A Medical Spoken Dialogue System Based on Integrated Task and Ontological Knowledge: Theory and Results

Martin Beveridge and John Fox

Advanced Computation Laboratory,
London Research Institute,
Cancer Research UK,
London WC2A 3PX
{martin.beveridge, john.fox}@cancer.org.uk

Abstract

Dialogue systems seem to have great potential for healthcare, but the state of the art in medical applications of dialogue systems is extremely limited, with dialogues typically hand-coded using languages such as VoiceXML. Apart from the limitations of hand-coding all possible dialogue states, dialogue systems also need to be better integrated with the healthcare domain so that the dialogue can be closely related to the underlying clinical context and goals. This paper presents some current research into the ways in which knowledge of underlying task structure (a medical guideline) and ontological knowledge (medical semantic dictionaries) can be integrated with dialogue models in order to provide flexible and reconfigurable dialogue. A practical implementation of this approach using a 3-layer agent architecture is also described, along with the results of an initial evaluation of the system as applied to the domain of breast cancer referrals.

Introduction

The cost of managing patients with chronic diseases is progressively increasing. In order to improve the capability of healthcare institutions to manage chronic diseases, it has been recognised that medical guidelines must be implemented more effectively (Shiffman et al., 1999). A cost-effective way of achieving this may be to use communication technologies such as the internet and telephone to allow healthcare practitioners to provide data (e.g. blood pressure measurements for hypertension patients) and to receive advice based on current best practice (e.g. whether or not a patient should be referred to a medical specialist, or whether they have an increased genetic risk of developing a disease). If extended to patients, then this approach could also offer them more control of their own care by allowing them to obtain advice whenever and as often as they need it. Voice interaction, in particular, seems an attractive option as speech is a natural way for people to communicate and does not require the use of any technology other than a telephone. Hence, the technological barrier is reduced and the potential reach of the system is increased.

Natural language dialogue systems therefore seem to have great potential for assisting in delivery of healthcare services, but the state of the art in medical applications of dialogue systems is extremely limited. In most cases, the dialogue is hand-coded by specifying all the possible dialogue states and transitions (e.g. Azzini et al., 2001). Whilst this approach may be suitable for small systems (e.g. information lookup for route planning), it becomes impractical for complex systems requiring large numbers of states. In the medical domain, the knowledge structures are particularly complex and the range of dialogues will typically be more varied than in information-seeking tasks often used in non-medical domains, such as ticket booking or route planning. At the same time, clinical services must also make use of expert knowledge and be able to perform complex reasoning and decision-making in responding to the user. It therefore seems necessary to better integrate technologies such as medical guidelines and advice systems with the dialogue system so that the dialogue can be closely related to the

underlying clinical context and goals. This problem is currently being addressed by the HOMEY project[1] and this paper describes one approach which attempts to relate representations commonly employed in discourse analysis and dialogue systems to those used in medical knowledge representation schemas. It is argued that this approach allows the dialogue to be derived from the domain representation, rather than needing to be authored directly, and that the dialogue is then integrated with other non-dialog domain tasks such as decision-making, controlling external devices, database access etc. Furthermore, the dialogue can be reconfigured for other similar domains without making significant changes to the dialogue manager itself.

This paper is structured as follows. First, some background on current approaches to representing discourse and, in particular, dialogues, is provided. Next, the theoretical framework adopted in the design of the dialogue system described here is discussed, followed by a description of its implementation. Finally, an evaluation of a specific application of the dialogue system in the domain of breast cancer referrals is described.

Background

Discourse Structure

Grosz and Sidner (1986) suggest that the *linguistic* structure of discourse, i.e. the sequence of actual utterances in a monologue or dialogue, reflects an underlying *intentional* structure, which, in the case of dialogue, is shared by each participant in the conversation. This consists of a series of intentions, each of which may serve as the purpose of a discourse segment in the linguistic structure, and which defines how the related segment contributes to the overall discourse purpose (the term "intention" here has more or less the same meaning as "goal" elsewhere in AI). Grosz and Sidner (1986) furthermore claim that the linguistic structure can be explained in terms of just two relations between intentions: *satisfaction-precedence* and *dominance*. These are defined (where I1 and I2 denote intentions) as: I1 satisfaction-precedes I2 if I1 must be satisfied before I2, and I1 dominates I2 if I2 provides part of the satisfaction of I1. Importantly, these only represent structural relations between intentions (similar to the constituent structure of sentences) and not semantic relations.

It has also been argued, however, that coherent discourse can be described in terms of *informational* (semantic) relationships between segments. This approach has particularly been applied to text generation (Hovy, 1993b). For example, in order to generate the sentence "mammography is preferred because it is a non-invasive procedure" it is necessary to know about the semantic relation of causality between the notion of "non-invasive procedure" and "being preferred" in order to generate the appropriate linking word "because" (rather than "unless", "although" etc). It is difficult, however, to determine what set of such relations is required. In order to address this problem, Maier and Hovy (1993) carried-out a study comparing all the various relations used in the literature and classifying them into a taxonomy according to their function in the text. The number of researchers that had used each relation was also noted, thereby providing a rough confidence score. The proposal was that relations could be used at whatever level of the taxonomy was appropriate to the domain (with the lower levels describing increasingly subtle distinctions). The most common top-level relations were:

- *Elaboration*: one clause presents additional detail regarding the situation described in the other clause, e.g. [I normally recommend mammography.]C1 [It is a widely available service.]C2, where C2 elaborates C1.
- *Cause/Result*: one clause presents the cause of the result described in the other clause, e.g. [Mammography is preferred]C1 [because it is a non-invasive procedure]C2, where C1 is the result of C2.

[1] Home Monitoring through an Intelligent Dialogue System, EC project IST-2001-32434.

- *Sequence*: the situations described in the clauses occur in sequence, e.g. [I took a patient history]C1 [then I decided to carry out mammography]C2, where C1 occurred before C2.
- *Circumstance*: one clause describes the circumstances of the situation described in the other clause, e.g. [It was more than two weeks later]C1 [when the patient was finally seen]C2, where C1 describes the time at which C2 occurred.

Using such relations, it is possible to describe the relationships between the propositional contents of utterances in a discourse, i.e. "an informational structure, imposed by domain relations among the objects, states and events being discussed" (Moser and Moore, 1996, p. 416).

Another aspect of discourse structure that has been studied is the way in which the discourse unfolds over time, e.g. the way in which the participants' focus of attention shifts and the salience of entities under discussion varies. Grosz and Sidner (1986) represent this via the notion of a dynamic attentional state, which describes all the objects, properties and relations that are salient at a particular point in a discourse. In their approach, the attentional state is an abstraction of a 'focusing structure' - a stack of focus spaces each relating to a discourse segment, and representing the underlying intention for that segment along with descriptions of the salient entities and relations. A key function of this component is to coordinate the intentional and linguistic structures, allowing intentions to be mapped to utterances and vice versa. It is therefore in some respects an interface representation between the linguistic structure and essentially non-linguistic representations such as intentions (which are primarily related to task goals) and information relations (which are primarily related to domain relations between entities under discussion).

Models of discourse such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) suggested that only one relation should be assigned to consecutive discourse segments, effectively forcing a choice between intentional and informational relations (Moser and Moore, 1996). Moore (1995), however, suggests that recently there is a growing consensus that all three of the structures described above (intentional, attentional, and informational) are required to describe the linguistic structure of discourse. For example, Moore and Pollack (1992) have demonstrated that natural language interpretation and generation require parallel representations of both intentional and informational relations between discourse segments. These two structures have been characterized as representing "what is being talked about (informational)…[and] why we are talking about it (intentional)" (Moser and Moore, 1993, p.94) and can provide quite different accounts of discourse. In particular, recognizing one type of relation can aid the hearer in recognizing the other (Moore and Pollack, 1992; Hobbs, 1996).

As a result of this, there have been various attempts to combine intentional and informational approaches to discourse, e.g. (Marcu, 2000; Moser and Moore, 1996). In addition, Hovy has argued that "the inclusion of control information [e.g. attentional state] in discourse planning systems has not received the attention it deserves" and argues for an even broader notion of 'rhetorical structure' which "differs from the semantic and the intentional structure" in that it incorporates "the effects of both, as well as of other constraints on the discourse" (Hovy, 1993a, p.37). The end result of these studies has been that much of the remaining debate on discourse structure "centres around which of these three structures are primary and which are parasitic" (Moore, 1995), and the different roles they play in practical interpretation and generation tasks. The different points of view seem largely to depend on the type of discourse in question (e.g. monologue or dialogue) and the task to be achieved (e.g. interpretation or generation).

