Empirical analysis of a discourse model for natural language interfaces

Nils Dahlbäck
Department of Computer and Information Science
Linköping University
S-581 83 Linköping, Sweden
nda@ida.liu.se

Abstract

A structural model of discourse for natural language interaction developed for the LINLIN-system is evaluated using the Wizard-of-Oz metod. 21 dialogues were collected using five different background systems, making it possible to vary the type and number of tasks possible to perform by the users. The results indicate that the structural complexity of the discourse in man-machine dialogues is simpler than most human dialogues, at least for information retrieval and some types of ordering systems, suggesting that computationally simpler discourse models can be used in these domains.

Introduction

The aim of the LINLIN-project is to develop a natural language interface for Swedish. One important characteristic of the project is that it is based on a sublanguage approach (Grishman & Kittredge, 1987). The system is viewed as a shell that needs to be customized to different applications. The general parts are the different processing modules, whereas the static knowledge bases used by the processing modules need to be adapted to the different applications. The purpose of the present paper is to describe the empirical foundations of the project's approach to discourse structure. The linguistic and computational aspects of the project have been described elsewhere (Ahrenberg, 1989, Ahrenberg, Jönsson & Dahlbäck 1990, Jönsson 1991). Other aspects of the empirical work are presented in Jönsson & Dahlbäck (1988), Dahlbäck & Jönsson (1989).

The LINLIN-approach to discourse differs in some important respects from the 'standard model' of discourse within computational linguistics (Grosz&Sidner 1986). Our intention is not to create a general discourse theory. It is only concerned with man-machine interaction in natural language; there is no claim that it is applicable to all types of discourse, or even all types of dialogue. In a sense, computational efficiency within a limited domain is preferred to generality with unknown properties. It is thus a computational theory in the sense that its focus is on interaction with computers, but also in the sense that it is implemented (albeit not as yet integrated with the other modules in the system) (Jönsson, 1991).

This work was supported by the Swedish Council for Research in the Humanities and Social Sciences (HSFR).

The Dialogue Model

The dialogue model of the LINLIN-system (called LINDA, for LINköping DiAlogue) is an abstract characterization of the dialogue structure in NLI-dialogues. In its pure form, the model is only about the transfer of content between interlocutors in so-called 'transactional dialogues' (Brown & Yule, 1983, p 1). I will first describe this pure model, and then return to the necessary amendments. The description below is based on Ahrenberg, Jönsson & Dahlbäck (1990).

The dialogue objects are divided into three main classes on the basis of structural complexity. There is one class corresponding to the size of the dialogue, one corresponding to the size of a discourse segment, and a third corresponding to the size of a single speech act or dialogue move. The dialogue is thus structured in terms of discourse segments, and discourse segments in terms of moves and embedded segments. Utterances are not analyzed as dialogue objects, but as linguistic objects which function as vehicles of one or more moves.

The basic discourse segment unit is an initiative-response-unit (*IR-unit*). i.e. an initiative (which opens a segment) and a response (which closes the segment). Initiatives are basically directives in standard speech-act terminology. An initiative can be seen as an act where the speaker introduces a goal and the expectation that it will be satisfied. A response is an act where the goal is satisfied, i.e. the value is given, or the expectation is eliminated by other means. The prototypical case of this is where a question is followed by a content answer or saying that the speaker does not know the answer (elimination of expectation by other means). The response closing the segment need not be adjacent to the initiative, since interruptions and clarification sub-dialogues may intervene, but it must follow at some point. The IR-units are

then sub-classified into different types, such as Q/A (question-answer) *do you have any books on AI? / No*, Q/AS (question-assertion) *who is the author of Winnie-the-Pooh? / PUB has no information about that book*, D/ACK (directive-acknowledgement) *Luxor Symphony ordered / OK*, etc.[1]

The basic categories in the dialogues are turn and move. A turn is a continuous period of the dialogue where one participant "speaks". A turn is co-extensive with an utterance by user or system, since mid-turn interruptions cannot occur. Each turn consists of one or more moves. A move consists of one (speech) act.

