Identification of Speaker Actions in Mixed Initiative Dialogue

Dimitris Papazachariou, Niels Ole Bernsen, Laila Dybkjær, Hans Dybkjær
{dimitris, nob, laila, dybkjaer}@cog.ruc.dk
phone (+45) 46 75 77 11, fax (+45) 46 75 45 02
Centre for Cognitive Science, Roskilde University
P.O.Box 260, DK-4000 Roskilde, Denmark

Abstract

Today's state-of-the-art spoken language dialogue systems (SLDSs) are mainly system-directed leaving very little initiative to the user. This paper discusses how to enable controlled steps in the direction of mixed-initiative SLDSs. The discussion is based on experience from the design, implementation and test of system-directed dialogue for a spoken language dialogue system and on first Wizard of Oz experiments towards achieving implementable mixed-initiative dialogue. To this end, we present a categorisation of utterances in simulated human-machine dialogue based on an identification of the specific actions users perform with them. The categorisation builds on Speech Act Theory and holds the promise that limited mixed-initiative human-machine dialogue may be implemented in current SLDSs.

1 Introduction

Today's state-of-the-art spoken language dialogue systems (SLDSs) are mainly system-directed leaving very little initiative to the user. Such systems are acceptable for solving tasks which are well-structured in the sense that there is a prescribed amount of information which has to be exchanged between user and system and preferably in a certain order. However, many tasks are not well-structured and even those which are may drive system-directed dialogue close to its limits in some situations.

As discussed in [Dybkjær et al. 1995a] the reservation or ordering task in its pure form is well-structured but the task of making a reservation often includes the sub-task of seeking and providing information about that which is being reserved or ordered. To properly complete the task of booking a flight ticket, for instance, one often needs on-the-spot specific information about fares, timetables or other aspects of the airline travel domain. The task of reservation, in other words, is in many cases a task of *informed reservation.* System-directed dialogue is incapable of enabling the informed reservation task. The reason is that the system is unable to predict when, during the reservation dialogue, the user (interlocutor) might suddenly want to ask for some piece of information in order to be able to continue making the reservation. To ask for information means taking over the initiative from the system, so that the dialogue effectively becomes a mixed initiative dialogue.

The pure airline ticket reservation task belongs to the class of large well-structured tasks which can be handled through system-directed dialogue. The informed airline ticket reservation task belongs to a different and more complex task category, namely that of large *ill-structured* tasks. Such tasks are characterised by having a large number of optional sub-tasks. Each of these sub-tasks may be well-structured in itself but the overall task becomes ill-structured because of the optional character of the many sub-tasks it includes. This means that the system cannot have a valid stereotype that tells which sub-tasks the user wants to accomplish and in which order [Bernsen et al. 1994, Dybkjær et al. 1995a]. In the absence of such a stereotype, and to some extent also because the task size is large, system-directed dialogue is too inefficient for the negotiation of ill-structured tasks. In a nutshell, if you want to ask me (or the system) about something and if I have no idea about what you want to ask me about, it is infinitely more efficient that you pose me the question than that I have to question you to find out what you want to ask me about.

This leads to the main question of this paper: how is mixed initiative SLDSs possible for large, ill-structured tasks and, in particular, how may speech acts be used to enable the construction of mixed initiative SLDSs? A mixed initiative SLDS for informed reservation would require relaxation of the technological constraints of our existing dialogue system, P2 [Bernsen et al. 1995]. We shall assume that a limited

enlargement of system focus and user utterance length, sufficient for the approach to be presented below, will be possible.

However, mixed initiative SLDSs are not currently feasible for large, ill-structured tasks in the general case. Current exploratory design projects include mixed initiative systems for small ill-structured tasks [Kanazawa et al. 1994, Smith 1991], which are not really relevant to our problem. The reasons why systems like those described in Kanazawa et al. [1994] and in Smith [1991] can allow mixed-initiative without restriction are that the task is small and that the vocabulary is small. This allows all possible sub-tasks to be in system focus at the same time, especially when word-spotting is being used instead of full syntactic-semantic representation of user utterances. The use of word-spotting makes it less relevant to consider limiting the length of user utterances.