Research on developing practical text generation systems, e.g. (McKeown, 1985; Moore and Swartout, 1990; Paris, 1990), has tended to rely primarily on information relations, which are particularly useful for determining text structure (e.g. relations of clauses, cue words etc.). For example, Hobbs (1996), whilst supporting the integration of intentional and information structures, has argued that intentions are often indirect or uninformative or, in

the case of written texts and monologues, unimportant, with interpretation relying more on appreciating the information contained in the discourse. This does not mean, however, that such systems do not make use of intentions at all, but rather that intentional relations are often effectively compiled-out into schemas which are sufficient for a particular domain and task, or conflated with the informational relations (Mittal and Paris, 1993).

On the other hand, it has been shown that informational relations alone are not sufficient to describe dialogue (Mittal and Paris, 1993; Carberry et al., 1993; Traum, 1993). Here, handling misunderstandings or communication failures relies on knowing the original intention in order to propose follow-up questions for example (Moore and Paris, 1992). For these reasons, dialogue systems have been primarily based on notions of intentional structure, with relatively little use of informational relations (although Mittal and Paris, 1993, suggest that they "are a useful computational tool to represent constraints we currently don't totally understand, avoid duplicating reasoning from first principles, and provide an appropriate level of interface with the realization component"). Hence, Moser and Moore (1993), in their discussion of the relation of intentional and informational structures, considered "only monologic discourse…believing generalizations between this and multi-agent discourse to be premature" (p. 94). Recently, however, there has been some renewed interest in applying informational relations to dialogic discourse analysis, e.g. (Stent, 2000). In the case of medical dialogue systems, the domain of discourse is particularly rich, consisting of complex task structures and large numbers of concepts with a wide range of ontological relations between them. In order to maintain a coherent dialogue in such a domain it seems necessary to make use of both these knowledge sources in determining the structure underlying discourse contributions.

Approaches to Dialogue

One of the first approaches to representing dialogue was the development of a prescriptive dialogue grammar that described commonly occurring sequences of utterances such as adjacency pairs (e.g. question followed by answer), and in some cases the entire dialogue. In this approach the linguistic structure of a dialogue is specified directly, without any reference to non-linguistic notions such as intentions or informational relations.

By employing dialogue grammars, a dialogue management system need only be a graph or finite state machine, where each node represents a prompt to the user with a set of options, and the user's response causes a transition to a new node. Such approaches have been useful for systems where the dialogue structure closely matches the task structure, e.g. automated bill payment services. In particular, since the system always takes the initiative it can restrict the number of options presented to the user and, to an extent, induce a valid user response via the priming effect of the prompts. In order to allow some mixed-initiative, extensions to graph systems have been developed such as frame-based systems. In these, entities are defined (e.g. a journey) which have slots to be filled (e.g. departure time, departure location etc) and at each node in the graph the dialogue manager has to ensure all mandatory slots are filled. This might be achieved by the system taking the initiative and prompting the user until all information has been gathered, or the user might take the initiative and fill more than one slot at once providing all the relevant information.

In contrast to dialogue grammars and frame-based systems, plan-based approaches to representing dialogue allow for much greater complexity in the dialogue. They take the approach that dialogue is goal-driven and so the aim of the dialogue manager is to infer these goals and respond appropriately. This approach allows for more complex phenomena such as indirect communicative acts where what is meant is not the literal interpretation of what is said. For example, in the case where a user asks a train timetable system "can you tell me when the last train to London leaves?", the correct response is for the system to inform the user of the departure time for the requested train and not to answer "yes" or "no".

In order for a dialogue system to be able to reason about goals and their connection to utterances, a model of an agent's 'mental state' is required so that speech acts can be related to these mental states in the conversational participants. The model of mental state that was

originally proposed (Cohen and Perrault, 1979) involved describing the configuration of beliefs, desires, and intentions of an agent, and is therefore often referred to as the BDI model. There are many dialogue phenomena, however, which do not fit into the BDI framework, for example: dialogue control phenomena such as acknowledgements, pause-fillers, indications of turn-taking etc. These maintain the dialogue and coordinate participants. More importantly the BDI model doesn't capture the notion of obligations (Traum and Allen, 1994; Kreutel and Matheson, 2000) which seem to arise from social convention and include, for example, the fact that if someone asks a question, it is considered unreasonable (or at least highly marked) for the other conversational partner not to answer. In fact speech act theory (and therefore the BDI formulation) only deals with a single utterance and so cannot distinguish between, for example, an answer to a question, and a standard declarative used to initiate a conversation (Pulman, 1997; Lewin, 2000). This inability to capture the local context of an utterance, and represent its function given that context, means that there is no way to capture the convention that answers follow questions or that people don't walk away in the middle of a conversation – things that, in fact, can be captured in dialogue grammars by distinguishing grammatical and ungrammatical dialogue structures.

Another approach which has developed largely in parallel with BDI approaches, and which can be seen as addressing some of the problems of speech act theory (Pulman, 1997), is the use of conversational games (Power, 1979; Houghton and Isard, 1987; Kowtko and Isard, 1993; Carletta et al., 1996). Conversational games provide a descriptive approach to dialogue rather than a theory of 'rational agency' as the BDI approach is intended to be, and so circumvents some of the problems encountered by BDI. It does this by representing dialogue at two functional levels: at the *plan-based* level are conversational games which are associated with the mutual goals of the participants, and at the *structural* level are sequences of conversational moves which are intended to achieve those goals (Kowtko and Isard, 1993).

The notion of 'move' employed here extends speech acts to include acts such as reply, acknowledge, clarify etc. Moves are either initiating moves of games (i.e. rather similar to speech acts) or responding moves. Dialogues are thought-of as being comprised of a series of games each aiming to achieve some sub-goal of the dialogue. Each game itself consists of a series of moves starting with an opening move and finishing with an end move. Importantly the definition of a game includes moves by both participants, e.g. a request game includes a request by the initiating participant and a reply by the other participant, hence conventional links such as question-answer are captured by using a unit of discourse that spans multiple utterances (Pulman, 1997). The internal structure of a game is typically represented in a similar way to dialogue grammars. For example, a request game may consist of a request move from the speaker, followed by a reply move by the hearer and optionally a final acknowledgement from the speaker to indicate that the information in the reply is grounded (i.e. mutually believed by both conversational particpants). This can be represented as a finite-state network. Additionally, a game can have nested sub-games or a break. Sub-games account for phenomena such as clarifications, side sequences etc in which the sub-game contributes to the goals of the parent game. Breaks account for misunderstandings and indicate that either repair is needed in order to continue, or that the current game may have to be abandoned (Kowtko and Isard, 1993).

The notion of viewing dialogue in terms of games and moves therefore captures the fact that most conversations to achieve a task follow standard scripts (e.g. question-answer) to achieve a limited set of goals (e.g. getting some information, instructing someone) and so generates quite specific expectations regarding a participant's response to a conversational move (Poesio and Mikheev, 1998). At the same time the recursive structure of games and sub-games allows complex mixed-initiative dialogues to be modelled. This approach therefore combines aspects of plan-based approaches with aspects of dialogue grammars, with moves providing a model of the conventional structure of dialogue, and the higher-level model of plans and goals being represented in terms of games, hence allowing more complex reasoning about the motivations of the dialogue and conversational cooperation.

Current Approach

As described above, finite-state network approaches to specifying dialogue, such as dialogue grammars, attempt to directly represent the linguistic structure of dialogue. The interpretation and execution of this (low-level) dialogue specification by a client system then gives rise to the actual dialogue. This is probably the most common approach taken in building dialogue systems and recently standards such as VoiceXML (McGlashan et al., 2004) have been developed to assist in the development of such systems. Hence, a spoken dialogue is the result of the interpretation of a VoiceXML specification by a voice-enabled browser, which controls automatic speech recognition (ASR) and text-to-speech (TTS) components to realise that specification.

Whilst it is obviously possible to author dialogues directly using such low-level specification languages, it was argued previously that this approach is unable to handle certain dialogue phenomena, is time-consuming, and is difficult to reconfigure for different domains. On the other hand, simply replacing the low-level description with a high-level one introduces its own problems. For example, high-level approaches such as BDI allow complex phenomena to be accounted for, but at the expense of being able to easily represent many of the phenomena that low-level approaches can handle. The solution to this problem is that both low-level (structural) and high-level (plan-based) representations are required, e.g. (Traum, 1996; Pulman, 1997; Hulstijn, 2000). Crucially then, it is the "insight that pattern-directed approaches need to be combined with higher-level notions like plans and goals" (Hulstijn, 2000) that is important. Similarly, Dahlbäck and Jönsson (1999) advocate developing approaches that find a middle ground between the conflicting demands of generality (typically provided by high-level representations) and computational efficiency (as provided by low-level representations).

