Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

The dialog component in the Waxholm system

Carlson, R.

journal: TMH-QPSR
volume: 37
number: 2
year: 1996
pages: 109-112

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

The dialog component in the Waxholm system

Rolf Carlson
Department of Speech, Music and Hearing, KTH

Abstract

In this paper we will give a short overview of the dialog component in the Waxholm spoken dialog system. Dialog management based on grammar rules and lexical semantic features is implemented in our parser, STINA. The notation to describe the syntactic rules has been expanded to cover some of our special needs to model the dialog. The parser is running with two different time scales corresponding to the words in each utterance and to the turns in the dialog. Topic selection is accomplished based on probabilities calculated from user initiatives. Results from parser performance and topic prediction are included in the presentation.

Introduction

Our research group at KTH[1] has, for some years, been building a generic system in which speech synthesis and speech recognition can be studied in a man-machine dialog framework. The demonstrator application, Waxholm, gives information on boat traffic in the Stockholm archipelago. It references time tables for a fleet of some twenty boats from the Waxholm company which connects about two hundred ports. The system has been presented on several occasions, for example, the Eurospeech '93 conference (Blomberg et al., 1993), the ARPA meeting '94 (Carlson, 1994) and the ETRW on Spoken Dialog Systems (Bertenstam et al., 1995a).

The application has great similarities to the ATIS domain within the ARPA community, the Voyager system from MIT (Glass et al., 1994) and similar tasks in Europe, for example SUNDIAL (Peckham, 1993), the systems for train timetables information developed by Philips (Aust et al., 1994; Oerder and Aust, 1994) and CSELT (Clementino, and Fissore, 1993; Gerbino E and Danieli M, 1993) and flight information in the Danish Dialog Project (Dalsgaard and Baekgaard, 1994).

Spoken dialog management has attracted considerable interest during the last years. Special workshops and symposia, for example the special workshop at Waseda University, Japan 1993 (Shirai and Furui, 1995), the AAAI 1995 Spring Symposium: Empirical methods in discourse interpretation and generation, Stanford University, USA and the 1995 ESCA workshop on Spoken Dialog Systems in Vigsø, Denmark, have all been arranged to forward research in this field.

We will not attempt to review this growing field in this paper. We will, however, describe in some detail the current effort to model the dialog in the Waxholm system. Our objective is to develop a dialog management module which can handle the type of interaction that can occur in our chosen domain. The Waxholm system should allow user initiatives, without any specific instructions to the user, complemented by system questions to achieve the user's goal.

Natural language modelling

Our parser, STINA, is knowledge based and contains a context-free grammar which is compiled into an ATN. Lexical semantic information combined with the grammar rules describe the system constraints. Probabilities are assigned to each arc after training. These probabilities are in the first hand used to reduce the search time and to improve the pruning. The parsing is done in two steps. The first step makes use of broad categories such as nouns, while the last step expands the nouns in more detailed solutions. Thus, the choice of semantic features and preterminal nodes will automatically turn the general grammar into a subgrammar based on the domain. Smoothed n-gram models are used on the node level in the grammatical analysis in addition to the regular transition probabilities. In this context we might note that a simple phrase rule "TO noun" in our domain is mostly turned into the phrase "TO PORT" since the terminal PORT is part of the noun class and all other solutions are rather un-

[1]The Waxholm group consists of staff and students at the Department of Speech, Music and Hearing, KTH. Most of the efforts are done part time. The members of the group in alphabetic order are: Johan Bertenstam, Jonas Beskow, Mats Blomberg, Rolf Carlson, Kjell Elenius, Björn Granström, Joakim Gustafson, Kåre Sjölander, Sheri Hunnicutt, Jesper Högberg, Roger Lindell, Lennart Neovius, Lennart Nord, Antonio de Serpa-Leitao and Nikko Ström.

likely. As an additional example we find that the utterance "I want to go from X to Y" is more probable in our application than "I want to go to X from Y" as reflected in the node trigram probabilities. With this approach we can formulate a general grammar and make it domain specific with the help of lexical specifications. In our application, then, the lexicon defines that there are nouns with a specific function, PORTS, and is able to separate them from other nouns.

Dialog modelling

Two major ideas have been guiding the work on the dialog model. First, the dialog should be described by a grammar. Second, the dialog should be probabilistic. In our system, dialog building blocks are described by nodes. Each node has specifications concerning, for example, dialog action, constraint evaluation and system response. Topic selection is accomplished based on probabilities calculated from user initiatives (Carlson, 1994; Carlson and Hunnicutt, 1995; Carlson, Hunnicutt and Gustafson, 1995). Lexical semantic information combined with semantic grammar nodes are used as factors in this calculation. A modification of the domain implies an addition of how to handle a new topic, but it is our ambition that the implementation and the training procedures should, as much as possible, be kept the same.

Topic selection

The topic selection based on probabilities in our system has similarities with the effort at AT&T (Gorin, 1994; Gorin et al., 1994). A different approach, also based on training, has been presented by Kuhn and De Mori (1994) in their classification approach. A special session in the Eurospeech 1995 conference was devoted to word spotting including topic spotting based on keywords. The work by Nowel and Moore goes one step further exploring non-word based topic spotting based on a dynamic programming technique (Nowel and Moore, 1995).