Mixed initiative dialogue on large ill-structured tasks has been marginally realised in the SUNDIAL system [Peckham 1993]. The full-fledged approach adopted in the Esprit PLUS project would seem to have failed [Grau et al. 1994]. The latter project demonstrates that the problems involved in solving the general case of mixed initiative SLDSs for large, ill-structured tasks not only derive from technological constraints on system focus and user utterance length, but derive as much from unsolved scientific problems in natural language processing.

In system-directed dialogue users' speech acts are by definition limited to answers to the system's questions and, optionally, to issuing commands by using keywords to initiate meta-communication as in P2. However, when the user may also take the initiative and not only by using keywords there is a need to determine the user's speech acts in order to make the system behave appropriately. Unfortunately, speech act identification remains an unsolved problem in realistic applications, mainly because of the existence of indirect speech acts. For example, the user utterance "Can I have this flight?" is not a question about whether the customer is allowed to book a certain flight, but is an actual booking of the flight. Conversely, the user utterance "Is there a flight in the morning?" is not necessarily a booking of this flight even if it exists.

The aim of this paper has been on the basis of a simulated spoken human-machine mixed initiative corpus to define the speaker actions of users in a way which would allow the handling of some degree of mixed initiative dialogue in an implemented SLDS. As a first step in this direction we tried to find an existing corpus of this kind, preferably within the ATIS (air travel information systems) domain. However, this was not possible (Section 2) and we eventually decided to make one ourselves by simulating a system which allowed some degree of user initiative. Section 3 describes the collection and analysis of this corpus. Finally, Section 4 concludes the paper.

2 Existing corpora

In order to study speech acts in mixed-initiative human-machine dialogue we tried to find a suitable corpus. It appears, however, that publicly available simulated human-machine mixed-initiative corpora are non-existing. We considered each of the three corpora described below. However, each of them were found inappropriate for our purposes.

A number of mixed-initiative human-human dialogue corpora are available. One of these is the American Express corpus which contains dialogues between a travel agent from the American Express Card and his/her customers [Sidner 1992]. The topics of the dialogues are reservations, information, help to customers in planning their journeys, cancellations, and changes of specific reservations. However, since the dialogues are conducted freely between humans they show none of the limitations and constraints which are the constant problem in human-machine dialogue and hence are far beyond what can be realised by today's machines. We did not, therefore, find them well-suited for our study since many of the contextual elements used by the human interlocutors are absent from, or at least different in, human-machine dialogue. And if they were present, current machines could not use all of them.

We also considered a simulated human-machine corpus called the "TRAINS" corpus [Gross et al. 1993]. It is a collection of 91 dialogues (only 16 of which are available via ftp) between a human and a system simulated by a human. The dialogues are planning dialogues. The human is a manager who is supposed to construct a plan for the delivery of goods by railway through help from the system. The manager knows only the final goal, and has to obtain all the necessary information from the system in order to successfully plan the necessary stages of delivery. The system has and is able to provide all the necessary information related to the freight problem, and it can check the feasibility of the manager's plan but cannot propose any solution. The human who simulated the system maintained all its functional limitations, but his/her linguistic behaviour was absolutely human and could not be realised in an implemented system.

The simulated human-machine corpus we had collected ourselves when developing our implemented dialogue system, P2, was also considered but could not be used because the dialogues were system-directed, cf. [Dybkjær et al. 1995b] and hence were not well-suited for the study of different user speech acts.

As it appeared impossible to locate and obtain a suitable mixed initiative human-machine corpus for our purpose, we finally decided to collect one ourselves.