The approach taken here is to employ a *high-level dialogue specification* that can be used to dynamically generate low-level (e.g. VoiceXML) descriptions for individual dialogue segments as and when they are needed. The purpose of the high-level specification is to capture those levels of description that are required to account for discourse structure (Moore, 1995), namely: intentional structure, informational structure, and attentional state. This approach is consistent with Hovy's (1993a) argument, by analogy with approaches to syntactic structure, that "the content of a discourse derives from several sources, and that a common, surface-level-ish structure is needed to house them all". In this case, the high-level dialogue specification provides the common structure, represented in terms of conversational games since they allow dialogue to be described at both plan-based and structural levels, hence bridging the gap between high- and low-level dialogue specifications.

Importantly, it is not intended that the high-level specification should be authored directly. Instead, it is itself generated from medical domain knowledge provided by existing medical technologies, in particular a *domain plan* (specifying the tasks required to carry-out a process) and *domain ontology* (specifying the relevant medical concepts and their inter-relations). Hence, it should be easy to change the domain of the dialogue by authoring a new plan and, possibly, ontology (depending on the extent to which it is situated in a particular domain).

This distinction between dialogue specification and domain specification avoids the problem of a mismatch between representations suited to the task domain, e.g. clinical knowledge, and representations suited to language (Hovy, 1993a). For example, Dahlbäck and Jönsson (1999) distinguish two senses of the notion of 'task' as used in dialogue systems: firstly "some real-world non-linguistic activity that is directed towards achieving a particular goal" and secondly "the sequence of information that needs to be collected by the information providing system….[I]n the former case this knowledge is a separate structure, whereas in the latter it is intertwined with other aspects of the dialogue model". Similarly, Flycht-Eriksson (2000) argues that "domain knowledge reasoning should be clearly separated from dialogue management and performed by a separate module".

The following sections describe the intentional, informational and attentional structures employed in the high-level dialogue specification in more detail. In particular, their

representation in terms of conversational games, and their derivation from the underlying domain plan and ontology, are discussed.

Intentional Structure

Intentions can be captured in a conversational game framework by relating them directly to the goals associated with games (Kreutel and Matheson, 2000). In this way intentions are implicitly captured by the structure of games (Maudet and Evrard, 1998), and dominance and satisfaction relations between intentions can be treated simply as relations between games. The set of game types employed here for describing dialogue are based on those proposed by Carletta et al. (1996) and includes games whose initiating (forward-looking) moves are inform, instruct, check, propose, query-yn or query-w. The implicit intentions associated with these games are described informally below (S is the speaker, H is the hearer, *P* is a proposition and *A* is an action).

- *Query-yn*(*P*): S intends to know if H believes *P*
- *Query-w*(P_x): S intends to know what *x* H believes is the referent of P_x
- *Inform*(*P*): S intends that H believe *P*
- *Instruct*(*A*): S intends that H has done *A*
- *Check*(*P*): S intends to know if H believes *P* (where *P* represents previously requested information)
- *Propose*(P_0, $\{P_0,\ldots,P_n\}$): S intends that H believe that S believes P_0 and S intends to know which of $\{P_0,\ldots,P_n\}$ H believes.

In addition to these initiating moves, the games they initiate will also have response (backward-looking) moves. These constitute moves towards satisfying the intention underlying the game. The response moves employed here are:

- *Acknowledge*: S intends to acknowledge H's last utterance
- *Reply-yn*(*P*): S intends that H know if S believes *P*
- *Reply-w*(P_x): S intends that H know what *x* S believes is the referent of P_x

Since the domain plan determines the overall process to be followed and the individual tasks required, it can quite naturally be seen as the basis for deriving the intentional structure of a dialogue concerning that domain. In fact, the task specification may be seen as imposing certain obligations on the dialogue system in order that the plan execution system can achieve successful completion of the plan[2]. These obligations will then give rise to intentions on the part of the dialogue system to engage in particular dialogues with the user in order to meet those obligations. Such an approach is consistent with claims that dialogue structure is largely determined by task structure (Grosz and Sidner, 1986), or to put it another way: "engaging in a dialogue is typically not a goal in itself, but is motivated by some underlying task or goal one wants to achieve, and for which the dialogue is instrumental" (Bunt, 1996).

Furthermore, intentional relations such as dominance and satisfaction-precedence (Grosz and Sidner, 1986) can be derived from relations between tasks in the domain plan (Young and Moore, 1994). For example, task preconditions in the plan can be considered to give rise to satisfaction-precedence relations in the intentional structure. Hence if task T2 has preconditions such that it cannot be started until task T1 has completed then a satisfaction-precedence (SP) relation can be inferred between G1 (the game whose associated intention derives from the obligation imposed by T1) and G2 (the game whose associated intention derives from the obligation imposed by T2) such that G1 SP G2. Such dependencies can also arise more indirectly through the data flow rather than control flow of the plan. For

[2] Note that, consistent with the desire to find a middle ground between the generality of AI-oriented approaches and computational efficiency (Dahlbäck and Jönsson, 1999), only a plan execution system is assumed here rather than a full-blown AI planner (cf. Young and Moore, 1994).

example, if task T2 has preconditions such that it cannot be started until a data item has a (particular) value and the goal of task T1 is to acquire a value for that data item, then a satisfaction-precedence relation, G1 SP G2, can again be inferred (where G1 and G2 are defined as before).

Dominance relations can similarly be inferred from task decomposition. For example, if task T1 is decomposed into tasks T2 and T3 then the successful completion of G1 (the game whose associated intention derives from the obligation imposed by T1) requires the successful completion of G2 and G3 (the games whose associated intentions derive from obligations imposed by T2 and T3 respectively). Hence, a dominance (DOM) relation[3] can be inferred between G1, G2 and G3 such that G1 DOM G2 and G1 DOM G3.

Note that, whilst obligations imposed by the task specification will provide some of the dialogue intentions – "those represented by the acquisition of agreed propositions" (Pulman, 1997) – others will arise as a result of the dialogue itself in terms of obligations imposed by the user, e.g. to reply to a clarification request. These "communicative subgoals may also arise locally in the dialogue because of unanticipated responses and because of the complexity of the perceptual, understanding, evaluation, and other cognitive processes involved in interpreting and generating communicative behaviour" (Bunt, 1996). Moreover "it is not plausible to assume that such moves are planned: rather, they arise as an immediate response to the current state of the dialogue" (Pulman, 1997). These obligations imposed by the user will also lead to particular intentions in the dialogue system, which must then be balanced with those deriving from obligations imposed by the task specification. It is assumed that obligations imposed by user moves should be processed before those derived from the task specification (Traum and Allen, 1994; Traum, 1996). Hence, the system must respond to clarification questions or meta-level questions from the user before it can continue to pursue domain goals.

Informational Structure

The informational structure of discourse was described above as being "imposed by domain relations among the objects, states and events being discussed" (Moser and Moore, 1996, p. 416). Dahlbäck and Jönsson (1997) similarly argue that task-specific knowledge must be augmented with a conceptual model that describes general information concerning the relationships between objects in a domain. For example, in a library system they suggest a conceptual model in which 'book Is-a publication', 'author is-aspect-of publication' etc.

It is assumed here that such information should be derivable from the concepts and relations specified in the domain ontology. The informational relations useful for language, however, are generally at a more abstract level than domain ontological relations. For example an 'elaboration' relation between two concepts, C1 and C2, in the informational structure, such that C1 elaborate C2, might arise from various ontological relations between C1 and C2 such as C1 is-a C2, C1 is-part-of C2, C1 is-attribute-of C2 and so on (Mann and Thompson, 1988). Indeed, even these subtypes such as 'part-of', may be too abstract to have a direct correlate in the domain ontology. For example, a medical ontology might typically require more fine-grained notions of 'part-of' such as 'is-anterior-of', 'is-linear-division-of', 'is-layer-of' and so on. Hence, the informational structure is an abstraction that is isomorphic to (but not the same as) the more fine-grained ontological relations between domain entities.

One of the problems associated with applying information relations to dialogue is determining the appropriate units that such relations should apply to. In text generation, they are applied to successive utterances, but in dialogue successive utterances may be made by different participants, i.e. they may be different turns in the conversation. Furthermore these turns are linked to each other in terms of their functions, e.g. question-answer, inform-acknowledge, as represented by conversational games. It might therefore seem, as Stent (2000) points-out, that the simplest approach would be to only describe informational relations within turns and use conversational games to describe relations between turns. This

[3] More precisely, task decomposition yields *immediate* dominance, which defines a total ordering over games, rather than the partial ordering described by Grosz and Sidner (1986).

is not sufficient, however, as there are in fact many cases of information relations spanning turns. For example, consider the two dialogues given below (adapted from Stent, 2000). Conversational moves are shown in square brackets at the end of each turn.

