EFFECTIVE HUMAN-COMPUTER COOPERATIVE SPOKEN DIALOGUE: THE AGS DEMONSTRATOR

M.D. Sadek, A. Ferrieux, A. Cozannet, P. Bretier, F. Panaget, and J. Simonin

France Télécom — CNET — LAA/TSS/RCP
Technopole Anticipa - 2, avenue Pierre Marzin - F22307 Lannion Cedex France
Emails: {sadek, ferrieux, cozannet, bretier, panaget, simonin}@lannion.cnet.fr

ABSTRACT

This paper describes a spoken dialogue system based on a generic specification of a cooperative communicating rational agent. We present some theoretical and practical apects of the overall approach, along with the speech-specific and natural language related issues raised by the effective implementation of the system. An account is also given of an evaluation of the system with naive users on the task of voice services directory (AGS) inquiry.

1. INTRODUCTION

Among the fields of application of intelligent systems, man-machine dialogue plays a special role: a system designed ultimately to replace a human operator must satisfy stronger requirements than most computer-based tools. For example, cooperative behaviour such as providing suggestive answers, is an obvious prerequisite for acceptance by humans. This alone favours a generic intelligent-agent-based approach to automated dialogue. The need for such an approach appears more strongly in the special case of *spoken* man-machine dialogue not only because of the current limitations of speech recognition technology at the signal level, but more fundamentally because the natural freedom to which people are accustomed in spoken exchanges introduces a source of complexity at its own level. Handling such spontaneous behaviors is also a target for highly generic approaches.

Guided by our conviction that generic intelligent systems are particularly useful in this context, we designed a complete spoken dialogue system which satisfies the external requirements mentioned above. The first two sections of this paper give the context of the application, illustrated by a sample dialogue. Section 3 describes the architecture we implemented. Since some of the problems of natural language input (for example) do not have generic solutions yet, we had to build specialized components to handle the corresponding tasks. These "black boxes" are described in section 4. The heart of the system, however, is a generic logical theory implemented with a theorem prover. This *rational unit*, which embodies a real demonstration of the applicability of the generic AI approach to dialogue, is described in section 5. Finally, we give the results of an evaluation of the system by naive users.

2. APPLICATION CONTEXT

Our system is demonstrated on the AGS application [14], i.e., the vocal query, over the telephone, of the shared-revenue information voice services directory hosted by France Télécom (Audiotel services). We have chosen to begin with the following restricted domain and semantic representation level: (1) the system only handles two major classes of services: weather forecasts and employment, and (2) the relevant attributes for each server are macroscopic ones: localization, target occupations, and so on. With this kind of semantic information as a background, the system accepts nearly unconstrained spontaneous language as input, and reacts in a cooperative way with natural-sounding output. The speech-specific components are described in section 3; below is a sample dialogue transcription, which illustrates both speech- and non-specific challenges.

3. SAMPLE DIALOGUE

Below is an example of dialogue with the AGS demonstrator (*U: user, R: output of the speech recognition system (X: "semantic garbage"), S: system*):

S: Welcome to the AGS server. How can I help you?
Bienvenue sur le serveur A.G.S. Que puis-je pour vous ?

U: Hello, I would like to know the sea weather forecast for the Finistère area.
Bonjour, je voudrais connaître la météo marine du Finistère.

R: ain X savoie meteo marine du finistere.

S: There is no sea weather forecast server for the Finistère area. But there are some for the Channel and for the Atlantic. Are you interested in one of these?
Il n'y a pas de serveurs de météo marine pour le Finistère. Par contre, il y en a pour la Manche ou l'océan Atlantique. L'un d'eux vous convient-il ?

U: Yes euh the Channel.
Oui euh la manche.

R: oui X manche.

S: There are two servers for coastal weather forecast for the Channel. Their phone numbers are 36 70 12 34 and 36 68 08 08. Would you like more information?
Il y a 2 serveurs de météo marine pour la Manche. Leurs numéros sont le 36 70 12 34 et le 36 68 08 08. Désirez-vous un autre renseignement ?

