ARGOT: The Rochester Dialogue System

James F. Allen, Alan M. Frisch, and Diane J. Litman

Computer Science Department
The University of Rochester
Rochester, NY 14627

Abstract

We are engaged in a long-term research project that has the ultimate aim of describing a mechanism that can partake in an extended English dialogue on some reasonably well specified range of topics. This paper is a progress report on the project, called ARGOT. It outlines the system and describes recent results as well as work in progress.

1. Introduction

Consider Dialogue 1, a slightly cleaned up version of an actual dialogue between a computer operator and a user communicating via terminals.

(1)	User:	Could you mount a magtape for me?
(2)		It's T376.
(3)		No ring please.
(4)		Can you do it in five minutes?
(5)	Operator:	We are not allowed to mount that magtape.
(6)		You will have to talk to the head operator about it.
(7)	User:	How about tape T241?

Dialogue 1.

We are building a computer system called ARGOT that plays the role of the operator in extended dialogues such as the above. This dialogue illustrates some of the many issues that must be addressesed in building such a system. For instance, the first utterance taken literally is a query about the system's (i.e., the operator's) abilities. In this dialogue, however, the user intends it as part of a request to mount a particular magtape. Thus, the system must recognize an indirect speech act. Utterance (2) identifies the tape in question, and (3) and (4) add constraints on how the requested mounting is supposed to be done. These four utterances, taken as a unit, can be summarized as a single request to mount a particular magtape with no ring within five minutes.

Once the system makes the above inferences, it generates (5), which denies the request, as well as (6), which provides additional information that may be helpful to the user. The system believes that talking to the head operator will be of use to the user because it has recognized the user's goal of getting a tape mounted. Utterance (7) taken in isolation is meaningless; however, in the context of the entire dialogue, it can be seen as an attempt to modify the original request by respecifying the tape to be mounted.

Allen's [1979] model of language as cooperative behavior provides answers to several of the difficulties suggested by Dialogue 1. The basic assumption of that approach, which is adopted in ARGOT, is that the participants in a dialogue are conversing in order to achieve certain goals. As a consequence, a major part of understanding what someone said is recognizing what goals they are pursuing. In purposeful dialogues this model accounts for helpful responses, as well as for responses to indirect speech acts and some sentence fragments. However, since his model has no knowledge of discourse structure it cannot partake in an extended dialogue.

One of the major advances made in ARGOT is that it recognizes multiple goals underlying utterances. For example, consider the user's goals underlying utterance (2). From the point of view of the task domain, the user's goal is to get the tape mounted (by means of identifying it). From the point of view of the dialogue, the user's goal is to elaborate on a previous requests, i.e. the user is specifying the value of a parameter in the plan that was recognized from the first utterance. In the ARGOT system, we recognize both these goals and are investigating the relationship between them. The need for this type of analysis has been pointed out by many researchers (e.g., [Levy, 1979; Grosz, 1979; Appelt, 1981; and Johnson and Robertson, 1981]).

2. Organization of ARGOT

Currently, the ARGOT system is divided into many subsystems, each running concurrently. The three subsystems we shall consider in this paper are the *task goal* reasoner, the *communicative goal* reasoner, and the *linguistic* reasoner. Each of these levels is intended to perform both recognition and generation. In this paper we consider only recognition, since the generative side of the system is not currently being implemented.

The task goal reasoner recognizes goals in the domain of discourse, such as mounting tapes, reading files, etc. The communicative goal reasoner recognizes goals such as introducing a topic, clarifying or elaborating on a previous utterance, modifying the current topic, etc. Allen's earlier system had parts of both types of analysis but they collapsed into one level. A result of this was that it was difficult to incorporate knowledge of the dialogue structure into the analysis.

