THE MINDS SYSTEM: USING CONTEXT AND DIALOG TO ENHANCE SPEECH RECOGNITION

Sheryl R. Young[1]
Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

Contextual knowledge has traditionally been used in multi-sentential textual understanding systems. In contrast, this paper describes a new approach toward using contextual, dialog-based knowledge for speech recognition. To demonstrate this approach, we have built MINDS, a system which uses contextual knowledge to predictively generate expectations about the conceptual content that may be expressed in a system user's next utterance. These expectations are expanded to constrain the possible words which may be matched from an incoming speech signal. To prevent system rigidity and allow for diverse user behavior, the system creates layered predictions which range from very specific to very general. Each time new information becomes available from the ongoing dialog, MINDS generates a different set of layered predictions for processing the next utterance. The predictions contain constraints derived from the contextual, dialog level knowledge sources and each prediction is translated into a grammar usable by our speech recognizer, SPHINX. Since speech recognizers use grammars to dictate legal word sequences and to constrain the recognition process, the dynamically generated grammars reduce the number of word candidates considered by the recognizer. The results demonstrate that speech recognition accuracy is greatly enhanced through the use of predictions.

OVERVIEW

One of the primary problems in speech recognition research is effectively analyzing very large, complex search spaces. As search space size increases, recognition accuracy decreases. Previous research in the speech recognition area illustrates that knowledge can compensate for search by constraining the exponential growth of a search space and thus increasing recognition accuracy [12, 4, 8]. The most common approach to constraining a search space is to use a grammar. The grammars used for speech recognition dictate legal word sequences. Normally they are used in a strict left to right fashion and embody syntactic and semantic constraints on individual sentences. These constraints are represented in some form of probabilistic or semantic network which does not change from utterance to utterance [2, 8].

Today, state-of-the-art speech recognizers can achieve word accuracy rates in excess of 95% when using grammars of perplexity 30 - 60. As the number of word alternatives at each point in time increases (or as perplexity increases) performance of these systems decreases rapidly. Given this level of performance, recently researchers have begun using speech in computer problem solving applications. Using speech as an input medium for computer applications has resulted in two important findings. First, the grammars necessary to ensure some minimal coverage of a user's language have perplexities which are an order of magnitude larger than those used in today's high performing speech systems [18]. Second, the use of speech in problem solving tasks permits knowledge sources beyond the sentence level to be used to compensate for the extra search entailed by the increased perplexities. There are two reasons why higher level, contextual knowledge sources can be used to reduce the effective search space: first, the input does not consist of isolated, spoken sentences; second, all input is goal directed. There are many knowledge

[1] This research was conducted with the help of Alexander G. Hauptmann, Wayne H. Ward, Edward T. Smith and Philip Werner. This research was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 5167, under contract number N00039-85-C-0163. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

sources above the sentence level. Some of these include inferring plans, using context across clausal and sentence boundaries, determining local and global constraints on utterances and dealing with definite and pronominal reference. Work in the natural language community has shown that these knowledge sources are important for understanding language. The representation and use of goals and the plans to accomplish them have received much attention in the artificial intelligence literature [13, 14, 5]. Furthermore, natural language research has demonstrated that goals, plans and context are important for understanding implied and unexpected information as well as for providing helpful responses [16, 3, 1, 11, 7].

While the importance of contextual, dialog-level knowledge sources has been recognized in the natural language community, these knowledge sources have never been applied to the actual speech recognition process. In the past, contextual, dialog level knowledge sources were used in speech to either correct speech recognition errors [6] or to disambiguate spoken input and perform inferences required for understanding [10, 15, 17]. In these systems, dialog knowledge was applied only to the output of the recognizer.

In this paper, we describe the use of contextual, dialog based knowledge sources to reduce the effective search space for words in a speech signal and report results which illustrate their effect on recognition accuracy. Our approach uses predictions derived from the contextual knowledge sources to delimit the possible content of an incoming utterance. These knowledge sources are integrated into a robust speaker-independent, large-vocabulary speech recognition system. The knowledge sources are used predictively, and are used in conjunction with traditional syntax and semantics to constrain the recognition process by eliminating large portions of the search space for the earliest acoustic phonetic analysis. At each point in a dialog, we predict the concepts which are likely to occur. The concepts are expanded into possible word sequences which are combined with syntactic networks to produce a semantic grammar. To avoid system rigidity which could result from unfulfilled predictions, we generate predictions at different levels of specificity by using our knowledge sources opportunistically. This results in a flexible robust system which displays graceful recognition degradation. Our approach is demonstrated in a system called MINDS. MINDS works in a resource management domain, featuring information obtained from a database of facts about ships in the United States Navy. The basic problem scenario involves determining what to do about a disabled ship which is performing a specific mission. The user must determine the impact of the damage on the mission and then determine whether to replace the ship, replace the damaged equipment, delay the mission, etc. These decisions are made based upon the availability of other ships, which is determined by querying the database.

