A Generic Template to evaluate integrated components in spoken dialogue systems

Gavin E Churcher and **Eric S Atwell** and **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 and eric@scs.leeds.ac.uk and 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 (eg Dybkjaer et al 1995, Vilnat 1996, EAGLES 1994, Fraser 1995); it follows more closely emerging qualitative evaluation techniques for NL grammatical parsing schemes (Leech et al 1996, Atwell 1996).

KEYWORDS: evaluation, comparisons, generic model, standards.

1 Background

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. Dybkjaer et al (1995) conducted interviews after each session and asked whether the dialogue seemed natural and pleasant. 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 Vilnat (1996) argues, there is no clear consensus of what comprises a good dialogue. When asking the user, the designer has to make sure that the user is representative of the end user in terms of background and frequency of use. 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 Fraser 1995a) 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, from flight information services (Fraser 1995b) and appointment scheduling in Verbmobil (Alexander and Reithinger 1995, Maier 1996, Alexandersson 1996) to theatre ticket booking (Hulstijn et al. 1996) and virtual space navigation (Nugues et al. 1996), 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. (Sutcliffe et al 1996). However, rival parsing schemes include varying 'levels' of syntactic information, as shown in EAGLES recommendations (Leech et al 1995). Atwell (1996) 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; for example, the ENGCG parser (Voutilainen and Jarvinen 1996) is very fast and accurate BUT its underlying parsing scheme instantiates only a small subset of the EAGLES 'template', which moderates an overall 'score'. 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 (Churcher et al 1997), 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 (Atwell 1996).

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 (Churcher et al 97) 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. By incorporating anaphora, a speaker can reduce redundancy and economise their speech.

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. Taleb (1996) 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. Where the user has the initiative throughout the dialogue such as in command and control applications, the user has greater expressibility and freedom of choice. The most difficult dialogues to model are those where the initiative can be taken be either the system or the user at various points in the dialogue. As noted by Eckert (1996), mixed initiative systems involve dialogues which approach the intricacies of conversational turn-taking, utilising strategies which determine when, for example, the system can take the initiative away from the user. 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. A simpler measure is to count the number of semantically distinct tasks a user can perform.

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. The level of constraint should not be measured when the system is recovering from deviations in the dialogue, since focussing the user may be necessary for recovering from the deviation in as few steps as possible.

G: Over-informativeness

There are two interpretations of over-informativeness, system and user orientated. system orientated over-informativeness allows the dialogue manager to present more information to the user than was actually explicitly requested. 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. As an example, consider the following dialogue between the system and user where the user responds with a reply which is over-informative:

User: I'd like to make an appointment.

System: Who would you like to make an appointment with?

User: John Smith at 2pm.

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, 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.

Table 1: 6 DMSs

[1] Daimler-Benz Generic DMS (Heisterkamp 1993, Heisterkamp and McGlashan 1996, Regel-Brietzmann et al. (forthcoming))
[2] LINLIN (Ahrenberg et al. 1990, Jonsson 1993, 1996)
[3] EVAR German Train-Timetable Spoken Dialogue Information System (Eckert et al. 1993, Boros et al. 1996)
[4] VERBMOBIL dialogue component (Alexandersson et al. 1996,1997)
[5] The Slovenian Dialog System for Air Flight Inquiries (Ipsic et al. 1997, Pepelnjak et al. 1996)
[6] SAPLEN - Sistema Automatico de Pedidos en Lenguaje Natural (Lopez-Cozar(forthcoming))

Table 2: Features ranked in 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
Mean	2.0	2.0	2.5	3.7	3.6	2.5	2.4

Note that where '-' occurs, the feature was not ranked, and so is omitted from the mean. It is interesting to note that different respondents interpreted the ranking differently. Whilst some understood the points system to indicate the order of importance of each feature, others, such as [6] considered the points to be an indication of how important the feature was to their system.

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.

References

L. Ahrenberg, A. Jönsson and N. Dahlbäck, "Discourse Representation and Discourse Management for Natural Language Interfaces", in Proceedings of the 2nd Nordic Conference on Text Comprehension in Man and Machine, Täby, Sweden. 1990.

J. Alexandersson and N. Reithinger, "Designing the dialogue component in a speech translation system - a corpus-based approach", in Andernach et al. (eds.) 1995.

