EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

SPOKEN DIALOGUE SYSTEM EVALUATION: A FIRST FRAMEWORK FOR REPORTING RESULTS

Norman M. Fraser
Department of Linguistic and International Studies
University of Surrey, Guildford,
Surrey GU2 5XH, United Kingdom
E-mail: n.fraser@surrey.ac.uk

ABSTRACT

There are no agreed standards for reporting the performance of spoken dialogue systems. This paper proposes a core set of metrics to be used for this purpose. For this set, operational definitions are supplied, to regularise their application. The intention in proposing this framework is not that it should be exhaustive, nor that it should be perfect, but rather that it should provide a practical starting point, thereby allowing initial system comparison to be achieved quickly and with some measure of confidence.

1. INTRODUCTION

Spoken dialogue systems are complex interactive artefacts. This makes it very difficult to compare the performance of two systems, even if they address similar tasks in similar domains. Comparison would be rendered at least a bit easier if all researchers focused on the same system aspects every time they published a system evaluation. Unfortunately, there are currently no accepted standards for reporting the performance of interactive dialogue systems. The absence of such standards obscures the significance of each new system, and renders scientific comparison of alternative systems as reported in the literature virtually impossible. This unfortunate situation results from two main deficits:

1. There is no agreement on *what* to report about spoken dialogue systems. One researcher thinks the temporal course of a dialogue is important and meticulously records it and reports on it. Another overlooks the temporal structure and focuses instead on the number of understanding errors. A third researcher may overlook both of these and concentrate primarily on human factors issues relating to the usability of the system. In practice, most researchers characterise their systems with an assortment of specially devised metrics which selectively cover a range of aspects of system performance, from objective system performance to subjective perception of its quality. It is virtually unheard of for two systems to be reported using exactly the same set of metrics.
2. Even if members of the spoken dialogue community agreed on what the set of items to measure and report should be, there is no agreement on *how* to do this. For example, most researchers currently offer a measure of how successful their system is at completing whole dialogues and the tasks associated with them. This is usually identified by some term such as 'dialogue completion', 'dialogue success', 'transaction success', 'task completion, etc. Sadly, the differences run deeper than these superficial terminological variations. Are all dialogues included, or are some excluded from the measure (e.g. apparently deliberate attempts to 'break' the system)? What happens if the system fails to carry out a task but correctly recognises that the task is ill-formed? How should success be measured if a dialogue consists of two tasks, one of which succeeds while the other one fails? Questions such as these have not been answered consistently by researchers in the field.

This paper primarily addresses the first of these problems. It builds on work carried out in connection with the EAGLES project, which offers some leverage on the second problem [1,2].

2. A STANDARD REPORTING FRAMEWORK

We propose a simple and practical reporting framework for spoken dialogue systems. The intention is not that it should be exhaustive, nor that it should be perfect, but rather that it should provide a minimal common standard for reporting system performance (and evaluation conditions), thereby allowing initial system comparison to be achieved with some measure of confidence.

The proposed framework is derived from analysis of a broad corpus of spoken dialogue systems described in the literature and from direct experience of evaluating spoken dialogue systems. The approach provides a table of parameters and associated values. Users provide a partial characterisation of their systems by supplying values for this closed set of parameters. Parameters define key aspects of the System, the Test Conditions, and the Test Results:

```
SYSTEM
  Input type
  Input vocabulary
  Input perplexity
  Output type
  Dialogue type
TEST CONDITIONS
  Type of users
  Number of users
  Number of dialogues
  Number of tasks
TEST RESULTS
  Average turns per dialogue
  Average dialogue duration
  Average turn delay
  Dialogue success rate
  Task success rate
  Crash rate
```

It is very important for the success of a reporting framework such as this that it be handled in a consistent fashion by all its users. In the next section we describe the purpose of each parameter and identify the range of values which may legitimately be used to fill each slot. On some occasions it may prove difficult to provide a value which *exactly* meets the specification laid down in the standard. Under these circumstances it is appropriate to fill the slot with a measure which closely approximates to what is required by the standard, but this *must* always be noted by the presence of a marker next to the value in the slot. This marker should reference a note which details what exactly is being reported. Where the values reported are entirely standard there is no need for this.

3. PARAMETERS AND VALUES

In this section we define the purpose of the parameters in the reporting framework and specify the range of possible values. Some parameters allow modifiers to be added to the basic values. These simply serve to add extra detail to the information provided by the value itself. Modifiers should be used where possible, but may be omitted where necessary. Modifiers are written in round brackets after the value they modify. More than one modifier may be attached to a single value, where appropriate.

It must be stressed that the choice of parameters cannot be considered *optimal*, since current levels of understanding do not allow us to make any such claims. Neither can they be claimed to be exhaustive—the position adopted here is that a brief list is preferable to a lengthy one, since it is more likely to be used. Though the set of parameters and values proposed is intended to reflect the informed suggestions of a broad cross-section of practicioners in the field, there is a sense in which *any* reasonable proposal would suffice, so long as it comes to be adopted as a common reporting standard.

The onus, then, in the following discussion of the proposed set of parameters and values is on establishing a reporting framework which is simple to understand and easy to apply, but which facilitates improved communications in the area of interactive dialogue systems.