The paper begins with a description of the knowledge sources used in the MINDS system and their representation. Then the use of the knowledge sources by the recognizer is addressed. Finally, results which show the effect of the knowledge sources on speech recognition accuracy are presented.

KNOWLEDGE SOURCES AND THEIR REPRESENTATION

The MINDS system relies upon four knowledge sources to generate layered predictions about the content of an incoming utterance. These are dialog structure, general world knowledge, a domain model or model of task semantics, and models of individual users.

DIALOG STRUCTURE

One of the ideas underlying the MINDS system is that tracking all information communicated (user questions and database answers) enables a system to infer a set of user goals and possible problem solving plans for accomplishing these goals. Once goals and plans are identified, progress can be tracked and the system can generate predictions about what goal or plan steps could be executed next. In the convention of Newell and Simon (1972) MINDS represents goals and plans as a hierarchically organized AND-OR tree. This tree represents all possible abstract goals a user may have during a dialog. *For example, in the domain of dealing with disabled ships, a goal would be finding a replacement ship.* Each node in the tree is characterized by the possible subgoals into which it can be decomposed and a set of domain concepts involved in trying to achieve the goal. The concepts associated with each node can be single or multiple use as well as optional or mutually exclusive. The rationale for representing the combinations of concepts which may be involved in trying to achieve a goal or plan step is that speech systems are guided by grammars. If one can identify possible concepts, the traditional understanding process can be reversed and word sequences which express the concepts can be output.

The goal tree not only defines the goals, subgoals and domain concepts, but also the traversal options available to the user. Additionally, the dialog structure permits constraints derived from previously communicated information to be propagated to related portions of the tree. These constraints restrict either the concepts associated with various goals or the expansion of concepts associated with goals. Thus, by tracking progress through the goal tree, one can identify a list of possible next states and use them to generate a set of possible concepts which could be spoken in the next utterance.

TASK SEMANTICS

The second important knowledge source is a knowledge base of domain concepts. In this data structure we represent all domain objects and their attributes. The representation uses standard frame language which allows inheritance and multiple relations between frames and frame slots. The domain concepts represent everything that can be expressed by a user as well as all the relations and interrelations and default assumptions about domain objects. Each utterance can be mapped into a combination of domain concepts which constitute the meaning of the utterance. The domain theory or task semantics are used to perform inferencing, as well as to restrict the concepts and concept expansions that are associated with various nodes in the goal tree. *For example, if a helicopter is disabled, it is possible for the user to locate a replacement helicopter as opposed to locating a replacement ship, while it is not possible to borrow equipment if a radar device is damaged.*

GENERAL WORLD KNOWLEDGE

Our third knowledge source is a collection of domain independent, general world knowledge sources that are represented as a procedures and rules. In the MINDS system, this knowledge includes determination of what objects are in "focus" and are relevant, procedures for determining which constraints are propagated to additional nodes in the goal tree given their relative embedding. MINDS also has procedures for determining which objects or attributes could be used in an incomplete sentence in the next utterance, rules for detecting when a plan fails and principles for determining when a clarification dialog can be pursued as well as its content. Additionally, procedures are included for determining the types of anaphoric references which can be used as well as the object available for such reference. These knowledge based procedures are intended to limit the concepts which can be expressed in a next utterance, to limit the syntactic methods of expressing concepts and to limit concept expansions. Thus the set of concepts associated with a particular state in the goal tree is dynamically computed using the above described rules and procedures. The restrictions on concept expansions are computed each time a concept is predicted to be included in a future utterance.

USER KNOWLEDGE

Knowledge about individual user's domain knowledge is contained in a user model. Specifically, the user models contain domain concepts and relations between domain concepts that a user knows. These models are represented as control structures attached to individual goal nodes in the goal tree. The control structures further refine goal tree traversal options by specifying mutually exclusive goal states as well as optional goal states for a particular user. Essentially they specify what information can be inferred from other information if a user knows certain facts, and what information a user is unlikely to ask when the concepts are foreign to the individual user.