The entire dialogue is thus structured by the recursive use of IR-units, which give the dialogue a tree structure. In contradistinction to some other models intended for computational use (e.g. Wachtel, 1986), there is no intermediate topic-based descriptive level. Instead a number of topical parameters are distinguished, e.g. CurrentObject, CurrentAttribute, and CurrentSet, which may vary independently of each other. The managemet of topic and topic structure is thus more elaborate than in the Grosz and Sidner (1986) model. A basic distinction is made between different topic domains, e.g. task-related (T) and discourse related (D) moves and IR-units. Other topic domains are also used, but the types and numbers vary between different dialogue domains.

It should be noted that Levinson (1981) has convincingly argued against the possibility of basing general structural models of dialogues on speech act categories and an adjacency pairs structure, i.e. that there can be no syntax of conversational structure. Two important parts of the argument is that there are no means of determining what counts as what in a dialogue independently of the situation in which it takes place, and that different speech-acts within an utterance need not be separated into different sub-parts, but that the same part of an utterance can perform more than one such act. A rather strong assumption of the LINDA model is that this need not be true for man-computer dialogues.

Empirical analysis

The LINDA-model has some obvious computational advantages, and some obvious deficiencies as a general computational model of discourse. The question, then, is to what extent the LINDA model is sufficient for this application domain in its entirety, or to some subsets thereof. Since we lack an adequate sub-language description of this domain[2], this question can only be answered by an empirical analysis, in the present case we have used so-called Wizard-of-Oz experiments. (For a review, see Jönsson & Dahlbäck (1988), and Fraser and Gilbert (in press))

The basic idea is simple. We have our subjects sit down at a terminal to solve some problem which requires them to use the computer. But instead of interacting with the system through our "newly developed natural language interface", we have one person sitting at the other end of the line, simulating such a device.

The idea might be simple. But to actually run such an experiment is not quite that simple, since you want the output from the simulated 'system' to be everything that the person simulating it is not — providing quick, consistent answers without spelling mistakes and other signs of human fallibility. To achive this, we use a simulation environment with pre-stored answers and other help systems. We consider the quality of the simulation environment a crucial factor behind the ecological validity of the dialogues collected and the concomitant applicability of conclusions obtained from analysing them. It can also be noted that researchers not having such a method at their disposal often comment on the need for it when discussing their work (e.g. Mc Kevitt, 1990). None of our subjects have realized that the 'system' was not a system.

Background systems

To circumvent the risk of drawing general conclusions that in fact are only a reflection of the specific experimental setting used, we have used five different background systems. We have varied not only the content domain, but also the 'intelligence' of the systems, and the number and types of tasks possible to perform by the user.

We have used two database systems. *PUB* is a library DB in use at our department. *C-line* is a simulated DB containing information about the computer science curriculum at Linköping University.

In the *HiFi*-system the user can order HiFi-equipment after having queried a (simulated) DB containing information about the available equipment. The *Travel* system simulates an automated travel agency offering charter holidays to Greek islands. These systems differ from the two above in two respects; the system is more 'cognitively' advanced, and there are more actions that can be performed by the user, i.e. not only asking for information but also order something.

The *Wine* system is a simulated advisory system, capable of suggesting suitable wines for different dishes, if necessary within a specific price range.

The total number of dialogues is 21; PUB: 4, HiFi: 5, C-line: 5, Wine: 4, Travel: 3. The total number of utterances is 1055.[3] This gives us an average of 50 utterances/dialogue. The longest are in the travel

[1] For at least some spoken dialogues, the minimal segment seems to consist of three units; (I)nitiative, (R)esponse, (C)onfirmation, cf Stubbs, 1983, Bilange, 1991.

[2] For a theoretical analysis of the differences between human dialogues and NLI dialogues, see Jönsson&Dahlbäck (1988), Dahlbäck (1991). An experimental investigation of this is currently in progress at our lab.

[3] Apart from the dialoges analyzed here, we have collected more than 60 others, using four other real or simu-

domain, where the average dialogue is 92 utterances long, and the shortest are the PUB dialogues with an average of 25 utterances. (The dialogues are available on file or in print on request. All are in Swedish.)

Analysis criteria

Our analysis is based on the same IR categories that are used in the LINLIN system. One important reason for this is that the categories are domain independent. We can therefore compare dialogues from different domains. Another advantage is that the categories are (fairly) simple to define and identify, making it possible to code the dialogues with sufficient inter-rater reliability (97.1 %). The computational advantage of not having to infer the invisible intentions behind an utterance in order to structure the dialogue is thus parallelled by a similar advantage from an analysis point of view.

