# Generic Template for the evaluation of Dialogue Management Systems

*Gavin E Churcher, Eric S Atwell, Clive Souter*
Centre for Computer Analysis of Language And Speech (CCALAS)
Artificial Intelligence Division, School of Computer Studies
The University of Leeds, LEEDS LS2 9JT, Yorkshire, England
gavin@scs.leeds.ac.uk eric@scs.leeds.ac.uk cs@scs.leeds.ac.uk
WWW: http://agora.leeds.ac.uk/amalgam/

## ABSTRACT

We present a generic template for spoken dialogue systems integrating speech recognition and synthesis with 'higher-level' natural language dialogue modelling components. The generic model is abstracted from a number of real application systems targetted at very different domains. Our research aim in developing this generic template is to investigate a new approach to the evaluation of Dialogue Management Systems. Rather than attempting to measure accuracy/speed of output, we propose principles for the evaluation of the underlying theoretical linguistic model of Dialogue Management in a given system, in terms of how well it fits our generic template for Dialogue Management Systems. This is a measure of 'genericness' or 'application-independence' of a given system, which can be used to moderate accuracy/speed scores in comparisons of very unlike DMSs serving different domains. This relates to (but is orthogonal to) Dialogue Management Systems evaluation in terms of naturalness and like measurable metrics; it follows more closely emerging qualitative evaluation techniques for NL grammatical parsing schemes.

## 1. INTRODUCTION

Dialogue management systems, particularly those which replace a graphical user interface with a spoken language one, have become increasingly popular. Speech recognition is gradually becoming robust enough to be employed in the commercial market place, and because of this many companies are realising the value of a spoken interface to their products and services. The research community provides a number of methodologies to the representation of dialogue and its implementation on a computer. Correspondingly, there are a number of design methodologies for building such a system. Despite there many differences, every one contains a common process: an evaluative cycle. Evaluating a dialogue management system is a difficult and often subjective experience. Whilst it is possible to objectively measure recognition performance, evaluation of a dialogue is not as straightforward. Even those systems which exhibit appalling speech recognition performance can nevertheless lead to "successful" dialogues.

## 2. QUANTITATIVE AND QUALITATIVE EVALUATION

There are two approaches to evaluating a dialogue management system: to use a qualitative or a quantitative measure. A qualitative evaluation would rely on the user's opinion of the system. Such a subjective evaluation is fraught with problems. For example, the user may learn after the first attempt how to address the system and which words to use or avoid. Subsequent evaluations of the same system may then vary even though the system has not changed. Some users may find the system difficult to use whilst other will find it effortless. "Pleasantness" differs from person to person, too. As [1] argues, there is no clear consensus of what comprises a good dialogue. Because of these problems, many researchers have tried to provide a means of objectively evaluating a system.

The two methodologies for quantitative evaluation, black and glass box, are concerned with input and output behaviour and the behaviour of each of the components in the system, respectively. Glass box evaluation can rely on a comparison between the output of a component and a retrospective reference. By directly comparing the two it is possible to measure the accuracy of that component. The black box approach, on the other hand, cannot use this method to evaluate a dialogue since there is no "correct" dialogue to compare it with. Despite this, objective evaluation of the dialogue is necessary in order to compare the performance of different systems. Initial efforts have been made to standardise this (for example in EAGLES, see [2]) but remain work in progress.

## 3. COMMON COMPONENTS IN PRACTICAL DIALOGUE MANAGEMENT SYSTEMS

Our recent survey of a number of dialogue management systems has led us to identify those features and components which occur in many of the systems. By examining a range of successful systems (see [3]), from flight information services and appointment scheduling to theatre ticket booking and virtual space navigation, a template for a generic dialogue management system has been drafted. A number of features are incorporated, including a pragmatics interpreter dealing with discourse phenomena such as anaphoric resolution and ellipsis, a

model of the task structure and how it relates to the dialogue structure, a model of conversation incorporating an interaction strategy and a recovery strategy, and a semantic interpreter which resolves the full interpretation of an utterance in light of its context. This generic template can be used in the design of future dialogue management systems, highlighting important features and the mechanisms required to implement them. The template also provides an application-independent method for assessing systems according to the features they exhibit.