3.1 System metrics

System metrics are used to characterise some basic features of the spoken dialogue system to be evaluated.

3.1.1 Input type

Values: SPEECH, TEXT, DTMF, PULSE, OTHER

Modifiers:

DIMENSION	VALUES	MODIFIER NAMES
channel	SPEECH	MICROPHONE, TELEPHONE
quality	SPEECH (MICROPHONE)	STANDARD, CLOSE, ARRAY
	SPEECH (TELEPHONE)	MOBILE, PBX, PSTN
linguistic complexity	SPEECH	ISOLATED, WORD-SPOTTING, CONTINUOUS DIGIT, PHRASE, NATURAL

This parameter characterises the way in which the user's dialogue contributions are input to the system and supplies modifiers of three different kinds. There are two values to characterise linguistic input, SPEECH, and TEXT. Values are supplied for the common telephone-based non-linguistic signalling systems DTMF and PULSE. Modalities which are not covered by these values should be characterised using the OTHER value accompanied by a note supplying more details. There is an intentional bias towards speech in this early version of the reporting framework, with some other aspects left underdeveloped.

Channel modifiers supply further information about the kind of channel by which speech input enters the system. There are two options: MICROPHONE and TELEPHONE.

Quality modifiers supply further information about the input channel, which give an indication of the quality of the input material which the system has got to work with. For MICROPHONE speech, use STANDARD for an ordinary microphone, CLOSE for a high quality close-talking microphone, ARRAY for a microphone array. For TELEPHONE speech, use MOBILE when the input comes from mobile telephones (augmented, where appropriate with useful additional information on the type of mobile signalling, e.g. analogue, GSM, etc.), PBX when it comes over private branch exchange lines only, and PSTN when the input is relayed over the public switched telephone network, including local, national and international lines.

The modifiers ISOLATED, CONNECTED DIGIT, PHRASE, and NATURAL are used to indicate the linguistic complexity of spoken input *from the system's point of view*. The modifier is used to characterise what the

system is capable of, not what the user optimistically attempts to use, even if successful. ISOLATED is used for systems which are capable of accepting only single words or multi-word fixed expressions as input. If it is possible to embed the isolated words in surrounding acoustic material, the WORD-SPOTTING modifier should be used. Use CONTINUOUS DIGIT if multiple digits may be spoken in a single utterance, unless the more general PHRASE or NATURAL are available. PHRASE should be used if the system supports input of task-oriented phrases such as date, time, or money expressions, while NATURAL should only be used when more general multi-word utterances are possible. Where a dialogue system allows different levels of linguistic complexity at different points in a dialogue, characterise the system using the modifier which accounts for the largest number of inputs in the test corpus, and supplement this with a note explaining what has been done. Use more than one modifier where appropriate.

3.1.2 Input vocabulary
The system's overall vocabulary size should be indicated.

3.1.3 Input perplexity
List the average preplexity of the recognition vocabulary, supplemented where appropriate by the perplexity of any language model used.

3.1.4 Output type
Values: SPEECH, TEXT, OTHER
Modifiers:

DIMENSION	VALUES	MODIFIER NAMES
quality	SPEECH	CANNED, SYNTHESISED

This parameter characterises the system's output to the user. Use SPEECH when the system produces spoken language output, and TEXT when it produces orthographic text. In all other conditions, use OTHER. The exact nature of OTHER should always be described in a note.

In conjunction with the SPEECH value, use the CANNED modifier for pre-recorded system messages or message fragments, and SYNTHESISED when toutput is synthesised on the fly.

3.1.5 Dialogue type
Values: MENU, SYSTEM-LED, MIXED-INITIATIVE
This parameter provides a clue to the level of dialogue complexity supported by the system being evaluated. A MENU dialogue provides the user with an explicit list of valid next moves at each point in the dialogue; a SYSTEM-LED dialogue leads the user in a highly structured fashion through the dialogue, but does not provide an explicit list of valid next moves before each user turn; in a MIXED-INITIATIVE dialogue, either the user or the system may direct the flow of the dialogue. Characterise mixed systems according to the predominant type, and add a note to explain what has been done.

3.2 Test conditions

Test condition metrics are used to characterise some basic features of the evaluation exercise.

3.1.1 Type of users
Values: PROJECT, EXPERT, NAÏVE
Modifiers:

DIMENSION	VALUES	MODIFIER NAMES
demography	EXPERT, NAÏVE	STAFF, STUDENT, PANEL, PUBLIC
motivation	EXPERT, NAÏVE	INTEREST, REWARD, NEED

For any evaluation exercise, the kind of users should be characterised. The basic values are arranged along a dimension of expertise. PROJECT users are those who have been intimately involved in designing and/or building the system being tested and who are familiar with the task domain. EXPERT users are those who are familiar with the domain. NAÏVE users are completely unfamiliar with the domain or encounter it very infrequently. Where users are drawn from more than one category, all of the relevant categories should be used, with associated percentages.