U: Does it have specific weather forecast for sailing?
Est ce que il délivre la météo dédiée à la voile ?

R: est ce que X X meteo dediee X voile.

S: No, it doesn't. There is no weather forecast server specialized for sailing for the Channel. But there are some for France. Is that ok?
Non. Il n'y a pas de serveur de météo dédiée pour la voile pour la Manche. Par contre, il y en a pour la France. Est-ce que ça vous convient ?

U: Yes.
Oui.

R: oui.

S: There are two servers (...) Would you like more information?
Il y 2 serveurs (...) Désirez-vous un autre renseignement ?

U: Yes. I am looking for a job in the Ile-de-France area. Could you give me some numbers for job offers?
Oui je cherche du travail en Ile-de-France pourriez vous me donner des numéros de serveurs de petites annonces ?

R: oui le cher du travail en X provence X me donner numero le serveur X annonce.

S: There are two servers for job offers in Provence (...) Would you like more information?
Il y a 2 serveurs de petites annonces d'emplois pour la Provence (...) Désirez-vous un autre renseignement ?

U: No, not Provence but Ile-de-France?
Non, pas la Provence l'Ile-de-france ?

R: non pas provence ile de france.

S: A server for the Ile-de-France area. What type of employment are interested in?
Un serveur pour l'Ile-de-France. Vous êtes intéressé par quel métier ?

U: What can you propose?
Que me proposez-vous ?

R: X proposez vous.

S: There are, for example, the following items: agriculture, entertainment, communication and hotel employment.
Vous avez par exemple l'agriculture, l'animation, la communication et l'hotellerie.

U: Yes, for hotels.
Oui l'hôtellerie.

R: oui hotellerie.

S: There are two servers (...) Would you like more information?
Il y a 2 serveurs (...) Désirez-vous un autre renseignement?

U: (...)

4. ARCHITECTURE

The system runs on a single Sun Ultra-SPARC1 workstation. An ISDN interface handles phone calls, getting the standard 8kHz μ-law speech input. CNET's HMM-based speaker-independent, continuous-speech recognition software, using a bigram language model (PHIL-DYANE), computes the best word sequence [6, 8]. The natural language input subsystem finds the best coherent interpretation of the sentence. This interpretation is formulated as dialogue acts. The kernel of the system, a rational unit, cooperatively reacts to this input by producing dialogue acts. The generation subsystem verbalizes these acts, producing sentences in a natural language. CNET's PSOLA-based TTS synthetizes these sentences [2]. The resulting signal is fed back into the telephone.

5. SPECIALIZED COMPONENTS

The above description includes two isolated subsystems to deal with natural language input and output. Today, algorithmic bottlenecks (esp. the size of the search space for the input side) prevent the direct logical specification of these processes in our forward-chaining theorem-proving framework. However, both are meant to be eventually integrated as logical theories in the rational unit.

5.1. Natural Language Input

The natural language interpretation subsystem features both syntactic and semantic robustness using island-driven parsing and semantic completion. Island-driven parsing simply means to spot small syntactic structures in the text, with as few long-range dependencies as possible.

Example:

Input sentence:
I'd like to know the weather forecast for the Lannion area

Recognized:
I_d_like X weather forecast X X lannion area

Concepts:
I_u $\quad weather\_forecast \quad$ $lannion$

The result is a set of mentioned concepts, or a list of possible alternatives when overlapping phrases yield nondeterminism. Each of these hypotheses is then fed into the semantic completion process, which builds a well-formed logical formula out of the mentioned items. The fundamental hypothesis here is that of *semantic connectedness*: each of the user's utterances can be described by a connected formula, i.e., the user never mentions two concepts without (at least implicitly) linking them by some relation(s) or other concept(s).