3 Collection and analysis of a corpus

We created a small corpus by using the Wizard of Oz (WOZ) method. The task of the dialogues was informed reservation. This means that reservation was the central topic of the conversation and the backbone of the system would be much like the P2 system in which the system asks for the information needed for the successful completion of the reservation task. However, users could take the initiative and ask for information whenever they needed it. In such cases, the system played the role of the expert that has all the necessary information and details about domestic flights, e.g. departures, arrivals, flight numbers and fares.

The corpus we collected was a very small one (typically 2-4 dialogues per iteration and totally about 10 dialogues). Only two different scenarios were used throughout the experiments, and each subject always performed both scenarios. The scenarios forced users to make questions in order to find the optimal solution for the successful completion of their task.

In the beginning more initiative was left to the user than would be possible in an implemented system. The opening system phrase invited the user to take the initiative and the wizard would understand even very long user phrases. Having examined the first dialogues, we realised that we had to impose additional constraints on the simulated system's understanding. Therefore, user initiative was reduced, i.a. in order to decrease the length of the user's first phrase. The system would state its restrictions in its introduction to the dialogue, and if users did not adhere to its admonitions the system would not understand them or only understand the first part of what they said. In the last experiment, for instance, the system would not understand the last part of the following user answer:

1 S: Please state your business.
2 U: I'd like to make a reservation for a flight to Aarhus, for this week-end.

The idea which eventually emerged with respect to what would be feasible in an implemented system, was the following: at the general level we assume that the user's goal is to make a reservation. This allows us to maintain the stereotypical structure of the reservation task as a 'backbone' for dialogue design. This means that the task context will (still) strongly constrain the dialogue behaviour of co-operative users. They can be expected to follow the overall system-directed course of the dialogue and to only take over the initiative when they need information from the system in order to proceed in making the commitments needed for reservation. Users are assumed to take the initiative through asking questions. These questions, moreover, can be expected to primarily concern sub-tasks which are closely associated with the question in current system focus. There remains, however, a number of important unknowns. Firstly, we need additional constraints to ensure limited user utterance length in the cases where users take over the initiative from the system. Secondly, we probably need additional constraints to ensure that users will not be asking for arbitrary pieces of domain information at arbitrary points during the dialogue. And thirdly, we must make sure that the system has the linguistic capabilities to detect the shift in initiative which occurs when, at arbitrary points during dialogue, users request domain information.

When users have more initiative the utterance length can be expected to grow compared to system-directed dialogue in which an elliptical or otherwise brief answer typically will be made. Since utterance length is a critical parameter we must ensure limited growth. Terse system language is known to have a positive effect on utterance length [Zoltan-Ford 1991], and the mainly system-directed dialogue which proceeds through non-open questions, i.e. questions which do not offer users the initiative, that are only interrupted by user requests for information probably also will influence user utterance length in the right direction. In addition, the system will admonish users to express themselves briefly in order to be understood by it and to ask only one question at a time.

The system will not be able to handle arbitrary requests for information at arbitrary points during dialogue, because of the uncontrolled growth in the focus set this would entail. Users should not ask, e.g., about departure times when the system addresses the travel destination. For some sub-tasks, such as number and names of travellers, we would expect no questions at all. Users can be expected to know who is going to travel without

having to negotiate this with the system. For other sub-tasks, however, it may be highly relevant to ask for information. For instance, users often do not know the precise departure and arrival times and must be informed on these by the system. Or users will want to know about reduced-fare departures before committing to a specific departure time or even departure date. We assume that it will be possible to 'cluster' such dependencies between system questions and relevant user questions in such a way that the system focus set can still be kept limited.

When requests for information are allowed during reservation dialogue, more than one type of user dialogue act is allowed as well. The important point is that, due to the informed reservation task context, only two different basic types of dialogue act seem relevant and must be distinguished by the system, i.e. reservation commitments and requests for information (or questions). The system must be capable of detecting, on the one hand, when the user wants information and, on the other, when the user provides a piece of information which should fill a slot in the reservation record. Two general cases may be distinguished.