1) a. A: So that takes care of the ill guy. [*Explain*]
   b. B: Okay… [*Acknowledge*]
   c. B: And that takes two hours. [*Explain*]
   d. A: Right. [*Acknowledge*]

2) a. A: First they can take out the power. [*Explain*]
   b. B: Right… [*Acknowledge*]
   c. B: And then we have to wait. [*Explain*]
   d. A: Yup. [*Acknowledge*]

Example (1) above illustrates a cross-speaker elaboration relation in which (a) and (b) form an Explain game initiated by participant A, and (c) and (d) form a second Explain game initiated by participant B, and the propositional content of (c) is in an elaboration relation with that of (a). Similarly, example (2) above illustrates a cross-speaker sequence relation in which (a) and (b) again form one game and (c) and (d) form another and (c) is in a sequence relation to (a). Both the above examples therefore involve relations between statements made by different speakers and in different games. Stent (2000) explains that "in a DAMSL-tagged set of 8 dialogs in our corpus, 40% of the utterances were statements, and many of these appeared in sequences of statements. The relationships between many of these statements are unclear without a model of rhetorical structure". On the basis of such examples, it might appear that information relations should be defined between the initiating moves of games. However, Stent also describes relations involving question-answer pairs, such as the example given below (adapted from Stent, 2000).

3) a. A: We have to send buses to the lake. [*Explain*]
   b. B: Ok… [*Acknowledge*]
   c. B: How many are we sending? [*Query-w*]
   d. A: Two. [*Reply-w*]

In this example the topic of both utterances (c) and (d) (i.e. the whole Query-w game) is in an object-attribute elaboration relation with utterance (a) (which introduces the topic of the Explain game). This suggests that the appropriate level for defining these information relations is that of conversational games, as shown in Figure 1. That is to say that informational relations arise as a result of domain relations between the topics of games, and it is this that leads to cross-speaker relations because subsequent games may be initiated by different speakers.

A further example is given below in the context of a medical dialogue system which is trying to determine whether a patient with suspected breast cancer should be referred to a specialist or not.

4) a. S: Does the patient have a lump? [*Query-yn*]
   b. U: Yes [*Reply-yn*]
   c. S: And is it a nodularity? [*Query-yn*]
   d. U: No [*Reply-yn*]

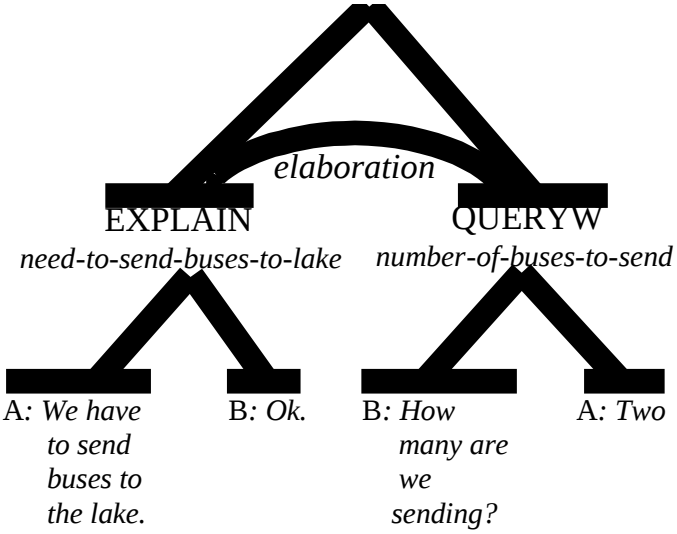

Figure 1: Sample analysis of the informational structure of a dialogue

In this example, the second Query-yn game (utterances c and d) seeks to elaborate the information provided in the first game. This relation arises because the topic of the second game (nodularity) elaborates the topic of the first game (lump). This relation is important for the purposes of dialogue management because it is the basis for the selection of the cue word "and" in utterance (c), and also licenses the use of an anaphor to refer to the topic being elaborated. As will be described in the next section, it is also important because it can be used to determine the order in which games are played, in order to ensure the semantic coherence of the dialogue.

Attentional State

The high-level dialogue specification needs to provide a description of attentional state in addition to intentional and informational structure. In particular, since there are likely to be several playable moves at any point in the dialogue, the dialogue specification should indicate which is the preferred move. Maudet and Evrard (1998) describe this as the "strategy problem", which they suggest is game-independent, depending instead on agent behaviour and goal-dependent heuristics. It must therefore take into account at least the satisfaction-precedence constraints imposed by the intentional structure (in order to conform to the underlying task) and possibly relations in the informational structure (in order to ensure semantic coherence in the sequence of moves). In the approach taken here, the choice of next move is made by marking a particular game as focused in the high-level dialogue specification. The move associated with this focused game therefore becomes the next system move.

The generation component, however, will need to know more than just the next move. In fact, it will need to know the whole set of conventionally and rationally acceptable dialogue moves for either participant (Maudet and Evrard, 1998). This is necessary because both system and user may wish to make moves in more than one game simultaneously. For example, the user may make reply moves in several query games (i.e. answer more than one question) with a single utterance, e.g. "the patient is thirty-five, female with a breast cyst". Similarly, the system may wish to initiate several games with a single utterance by aggregating the initiating moves, e.g. "what are the patient's age, sex and weight?"[4].

The notion of attentional state required here is therefore different to the approach taken by Grosz and Sidner (1986). They describe a stack of focus spaces, each of which maps

[4] Similarly, in the case of a multimodal system, the full set would be required for display in the visual modality, in order to generate a form for the user to complete and submit.

an intention onto a dialogue segment (the topmost element representing the current dialogue segment), with focus spaces only added to the attentional state when needed, and removed when the associated dialogue segment is complete. The approach taken here is instead to have an attentional state in which all possible games are represented simultaneously, hence allowing participants to make moves in several games in parallel and to move between games (similar to the approach taken by Burton and Brna, 1996).

In addition, it is useful to distinguish those games in the attentional state that are currently playable by the system, and those which the system is not currently planning to play, but which the user might. This distinction is captured by marking games as 'foreground' if they are playable by the system and as 'background' otherwise. Whether or not a game is playable by the system depends on various constraints in the intentional and informational structures. For example, if game G1 satisfaction-precedes game G2 in the intentional structure, then G1 will be marked as foreground and G2 will be marked as background, hence G1 is playable by the system but G2 is not. Similarly, if a game has already been played then it will no longer be in the set of playable games for the system. However, it will still be accessible to the user, who may wish to make further moves in it in order to correct errors. Information relations also play a part in the strategy for choosing the next move. This is necessary in order to preserve dialogue coherence by ensuring that the next move made by the system is as semantically relevant as possible to previous moves in the dialogue history. For example, suppose the high-level dialogue specification at a particular point describes an intentional structure such as that below (expressed in terms of games):

G1 = Query-yn(NIPPLE DISCHARGE)
G2 = Query-yn(BLOODSTAINED NIPPLE DISCHARGE)
G3 = Inform(…)
G1 SP G3
G2 SP G3

From this it can be inferred that G1 and G2 must both be completed before G3 can be initiated, but it does not specify which of G1 and G2 should receive focus first in the attentional state. Ontological knowledge, however, allows the inference of an informational elaboration relation between G1 and G2 where G1 is the nucleus of the relation and G2 is the satellite - hence G1 should receive focus first, followed by G2, in order for the dialogue to be coherent. This has two important ramifications. First, the question concerning bloodstained nipple discharge will not be asked until the question regarding nipple discharge has been asked (and may then be discarded, depending on the answer). Hence, discourse topics will "by preference be 'fitted' to prior ones – topics therefore often being withheld until such a 'natural' location for their mention turns up" (Levinson, 1983, p. 313). Secondly, once the question regarding nipple discharge has been asked, the question of bloodstained nipple discharge will have high priority for being considered next. This means that related topics are pursued as soon as possible, whilst they are still relevant, in order to avoid "unlinked topic 'jumps'" later on (Levinson, 1983, p. 313). This is demonstrated in the following example in which the user, in replying to the system's question, takes the initiative and supplies the additional information that the patient has nipple discharge. The system then immediately follows-up this information with the elaborating question regarding bloodstained nipple discharge.

5) a. S: What is the patient's sex?
   b. U: She is female and she has some nipple discharge
   c. S: Ok. And is it a bloodstained nipple discharge?
   d. U: No.
   e. S: Ok. What is the patient's age?

Note that if information relations were not taken into account in this way then the user's introduction of the topic of nipple discharge would not influence the system's plan and the

follow-up question regarding bloodstained discharge might not be raised until much later in the dialogue, possibly requiring a marked construction to introduce it (e.g. “returning to the nipple discharge…”). Levinson points-out that “the relative frequency of marked topic shifts of this sort is a measure of a ‘lousy’ conversation” in human-human dialogue (Levinson, 1983, p. 313).