We use only two basic types of moves, initiatives (I) and responses (R). The definiton of the categories is based on local information only. If the move is seen as introducing a goal it is scored as an initiative, if it is a goal-satisfying move, it is scored as a response. In the analysis of the dialogues we do not take into consideration the fact that in the LINDA model only segments of the I-R type are allowed, but not for instance I-R-R. It then becomes an empirical question, which we will return to, as to what extent the dialogues follow the structural constraints postulated by the model, e.g. its tree structure.

In most cases the I and R categories of the LINDA model can be used in a fairly straightforward way. There are however some cases requiring further consideration. One case is where the system cannot answer the question and replies with phrases such as "Don't understand", or "What do you mean by X?". We code the first case, where the utterance concerns the entire phrase, as a response, and the latter as an initiative. Thus we only identify subdialogues of clarification if part of the previous input is understood by the system. When the system replies that it cannot provide an answer because the parameter or value can not be found in the background system,we code this as a response, and not an initiative for clarification, even if the system supplies information about what value or field was missing.

To illustrate these definitions, consider the dialogue below (PUB1:17-20). Note that the second move spans two sentences. The dialogue is scored I, I, R, R. The reason for not scoring the first part of the second turn as a response, is that it does not in a strict sense answer the question about *where* these books can be found. Nor is it a response giving reasons for why an answer cannot be provided, but rather a question about how the answer should be presented to the user, or if any further processing should be done before presenting it.

lated background systems. Dahlbäck (1991) describes some of these in more detail.

```
U> Where can I find books about Artificial
Intelligence?
S> There are five books on Artificial
 Intelligence. Do you want me
 to find out where all are?
U> Yes
S> (LIST OF BOOKS)
```

Discourse management moves such as "Can I help you with anything more?" are all scored as initiatives. We subcategorize them as DO (discourse opening), DC (discourse continuation), and DE (discourse ending), to make it possible to exclude them from some of the analysis presented below. (Responses to these kind of initiatives are optional in the model). Only one type of move cannot be coded as initiative or response, namely acknowledgements of input (AI) where the system replies "PUB searching" or something similar.

Results

The results of the analysis will be presented in two main parts. This section gives an overview of the discourse in the different domains, and at the same time discusses the applicability of the LINDA model to them. The next section raises some general issues emerging from the analysis.

Database dialogues

The PUB dialogues can be characterized as a sequence of isolated question-answer sequences with no global structure. A large number of the utterances do not, however, follow the simple pattern of I followed by R. The reason for this is that the system, when asked about e.g. books about a specific topic, often does not print all items that have been found, but instead tells how many books have been found and asks if the user wants to see all of them, as in the example above.

These dialogues all fit the LINDA model very well. In fact, they all consist of sequences of only three types; (I R), (I (I R) R), or (I (I R) (I R) R), with only one instance of the last type. Almost all of the embedded segments concern the presentation of the information found by the system. The sub-dialogues never go more than one level down, the longest being six moves long.

The C-line dialogues, also database dialogues, have a clear and simple IR structure. Apart from the ending of the dialogues, where the system does not always send some kind of good-bye response when the user declares he has finished, there are only three cases where the dialogues deviate from being a sequence of pure IR-units with no embeddings. One occurs because of a mistake by the Wizard (or malfunctioning of the system). The other two occur because of requests of clarification from the system. In one of these user supplies the requested information, in the other the user rephrases the entire question, and thus aborts the current IR-segment and creates a new one. The sub-dialogues are never more than two moves long, and their maximum depth is consequently one level down.

Summary of the database dialogues Both sets of database dialogues show a simple and clear discourse structure, characterized by only local and no global coherence. Or, if you so wish, there is one global structure, but no intermediate structures, something which was also found in Grosz' (1977) analysis of her DB-dialogues. The sub-dialogues are of a simple character, being concerned either with requests for clarification or questions of how the result should be presented on the screen. These sub-dialogues are the only cases where the system takes over the initiative; for the major part the dialogues are user driven.