In the first general case, the contents of a user utterance *cannot* be used to fill a slot in the reservation record, e.g.:

S: When would you like to leave?
U: Which flights are there on Friday night?
S: On Friday night there is a flight at 19:30 and another one at 21:30. Would you like one of these?

or the contents only provide *partial* information for a slot, e.g.:

S: When would you like to leave?
U: On Friday night.
S: On Friday night there is a flight at 19:30 and another one at 21:30. Would you like one of these?

In such cases, the system should treat the user utterance as a request for information no matter whether it is phrased as a question or not. This rule actually solves the problem that some requests for information may be hard to detect because their status as questions is mainly expressed through intonation. Intonation has not yet been exploited in realistic SLDSs although this possibility is the subject of ongoing research [Buchberger, this volume].

In the second general case, the user's utterance provides information that *could* fill a slot in the reservation record. In this case there are two possibilities. The obvious possibility is that the utterance is intended to fill a slot in the reservation record, e.g.:

S: When would you like to leave?
U: On Saturday at 8:15.

However, the utterance might ask for information instead, such as the following:

S: When would you like to leave?
U: Is there a plane on Saturday at 8:15?

The identification of the user's intended action is in such cases essential to the successful completion of the reservation task. We carefully analysed the corpus to see how utterances belonging to each of the two sub-groups of the second general case could be expressed.

In terms of Speech Act Theory [Searle 1969], we basically only found two categories of speech act, namely directives and assertives. Although there are surface speech acts that do not belong to any of these two categories, the indirect speech act expressed in such utterances will belong to one of the two categories. For example, in our corpus we found the surface expressive:

U: I would like to know which flights there are on Friday evening, tonight.

Although this utterance is a surface expressive speech act, it can only function as an indirect request for information (a directive) in the specific context of the informed reservation dialogue. In other situations, for example, in communication with friends none of whom is an expert in this field, and when the purpose of communication is not information exchange but social conversation, the same utterance could hardly be a request for information. It would remain a clear statement. In our small corpus we did not find any indirect expressives, commissives or declaratives.

We found two actions which are more related to the flow of conversation than to the successful completion of the task:

The first type of action was the phatic action which shows that the user understood the information which was provided or agrees with what was said. This action was expressed with minimal expressions like "OK", "hm-hm", "yea" etc., expressions that could be characterised as assertives. We did not analyse such actions in detail since they are not necessary for successful task completion, and if the system does not listen while it talks as is the case for P2, such user speech acts do not play any role at all.

The second type of action makes comments on the piece of information provided by the system. Such comments typically function as an introduction to the user's next move which is to either select a usable piece of information from what was provided by the system, or make a new request for information, for example:

S: When will you return?
U: Are there any flights on Sunday afternoon?
S: There are flights at 15:00, 16:00 and 18:00.
U: Oh, that sounds good. Can I have the 18:00 flight?

or

U: When is the first flight on Saturday morning?
S: The first flight on Saturday morning is at 9:30.
U: Hm, it's no good then. When is the last flight on Friday evening?

Although these comments are important to the explanation of the next user action they do not replace the actions that perform the exchange of information for the completion of the task. In the later dialogues which included more system restrictions and less user initiative, the number of such comments were much reduced.

The simulations focused on domain communication. Although meta-communication is important and should be allowed, we did not simulate misrecognitions but plan to return to this topic on a later occasion and make a more detailed study of meta-communication mechanisms.

3.1 Information

There are several ways in which users can utter a request for information or try to complete part of the reservation task. In particular, users made their requests for information in the following ways:

a) Direct polar questions:
Are there any discounts?
Is there any other flight after that, before noon?

b) Direct WH questions:
When is the earliest flight?
Which flights are there on Friday night?

c) Questions that refer to the ability of the system to provide information:
Could you tell me when is the last flight on Friday night?
Do you know the flights on Friday evening?

d) Statements that refer to the will of the user to get the information from the system:

I would like to know which flights there are on Friday evening, tonight.

e) Statements that show the user's attempt to make the system provide the information:
Tell me when is the first flight on Saturday morning.

f) Tag questions -which are interpreted as direct polar questions-:
It must be with this combination price, isn't it?

g) Intonation questions:
And on Saturday morning before 12?