As can be seen from the previous discussion, the attentional state draws on both intentional and informational relations in order to determine which games can be initiated by the system (i.e. foreground) or by the user only (i.e. background), and which of the foreground games should be chosen next (i.e. focused). It is clear, however, that the strategy employed in the attentional state may be influenced more or less by either intentional or informational structures at different times. For task-oriented dialogues, such as those discussed by Grosz and Sidner (1986), it is clear that the system must have a representation of the shared non-linguistic tasks to be accomplished and the order in which they should be performed, and so the dialogue will be driven primarily by intentions derived from these underlying tasks. For information-seeking dialogues, however, the system need not have a complex task structure but merely a definition of the set of information that needs to be collected at a particular point. The sequence in which the different items of information are collected (and hence the dialogue structure) will then be derived primarily from informational relations between those items. Recent examples of this category include ‘information-seeking chat’, in which the dialogue wanders from topic to topic following associations between them (Stede and Schlangen, 2004) and home-control systems where the dialogue is primarily based on relations between objects (e.g. rooms, devices etc.) in the home (Montoro et al., 2004).

Implementation

The approach described above has a parallel in current trends toward hybrid architectures for artificial agents. Early agent architectures were essentially deliberative, relying on planning and world-modelling, but, because these turned out to be more complex problems than expected, it was realised that such an approach was inadequate when faced with an uncertain and unpredictable environment (Gat, 1998). Brooks instead proposed a different approach, the Subsumption Architecture (Brooks, 1990), which involved no reasoning but instead wired together many small finite state machines in a series of layers. Complex tasks could then be accomplished reactively, e.g. by simply coupling sensors to actuators through a simple transfer function. Wooldridge (2002) points-out, however, that this architecture also has limitations: it relies only on local information and is difficult to engineer because the combined interactions between all the individual reactive components are difficult to understand.

Recently, there has been research on combining reactive and deliberative approaches in so-called hybrid architectures. This work seems to parallel recent approaches to dialogue, which Hulstijn (2000) argues have led to the conclusion that “the perceived opposition between a plan-based and a pattern-based approach to natural language dialogue is false. … The smallest recipes for joint action are precisely the exchanges described by dialogue game rules. On the other hand, plans and goals may function as a semantics for dialogue game rules.” Hence, Hulstijn (2000) argues that “[t]he insight that pattern-directed approaches need to be combined with higher-level notions like plans and goals, is compatible with a general trend towards hybrid architectures for agents” where “[t]he general principle that underlies the trend seems to be that frequently occurring activities that can be ‘automated’ are often dealt with by fixed pattern-directed protocols, recipes or rules. Infrequent activities or failure and misunderstanding require higher-level deliberation. It seems that dialogue is no exception to this principle.”

Traum (1996b) takes a similar approach in developing a reactive-deliberative agent to manage the dialogue. Traum (1996b) states that for the dialogue they chose “a reactive approach, in which the agent is constantly making local decisions as to what to do next” partly because “timely behavior is critical: the same response can have a very different connotation if it is delayed”. When there are no local dialogue decisions to be made Traum’s

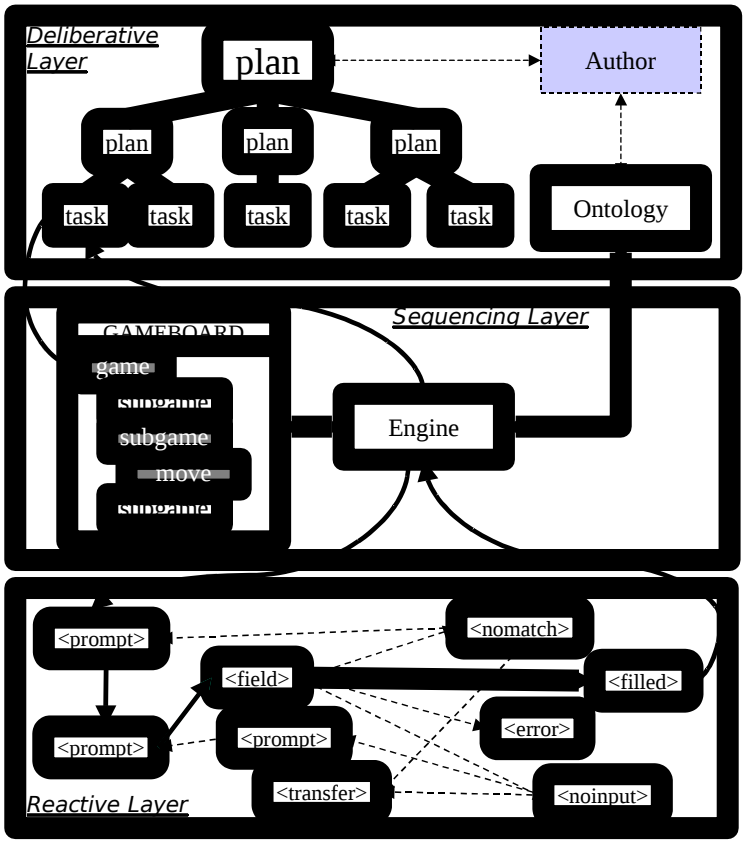

Figure 2: The architecture of the implemented dialogue system.

(1996b) agent returns to deliberation about high-level discourse goals and domain plan negotiation. Between these two extremes he also describes a series of middle-level tasks such as planning speech acts. The balance between tasks at different levels is established by assigning priorities to different levels.

A similar architecture that has arisen from research in robotics is the Three-Layer Architecture (Gat, 1998). This consists of a deliberative layer, which carries-out inference in order to determine future courses of action, and a reactive layer with minimal state that can react quickly to stimuli by following predefined patterns. A key aspect of this architecture, however, is that the reactive and deliberative layers are connected by a sequencing layer. This layer defines a sequence of actions to be carried-out to achieve the goals defined by the deliberative layer. The sequencer then follows this specification in order to select which primitive behaviours the reactive layer should use at a given time.

In developing a practical implementation of the approach described in this paper, a 3-layer hybrid agent architecture (based on the 3T architecture described in Bonasso et al., 1998) proved to be particularly appropriate. In this architecture, as shown in Figure 2, the deliberative layer comprises a plan, decomposed into sub-plans and eventually tasks (e.g. acquiring values for a set of data items, negotiating a proposal and so on) together with a world model (provided by a domain ontology). In this case, a human “author” is also part of the deliberative layer since the plan hierarchy and ontology are manually constructed (and simply interpreted at runtime). The reactive layer consists of a finite-state machine (expressed in VoiceXML) that can handle some small unit of interaction (the sequence of utterances associated with a segment of dialogue) without need for any reasoning or planning. The high-level dialogue specification described previously fits naturally into the sequencing layer, which bridges the gap between the reactive and deliberative components. It does this by generating a set of dialogue games, based on the current state of the domain plan (together with ontological knowledge) and then mapping the moves associated with the currently selected game into a VoiceXML document and speech grammar. Note that the VoiceXML

and grammar are generated dynamically at each system turn, based on the current dialogue state. This means that the low-level reactive layer is constantly adapted according to the current high-level context.

Evaluation

In other work, Cancer Research UK has developed a system (ERA) for advising doctors on whether patients require urgent referral for suspected cancer (Bury et al., 2001). The system is currently accessed by a standard web browser that generates web pages for collecting patient data and reporting on results (see http://www.infermed.com/era). As part of the HOMEY project, it was decided to evaluate the generic dialogue system described previously by applying it to this domain in order to create a speech-enabled version of the ERA system. This section describes a preliminary evaluation of the final ERA dialogue application[5], and discusses the implications for the overall approach to dialogue described in this paper.

Method

The physical architecture of the ERA dialogue application was as follows. The domain plan was provided by the ERA process specification previously developed by Cancer Research UK (Bury et al., 2001) using the PRO*forma* language and toolset (Fox et al., 2003; Sutton and Fox, 2002; see www.openclinical.org/gmm_proforma.html). The domain ontology was provided by a breast cancer ontology supplied by Language & Computing NV (Ceusters et al., 2001; see www.landcglobal.com). The VoiceXML interpreter and automatic speech recogniser (ASR) were provided by ITC-Irst (Azzini et al., 2001; see www.itc.it/irst) and were integrated with the Loquendo Actor text-to-speech (TTS) system (see www.loquendo.com) in an interactive voice response (IVR) platform provided by Reitek (see www.reitek.com). The dialogue system and medical technologies were installed on a server at CR-UK's premises in London, UK, and the IVR platform with VoiceXML interpreter, ASR and TTS were installed on a server at Reitek's premises in Milan, Italy. The two components communicated via HTTP over the Internet, following a web-service model, with the IVR platform as client and the dialogue system as document server. Note, however, that the dialogue system is not tied to these specific technologies, but instead simply specifies interfaces for the domain plan and ontology components (which can then be implemented by any available backend technology that has the required functionality) and generates VoiceXML for the client (which can therefore be any VoiceXML-enabled browser).