The decision about which topic path to follow in the dialog is based on several factors such as the dialog history and the content of the specific utterance. The utterance is coded in the form of a "semantic frame" with slots corresponding to both the grammatical analysis and the specific application. The structure of the semantic frame is automatically created based on the rule system.

Each semantic feature found in the syntactic and semantic analysis is considered in the form of a conditional probability to decide on the topic. The probability for each topic is expressed as: p(topic|F), where F is a feature vector including all semantic features used in the utterance. Thus, the BOAT feature can be a strong indication for the TIME_TABLE topic but this can be contradicted by a HOTEL feature. The topic prediction has been trained using a labeled set of utterances taken from the Waxholm database. Only utterances indicating a topic (about 1200) have been included in this set. The probability is calculated according to: $p = (n+1)/(N+2)$, where N = number of times a feature can be a preterminal node in the feature tree, and n = number of times a feature actually is a preterminal node.

The Waxholm database

We have been collecting speech and text data using the Waxholm system. Initially, a "Wizard of Oz" (a human who simulates part of a system) replaced the speech recognition module. A full report on the data collection and data analysis can be found in Bertenstam et al., 1995b, and Bertenstam et al., 1995c.

The database used in this study includes some 68 subjects and 1900 dialog utterances containing 9200 words. The total recording time amounts to 2 hours and 16 minutes. The most frequent 200 words out of the total of 720 words cover 92 percent of the collected transcribed data. About 700 utterances are simple answers to system questions while the rest, 1200, can be regarded as user initiatives.

We can find a few examples of restarts in the database due to hesitations or mistakes on the semantic, grammatical or phonetic level. However, less than 3% of the utterances contain such disfluencies. Some of the restarts are exact repetitions of a word or a phrase. In some cases a preposition, a question word or a content word is changed. The average utterance length was 5.6 words. The average length of the first sentence in each scenario was 8.8 words.

Results

The Waxholm database was collected using preliminary versions of each module in the Waxholm system. This procedure has advantages and disadvantages for the contents of the database. System limitations will already from the beginning put constraints on the dialog, making it representative for a human-machine interaction. However, since the system was under development during the data collection, it was influenced by the system status at each recording time. After about half of the recording sessions, the system was reasonably stable, and

the number of system "misunderstandings" had been reduced. In this section, we will discuss parser performance, topic selection. As research on dialog systems develops, it becomes more important to develop new methods to evaluate human-machine interaction (Hirschman & Pao, 1993).

Parser evaluation

The parser has been evaluated in several different ways. Using about 1700 sentences in the Waxholm database as test material, 62 percent give a complete parse, whereas if we restrict the test data to utterances containing user initiatives (about 1200), the result is reduced to 48 percent. This can be explained by the fact that the large number of responses to system questions typically have a very simple syntax.

If we exclude extralinguistic sounds such as lip smack, sigh and laughing in the test material based on dialog initiatives by the user, we increase the result to 60 percent complete parses. Sentences with incomplete parses are handled by the robust parsing component and frequently effect the desired system response.

The perplexity on the Waxholm material is about 26 using a trained grammar. If only utterances with complete parses are considered we get a perplexity of 23.

Evaluation of topic selection

We have performed a sequence of tests to evaluate the topic selection method. The evaluation has used one quarter of the material, about 300 utterances, as test material, and the rest as training material, about 900 utterances. This procedure has been repeated for all quarters and the reported results are the mean values from these four runs. The first result, 12.9% errors in Table 1, is based on the unprocessed labeled input transcription. The eight possible topics have a rather uneven distribution in the material with TIME_TABLE occurring 45% of the time. One of the topics, labeled "no understanding," is trained on a set of constructed utterances that are not possible to understand, even for a human. This topic is then used as a model for the system to give an appropriate "no understanding" system response. It should be noted that, in principle, this is not a question of utterances that do not get a reasonable parse. However, the topic prediction is certainly influenced by this fact. It seemed reasonable to exclude the "no understanding" prediction from the result since the system at least does not make an erroneous decision. The accuracy model in word recognition evaluation has the same underlying principle. By excluding 55 utterances, about 5% of the test corpus, predicted to be part of the "no understanding" topic, we reduce the error by about 4%.

In the next experiment, we excluded all extralinguistic sounds, about 700, in the input text. This will increase the number of complete parses with about 10% as discussed earlier. The prediction result was about the same compared to the first experiment.

The final experiment included only those utterances that gave a complete parse in the analysis. The errors were drastically reduced. This means that the utterances with a syntax covered by our grammar also were semantically easier to interpret. On the other hand, we do not yet know if an increased grammatical coverage also will reduce the topic prediction errors.

Final remarks

The natural language and dialog components in the Waxholm system are modeled with the help of general and domain-specific rules. The domain-specific aspects of the system are reflected in the choice of lexical entries, feature system, topic selection and rule design. Empirical data play an important role for the system function, since nearly all involved knowledge sources have associated probabilities.