All the above different ways (a-g) in which users expressed themselves to make a request for information, refer to the 'felicity conditions' of the speech act, i.e. the pre-conditions which are responsible for the characterisation of a speech act as a request for information.

In particular, the direct polar and WH questions express the propositional content condition of requests for information. They are the most typical and simple cases of requests for information and, when performed by users in the given context, can only be requests for information. In the same category we also put the tag questions, as well as the intonation questions. The questions about the ability of the system to provide a piece of information refer to one of the preparatory conditions, i.e. that the hearer (the system) is able to provide information, which is related to the expert role of the system. The statements which refer to the will of the user to obtain a piece of information from the system express the sincerity condition of requests for information, i.e. that the speaker wants to obtain information from the hearer. Finally, the statements that show the speaker's attempt to make the hearer provide a piece of information express the essential condition of requests for information, i.e. that the speaker's action is an attempt to make the hearer provide information.

Searle [1975] argues that the speaker can make an indirect directive (requests are directives) requesting or stating the propositional content condition, the preparatory condition that refers to the ability of the hearer, and the essential condition, as well as stating the sincerity condition. In our corpus we only found examples of some of Searle's groups: the questioning of the ability of the system (e.g.: "Could you tell me the flights on Friday evening?"), the stating of the sincerity condition (e.g.: "I would like to know which flights there are on Friday evening, tonight"), and finally, the essential condition of the requests for information (e.g.: "Tell me when is the first flight on Saturday morning").

The background context as defined by the specific task and the roles of the speakers (user/system) can also predict the possible and impossible utterances that will refer to the felicity conditions of the requests for information. For instance, it would not be logical and hence not co-operative that somebody would ask for information while at the same time questioning his/her will (the sincerity condition) and/or the nature of his/her act (i.e. the essential condition). Also, it would not be logical and hence not co-operative that the same person would start this specific communication if s/he questioned his/her belief in the abilities of the expert/system (one of the preparatory conditions). Moreover, it would seem redundant for the speaker to state either the system's ability or his/her belief in the system's ability, because these form the basis of the respective roles of user and system. However, a study of a large corpus is necessary in order to test the above predictions.

3.2 Reservation

In our corpus, the successful completion of a part of the reservation task was performed with the following three different types of formulation:

a) Statements:
S: Where are you going?
U: To Aarhus

b) Questions about the ability of the speaker to book a flight with a specific characteristic:
U: Could I have the Saturday morning flight, at 9:20?

c) Questions about the ability of the system to make a particular reservation:

U: Can you give me an earlier flight, between midnight and 9:00?

Statements (a) are the typical type of formulation of a simple answer. The other two types of formulation (b and c) refer to the felicity conditions of the request for reservation. In particular, the questions about the ability of the speaker to book a flight with a specific characteristic refer to one of the preparatory conditions of the request for reservation, and the questions about the ability of the system to make a particular reservation refer to another preparatory condition of the requests for information.

Summarising the above, it seems that we can determine and distinguish the speech acts which users can produce in this type of dialogue, just from the semantics of their sentences. Hence the system can use the semantics of the sentence to identify the cases in which it has to provide information. When the semantics of the sentence do not refer to the felicity conditions of requests for information, then the system may safely accept the utterance as a reservation commitment. In this situation, there will be no problem in identifying the following utterance as an answer and not as a request for information:

Could I have the Saturday morning flight at 9:20?

Speech act theory supports the theoretical justification of this determination, not only by its presentation of the felicity conditions of each speech act, but also by its description of the important roles of the general principles of co-operative conversation, the shared knowledge of the world and the background knowledge, as well as the ability of humans to make inferences.