The validation of this application was based on thirty dialogues by six users. These ranged from people very familiar with the ERA domain through to people with no specific knowledge of the domain or any wider knowledge of medicine or healthcare. Due to the technical nature of the domain, the users were provided with the script shown in Appendix A, rather than a more general description of the task to perform. Users were asked, however, not to just blindly follow the script, but to ensure that the information acquired by the system was correct according to the described scenario (the script itself contained examples of correcting misunderstandings, using help etc. so users could quickly see how to use the system). Hence, the script provided both a scenario to be communicated and an example of how the system could be used.

In order to evaluate the overall competence of the dialogue system (e.g. dialogue flexibility, functionality etc.), the following metrics were used:

1. *TRINDI*[6] *Tick-List*: three sets of questions that are intended to elicit explanations describing the extent of a system's competence (Bohlin et al., 1999).

[5] A video and interactive demonstration of this dialogue application are available at www.acl.icnet.uk/lab/homey.

[6] Task Oriented Instructional Dialogue, European Telematics Applications Programme LEA-8314.

2. *DISC*[7] *Dialogue Management grids*: a set of questions, similar to the Trindi tick-list above, that are intended to elicit some factual information regarding the potential of a dialogue system (Heid et al., 1998).

The following standard measures were used to evaluate the performance of the speech recogniser:

1. *Word Accuracy*: the accuracy of the system in recognising individual words, measured in terms of the number of word substitutions, deletions and insertions, relative to the total number of words in the actual spoken utterances.
2. *Sentence Recognition*: the percentage of sentence strings that were completely correctly recognised (i.e. where every word in the sentence was correctly recognised).
3. *Concept Accuracy*: the accuracy of the system in acquiring concepts (i.e. degree of semantic understanding) measured in terms of the number of substitutions, insertions, and deletions of semantic units relative to the total number of semantic units uttered.
4. *Semantic Recognition*: the percentage of completely correctly understood sentences (i.e. where every concept in the input utterance was correctly acquired).

The following metrics were used to analyse the performance of the dialogue system:

1. *Task Success*: the degree of success in achieving the desired task, including
    a. The number of users who managed to complete a dialogue
    b. The correctness of data provided by the system (transaction success)
    c. The correctness of data acquired from user
2. *Dialogue Cost*: the cost of successful completion, including
    a. The system response time
    b. The amount of time required to complete a dialogue
    c. The number of turns required to complete a dialogue
    d. The proportion of turns that were spent correcting errors
3. *Usability*: the overall usability of the system, including
    a. The number of times a user made use of 'help'
    b. The quality of system responses: whether the response is appropriate, inappropriate or incomprehensible as measured by the SUNDIAL[8] 'Contextual Appropriateness' metric (Simpson and Fraser, 1993)
    c. The quality of user responses: the degree to which user answers were responsive, usable and/or concise as measured by the Behavioural Coding Scheme (Sutton et al., 1995)
    d. User report (users' own impressions of usability)

Results

It was originally anticipated that scripted interactions would only provide data relevant to evaluating speech recogniser performance. However, the effect of speech recogniser misrecognitions, combined with the instructions to users to correct misunderstandings, led to a lot of variation between dialogues (for instance, the number of turns per dialogue ranged between 32 and 78). Because of this, it was possible to carry out a much wider evaluation, including speech recogniser performance, dialogue manager competence, and dialogue manager performance.

As can be seen from the script in Appendix A, the dialogue system has a high degree of competence (i.e. handles a wide range of dialogue phenomena). For example, in utterance (4) the user provides more information than was requested. This additional information is accommodated and the system then follows-up with questions, in (5) and (7), that elaborate the supplied information (in this case, the fact that the patient has bilateral nipple discharge).

[7] Esprit Long-Term Research Concerted Action No. 24823.
[8] Speech UNderstanding in DIALogue, ESPRIT contract P 2218

This use of ontological information to dynamically re-order tasks for maximal coherence is novel to this approach (Milward and Beveridge, 2004) and is not supported by other state-of-the-art dialogue managers such as HMIHY (Gorin et al., 1997), GoDiS (Larsson et al., 2000) or TRIPS (Allen et al., 2001). In utterance (10) the user takes the initiative and requests clarification of the system's previous question regarding acquired nipple deformity, and so the system responds by providing more specific examples of this concept in (11). Utterance (12) demonstrates the use of negations to pre-empt questions that the user expects the system to ask whilst utterances (21) to (26) demonstrate the verification strategy, which was to wait until all data had been collected and then confirm all the acquired values together in a single batch. The user is then able to correct or retract values and the dialogue manager will not proceed to the next task in the domain plan until all the required data items have correct values. Finally, utterances (28), (30), and (32) show how the user is able to take the initiative and ask for explanations of the system's advice (in terms of arguments for and against the various options). A more formal analysis of competence using the TRINDI Tick-List and DISC dialogue management grids is presented by Beveridge et al. (2004).

Our evaluation of the overall dialogue system performance suggests that the competence reported above has not been gained at the expense of performance. In fact, the ERA application demonstrates good speech recognition performance (word accuracy: 71.82%, sentence recognition: 59.24%) with the majority of word recognition errors involving short function words (e.g. "the", "a", "an") and inflections (e.g. "cyst" misrecognised as "cysts") rather than content words (Beveridge et al., 2004). This means that concept recognition accuracy is better than individual word accuracy (concept accuracy: 77.95%, semantic recognition: 76.06%). The system also demonstrates good performance in acquiring data (97.6% of data items were assigned correct values). The data acquisition accuracy is high because all data were verified by the dialogue system before being confirmed to the plan execution engine. In the few cases where incorrect values were acquired, this seemed to be due to the user mishearing or not properly attending to the verification prompt and confirming that data was correct even though there were errors. This may be due to the verification prompts being overlong or because of unclear pronunciation of some values by the speech synthesizer. Overall, the system achieved a high proportion of successfully completed transactions (80.77% of dialogues completed, of which 85.71% achieved the dialogue goal). Moreover, this was achieved whilst maintaining a good response time from the dialogue manager (on average 531 ms from request to response), and an efficient dialogue from the user's point of view (on average 6.95s to acquire a concept, including both system and user turns) with a low correction rate (on average 8.2% of turns spent correcting errors).

In terms of usability, the majority of system responses were found to be contextually appropriate (79.16%) with a further small proportion (4.63%) of borderline cases. These borderline appropriate responses were ones that were appropriate according to local context (e.g. the user's last utterance) but not appropriate according to the wider context. This is demonstrated in the example below (taken from an actual dialogue with the system).

6) a. S: What kind of lump does the patient have? [*IA*]
b. U: no lump
c. S: And is it a breast nodularity? [*IA*]
d. U: wh...what do you mean?
e. S: For example: pre-existing breast nodularity, or symmetrical breast nodularity [*AI*]
f. U: neither
g. S: Does the patient have a skin ulcer? [*AP*]

In this example the system has misinterpreted a prior user utterance as meaning that the patient has a lump, hence leading to an inappropriate system response in (a) (indicated by IA in square brackets at the end of the utterance). The user then tries to correct the data (b), but their utterance is again misrecognised (the string returned by the speech recogniser in this case was "a lump") leading to another inappropriate utterance in (c) trying to clarify the type of lump (this time asking directly for the type of lump that the system is interested in – a

breast nodularity). The user then asks for help in (d) and so the system response in (e), providing an explanation, is appropriate in the local context of the user's last utterance but inappropriate in the wider context in which the user is trying to indicate that the patient does not have any kind of lump (marked as AI here). The user utterance in (f) finally ends the digression regarding whether or not the patient has a breast nodularity, and so the next system prompt in (g) is fully contextually appropriate (marked as AP).

In addition, it was found that system prompts were sufficient to elicit user responses were almost all (94.18%) characterised as 'concise and responsive' (e.g. "S: what is the patient's age? U: 35") or 'usable but not concise' (e.g. "S: what is the patient's age? U: her age is 30"). The remaining responses (5.82%) were characterised as 'responsive but not usable' (e.g. "S: what is the patient's age? U: she's middle-aged"), 'not responsive' (e.g. "S: what is the patient's age? U: I don't know"), or as containing no speech (e.g. just noise or silence). Perhaps for these reasons, users did not, on average, make use of system help, and when they did it was only for a small proportion (5.3%) of turns. Although users' own assessments of usability were not formally investigated (e.g. via user questionnaires), anecdotal report suggested that users generally found the system easy to use, although it was also clear that there are areas in which the system needs to be improved, in particular: verification of decisions, and more robust handling of repeated speech misrecognitions of the same concept.