Splitting the analysis of intention into the communicative and task levels brings about the problem of identifying and relating the high-level goals of the plans at each level. The high-level goals at the task level are dependent on the domain, and correspond to the high-level goals in the earlier model. The high-level communicative goals reflect the structure of English dialogue and are used as input to the task level reasoner. In other words, these goals specify some operation (e.g., introduce goal, specify parameter) that indicates how the task level plan is to be manipulated. Our initial high-level communicative goals are based on the work of Mann, Moore and Levin [1977]. In their model, conversations are analyzed in terms of the ways in which language is used to achieve goals in the task domain. For example, bidding a goal is a communicative action which introduces a task goal for adoption by the hearer.

Given the communicative goals, we must now be able to recognize plans at this level. Neither Mann et al. [1977] nor Reichman [1978] have described in detail the process of recognizing the communicative goals from actual utterances. Currently, we adapt Allen's [1979] recognition algorithm, which finds an inference path connecting the observed linguistic action(s) to an expected communicative goal. This algorithm uses the representation of the utterance from the linguistic level and a set of possible communicative acts predicted by a dialogue grammar which indicates what communicative acts are allowed at any particular time for both participants, and is modeled after Horrigan [1977].

The work at SRI [Walker, 1978] in expert-apprentice dialogues monitored the goals of the user at the task level. The only analysis at the communicative goal level was implicit in various mechanisms such as the focusing of attention [Grosz, 1978]. Their work ties the task structure and communicative structure too closely together for our purposes. Appelt [1981] also views utterances as actions which satisfy goals along various explicit dimensions--a social dimension as well as what would correspond to our task and communicative levels. However, his communicative dimension is again mainly concerned with focusing.

The linguistic level is responsible for providing input to the other levels of analysis that reflects the content of the actual utterances. This parser will be based on the Word Expert Parser system of Small and Reiger [1981]. As the linguistic analysis progresses, it will notify the other levels of the various noun phrases that appear as they are analyzed. This allows the other levels to start analyzing the speaker's intentions before the entire sentence is linguistically analyzed. Thus, an interpretation may be found even if the linguistic analysis eventually "fails" to find a complete sentence. (Failure is not quite the correct word here, since if the utterance is understood, whether it was "correct" or not becomes uninteresting.) We are investigating other information that could be useful for the rest of the system during parsing; for instance, the recognition of *clue* words to the discourse structure [Reichman, 1978]. If a user utterance contains the word "please," the communicative level should be notified so that it can generate an expectation for a request.

In addition, the rest of the system may be able to provide strong enough expectations about the content of the utterance that the linguistic level is able to construct a plausible analysis of what was said, even for some ungrammatical sentences.

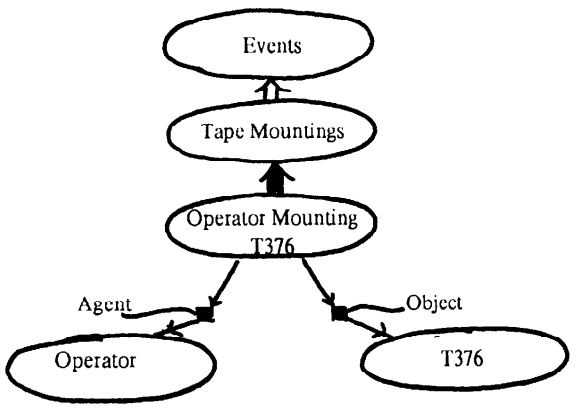
3. Issues in Knowledge Representation

All levels of analysis in ARGOT make use of a common knowledge representation and a common knowledge base module (KB). The KB stores a set of sentences of the representation language and provides retrieval facilities for accessing them. These retrieval facilities are used extensively at all levels of analysis. Because of this, the sentences stored in the KB not only represent the knowledge associated with magtapes and the doings of a computer room, but also the knowledge necessary to deal with language. This includes knowledge of physical, mental, and linguistic actions, how these actions are involved in plans, what computer users do, and what they expect of computer operators.