Example: $(mentioned)$ $weather\_forecast \quad lannion$
$(default)$ $\quad phones(y)$
$(inferred)$ $\quad phone(x,y) \quad servers(x)$
$theme(x, weather\_forecast)$
$domain(x, lannion)$

This minimal extension of the mentioned material towards connectedness is a case of Steiner's problem, where the compulsory nodes are the given items, and the underlying graph is the semantic network describing knowledge about the domain (e.g., the relations defined on a certain kind of object). We implemented a heuristic solution to this problem given by [7], and extended it to the case where the underlying network is not known in advance, but generated (instances of general typing rules are repeatedly built on the given nodes, yielding new (existentially quantified) nodes, and thus potentially generating an infinite graph). What we get is a minimal contextually consistent first-order formula containing the mentioned objects. We then obtain a full-fledged dialogue act (hopefully) conveying the user's request whose logical form is suited for direct use by the rational unit:

$$< u, inform(s, I_u\, Kref_u(\iota y\ phones(y) \wedge \exists x\ server(x) \wedge phone(x,y) \wedge theme(x, weather\_forecast) \wedge domain(x, lannion))) >$$

Informally, this formula means that the user (u) informs the system (s) that she/he wants (I_u) to know the ($Kref_u$, ιy) phone number of a weather forecast server for Lannion.

5.2. Natural Language Output

The natural language generation subsystem [10, 9] contains a *linguistic acts planner* and a *linguistic realizer* (see also Appelt's system [1]). The former is concerned with how the intentions of the system (i.e., dialogue acts) are communicated to the user. The latter is concerned with taking the

acts specified by the planner and realizing them in a well-formed utterance.

Linguistic acts serve as an abstract representation of utterances or parts of utterance. Two families are distinguished: *surface speech acts* and *referring acts*. There are three types of surface speech acts, corresponding to the three fundamental sentential modes in French: imperatives, interrogatives and declaratives. The aim of the use of surface speech acts is to deal with the fact that: (1) different dialogue acts can be realized by the same utterance and (2) a single utterance can verbalize a complex sequence of dialogue acts. There are also three types of referring acts corresponding to noun phrase, pronoun and proper name, respectively. Linguistic acts are defined in terms of their effect and their preconditions of feasibility. The effect of a surface speech act is the verbalisation of dialogue acts whereas the effect of a referring act is the reference to objects of the world. Preconditions are used to express that: (1) the accomplishment of a linguistic act must be relevant to the context in which the utterance will be used and (2) there exists a well-formed part-of-utterance. An interesting point is that the second type of precondition establishes explicitly the communication channel from the planner to the realizer.

The linguistic realizer satisfies the second type of precondition by constructing the most relevant parts-of-utterance and utterances according to the context and to its linguistic knowledge. During the construction, when the realizer requires a noun phrase, a pronoun or a proper name, it asks the planner to select a referring act. This mechanism establishes the second direction of the communication channel.

For example, if the system wants to inform the user that there is a "provider" relation between the Côtes-d'Armor weather forecast server and Météo-France, namely if the dialogue act to verbalise is the following:

$$< s, inform(u, provider(\iota x\ server(x) \wedge theme(x, weather\_forecast) \wedge domain(x, cotes\_d\_armor), meteo\_france)) >$$

the generator can produce, according to the context, either a simple declarative sentence with a proper name and a noun phrase, a positive answer with a pronoun and a proper name or an elliptical sentence:

"Météo-France is the provider of the Côtes-d'Armor weather forecast server."
"Yes, it is provided by Météo-France."
"Météo-France."

An interesting property of the generation subsystem is the ease of moving to a new application or a new output language. For instance, the specification of the linguistic knowledge needed to produce English sentences (in the context of the AGS application) was achieved by one person in 2 weeks.