## 4. ADVANTAGES OF QUALITATIVE ASSESSMENT AGAINST A STANDARD

Speech And Language Technology researchers are used to thinking of evaluation in terms of speed and accuracy of system outputs, for example 'success rate' of a speech recogniser or syntactic parser in analysing a standard test corpus. However, 'Dialogue Management' is a high-level linguistic concept which cannot be measured so straightforwardly for several reasons:

- existing DMSs are very domain-specific, and we need to compare dialogue systems across domains; so it makes no sense to look for a common standard 'test corpus';

- the boundary between 'good' and 'bad' dialogue is very ill-defined, so it makes little sense to try to assess against a target 'correct output', or even by subjective assessment of 'pleasantness' of output;

- the structure of dialogue (and hence a DMS) is complex, multi-level, and non-algorithmic, making a single overall 'evaluation metric' meaningless without consideration of component behaviours;

- we need to evaluate the integrated system holistically, as opposed to measuring speed or accuracy of individual components;

- alternative dialogue systems use a wide range of alternative component technologies; only by fitting these against a generic template can we discriminate between superficial and substantive differences in component assumptions and functionalities.

There is a useful analogy with evaluation of NL parsers; typically, rival parsers are compared by measuring speed (sentences-per-minute) and/or accuracy (e.g. percentage of sentences parsed) - e.g. [4]. However, rival parsing schemes include varying 'levels' of syntactic information, as shown in EAGLES recommendations ([5]). [6] proposes an orthogonal evaluation of parsing schemes against the generic EAGLES 'template' of syntactic levels, so that a given parser speed/accuracy measure should be moderated by a 'genericness' weight. In much the same way, we propose that very unlike rival DMSs can be meaningfully compared by assessing how well they match our generic template for dialogue management architecture, and using this 'genericness' score to temper any measures of speed, accuracy, naturalness, etc.

Consider [3], which included a first attempt at an outline of a generic spoken language system. The model includes generic modules for syntactic, semantic, and speech act constraints; these constraints are integrated into spoken input interpretation to compensate for limitations in speech recognition components. The model constitutes a template tool for designing integrated systems; it specifies the standard components and how they fit together. As is the predicament of any generic system it is necessarily vague and since it attempts to combine components found in a variety of individual models, it may not fit all systems, if any in particular.

In our survey, we studied how this generic model mapped onto a range of existing real systems, by looking at the representation formats for the various linguistic features in the dialogue management schemes; as with grammatical analysis schemes, there is a need for a theory-neutral 'interlingua' standard dialogue representation scheme [6].

## 5. FEATURES OF NATURAL DIALOGUE

'Naturalness' in dialogue is difficult to define, but by examining phenomena which occur in human to human dialogue we can begin to draw some features which contribute to its definition. The proposed model in [3] reflects this to a certain extent by incorporating components for phenomena such as anaphora and ellipsis whilst abstracting away from those components which are domain specific, such as the model of task/dialogue structure. To begin with, seven such features are described below.

### A: Anaphora

Anaphora frequently occurs in dialogue. This form of deixis is applied to words which can only be interpreted in the given context of the dialogue. There are a number of different forms of anaphora including personal pronouns ("I", "you", "he/she/it" etc.), spatial anaphora ("there", "that" etc.) and temporal anaphora ("then"). Expressions relative to the current context often need to be interpreted into an absolute or canonical form. This form of anaphora includes expressions such as "next week" and "the next entry" which can only be resolved in relation to a previous expression.