J. Alexandersson, "Some ideas for the automatic acquisition of dialogue structure", in Luperfoy et al. 1996.

J. Alexandersson, N. Reithinger and E. Maier, "Insights into the Dialogue Processing of Verbmobil", in Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLP '97, Washington, DC. 1997.

E. Atwell, "Comparative evaluation of grammatical annotation models", in: Sutcliffe et al. 1996.

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

M. Boros, W. Eckert, F. Fallwitz, G. Hanrieder, G. Goerz, H. Niemann, "Towards Understanding Spontaneous Speech: Word Accuracy vs. Concept Accuracy", in Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-96), Philadelphia, 1996.

H. Dybkjeer, L. Dybkjeer, N. O. Bernsen, "Design, formalization and evaluation of spoke language dialogue", in J. A. Andernach, S. P. van de Burgt and G. F. van der Hoeven (eds), "Corpus-based Approaches to Dialogue Modelling", Proceedings of the 9th Twente Workshop on Language Technology , University of Twente, Enschede, Netherlands. 1995.

W. Eckert, T. Kuhn, N. Niemann, S. Rieck, A. Scheuer, E. G. Schukat-Talamazzini, "A Spoken Dialogue System for German Intercity Train Timetable Inquiries", in Proceedings of Eurospeech '93, Berlin

W. Eckert, "Understanding of Spontaneous Utterances in Human- Machine-Dialog", in Luperfoy et al. 1996.

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

N. Fraser, "Messy data, what can we learn from it?", in Andernach et al. (eds.) 1995.

P. Heisterkamp, "Ambiguity and uncertainty in spoken dialogue", in Proceedings of Eurospeech '93, Berlin, 1993.

P. Heisterkamp and S. McGlashan, "Units of dialogue management: an example", in Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-96), Philadelphia, 1996.

J. Hulstijn, R. Steetskamp, H. ter Doest, S. van de Burgt and A. Nijholt, "Topics in SCHISMA Dialogues", in Luperfoy et al. 1996.

I. Ipsic, F. Mihelic, K. Pepelnjak, J. Gros, S. Dobrisek, N. Pavesic, E. Noth, "The Solvian Dialog System for Air Flight Inquiries", in Proceedings of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs, Pilsen, 1997.

A. Jönsson, "Dialogue Actions for Natural Language Interfaces", in Proceedings of IJCAI-95, Montréal, Canada, 1995.

A. Jönsson, "A Model for Dialogue Management for Human Computer Interaction", in Proceedings of ISSD'96, Philadelphia, 1996.

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

R. Lopez-Cozr, P. Garcia, J. Diaz and A. J. Rubio, "A voice activated dialogue system for fast-food restaurant applications", Proceedings of Eurospeech '97, Rhodes. (forthcoming)

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.

E. Maier, "Context construction as subtask of dialogue processing - the VERBMOBIL case", in Luperfoy et al. 1996.

P. Nugues, C. Godereaux, P. El Guedj and F. Revolta, "A conversational agent to navigate in virtual worlds", in Luperfoy et al. 1996.

K. Pepelnjak, F. Mihelic, N. Pavesic, "Semantic Decomposition of Sentences in the System Supporting Flight Services", in Journal of Computing and Information Technology (CIT), Vol. 4, No. 1, Zagreb, 1996.

P. Regel-Brietzmann et al., "ACCeSS - Automated Call CEnter through Speech understanding System. A description of an advanced application. Proceedings of Eurospeech '97, Rhodes. (forthcoming)

M. Schillo, "Working while Driving: Corpus based language modelling of a natural English Voice-User Interface to the in-car Personal Assistant", MSc Thesis, School of Computer Studies, Leeds University, 1996.

M. Schillo, E. Atwell, C. Souter, T. Denson, "Language modelling for the in-car personal assistant", in C. Moghrabi (ed), "Proceedings of NLP+IA'96: International Conference on Natural Language Processing and Industrial Applications", Universite de Moncton, Canada, 1996.

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

L. Taleb, "Communicative Deviation in Finalized Informative Dialogue Management", in Luperfoy et al. 1996.

A. Vilnat, "Which processes to manage human-machine dialogue?", in Luperfoy et al. 1996.

A. Voutilainen, T. Jarvinen, "Using English Constraint Grammar to Analyse a Software Manual Corpus", in Sutcliffe et al. 1996.