The lexical entries have simple probabilities based on number of occurrences in the database. Class bigram probabilities are used in the speech recognition module. Trigram models, with back-off bigrams, are used on the node level in the grammatical analysis in addition to the regular transition probabilities. In this context we might note that a simple phrase rule

Table 1. Results from the topic prediction experiments.

Test material	All material		Excluding no understanding	
	% Error	N	% Error	N
woz input	12.9	1209	8.8	1154
no extralinguistic sounds	12.7	1214	8.5	1159
only complete parses	3.1	581	2.9	580

"TO noun" in our domain is mostly turned into the phrase "TO PORT" since the terminal PORT is part of the noun class and all other solutions are rather unlikely. Thus, the choice of semantic features and preterminal nodes will automatically turn the general grammar into a subgrammar based on the domain. As an additional example we find that the utterance "I want to go from X to Y" is more probable in our application than "I want to go to X from Y" as reflected in the node trigram probabilities.

Topic selection is, as already explained in detail, based on probability calculation, but the actual modeling of each topic is done by the system designer. Thus, a change of domain or addition of the current domain require manual lexical and rule additions, example sentences for bootstrapping the system followed by data collection of real data. All probabilities in the system can be trained during usage if necessary.

Acknowledgment

This work has been supported by grants from The Swedish National Language Technology Program.

References

Aust H, Oerder M, Seide F and Steinbiss V (1994). Experience with the Philips Automatic Train Timetable Information System. In: *Proc IEEE workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94)*, 67-72.

Bertenstam J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, Serpa-Leitao A de, Nord L and Ström N (1995a). The Waxholm system - a progress report. *Proc. Spoken Dialog Systems, Vigsø*, 81-84.

Bertenstam J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, de Serpa-Leitao A, Nord L and Ström N (1995b). The Waxholm Application Data-Base. In: *Proc Eurospeech '95, 4rd European Conference on Speech Communication and Technology*, Madrid, 833-836.

Bertenstam J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, Nord L, Serpa-Leitao A de and Ström N (1995c). Spoken dialog data collected in the Waxholm project. *STH-QPSR, KTH*, 1: 49-74. 49-74.

Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Lindell R and Neovius L (1993). An experimental dialog system: Waxholm. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology*, Berlin, 1867-1870.

Carlson R (1994). Recent developments in the experimental "Waxholm" dialog system. In: *ARPA Human Language Technology Workshop*, Princetown, New Jersey, 207-212.

Carlson R and Hunnicutt S (1995). The natural language component - STINA. In: *STL-QPSR, KTH*, 1: 29-48.

Carlson R, Hunnicutt S and Gustafson J (1995). Dialog mangement in the Waxholm system. In: *Proc. ESCA/ETRW on Spoken Dialog Systems*, Vigsø, Denmark, 137-140.

Clementino D and Fissore L (1993). A man-machine dialog system for access to train timetable information. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology*, Berlin, 1863-1866.

Dalsgaard P and Baekgaard A (1994). Spoken Language Dialog Systems. In: *Proc in Artificial Intelligence, Infix*. Presented at the CRIM/FORWISS workshop on 'Progress and Prospects of Speech Research and Technology, Munich.

Gerbino E and Danieli M (1993). Managing dialog in a continuous speech understanding system. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology*, Berlin, 1661-1664.

Glass J, Flammia G, Goodine D, Phillips M, Polifroni J, Sakai S, Seneff S and Zue V. Multilingual spoken-language understanding in the MIT Voyager System. In: *Speech Communication*. Vol. 17:1-2:1-18.

Gorin A (1994). Semantic associations, acoustic metrics and adaptive language aquisition. In: *Proc ICSLP, International Conference on Spoken Language Processing*, Yokohama, 79-82.

Gorin AL, Hanek H, Rose R, and Miller L (1994). Automatic call routing in a telecommunications network. In: *Proc IEEE workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94)*, 137-140.

Hirschman L and Pao C (1993). The cost of errors in a spoken language system. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology*, Berlin, 1419-1422.

Kuhn R and De Mori R (1994). Recent results in automatic learning rules for semantic interpretation. In: *Proc ICSLP, International Conference on Spoken Language Processing*, Yokohama, 75-78.

Nowel P and Moore RK (1995). The application of dynamic programming techniques to non-word based topic spotting. In: *Proc Eurospeech '95, 4rd European Conference on Speech Communication and Technology*, Madrid, 1355-1358.

Oerder M and Aust H (1994). A realtime prototype of an automatic inquiry system. In: *Proc ICSLP International Conference on Spoken Language Processing*, Yokohama, 703-706.

Peckham J (1993). A new generation of spoken dialog systems: results and lessons from the SUNDIAL project. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology*, Berlin, 33-40.

Seneff S (1992). TINA: A natural language system for spoken language applications. *Computational Linguistics*, 18/1: 61-66.

Shirai K and Furui S (1995). Spoken Dialog. In: Shirai and Furui, eds., *Special issue of Speech Communication*, 15: 3-4.