### B: Ellipsis

Ellipsis commonly occurs in a sentence where for reasons of economy, style or emphasis, part of the

structure is omitted. The missing structure can be recovered from the context of the dialogue and normally the previous sentences. Without modelling ellipsis, dialogue can appear far from natural.

**C: Recovery strategy**

Although misunderstandings often occur in conversations, speakers have the ability to recover from these and other deviations in communication. [7] presents an analysis of the type of communicative deviations which can occur in conversation and categorises them into content and role deviations. The inadequacies of speech recognition technology introduces additional potential deviations. A dialogue management system must be able to recover from any deviations which occur. Seldom in human to human conversation does the dialogue 'break down'.

**D: Interaction strategy**

At any stage in a dialogue, one participant has the initiative of the conversation. In everyday conversation, it is possible for either participant to take the initiative at any stage. Turning to dialogue management, the interaction strategy is important when defining the naturalness of the system. System-orientated question and answer systems where the system has the initiative throughout the dialogue are the simplest to model since the user is explicitly constrained in their response. The greater freedom the user has to control the dialogue, the more complicated this modelling strategy becomes. For systems using speech recognition, the ability to confirm or clarify given information is essential, hence system-orientated or mixed initiative should exist.

**E: Functional perplexity**

To a lesser extent, the range of tasks that can be performed by a particular dialogue is important. In human to human conversations, for example, an utterance can perform more than one illocutionary or speech act. In an analogous way, a dialogue can include more than one task, whether it is to book tickets for a performance, or to enquire about flight times. Looking to individual utterances, the greater the number of acts which can be performed, the more complex (or perplex) the language model becomes. In everyday conversation, humans are adept at marking topic boundaries and changes. For applications where more than one task is to be performed in a single dialogue, the dialogue manager needs to be able to identify when the user switches from one task to another. Functional perplexity is a measure of the density of the topic changes in a single dialogue and is accordingly difficult to calculate.

**F: Language perplexity**

The ability to express oneself as one wishes and still be understood is an important factor which contributes to naturalness in dialogue. This does not necessarily entail a very large vocabulary since corpus studies and similar language elicitation exercises can provide a relatively small, core vocabulary. The user's freedom of expression is implicitly related to the initiative strategy employed by the dialogue manager. For example, when the system has the initiative, the user's language can be explicitly constrained. In contrast a system which allows the user to take the initiative has less control of the user's language. Again, as with functional perplexity, the perplexity of a language in this sense is difficult to measure but it is helpful to look to the extent that the system attempts to constrain the user's language for performing a task.

**G: Over-informativeness**

User orientated over-informativeness is an important feature to have and is directly related to the degree of freedom of expression. In natural dialogue, a speaker can provide more information than is actually requested. Humans are able to take this additional information into consideration or ignore it depending on how relevant it is to the conversation. The information may have been volunteered in anticipation of a future request for information and as a result a dialogue manager which ignores it will not appear very natural.

## 6. A QUESTIONNAIRE

Whilst each of the above features are important, it is not obvious which are more important to 'naturalness' than others. Turning to the research community we asked those who had designed systems incorporating dialogue management for their experiences and opinions. The questionnaire asked the community to rank the features according to how important they thought they were to their particular dialogue manager and to comment on each one. Given the time constraints, it was not possible to ask more detailed questions about each feature, although the respondents were encouraged to give examples.

Table 1 shows the six systems detailed (see [3] for references to each system), table 2 a summary of the importance of the features to each system. The results range from 1 - the most important to 7 - the least important; the ratings were allowed to be tied.

| | |
|---|---|
| 1 | Daimler-Benz Generic DMS |
| 2 | LINLIN |
| 3 | EVAR German Train-timetable Spoken Dialogue Information System |
| 4 | VERBMOBIL dialogue component |
| 5 | The Slovenian Dialogue System for Air Flight Inquiries |

| | |
|---|---|
| 6 | SAPLEN - Sistema Automatico de Pedidos en Lenguaje Natural |