Following Brachman [1979b], our representation is constructed of two levels: the epistemological level and the conceptual level. The epistemological level provides a set of knowledge structuring primitives that are used to construct all conceptual level entities (e.g., action, time, and belief). Each level of the representation provides primitive symbols (particular predicate symbols, function symbols, and constant symbols) which can then be combined using the notation of FOPC. By inheriting the logical connectives and quantificational structure of FOPC, the resulting representation language is quite expressive.

3.1 The Epistemological Level of Representation

The epistemological level of the representation supplies a fixed set of predicates which are the knowledge-structuring primitives out of which all representations are built. The choice of knowledge-structuring primitives has been motivated by the study of semantic networks. For instance, where a semantic network such as Brachman [1979b] might have



we would have

SUBTYPE(Tape-Mountings,Events)
TYPE(Operator-Mounting-T376,Tape-Mountings)
ROLE(Operator-Mounting-T376,Agent,Operator)
ROLE(Operator-Mounting-T376,Object,T376)

Notice that the SUBTYPE predicate corresponds to the unshaded double arrow, the TYPE predicate to the shaded double arrow, and the ROLE predicate to the single arrow. Our constants are sorted into individuals (e.g.: Operator-Mounting-T376, Operator, T376), types (e.g.: Events, Tape-Mountings) and rolenames (e.g.: Agent, Object). These sorts somewhat correspond to the shaded oval, unshaded oval and shaded box of the network. Allen and Frisch [1982] have fully described and axiomatized the epistemological level of the representation and have compared it to semantic networks.

3.2 The Conceptual Level of Representation

The representation of actions is crucial to a dialogue participant for two reasons. The first is that the participant must be able to represent the meaning of utterances that refer to actions (e.g., "Can you mount a magtape for me?" refers to the action of mounting). The second, as previously discussed, is that it is advantageous to model the language comprehension and production processes as purposeful, planned action (e.g., uttering "Can you mount a magtape for me?" is a requesting action). However, existing models of action, most notably the state-space approach (e.g. [Fikes and Nilsson, 1971]), appear inadequate for the above purposes. Since a major deficiency with the existing models is an inadequate treatment of time, we first turn our attention to this issue.

An interval-based temporal logic and its associated inference processes have been defined [Allen, 1981a]. Rather than using a global time line, the representation employs a hierarchical set of reference frames. A particular interval is known by its location relative to the reference frames and other intervals. This is particularly important in a dialogue system for most temporal knowledge does not have a precise time. This representation of time has been used to produce a general model of events and actions [Allen, 1981b]. The occurence of an event corresponds to a partial description of the world over some time interval. Actions are defined as that subclass of events that are caused by agents. This is in contrast to the state-space view of an action as a function from one world state to a succeeding world state. Our approach enables the representation of actions that describe inactivity (e.g., standing still), preserving a state (e.g. preventing your television from being stolen), and simultaneous performance of simpler actions (e.g., talking while juggling).

Representing actions, particularly speech acts, requires the representation of beliefs. For example, the effect of the speech act of informing involves changing the beliefs of the hearer. A model of belief has been developed that treats BELIEVE as a predicate on an agent and a sentence of the representation language. To do this, there must be a name for every sentence in the language. Perlis [1981] and Haas [1982] have introduced naming schemes that provide enough expressiveness to deal with traditional representational requirements such as quantifying in. Haas [1982] has used this formulation of belief to predict an agent's action by constructing plans that can include mental actions. His treatment of belief and action does

not suffer from the problem of the possible worlds approach [Moore, 1979] that an agent believes all consequences of his beliefs.

3.3 The Knowledge Base Module

The Knowledge Base (KB) provides a set of retrieval facilities that is the sole access that the system has to the sentences stored in the KB. This retrieval facility corresponds to the matcher in a semantic network representation. Since retrieval must respect the semantics of the representation, it is viewed as inference. However, this inference must be limited because retrieval must terminate, and must do so in a reasonable amount of time. Frisch and Allen [1982] have shown how a limited inference engine suitable for knowledge retrieval can be given a formal, non-procedural specification in a meta-language and how such a specification can be efficiently implemented.