In addition to evaluating the ERA dialogue application, the reconfigurability of the generic dialogue system was tested by porting it to a new, larger, and more complex domain: genetic risk assessment based on family history. Despite the fact that the original system, and its application to the ERA domain, took about 2 person-years to develop, it took only 3 person-months of effort to extend the system to support this new domain (primarily extending the templates for generating speech grammars and prompts). Although, the new application has not yet undergone any performance evaluation, this alone suggests that the original framework generalises to different and more complex medical domains, and that the infrastructure of the existing system is sufficient to allow new domains to be rapidly implemented.

Conclusions

This paper has presented an innovative approach to building spoken dialogue systems in which the dialogue model is split into high-level and low-level representations, with the latter generated dynamically from the former. It can therefore make use of current voice-based standards such as VoiceXML, which are widely employed in commercial systems, whilst also utilizing high-level notions of intention, information and attention that form the basis of much research into discourse structure. It was further proposed that the high-level representation can be derived from a domain plan and ontology, hence removing the need to author dialogues directly, and providing reconfigurability, as well as allowing greater integration with the application domain and non-dialogue tasks.

An implementation of this approach based on 3-layer agent architecture was described, followed by an application of this system in the domain of electronic referrals for breast cancer, which made use of a pre-existing process specification and medical ontology. An evaluation of this application was reported, which, whilst preliminary in nature, suggests that the approach described here has been successful in providing a framework for rapid implementation of a particular class of dialogue systems based on pre-existing medical knowledge representation schemas. Furthermore, the resulting applications appear to exhibit a high degree of competence without loss of system performance or usability. It is suggested that this derives from basing the dialogue system on high-level knowledge representations, which allow more sophisticated reasoning about dialogue structure than the simple task lists employed in many other systems.

Acknowledgements

This work was funded by the European Union under the 5th Framework HOMEY Project, IST-2001-32434 (see http://www.acl.icnet.uk/lab/homey.html). Thanks to the project partners for many useful discussions and advice.

References

Allen, J., Ferguson, G., and Stent, A. (2001). An Architecture for More Realistic Conversational Systems. *Proc. Intelligent User Interfaces 2001 (IUI-01)*, Santa Fe, NM, Jan 14th – 17th.

Azzini, I., Falavigna, D., Gretter, R., Lanzola, G. and Orlandi, M. (2001). First Steps Toward an Adaptive Spoken Dialogue System in Medical Domain. In *Proc. Eurospeech 2001*, Aalborg, Denmark, Sept.

Beveridge, M. A., Giorgino, T., Falavigna, D., Gretter, R. (2004). *Validation Report*, Public Deliverable D19, HOMEY Project, IST-2001-32434 (http://www.acl.icnet.uk/lab/homey.html).

Bohlin, P., Bos, J., Larsson, S., Lewin, I., Matheson, C. & Milward D. (1999). *Survey of Existing Interactive Systems*. Deliverable D1.3, TRINDI Project, LE4-8314.

Bonasso, R. P., Kerr, R., Jenks, K., and Johnson, G. (1998). Using the 3T Architecture for Tracking Shuttle RMS Procedures. In *Proc. IEEE International Joint Symposia on Intelligence and Systems*, 21 – 23 May, Rockville, MD.

Brooks, R. A. (1990). Elephants Don't Play Chess. *Robotics and Autonomous Systems* 6:3-15.

Bunt, H. (1996). Dynamic Interpretation and Dialogue Theory. In M. Taylor, F. Neel and D. Bouwhuis (eds) *The Structure of Multimodal Dialogue*, vol. 2, John Benjamins, Amsterdam.

Burton, M., and Brna, P. (1996). Clarissa: an exploration of collaboration through agent-based dialogue games. In *Proceedings of the EuroAIED*, Lisbon.

Bury, J., Humber, M., and Fox, J. (2001). Integrating Decision Support with Electronic Referrals. In R. Rogers, R. Haux and V. Patel (Eds) *Medinfo*. IOS Press, Amsterdam.

Carberry, S., Chu, J., and Green, N. (1993). Rhetorical Relations: Necessary but Not Sufficient. *Proc. Workshop on Intentionality and Structure in Discourse Relations*. ACL-93, Columbus, OH.

Carletta, J., Isard, A., Isard, S., Kowtko, J., and Doherty-Sneddon, G. (1996). *HCRC Dialogue Structure Coding Manual*. Technical Report HCRC/TR-82, Human Communication Research Centre, University of Edinburgh, UK.

Ceusters, W., Martens, P., Dhaen, C., and Terzic, B. (2001). LinkFactory: an Advanced Formal Ontology Management System. *Proc. Interactive Tools for Knowledge Capture Workshop, KCAP-2001*, Victoria B.C., Canada.

Cohen, P., and Perrault, C. R. (1979). Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science* 3(3):177-212.

Dahlbäck, N., and Jönsson, A. (1997). Integrating Domain Specific Focusing in Dialogue Models. In *Proceedings of EuroSpeech '97*, Rhodos, Greece.

Dahlbäck, N. and Jönsson, A. (1999). Knowledge sources in spoken dialogue systems. In *Proceedings of Eurospeech '99*, Budapest, Hungary.

Flycht-Eriksson, A. (2000), A Domain Knowledge Manager for Dialogue Systems. *Proceedings of the 14th European Conference on Artificial Intelligence, ECAI 2000*. IOS Press, Amsterdam.

Fox, J., Beveridge, M. A., and Glasspool, D. (2003). Understanding Intelligent Agents: Analysis and Synthesis, *AI Communications*, 16, IOS Press, Amsterdam, pp. 139 – 152.

Gat, E. (1998). On Three-Layer Architectures. In *Artificial Intelligence and Mobile Robots*, D. Kortenkamp, R.P. Bonasso and R. Murphy (eds), AAAI Press, Menlo Park, CA.

Gorin, A. L., Riccardi, G., and Wright, J. H. (1997). How May I Help You? *Speech Communication*, 23(1):113-127.

Grosz, B., and Sidner, C. (1986). Attention, Intention and the Structure of Discourse. *Computational Linguistics* 12(3):175-204.

Heid, U., Bernsen, N., and Dybkjaer, L. (1998). *Current Practice in the Development and Evaluation of Spoken Language Dialogue Systems.* Deliverable D1.8, DISC Project, Esprit Long-Term Research Concerted Action No. 24823.

Hobbs, J. R. (1996). On the Relation between the Informational and Intentional Perspectives on Discourse. In *Burning Issues in Discourse: an Interdisciplinary Account*, E. Hovy and D. Scott (eds), Springer-Verlag, Berlin.

Houghton, G., and Isard, S. D. (1987). Why to Speak, What to Say and How to Say it: Modelling Language Production in Discourse. In *Modelling Cognition*, P. Morris (ed), John Wiley & Sons, pp. 249-267.

Hovy, E. H. (1993a). In Defense of Syntax: Informational, Intentional, and Rhetorical Structures in Discourse. *Proc. Workshop on Intentionality and Structure in Discourse Relations ACL-93*, Columbus, OH.

Hovy, E. H. (1993b). Automated Discourse Generation Using Discourse Structure Relations. In *Artificial Intelligence* 63, Special Issue on Natural Language Processing.

Hulstijn, J. (2000). Dialogue Games are Recipes for Joint Action. *Proceedings of Gotalog '00*, *4th Workshop on the Semantics and Pragmatics of Dialogues*. Gothenburg.

Kowtko, J. C. and Isard, S. D. (1993). *Conversational Games Within Dialogue*, Research Paper 31, Human Communication Research Centre, Edinburgh.

Kreutel, J., and Matheson, C. (2000). Obligations, Intentions, and the Notion of Conversational Games. In *Proceedings of Gotalog 00, 4th Workshop on the Semantics and Pragmatics of Dialogues*. Gothenburg.

Larsson, S., Ljunglof, P., Cooper, R., Engdahl, E., and Ericsson, S. (2000). GoDiS – An Accommodating Dialogue System, *Proc. ANLP/NAACL-2000 Workshop on Conversational Systems*, Seattle, May 2000.

Larsson, S., and Traum, D. (2000). Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering*, 6:323—340, Special Issue on Spoken Language Dialogue System Engineering.

Levinson, S. (1983). Pragmatics. Cambridge University Press, Cambridge, UK.

Lewin, I. (2000). A Formal Model of Conversational Game Theory. *Proceedings of Gotalog 00, 4th Workshop on the Semantics and Pragmatics of Dialogues*. Gothenburg.

