenarios per site. If we use the Carnegie Mellon wizard data as an example (about 10 utterances per scenario), this will yield a total of approximately (32 x 10 =) 320 utterances per site. If four sites participate in the evaluation, this will yield 124 scenarios and 1240 utterances.

Figure 2: Two 4x4 Latin squares (from Fisher & Yates, 6th ed., 1963)

```
c a b d        f h g e
d b a c        e g h f
b c d a        h e f g
a d c b        g f e h
```

3.3 Subjects

If possible, subjects should not have used a spoken language system (SLS) prior to this evaluation, for either data collection, demonstration, or other purposes. They should be as similar as possible to subjects used for data collection. For sites that may find it difficult to locate such individuals, it is permissible to use up to 4 subjects who have previously participated in ATIS data collection. These repeat subjects should be assigned to only one of the 4-scenario subsets (as described in section 3.2). The remaining subset can then provide for an unbiased comparison across sites.

The subjects used in this evaluation shall be computer-literate, in the sense that they have used a computer before and are comfortable with the use of a mouse (in cases where a mouse is used). Subjects should, however, have no direct experience in the fields of speech or natural language processing, as assessed by self-report. They should be drawn from a homogeneous population. Depending on the location of the the evaluation the population could be, e.g., graduate students or research staff. The group should be balanced for sex. In general, individual subjects should have the following characteristics:

1. completed high school
2. be native speakers of English
3. be between the ages of 20 and 60
4. never have been employed as a travel planner or in a travel agency.
5. may be experienced programmers but should not consider programming as their primary activity. (they should not identify with terms such as "hacker").

The experimenter will determine eligibility prior to scheduling a session. Sites may determine a maximum session duration (say 90 min) and have the option of eliminating subjects who take longer. The maximum allowed session time should be well in excess of the anticipated completion time for the experiment.

3.4 Procedure

All subjects will be given a standard set of instructions that explain the operation of the system and the nature of their task. These instructions are contained in a separate document, **Subject Instructions for End-to-End Evaluation**. Each site will produce an amended version of the **Instructions** that reflect local conditions. this amended version must be submitted to NIST prior to actual data collection.

In order to ensure that the subject understands the basic mechanism of the interaction, a single short practice scenario will be performed. All sites will use the same practice scenario, which will be determined by the logfile evaluation committee in advance of the test; it should be simple enough to be completed in a few utterances.

Experimenters will be given instruction on how to conduct a session so that their interactions with subjects are reasonably equivalent across sites. These instructions are contained in a separate document, **Experimenter Instructions for End-to-End Evaluation**. The instructions include a complete instruction script to be read to the subject. Other than for site-specific information, the script will be the same across sites.

The experimenter should be someone familiar with the operation of the SLS, but preferably not one of its developers. The experimenter's interaction with the subject will be limited to: explaining the input protocol (including how to interact with the system), performing adjustments to the microphone, etc. *NO* explanation of output or suggestions on how to proceed or explanations of unusual system behavior will be attempted.

If the system ceases to function for any reason ("crashes"), the experimenter will be responsible to restoring it to operational status; however, the clock will continue to "tick" during this period. If the experimenter believes that the system has become inoperable for an extended period, that session may be cancelled and the subject replaced.

A note will be made of such interventions.

After all problems have been completed, the Subjective Evaluation questionnaire will be administered. At the end of the session, subjects will receive a pre-determined "motivator" as their reward for participation in the evaluation. The nature of the motivator will be made known to the subject at the outset of the session (and may be used as an inducement for participation).

Document prepared by Alexander Rudnicky
412 268 2622
`air@cs.cmu.edu`