4 Conclusion

We have preliminarily discussed how to make a controlled step in the direction of mixed initiative dialogue exemplified by the informed reservation task. Due to the nature of this task we only needed to distinguish two basic types of user dialogue act expressing either a request for information or a commitment to reservation. It remains to be seen if the linguistic mechanisms proposed above will be sufficient for enabling limited mixed-initiative dialogue.

Possible next steps would be to allow even more user initiative in domain communication, e.g. by letting the system ask more open questions, and to allow more flexible forms of meta-communication that are not based on keywords as in P2. The increased complexity of domain communication due to increased user initiative can be expected to concern focus and utterance length but not the basic distinction between two types of dialogue act. If, however, we allow non-keyword-based meta-communication we can no longer take all information-requesting dialogue acts to mean requests for information at the domain level.

Acknowledgements

The work described in this paper was carried out under a grant from the Danish Government's Informatics Research Programme whose support is gratefully acknowledged. The EU's Human Capital and Mobility Programme supported Dimitris Papazachariou's work at the Centre for Cognitive Science. Thanks are due to SRI and Candy Sidner for the permission to use the American Express Corpus.

References

[Bernsen et al. 1994] Bernsen, N.O., Dybkjær, L. and Dybkjær, H.: A Dedicated Task-Oriented Dialogue Theory in Support of Spoken Language Dialogue Systems Design. *Proceedings of the ICSLP Conference,* Yokohama, Japan, September 1994, 875-78.

[Bernsen et al. 1995a] Bernsen, N.O., Dybkjær, H. and Dybkjær, L.: Exploring the Limits of System-Directed Dialogue. Dialogue Evaluation of the Danish Dialogue System. *Proceedings of EUROSPEECH '95,* Madrid, September 1995.

[Dybkjær et al. 1995a] Dybkjær, L., Bernsen, N.O. and Dybkjær, H.: Different Spoken Language Dialogues for Different Tasks. A Task-Oriented Dialogue Theory. To be published in Human Comfort and Security, *Springer Research Report 1995* (in press).

[Dybkjær et al. 1995b] Dybkjær, L., Bernsen, N.O., Dybkjær, H. and Papazachariou, D.: Contextual Elements in the Danish Dialogue System. *Proceedings of the HCM Dialogue and Discourse Workshop*, Dublin, April 1, 1995.

[Grau et al. 1994] Grau, B., Sabah, G. and Vilnat, A.: Control in Man-Machine Dialogue. *THINK*, Vol. 3, Tilburg, The Netherlands, May 1994, 32-55.

[Gross et al. 1993] Gross, D., Allen, J.F. and Traum, D.R.: The Trains 91 Dialogues. TRAINS Technical Note 92-1. The University of Rochester, Computer Science Department, Rochester, New York 14627.

[Kanazawa et al. 1994] Kanazawa, H., Seto, S., Hashimoto, H., Shinchi, H. and Takebayashi, Y.: A user-initiated dialogue model and its implementation for spontaneous human-computer interaction. *Proceedings of the ICSLP 94,* Yokohama, September 1994, 111-114.

[Peckham 1993] Peckham, J.: A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Project. *Proceedings of EUROSPEECH '93*, Berlin, September 1993, 33-40.

[Searle 1969] Searle, J.R.: *Speech Acts*. Cambridge: Cambridge University Press, 1969.

[Searle 1975] Searle, J.R.: Indirect Speech Acts. In Cole, P. and Morgan, J.L. (eds.): Syntax and Semantics. *Volume 3: Speech Acts*. New York, Academic Press.

[Sidner 1992] Sidner, C.: The American Express Corpus. SRI.

[Smith 1991] Smith, R.W.: A Computational Model of Expectation-Driven Mixed-Initiative Dialogue Processing. Ph.D. Thesis, Department of Computer Science, Duke University, Durham, NC 27706, USA, October, 1991.

[Zoltan-Ford 1991] Zoltan-Ford, E.: How to get people to say and type what computers can understand. *International Journal on Man-Machine Studies,* Vol. 34, 1991, 527-547.