USING THE KNOWLEDGE

The MINDS system uses the above described knowledge sources to dynamically derive constraints upon an incoming utterance. The basic processing loop involves analyzing the last utterance and database response to determine their significance and to track progress in the goal tree. Any constraints which can be inferred from the input information is stored and later propagated where appropriate. Next, the system determines the possible next goals the user might pursue given current positions in the goal tree. The list of possible next moves includes not only reasonable extensions to existing moves, but also clarifications and returns to previously abandoned goals. Once possible next goal states are determined, the constraints upon their associated concepts and their expansions are computed. Finally, the set of layered predictions are expanded into grammars and used to guide the search for words in an incoming speech signal. In this section we review the procedures involved in generating conceptual predictions and using the predictions to guide recognition.

PREDICTION GENERATION

The prediction generation process involves both processing the last utterance to update the record of which portions of the goal tree have been visited and/or completed and determining what options are available for the user to pursue next.

To process an input utterance and its database response, we first try to determine which goal states are targeted by the present interaction. Determination of activated goal states, or inferring a user's plan is by no means unambiguous. During one interaction many goals may be completed and many new goal states may be initiated. Similarly, it is possible that a previously assumed goal state is not being pursued by the user. To deal with these ambiguities, we use a number of algorithms. Goals that have just been completed by this interaction and are consistent with previous plan steps are preferred. If no goals have been completed, we prefer the activated goal states which are most likely to follow, given the active goals at higher, more abstract levels in the hierarchy. Based on this information we select the next set of plan steps which are most likely to be executed. This usually constitutes the second most specific set of predictions. The set of plan steps and actions are further pruned, if possible by applying any user knowledge represented in the control schemas attached to the goal states. The concepts associated with this set of information are used to generate the most specific layer of predictions. To generate additional layers of predictions beyond the two most specific described above, we maintain a list of all incomplete, active goal states -- regardless of their relative embedding the the hierarchy. These goal states are assessed to determine possible next moves and then used to generate additional, less restrictive layers of predictions. This procedure continues until all active goals are incorporated into a prediction set. Thus, goals are layered by the amount of constraint they provide as well as the reliability of the knowledge sources used to generate them. Hence, the least restrictive set of goals includes all domain concepts.

Once the prediction layers have been determined, restrictions on the concepts associated with each of the possible goal states are computed from the task semantics and procedures for applying prior context such as given their placement in the goal tree, and the general world knowledge procedures for propagating constraints and determining focus. Next, focus is used to determine objects and attributes available for references and use of pronominal references. Finally, objects and attributes available for inclusion in a partial, elliptical utterance are determined. This information is then used to generate the grammars and lexicons used by the speech recognizer, as described below.

PREDICTION EXPANSION AND USE

The idea behind the MINDS system is to use dialog knowledge to reduce the amount of search performed by a speech recognizer and thereby reduce the number of recognition errors caused by ambiguity and word confusion. Thus, once the layered predictions have been generated, they must be expanded into a form which can be used to guide the speech recognition module. Since the prediction layers are composed of sets of abstract concepts, we need to expand or translate these into sentence fragments or word sequences that signify the appropriate conceptual meaning. Additionally, since speech recognizers can be guided by a semantic grammar, we actually expand each layer of the predictions into a dynamically generated semantic grammar composed of different, precompiled rewrite rules. Because concepts themselves are also restricted by prior context, it is also necessary to supplement each grammar with a lexicon. *For example, a rewrite rule may allow any shipname but context may restrict the shipnames to include only a few, such as Whipple and Fox. In this case, the lexicon would only include the shipnames Whipple and Fox.*

Once the predictions have been expanded into a semantic grammar, we use the grammar to guide the speech recognition system, which in this case is a modified version of the SPHNIX system [9]. During recognition, the speech module performs a time synchronous beam search. We trace through the active nodes in each part of the finite state semantic grammar to control the word transitions. As the search exits a word, it forms a set of words to transit to given the successor states in the finite state network. The recognizer uses the grammars in order of most specific first. If no string of words is found that exceeds a predetermined goodness score, the signal is reprocessed with a less constraining grammar. This process continues until an acceptable sequence of words is found.