5.3. Constraint Relaxation Engine

Another "black box" is integrated in the design: the rational unit accesses the underlying database through a very fast metric-oriented engine, which is optimized to find satisfactory approximate answers when no real solutions exist. Indeed, if the contents of a database (like a directory) are viewed as points of a (product) space, it is a natural extension to give a metric to each of the coordinates (fields). The obtained product metric can then be used as a measure of proximity between two points, or between a point and a query (subspace). This allows us to homogeneously encode the kind of relaxation we expect in the relative scaling of the individual field metrics.

A nice additional feature of this approach is that it naturally produces a meaningful solution to the opposite problem: when the query is too weak and yields too many solutions, the most relevant constraint to instantiate can be defined as the field of maximal diameter on the set of selected solutions.

This approach allows us to represent the whole "relaxational background" of an application in an efficient form. This point stresses the wide applicability of such a relaxation engine, inside or outside the context of a rational unit.

6. THE RATIONAL UNIT

The central guideline of our approach is that an intelligent dialogue system has to be an intelligent system first of all. Intelligent behavior is clearly needed in application contexts requiring complex but still user-friendly interactions with human beings. Cooperative human-machine dialogue illustrates such a context. The most consensual achievement of intelligent behavior is rationality. In a simplified way, to behave rationally is to be permanently driven, at a certain representation level, by principles which optimally select the actions leading to those futures in conformity with a given set of motivations and desires (see, e.g., [5] and [11, 13]). It is at this (hypothesised) Knowledge Level that the concepts of mental attitude and intentional action are relevant.

As regard the formal approach for knowledge representation, the logic framework is adequate, for various reasons: its homogeneity, its genericity (due to its large coverage), its ability to intuitively account for mental attitudes (which makes it easy to maintain), and its potential usability both as a modelling and an implementation tool.

In this approach to dialogue, the main idea is that a dialogue process can be completely justified by rational behavior principles (which are more basic that discourse rules). Due to the genericity of its principles, this approach achieves the robustness required by an (intelligent) dialogue system: to soundly react to complex situations, possibly incompletely specified when the system has been designed.

This rational unit is the kernel of an intelligent agent. It gives to the system its dialogue ability, that results from explicit reasoning processes [11, 12, 13]. It involves a homogeneous set of generic logical axioms, which formalize basic principles for rational behavior and introspection, communication, and cooperation. The *rationality principles* characterize the agent's planning mechanisms. On the one hand, they allow an agent to infer the intentions which have motivated a communicative action he has observed. On the other hand, they specify the conditions which have to be satisfied for a

given action to be planned and performed. A central rationality principle involved in the agent planning process states that an agent cannot intend to bring about some proposition without intending, at the same time, that one of the actions leading to the desired effect be done. The *cooperation principles* express the agent's motivation to behave cooperatively, such as adopting the partner's intention whenever the agent has no reason not to do so.

The *coherence relationships* (such as belief consistency, or the fact that an agent cannot intend to achieve a property that he believes to be already satisfied) together with the rationality (and the cooperation) principles state a sound framework for the rational balance over all the agent's mental attitudes (i.e., beliefs, uncertainties, intentions, and plans).

In this framework, the planning process is a regular consequence of the rational behavior axioms, and is therefore completely deductive. Thus, communicative act plans are generated by inferring causal chains of intentions. Note that it is only because communicative acts are modelled in terms of rational actions involving an intention that the system is able to handle (cooperative) dialogue.

Let us take an example. Suppose that the system observes the user performing the following action (i.e., the natural language input subsystem identifies the corresponding dialogue act): the user informs the system that she/he wants to know if proposition p is true. On the basis of the rationality principles, the system infers the user's intention. Its cooperation principles lead it to adopt that intention, namely that the user will come to know if p is true. Also by the rationality principles, the system adopts the intention either to inform the user that p is true or to inform her/him that $\neg p$ is true (which are the actions expected to lead to the desired state). Then, the system will only select the one of the two actions which is currently feasible, for example informing the user that p, if it believes p. The selected act is then handled by the natural language generation subsystem.