The modifiers STAFF, STUDENT, PANEL, and PUBLIC may be used with the EXPERT and NAIVE categories. STAFF are members of the same organisation(s) as the PROJECT users, who have not had previous contact with the system. STUDENT users are full- or part-time students. PANEL users are drawn from a balanced sample of the population (provide a note describing its composition if necessary). PUBLIC users are ordinary members of the public who have not been selected in such a way as to be representative of anything.

The modifiers INTEREST, REWARD, and NEED may be used to characterise the users' motivations for participating in the trial. The INTEREST modifier is used for unpaid volunteers, REWARD is used for paid subjects, and NEED is used for people who are using the system as a means to achieving some goal in which they have a real interest.

3.2.2 Number of users
The credibility of the results achieved is related to the size of the sample on which the results were collected. In general, the significance of the results increases with sample size. However, a count of the number of dialogues is not, by itself, adequate. It is important to understand whether the corpus of test materials is comprised of, for example, small contributions provided by many people or major contributions provided by a small number of people.

3.2.3 Number of dialogues
This parameter is used to record the number of dialogues in the test corpus. A dialogue is a continuous session of interaction with the system, usually starting with an opening phase and terminating with a closing phase.

3.2.4 Number of tasks
This parameter is used to record the number of tasks in the test corpus. There may be a one-to-one mapping from dialogues to tasks. However, in some domains it is possible to have more than one task per dialogue. While it is straightforward to define a dialogue, it is much harder to define a task. Ideally, a report which includes a result for this parameter should also include a brief description of what constitutes a task in this domain.

3.3 Test results

Test result metrics are used to characterise some basic features of the system's performance collected during the evaluation exercise.

3.3.1 Average turns per dialogue
The average number of turns per dialogue is the total number of system *and* user turns in the test corpus divided by the number of dialogues in the corpus.

3.3.2 Average dialogue duration
Value: › 0 seconds
This parameter is used to describe the average dialogue duration in seconds, counted from the start of the first utterance to the end of the last one. If the dialogue times out, the end time should be calculated from the end of the last utterance plus the time-out constant used.

3.3.3 Average turn delay
Value: › 0 seconds
This parameter is used to describe the average time (in seconds) taken by the system to respond to a user input. Turn delay is counted from the end of a user's utterance to the beginning of the next system utterance. Since this figure is fairly hard to collect for a large corpus, it may be calculated on the basis of a representative sub-set of the test corpus. However, in this case a note should be provided describing the size of the sub-corpus. Results for the whole corpus should be used where possible.

3.3.4 Dialogue success rate
Value: any percentage › 0%
The dialogue success rate is the percentage of all dialogues in the corpus in which the system either succeeds in correctly satisfying all the user's tasks or it correctly identifies the fact that the tasks cannot be satisfied (e.g. they are ill-formed or fall outside the planned competence of the system) *and* the system succeeds in closing the dialogue gracefully.

A dialogue is deemed to have failed if any of the following conditions hold:

- the system provides the user with incorrect information (within the understanding of 'correctness' assumed in the project)
- the system fails to satisfy a task which it is designed to satisfy
- the system offers a solution to a task which it should not be able to satisfy
- the system behaves incoherently
- the system crashes (or times out, or enters some catatonic state)

3.3.5 Task success rate
Value: any percentage › 0%
This parameter is exactly the same as Dialogue Success Rate, except that it applies at the level of the task rather than the dialogue. In domains where there is a one-to-one mapping between dialogues and tasks, this parameter may be omitted.

3.3.6 Crash rate
Value: any percentage › 0%
This parameter records the percentage of all dialogues in which the system fails to complete a dialogue in a coherent fashion. Crashes are particularly undesirable in real world dialogue systems, so it is useful to draw this figure out. A crash is defined as an occasion on which:

- the system fails to respond within a pre-determined number of seconds (i.e. it times out)
- the system appears to enter a loop
- the system terminates without going through the normal dialogue closing formalities
- the system plays a 'debugging' or 'not-for-public-consumption' message on the normal output channel

Occasions on which the system terminates the dialogue against the user's wishes, but by means of some graceful closing procedure, should not be counted as a crash. For example, the system might tell the user *I'm sorry, I am unable to access the database to obtain the information you require. Please try again later. Good bye.* This does not constitute a crash.

4. APPLYING THE STANDARD

Standard reporting frameworks only have value if they are applied with scrupulous integrity. Wherever possible, results reported should conform exactly to the parameter value specifications given above. Where this is not possible, all deviations from the standard should be noted, thereby allowing independent observers to assess the significance of the results.

The easiest way to 'cheat' and achieve better performance than could otherwise be reported is to pre-select the material to include in the test corpus. Use of the standard described here implies that the test corpus has not been filtered in any way, unless expressly noted in the accompanying documentation.

5. REFERENCES

[1] N.M. Fraser, "Quality standards for spoken language dialogue systems: a progress report on EAGLES." *Proceedings of the ESCA Workshop on Spoken Dialogue Systems: Theories and Applications,* pp. 157-160. Vigsø, May 1995.

[2] D. Gibbon, R. Moore and R. Winski (eds) *Handbook of Standards and Resources for Spoken Language Systems.* Berlin: Mouton de Gruyter, 1997.