Ordering dialogues

When we come to the dialogues where the user can place an order, HiFi and Travel, the situation is somewhat more complicated. The main reason for this is that the user and the system in these cases have more tasks to perform; seek information, order, and tell where to send the ordered goods. As a consequence there are more kinds of topics to consider.

The HiFi dialogues all have a simple discourse structure that can easily be accommodated to the structural constraints of the LINDA model. In most dialogues the user discusses the available equipment of one kind, e.g. receivers, and thereafter orders one of these before continuing to the next kind of equipment. In all cases (but one) the move where the user orders something is immediately followed by an explicit confirmation of the order, either by telling which item that has been ordered, or simply by a confirmatory OK. The vast majority of dialogue segments consist of simple IR-units with no embeddings. There are in fact only two embeddings in the entire HiFi corpus, both caused by requests for clarification, making the sub-dialogues in HiFi as simple as in the previously discussed systems. Apart from the initiatives sub-classified DO, DC, and DE, there are only three moves which do not fit into the structural constraints of the LINDA model, in a total of more than 300 utterances and even more moves.

There is in fact one exception to the discourse pattern described here, namely HiFi1. The strategy of this user differs from the others as she uses a kind of "zooming in" strategy. Instead of choosing one item at the time, she asks initially for a set of models suitable for a small room. This set turns out to be too expensive, and the rest of the dialogue concerns which items to change to get an optimal set given the stipulated cost constraints. When she is satisfied, she declares this explicitly "this is what I want to order" (HiFi1:27), followed by the system's confirmation of the order, after which the sytem takes the initiative to obtain information about the delivery in the usual way.

This stategy is also seen in all the Travel dialogues. They contain a number of utterances where the user niether seeks information, nor orders something, but instead declares some intents or preferences. These moves have an important communicative function, since they constrain the set of alternatives to discuss further, but also the set of possible referents for referring expressions. After having declared the intent to go to Crete, a question like "are there any cheaper hotels?" is actually about hotels on Crete, not about hotels in Greece. Because of this gradual 'zooming in' on the order, there are no moves in the dialogue where the system explicily acknowledges what has been decided upon or ordered so far.

The way these moves are categorized and analyzed obviously affects the structural analysis of these dialogues. It is in fact possible to make the three dialogues all fit the structural constraints of the LINDA model. The way this can be done is to be somewhat 'liberal' in the assignement of responses to the initiatives where the user declares his intent. If we accept such an analysis strategy, there is actually only one case of an incomplete IR-unit in the entire corpus of Travel dialogues comprising some 300 moves, and this occurs because of a simulation mistake by the Wizard.

The problem with this analysis is that it fails to capture that which makes these dialogues differ from the database and HiFi dialogues (except for HiFi1). The problem of how to handle the confirmations of the orders or the bookings is from a system development point of view perhaps no large obstacle, since it concerns what the system does, and therefore its interpretation seems not to be all that critical. From an analysis point of view, the problem is that we seem to touch the limits of the model used.

Because of the difficulties with the structural analysis, it is somewhat difficult to assess the length and depth of the sub-dialogues here. If we restrict ourselves to the clear-cut cases, we find here such insertion sequences with a maximum length of 8 utterances and depth of three levels.

Summary of the ordering dialogues Perhaps the most interesting result emerging from the analysis of these dialogues is that, despite being conducted in a more complex situation than the pure database dialogues, they have a surprisingly simple discourse structure. The number of sub-dialogues are few and they are all rather short. There is one case going three levels down, but most of them are only one level deep.

The distribution of initiatives differs between the two dialogue domains. The HiFi dialogues are almost exclusively user-driven, whereas the opposite is more accurate for the Travel dialogues, though not to the same high extent. The important consequence of this is that there does not seem to be any simple connection between who has the initiative and the type of discourse strategy, since we find the 'zoom' strategy both in user-driven and system-driven dialogues.

Advisory systems

The Wine system is the only advisory system in our corpus of dialogues. This means that we have some difficulties in discerning how much of the patterns found

in the present corpus are reflections of the class of system it represents and how much it reflects the specific domain used.

These dialogues are by far the most difficult to analyze using our IR dialogue model. There are many reasons for this, on many different levels. Since the domain for these dialogues is not as clearly circumscribed as for the other systems under study, both as regards the competence of the system and as regards the mutual knowledge of the interlocutors, the users sometimes find it difficult to understand the limits of the linguistic and conceptual competence of the system.