**Table 1: 6 DMSs**

| Feature | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 2 | 3 | 2 | 2 |
| 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 |
| 3 | 2 | 1 | 1 | 1 | 3 | 1 | 1 |
| 4 | 1 | 1 | 1 | 6 | - | 2 | - |
| 5 | - | 3 | 6 | 6 | 5 | 3 | 2 |
| 6 | 3 | 5 | 5 | 5 | 5 | 5 | 5 |

**Table 2: Features ranked in 6 DMSs**

By taking the mean of the scores, the features can be ordered as follows, most important first:

A: Anaphora == B: Ellipsis
G: Over-informativeness
C: Recovery strategy == F: Language perplexity
E: Functional perplexity
D: Interaction strategy

## 7. COMMENTS ON APPROACH TAKEN

The initial, tentative ranking of features indicates that anaphora and ellipsis are important, whilst functional perplexity and interaction strategy are least important. Given that the systems surveyed performed just one or two tasks, it is not surprising that functional perplexity is not ranked highly. The low ranking of the interaction strategy reflects the application of the system. For example, system (4), Verbmobil, regarded the interaction strategy to be of low importance since it is a minimally intrusive system which facilitates the dialogue between two humans.

What is made clear is that we need to conduct further research into explicitly quantifying each feature for this approach to be worthwhile. Whilst features such as over-informativeness are either present or not, others are finer grained; the interaction strategy can be system-orientated, user-orientated or a combination of both. Language perplexity, in the sense meant here, needs to be quantified, too, before it can be considered a useful feature. In retrospect, the ranking of each feature needs to be made consistent.

## 8. CONCLUSION

Recent technological advances are bringing spoken dialogue systems closer to markets, to real applications. As the focus of this research field shifts from academic study to commercial reality, we feel it is important to maintain a theoretical underpinning: a generic model for independent qualitative assessment and comparison of practical Interactive Spoken Dialogue Systems.

We invite practical systems developers to help us assess their products against this generic template, allowing us in turn to maintain and refine the theoretical generic model to keep step with practical developments. The list of features can be used in two ways: to evaluate the 'genericness' of a dialogue manager, and to ascertain whether a dialogue manager is suitable to a particular application. In choosing between rival Dialogue Managment Systems, it is not sensible to try to use a simple metric of accuracy or naturalness applicable across all applications. Different applications require different DMS features. Prospective users hoping to re-use a DMS should first decide what they want from one; if they can frame their requirements in terms of our generic template, they can eliminate candidate systems which do not focus on the required features.

## 9. REFERENCES


[1] A. Vilnat, "Which processes to manage human-machine dialogue?", in [8], 1996.

[2] N. Fraser, "Quality Standards for Spoken Dialogue Systems: a report on progress in EAGLES", in Dalsgaard et al. (1995), pp 157-160. 1995.

[3] G. Churcher, E. Atwell, C. Souter, "Dialogue Management Systems: a survey and overview", Research Report 97.06, School of Computer Studies, Leeds University, 1997.

[4] R. Sutcliffe, H. D. Koch, and A. McElligott (editors), Industrial Parsing of Software Manuals, Rodopi. 1996.

[5] G. N. Leech, R. Barnett, P. Kahrel, "EAGLES Final Report and Guidelines for the Syntactic Annotation of Corpora", (EAGLES Document EAG-TCWG-SASG), Pisa. Italy. 1995

[6] E. Atwell, "Comparative evaluation of grammatical annotation models", in: [5], 1995.

[7] L. Taleb, "Communicative Deviation in Finalized Informative Dialogue Management", in [8], 1996.

[8] S. Luperfoy, A. Nijholt and G. Veldhuijzen van Zanten (eds.), "Dialogue Management in Natural Language Systems", Proceedings of the 11th Twente Workshop on Language Technology, University of Twente, Enschede, Netherlands. 1996.