The capabilities and limitations of the retriever can be thought of intuitively as follows. A set of axioms dealing solely with the epistemological primitives is built into the retriever. For example, three of these axioms are:

$$\forall\ t_1,t_2,t_3\ \ \mathrm{SUBTYPE}(t_1,t_2) \wedge \mathrm{SUBTYPE}(t_2,t_3) \rightarrow \mathrm{SUBTYPE}(t_1,t_3)$$

(SUBTYPE is transitive.)

$$\forall\ o,t_1,t_2\ \ \mathrm{TYPE}(o,t_1) \wedge \mathrm{SUBTYPE}(t_1,t_2) \rightarrow \mathrm{TYPE}(o,t_2)$$

(Every member of a given type is a member of its supertypes.)

$$\forall\ x,r,y,y'\ \ \mathrm{ROLE}(x,r,y) \wedge \mathrm{ROLE}(x,r,y') \rightarrow y=y'$$

(Role fillers are unique)

Through these built-in axioms, the retriever "knows about" all of the episemological primitives. The retriever's power comes from the fact that it can, for the most part, reason completely with the built-in axioms. Its limitations arise because it only partially reasons with the sentences stored in the KB. The retriever also has knowledge of how to control inferences with the built-in axioms. In this manner, the retriever only performs those inferences for which it has adequate control knowledge to perform efficiently.

4. A Simple Example

Let us trace a simplified analysis of utterance (1), "Could you mount a magtape for me?" The communicative acts expected at the start of a dialogue by the grammar are (in an informal notation)

user BID-GOAL to system, and
user SUMMON system.

Taking the utterance literally, the linguistic level uses both syntactic and semantic analysis to identify the linguistic actions (speech acts) performed by the speaker. For utterance (1) we have

user REQUEST that
system INFORM user if system can mount a tape,

which is sent to the communicative level. The plan recognition algorithm produces BID-GOAL acts for two possible goals:

(G.1) system INFORM user if system can mount a tape *(literal interpretation)*

(G.2) system MOUNT a tape *(indirect interpretation)*.

The indirect interpretation, (G.2), is favored, illustrating how goal plausibility depends upon what the dialogue participants know and believe. Most people know that operators can mount tapes, so the literal interpretation is unlikely. However, if the user did not know this, the literal interpretation would also have been recognized (i.e., the system might generate "yes" before attempting to mount the tape). It is important to remember here that the plan was recognized starting from the literal interpretation of the utterance. The indirect interpretation falls out of the plan analysis (see [Perrault and Allen, 1980] for more details). Thus, the linguistic level only needs to produce a literal analysis.

The communicative level sends the recognized BID-GOAL, (G.2), to the task reasoner. There, the user's task level goal to mount a tape is recognized, and the system accepts the user's goal as a goal of its own. Of course, since the task level reasoner is a general plan recognizer as well, it may well infer beyond the immediate effect of the specific communicative action. For example, it may infer that the user has the higher-level goal of reading a file.

The task level reasoner generates a plan for mounting a tape and then inspects this plan for obstacles. Assuming the user says nothing further, there would be an obstacle in the task plan, for the system would not know which tape to mount. The task level reasoner would generate the goal for the system to identify the tape and would send this goal to the communicative goal reasoner. This reasoner would plan a speech act (or acts), obeying the

constraints on well-formed discourse, that could lead to accomplishing the goal of identifying the tape. This speech act then would be sent to the linguistic level which would generate a response such as "Which tape?"

In Dialogue 1, however, the user identifies the tape in utterance (2), which the communicative level recognizes as a SPECIFY-PARAMETER action for the plan created by the initial BID-GOAL action.