Another important reason for the difficulties we encounter when analyzing the dialogues is that the dialogues have a mixed initiative structure. We saw previously that mixed initiative dialogues are not necessarily difficult to analyze using our model. The structures we find in the Wine dialogues are, however, of a different kind. They are not sub-dialogues of clarification, but rather of development of the topic discussed. This makes some utterances difficult to classify as consisting of separate moves consisting of pure initiatives or responses. There are 17 instances of the type of cases discussed here in the Wine dialogues. In other words they are quite frequent, covering approximately 40 % of the entire Wine corpus.

There is also another kind of move occurring in these dialogues that makes their structure deviate from some simple and 'pure' IR-pattern. This is when the system solicits information not requested by the user, and which shapes the continuation of the dialogue without being initiatives that makes the partner issue a response. These cases are the only ones in our dialogues where our simplified classificatory scheme is insufficient. There are five such utterances in the Wine dialogues. The surprising thing is not that we find these kind of sequences in our corpus, but rather that they are so few, and that they only occur in the Wine dialogues. In ordinary human discourse they are probably the rule rather than the exception

Using the criteria described above for the structural analysis, we find in this system sub-dialogues of a complexity similar to the one found in the Travel dialogues. The maximum length being 8 moves and the maximum depth three levels.

Evaluating the LINDA model

In its most simple form the dialogue structure is a sequence of IR-segments. It is one remarkable result of our analysis that such large parts of our dialogues actually fit this simple structure. One way of expressing this is to compute how much of the different dialogues that fit this pattern, using some simple complexity index. To do this we first exclude all initiatives that are coded DI, DC, or DE, and also all IA-moves, as the structure of the dialogues at these points is more a reflection of the design of the simulated system than anything else. We then count the number of times that the dialogue adheres to the prototypical adjacency-pair sequence of initiative followed by a response, divided by the number of pairs of moves(i.e. moves divided by two).

This complexity-index, which shows the extent to which the dialogues conform to the basic adjacency-pair sequence of an initiative followed be a response, gives the following figures: PUB 75.0; C-line 95.7; HiFi 98.0; Travel 87.5; Wine 78.1.

We can see that 75% or more of the dialogues consist of sequences of simple IR-units, a surprisingly high figure. This does not mean that the dialogues consist of a sequence of isolated questions and answers, as there is frequent use of anaphoric expressions. In fact 49 % of the *initiatives* contain some kind of anaphoric expression (Dahlbäck & Jönsson, 1989).[4] What the figures show is rather that in spite of being clear cases of connected discourse, these dialogues have a much simpler structural complextity than most other genres. It thus seems as if most man-machine dialogues in natural language, even when no restrictions on the users' way of expressing themselves, lack most of the complexity found in other types of discourse. Our corpus is admittedly of a limited size, but it covers some of the most typical possible applications for NLI technology, and is not tied to one particular topic domain. This gives us some confidence in believing that the results have some generalizability.

75 % is not 100 %, and a system that fails in one case out of four is obviously not sufficient. But the model is not as simple as this; recursive embedding of IR units is one of its central features. The low index figure for the PUB dialogues is due to the large frequency of simple clarification sub-dialogues. The index is thus misleading for this group, which means that for three of the dialogue systems the LINDA model can probably be used here without amendments.

The fit to the model, even when insertion sequences are taken into consideration, is not perfect even in these three sets of dialogues, but the causes for this are all rather simple. First, we have the opening and closings of the dialogues, but here we believe that careful design of these aspects of the dialogue will eliminate most of those cases. The other case of deviation from the pattern of strict and embedded IR-sequences is when the system initiates a clarification sub-dialogue which leads the user to abort this sequence and either rephrasing the question or even bringing up a new issue, leading to a ....R)(I (I))(I... sequence. What is interesting about these cases is that they seem to be more frequent in man-computer dialogues than in dialogues between people, the reason obviously being that there is no need to be polite to a computer!

If we expand the LINDA model to handle these cases

[4] Work is currently in progress on the relationship between the IR-structure and the use of anaphoric expressions in these dialogues.

too, we claim that our results support the notion that such a model can be used in an NLI for database applications.