Appendix 1: Questionnaire on Features of Natural Dialogue

Below are listed 7 generic features of natural dialogue. Please state whether your system deals with these features: insert your answer after the asterisk *

Also, please RANK the features in order of importance to your system: 1 - most important, 7 - least, entries can be tied. Please insert your ranking in the square brackets [ ].

Please name your system: *

Please give one or two References (published papers, URLs, tech reports) to cite, giving further details of your system: *

Does your system deal with:

Anaphora? [ ] YES / NO: *

- Which types? For example: Personal pronouns, relative expressions... - If you have time, some brief examples.

Ellipsis? [ ] YES / NO: *

- If you have time, some brief examples.

Recovery Strategy? [ ] Please comment: *

- What types of errors can the system detect and recover from? Can the system identify and cope with errors arising from speech recognition, domain ambiguity etc? See example below for more.

Interaction Strategy? [ ] SYSTEM / USER / MIXED INITIATIVE *

- Does the DMS force the system to take the initiative all of the time by prompting the user for input, or must the user take the initiative all of the time (eg. command and control applications)? Or does the DMS allow the user and the system to take the initiative when required, hence allowing mixed initiative?

Functional Perplexity? [ ] Please comment: *

- How many separate functions can the user get the system to perform in a dialogue? A function is a task or a general goal. For example, a theatre booking/ reservation system provides two functions which can be performed in one dialogue: theatre booking and ticket reservation.

Language Perplexity? [ ] Please comment: *

- Does the system strictly constrain the user's language, perhaps by explicit prompting of what the user can say? Or is a user free to use any language they wish for the task, and the system will attempt to cope with it?

Over-informativeness? [ ] YES / NO *

- Does the system cope with users' over-informative sentences? If a user provides more information than is strictly asked for, how does the system react?

Any Other Comments: *

=== END OF QUESTIONNAIRE===
=== START OF EXAMPLE ===
An example:

Please name your system: *

The In-car Personal Assistant

Please give one or two References (published papers, URLs, tech reports) to cite, giving further details of your system: *

Michael Schillo, "Working while Driving: Corpus based language modelling of a natural English Voice-User Interface to the in-car Personal Assistant", MSc Thesis, School of Computer Studies, Leeds University, 1996.

Michael Schillo, Eric Atwell, Clive Souter, Tony Denson, "Language modelling for the in-car personal assistant", in Chadia Moghrabi (ed), "Proceedings of NLP+IA'96: International Conference on Natural Language Processing and Industrial Applications", Universite de Moncton, Canada, 1996.

Does your system deal with:

Anaphora? [3] *YES

Can cope with the following anaphora:
- personal pronouns (eg. 'he', 'she', 'it' etc)
- relative expressions (eg. 'next Tuesday', 'tomorrow')
- other types (eg. 'that' referring to a diary entry -¿ 'delete that entry')

Ellipsis? [3] *YES

For example, when using the diary the user can ask:
User: When is my next appointment?
The system responds with:
System: Today at 2pm with Mr Smith
The user can then use ellipsis to ask:
User: And tomorrow?

Recovery Strategy? [1] Please comment: *

System deals with two types of error occurring:
- errors relating to speech recognition (misrecognition, speech ambiguity(eg. 'john' and 'tom' with similar confidence values), and low confidence in recognised speech leading to confirmation/clarification
- conflicts arising between the domain model and the user's request, for example, the user may ask the system: User: Delete the appointment with Mr Smith when there is no such appointment.

Interaction Strategy? [2] *MIXED INITIATIVE

The user has the initiative until the system needs to elicit or clarify something. For example, the user can ask the system to call somebody, but if the person is misrecognised then the system takes the initiative to elicit this.

Functional Perplexity? [2] Please comment: *

The system can perform any of 13 separate functions, including:
Telephone:
make a call
Diary: add an appointment;
retrieve an appointment;
retrieve all appointments between certain times;

delete an appointment;
etc.

Language Perplexity? [1] Please comment: *

The user is free to express his/herself as seen fit, since the user initially has the initiative and is not prompted what to say by the system.

Over-informativeness? [6] *YES

The user can give more information in a sentence than is actually asked for. For example, the following dialogue shows a user making an appointment, and providing information related to but inappropriate to the prompt:

User: Make an appointment.

System: Who do you want to make an appointment with?

User: Mr Smith at 2pm today.

System: The appointment has been made.

Any Other Comments: *

NOTE that these answers are not for a real, implemented system, but for the DMS we assumed would underly the simulation experiments.

=== END OF EXAMPLE ===