5. Current State and Future Directions

We have implemented the knowledge base, a simple dialogue grammar, and simple plan recognizers at both the communicative and task levels. Furthermore, we are currently incorporating a word-expert parser [Small and Rieger, 1981]. As discussed in the previous sections, further research on all aspects of ARGOT (i.e. the levels, the interactions between them, and the theoretical models) is still needed.

Acknowledgements

We would like to thank Bill Mann for providing us with the dialogues. We thank Lokendra Shastri and Marc Valain for their contributions to the development of ARGOT and Dan Russell for his helpful comments on an earlier version of this paper. This work has been supported in part by NSF Grant IST-8012418 and DARPA Grant N00014-82-K-0193.

References

Allen, J.F., "A plan-based approach to speech act recognition," Ph.D. thesis, Computer Science Dept., U. Toronto, 1979.

Allen, J.F., "An interval-based representation of temporal knowledge," *Proc.*, 7th Int'l. Joint Conf. on Artificial Intelligence, Vancouver, B.C., 1981a.

Allen, J.F., "What's necessary to hide?: Reasoning about action verbs," *Proc.*, 19th Annual Meeting, Assoc. for Computational Linguistics, 77-81, Stanford U., 1981b.

Allen, J.F. and A.M. Frisch, "What's in a semantic network?" *Proc., 20th Annual Meeting, Assoc. for Computational Linguistics*, U. Toronto, June, 1982.

Appelt, D., "Planning natural language utterances to satisfy multiple goals," Ph.D. thesis, Computer Science Dept., Stanford U., 1981.

Brachman, R.J., "On the epistemological status of semantic networks," in N.V. Findler (Ed.). *Associative Networks.* New York: Academic Press, 1979b.

Fikes, R.E. and N.J. Nilsson, "STRIPS: A new approach to the application of theorem proving to problem solving," *Artificial Intelligence, 2*, 189-205, 1971.

Frisch, A.M. and J.F. Allen, "Knowledge retrieval as limited inference," *Lecture Notes on Computer Science: 6th Conference on Automated Deduction Proceedings.* New York: Springer-Verlag, 1982.

Grosz, B.J., "Discourse knowledge," In [Walker, 1978].

Grosz, B.J., "Utterance and objective: issues in natural language communication," *Proc.*, 6th Int'l. Joint Conf. on Artificial Intelligence, Tokyo, 1979.

Haas, A.R., "Mental states and mental actions in planning," Ph.D. thesis, Computer Science Dept., U. Rochester, 1982.

Horrigan, M.K., "Modelling simple dialogues," *Proc.*, 5th Int'l. Joint Conf. on Artificial Intelligence, MIT, 1977.

Johnson, P.N. and S.P. Robertson, "MAGPIE: A goal-based model of conversation," Research Report #206, Computer Science Dept., Yale U., May 1981.

Levy, D., "Communicative goals and strategies: between discourse and syntax," in T. Givon (ed.). *Syntax and Semantics*, Vol. 12. New York: Academic Press, 1979.

Mann, W.C., J.A. Moore, and J.A. Levin, "A comprehension model for human dialogue," *Proc.*, 5th Int'l. Joint Conf. on Artificial Intelligence, MIT, 1977.

Moore, R.C., "Reasoning about knowledge and action," Ph.D. thesis, MIT, 1979.

Perlis D., "Language, computation, and reality," Ph.D. thesis, Computer Science Dept., U. Rochester, 1981.

Perrault, C.R. and J.F. Allen, "A plan-based analysis of indirect speech acts," *J. Assoc. Comp'l. Linguistics 6*, 3, 1980.

Reichman, R., "Conversational coherency," *Cognitive Science 2*, 1978.

Small, S.L. and C. Rieger, "Parsing and comprehending with word experts (a theory and its realization)", TR 1039, Dept. Computer Science, U. Maryland, 1981.

Walker, D.E. *Understanding Spoken Language.* New York: North-Holland, 1978.