The inference engine supporting the rational unit is a theorem prover for first-order modal logic, implemented using a "syntactical" approach [3, 4]. Inference is automated using an extended resolution method, where formulae are represented in their syntactical form and where the instantiation of axiom schemata uses sub-formulae unification. The inference process is supplied with the logical theory of rational interaction, briefly described above.

7. EVALUATION

A real test of the first prototype of the system (not involving a complete explicit rational unit but partially "emulating" its behavior) has been performed. 35 naive users (using speaker-dependent models), who were given informal scenarios, were asked to call the system and to request some particular information. A corpus of 600 dialogues was transcribed, and analysed. The mean duration of a dialogue session was 3.1 minutes and its mean length was 11 exchanges. The real time factor was approximately 3.5. The global goal success rate was 61% (and reached 73% when the goal was limited to finding out only the service phone numbers) despite the fact that only 49% of the user utterances were still understandable after the recognition process and that for more than 38%, the meaning was severely corrupted by the insertion of signifiant words leading to clarification sub-dialogues.

8. CONCLUSION

The AGS demonstrator we have described displays advanced human-computer user-friendly spoken dialogue abilities. The results of the first evaluation of the system are encouraging. This dialogue technology, based on the concept of communicating rational agent, is expected to be experimented, in the short or middle term, in context of real services.

9. REFERENCES

1. D. E. Appelt. *Planning English Sentences.* Cambridge University Press, 1985.
2. D. Bigorgne et al. Multilingual PSOLA text-to-speech system. *Proc. ICASSP'93*, Minneapolis, MN, 1993.
3. P. Bretier. *La communication orale coopérative : contribution à la modélisation logique et à la mise en œuvre d'un agent rationnel dialoguant.* Thesis diss., Univ. Paris 13, France, 1995.
4. P. Bretier & M. D. Sadek. Designing and implementing a theory or rational interaction to be kernel of a cooperative spoken dialogue system. In *Proc. AAAI Fall Symposium on Rational Agency*, Cambridge, MA, 1995.
5. P.R. Cohen & H.J. Levesque. Rational interaction as the basis for communication. In P.R. Cohen, J. Morgan, & M. Pollack eds, *Intentions in communication*, MIT Press, 1990.
6. P. Dupont. Dynamic use of syntactical knowledge in continuous speech recognition. *Proc. Eurospeech'93*, Germany, 1993.
7. L. Guyard-Daher. *Le problème de l'arbre de Steiner, modélisation par programmation linéaire et résolution par des techniques de décomposition.* Thesis diss., ENST Paris, France, 1985.
8. D. Jouvet, K. Bartkova, & J. Monné. On the modelization of allophones in an HMM-based speech recognition system. *Proc. Eurospeech'91*, 923-926, Genova, Italy, 1991.
9. F. Panaget. Using a textual representation level component in the context of discourse and dialogue generation. In *Proc. 7th International Workshop on Natural Language Generation*, Kennebunkport, MN, 1994.
10. F. Panaget. *D'un système générique de génération d'énoncés en contexte de dialogue oral à la formalisation logique des capacités linguistiques d'un agent rationnel dialoguant.* Thesis diss., Univ. Rennes 1, France, 1996.
11. M. D. Sadek. *Attitudes mentales et interaction rationnelle : vers une théorie formelle de la communication.* Thesis diss., Univ. Rennes 1, France, 1991.
12. M. D. Sadek. Towards a theory of belief reconstruction: application to communication. *Speech Communication Journal'94, special issue on Spoken Dialogue*, 15(3-4), 1994.
13. M. D. Sadek. Communication theory = rationality principles + communicative act models. In *Proc. AAAI Workshop on Planning for Interagent Communication*, Seattle, WA, 1994.
14. M.D. Sadek, A. Ferrieux, & A. Cozannet. Towards an artificial agent as the kernel of a spoken dialogue system: A progress report, *Proc. AAAI Workshop on Integration of Natural Language and Speech Processing*, Seattle, WA, 1994.