When we come to the class of ordering system, the issue gets a bit more complicated. The results from the HIFi system seem to indicate that the model can also accommodate these kinds of dialogues, but when we come to the Travel system we encountered some difficulties. The sample is too limited to make it possible to draw any definite conclusions on this issue, but as a working hypothesis we suggest that it is possible to allow for more than one task to be performed by the system, given that it is clear to which task each move belongs. From a system development point of view, there is another possible solution, namely to physically separate the two tasks into two windows on the screen. (These isssues are discussed further in Dahlbäck 1991)

So far we have not found any cases that cannot be handled by a slightly modified LINDA model. But in the advisory Wine dialogues we seem to touch the borders of the model, and perhaps even cross them. There are a number of interrelated issues here. Space limitations prohibit a detailed discussion of them here, but not mentioning them would make us paint too bright a picture of the model discussed.

Structurally, we found some cases of IRR-sequences and other deviations from the basic IR-structure. Expanding the model to allow for this kind of structure is perhaps possible to do and still preserve its basic assumptions. What is more troublesome is that we found some cases where it was difficult to classify an utterance as a clear case of an initiative or a response, it was rather a little of both. This is of course something quite natural in most dialogues, but it creates problems for the LINDA model. The difficulties above can be seen as illustrating the limitations of the LINDA model in its pure form. It was never meant to be a general discourse model, so the fact that it cannot handle some discourse types is not critical. But the interesting point is then in a sense not that cases of these kinds are found in our corpus, but rather that they are so few and confined to one type of dialogue.

Summing up

The most important result of the present study is the demonstration that man-machine dialogues in natural language do not show the structural complexity often found in human dialogues. A caveat is that a sublanguage approach to this domain suggests that the successful management of connected discorse in natural language interfaces need not be that far off, at least for some of the most important application domains.

Acknowledgements

The discourse manager was developed by Lars Ahrenberg and Arne Jönsson, and they also contributed to the conceptual and empirical analysis presented here.

References

Ahrenberg, L. 1989. On the integration of linguistic knowledge and world knowledge in natural language understanding. In Dahl&K. Fraurud (eds) *Papers from the First Nordic Conference on Text Comprehension in Man and Machine.* Institute of Linguistics, Stockholm University.

Ahrenberg, L. Jönsson, A. & Dahlbäck, N. 1990 Discourse Representation and Discourse Management for a Natural Language System. In Dahl&K. Fraurud (eds) *Papers from the Second Nordic Conference on Text Comprehension in Man and Machine.* Institute of Linguistics, Stockholm University.

Bilange, E. 1991 A Task Independent Dialogue Model. *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics (E-ACL '89)*, Berlin, Germany.

Brown, G. & Yule, G. 1983 *Discourse Analysis.* Cambridge: Cambridge University Press.

Dahlbäck, N. 1991 *Representations of Discourse; Cognitive and Computational aspects.* Forthcoming.

Dahlbäck, N. & Jönsson, A. 1989 Empirical Studies of Discourse Representation for Natural Language Interfaces. *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics (E-ACL '89)* Manchester, England.

Fraser, N. & Gilbert, N. (in press) Simulating Speech Systems. *Computer Speech and Language*

Grishman, R & Kittredge, R. (eds) 1987 *Analyzing Language in Restricted Domains* Hillsdale, N.J.: Erlbaum.

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

Jönsson, A. 1991 A Dialogue Manager Using Intiative-Response Units and Distibuted Control. *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics (E-ACL '89)*, Berlin, Germany.

Jönsson, A & Dahlbäck, N. 1988 Talking to a computer is not like talking to your best friend. *Proc. of the 1st Scandinavian Conference on Artificial Intelligence (SCAI' 88)* Tromsö, Norway.

Levinson, S. 1981 Some pre-observations on the modeling of dialogue. *Discourse Processes* 4, 93-116.

Mc Kevitt, P. 1990 Data acquisiton for natural language interfaces. Memoranda in Computer and Cognitive Science MCCS-90-178, New Mexico State University.

Stubbs, M. 1983. *Discourse Analysis.* Oxford: Blackwell.

Wachtel, T. 1986 Pragmatic sensitivity in NL interfaces and the structure of conversation. In *Proc. of the 11th International conference of Computational Linguistics* University of Bonn.