Maier, E. A., and Hovy, E. H. (1993). Organizing Discourse Structure Relations using Meta-Functions. In *New Concepts in Natural Language Processing: Planning, Realization, and Systems*, H. Horacek and M. Zock (eds). Pinter Publisher, London, pp. 69-86.

Mann, W. D., and Thompson S. A. (1988). Rhetorical Structure Theory: Towards a functional theory of text organization. *Text*, 8(3):243-281.

Marcu, D. (2000). Extending a Formal and Computational Model of Rhetorical Structure Theory with Intentional Structures a la Grosz and Sidner. In *Proc. COLING 2000*, pp. 523-529.

Maudet, N. and Evrard, F. (1998). A generic framework for dialogue game implementation. In J. Hulstijn and A. Nijholt (eds) *Proceedings of the second workshop on Formal Semantics and Pragmatics of Dialogue*, May 13-15, University of Twente, Enschede, Netherlands.

McGlashan, S., Burnett, D. C., Carter, J., Danielsen, P., Ferrans, J., Hunt, A., Lucas, B., Porter, B., Rehor, K., and Tryphonas, S. (2004). Voice Extensible Markup Language (VoiceXML) Version 2.0, W3C Recommendation, 16th March. Available from http://www.w3.org/TR/voicexml20/.

McKeown, K. R. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge.

Milward, D., and Beveridge, M. A. (2004). Ontologies and the Structure of Dialogue. In *Proceedings of CATALOG, 8th Workshop on the Semantics and Pragmatics of Dialogue*, 19th – 21st July, Barcelona, Spain.

Mittal, V. O., and Paris, C. L. (1993). On the Necessity of Intentions and (at Least) the Usefulness of Rhetorical Relations: A Position Paper. *Proc. Workshop on Intentionality and Structure in Discourse Relations*. ACL-93, Columbus, OH.

Montoro, G., Alamán, X., and Haya, P. A. (2004). A Plug and Play Spoken Dialogue Interface for Smart Environments. In *Proceedings of 5th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'04)*, 15-21 February, Seoul, Korea.

Moore, J. (1995). The Role of Plans in Discourse Generation, In *Discourse: Linguistic, Computational, and Philosophical Perspectives*, Daniel Everett and Sarah G. Thomason (Eds.).

Moore, J. D., and Paris, C. L. (1992). *Planning Text for Advisory Dialogues: Capturing Intentional, Rhetorical and Attentional Information*. Technical Report from the University of Pittsburgh, Department of Computer Science (Number 92-22) and from USC/ISI, #RS 93-330.

Moore, J. D., and Pollack, M. E. (1992). A Problem for RST: The Need for Multi-Level Discourse Analysis. *Computational Linguistics* 18(4).

Moore, J. D., and Swartout, W. R. (1990). Dialogue-Based Explanation. In *Natural Language in Artificial Intelligence and Computational Linguistics*, C.L. Paris, W.R. Swartout & W.C. Mann (eds). Kluwer, Boston, pp. 3-48.

Moser M., and Moore, J. D. (1993). Investigating Discourse Relations. *Proc. Workshop on Intentionality and Structure in Discourse Relations*. ACL-93, Columbus, OH.

Moser, M., and Moore, J. D. (1996). Toward a Synthesis of Two Accounts of Discourse Structure. *Computational Linguistics* 22(3):409-419.

Paris, C. L. (1990). Generation and Explanation: Building an Explanation Facility for the Explainable Expert Systems Framework. In *Natural Language in Artificial Intelligence and Computational Linguistics*, C.L. Paris, W.R. Swartout & W.C. Mann (eds). Kluwer, Boston, pp. 49-82.

Poesio, M. and Mikheev, A. (1998). The Predictive Power of Game Structure in Dialogue Act Recognition: Experimental Results Using Maximum Entropy Estimation. In *Proc. ICSLP98*.

Power, R. (1979). The Organization of Purposeful Dialogues. *Linguistics*, 17:107-152.

Pulman, S. G. (1997). Conversational Games, Belief Revision and Bayesian Networks, In *Proc. 7th Computational Linguistics in the Netherlands Meeting*, pp. 1-25.

Simpson, A., and Fraser, N. (1993). Black Box and Glass Box Evaluation of the SUNDIAL System. In *Proc. 3rd European Conference on Speech Communication and Technology (Eurospeech'93)*, Berlin, Germany.

Shiffman, R. N., Brandt, C. A., Liaw, Y., and Corb, G. J. (1999). A Design Model for Computer-Based Guideline Implementation Based on Information Management Services. *Journal of the American Medical Informatics Association (JAMIA)* 6(2):99-103.

Stede, M., and Schlangen, D. (2004). Information-Seeking Chat: Dialogue Management by Topic Structure. In *Proceedings of CATALOG, 8th Workshop on the Semantics and Pragmatics of Dialogue*, 19th – 21st July, Barcelona, Spain.

Stent A. (2000). Rhetorical Structure in Dialog. *Proc. 2nd International Natural Language Generation Conference (INLG'2000)*.

Sutton, D., and Fox, J. (2002). The Syntax and Semantics of the PRO*forma* Language, *Journal of the American Medical Informatics Association (JAMIA)*, 10(5):433-443.

Sutton, S., Hansen, B., Lander, T., Novick, D. G., and Cole, R. (1995). *Evaluating the Effectiveness of Dialogue for an Automated Spoken Questionnaire*. Technical Report CS/E95-12, Dept. of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology.

Traum, D. R. (1993). Rhetorical Relations, Action and Intentionality in Conversation. In *Proc. ACL SIG Workshop on Intentionality and Structure in Discourse Relations*, ACL-93, Columbus, OH, pp. 132-135.

Traum, D. R. (1996). A Reactive-Deliberative Model of Dialogue Agency. In J. P. Muller, M. J. Wooldridge, and N. R. Jennings, editors, *Intelligent Agents III - Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*.

Traum, D. R., and Allen J. F. (1994). Discourse Obligations in Dialogue Processing. *Proc. 32nd ACL*, Las Cruces, New Mexico, pp. 1-8.

Wooldridge, M. (2002). *An Introduction to MultiAgent Systems*. John Wiley & Sons Ltd, Chichester, UK.

Young, M. R., and Moore, J. D. (1994). Does Discourse Planning Require a Special Purpose Planner? *Proc. Annual Meeting of the American Association for Artificial Intelligence*, Seattle.

Appendix A: Evaluation Script

1. S: Hello, welcome to the breast cancer referrals system.
2. **U: Hello**
3. S: What is the patient's age?
4. **U: She is thirty, female, and she has a bilateral nipple discharge**
5. S: Ok. And is it a bloodstained nipple discharge?
6. **U: No**
7. S: Ok. And is it a cloth-staining nipple discharge?
8. **U: No**
9. S: Ok. Does the patient have an acquired nipple deformity?
10. **U: What do you mean?**
11. S: For example: gestational inversion of nipple
12. **U: No, and she doesn't have a lump, or any eczema**
13. S: Ok. Does the patient have a skin ulcer?
14. **U: Yes**
15. S: Ok. Does the patient have any intractable pain?
16. **U: Yes**
17. S: Ok. Does the patient have a breast abscess?
18. **U: Yes, and she has a cyst**
19. S: Ok. Do you mean a breast cyst?
20. **U: Yes**
21. S: Ok. So this patient is thirty, female, and has a skin ulcer, intractable pain, a breast abscess, a breast cyst, and a bilateral nipple discharge. Is that correct?
22. **U: No, she isn't thirty** *[or whatever age the system says]*
23. S: What is the patient's age?
24. **U: Thirty-five**
25. S: Ok. So this patient is thirty-five, female, and has a skin ulcer, intractable pain, a breast abscess, a breast cyst, and a bilateral nipple discharge. Is that correct?
26. **U: Yes**
27. S: I recommend urgent referral to breast specialist. Which action do you advise: urgent referral to breast specialist, support by general practitioner or non-urgent referral to breast specialist?
28. **U: Why do you recommend urgent referral?**
29. S: Urgent referral to breast specialist is indicated by the fact that this patient has a skin ulcer.
30. **U: What are the arguments for non-urgent referral**
31. S: Non-urgent referral to breast specialist is indicated by the fact that this patient has intractable pain and a breast abscess.
32. **U: Are there any arguments for support by general practitioner?**
33. S: There are no arguments for support by general practitioner.
34. **U: Ok, I advise urgent referral**
35. S: Ok. Do you mean urgent referral to breast specialist?
36. **U: Yes**
37. S: Ok. An urgent referral to breast specialist has been made for this patient.
38. **U: Ok**
39. S: Thankyou for using this service. Goodbye.
40. **U: Bye**