EVALUATION

To test the ability of our layered predictions to both reduce search space and to improve speech recognition performance, we used an independent test set. This means that the utterances processed by the system were never before seen by the system or its developers. Additionally, the test set did not include any clarification dialogs. We used ten speakers (8 male, 2 female) who had not been used to train the recognizer. Each speaker read 20 sentences from adapted (to be consistent with the CMU database) versions of three test scenarios provided by the Navy. Each of these utterances was recorded. The speech recordings were then run through the SPHINX recognition system in two conditions:

- using the system grammar (all legal sentences)
- using the grammar from the successful prediction layer merged with all unsuccessful layers

The results can be seen in Table 1. As can be seen, the system performed significantly better with the predictions. Error rate decreased by a factor of five. Perhaps more important, however, is the nature of the errors. In the "layered predictions" condition, 89 percent of the insertions and deletions were the word "the". Additionally, 67 percent of the substitutions were "his" for "its". Furthermore, none of the errors in the "layered predictions" condition resulted in an incorrect database query. Because both our database and the Navy's database shared the same fields and were implemented using Informix^{TM}, we could directly assess the accuracy of the SQL database queries to Informix. Hence, semantic accuracy, defined as a correct database query, was 100% in the "layered prediction" condition. Finally, we assessed the percentage of false alarms, where the recognizer output a sequence of words deemed acceptable from a prediction layer which did not contain a correct parse of the speech input. For the 30 utterances which could not be parsed at the most specific prediction layer, there were no false alarms.

Recognition Performance		
Constraints Used:	Grammar	Predictions
Test Set Perplexity	242.4	18.3
Word Accuracy	82.1	96.5
Semantic Accuracy	85%	100%
Insertions	0.0%	0.5%
Deletions	8.5%	1.6%
Substitutions	9.4%	1.4%

SUMMARY

The MINDS system was built to demonstrate the feasibility of using contextual, dialog-level knowledge sources to constrain the exponential growth of a search space and hence increase speech recognition accuracy. The results of our studies using the system indicate that such knowledge sources are effective for dynamically reducing the number of word candidates a speech recognition system must consider when analyzing a speech signal. As we move towards robust, interactive problem solving environments where speech is a primary input medium, use of these knowledge sources could prove important for enhancing system performance.

REFERENCES

1. Allen, J. F. and Perrault, C. R. "Analyzing Intention in Utterances". *Artificial Intelligence 15*, 3 (1980), 143-178.

2. Borghesi, L. and Favareto, C. Flexible Parsing of Discretely Uttered Sentences. COLING-82, Association for Computational Linguistics, Prague, July, 1982, pp. 37 - 48.

3. Cohen, P. R. and Perrault, C. R. "Elements of a Plan-Based Theory of Speech Acts". *Cognitive Science 3* (1979), 177-212.

4. Erman, L.D. and Lesser, V.R. The Hearsay-II Speech Understanding System: A Tutorial. In Lea, W.A., Ed., *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1980, pp. 340 - 360.

5. Fikes, R. E. and Nilsson, N. J. "STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving". *Artificial Intelligence 2* (1971), 189-208.

6. Fink, P. K. and Biermann, A. W. "The Correction of Ill-Formed Input Using History-Based Expectation With Applications to Speech Understanding". *Computational Linguistics 12* (1986), 13-36.

7. Grosz, B. J. and Sidner, C. L. "Attention, Intentions and the Structure of Discourse". *Computation Linguistics 12* (1986), 175-204.

8. Kimball, O., Price, P., Roucos, S., Schwartz, R., Kubala, F., Chow, Y.-L., Haas, A., Krasner, M. and Makhoul, J. Recognition Performance and Grammatical Constraints. *Proceedings of the DARPA Speech Recognition Workshop*, Science Applications International Corporation Report Number SAIC-86/1546, 1986, pp. 53 - 59.

9. Lee, K.F., Hon, H.W. and Reddy, R. "An Overview of the SPHINX Speech Recognition System". *IEEE Transactions on Acoustics, Speech and Signal Processing in press* (1989), .

10. Levinson, S. E. and Rabiner, L. R. "A Task-Oriented Conversational Mode Speech Understanding System". *Bibliotheca Phonetica 12* (1985), 149-196.

11. Litman, D. J. and Allen, J. F. "A Plan Recognition Model for Subdialogues in Conversation". *Cognitive Science 11* (1987), 163-200.

12. Lowerre, B. and Reddy, R. The Harpy Speech Understanding System. In Lea, W.A., Ed., *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1980, pp. 340 - 360.

13. Newell, A. and Simon, H. A.. *Human Problem Solving*. New Jersey: Prentice-Hall, 1972.

14. Sacerdoti, E. D. "Planning in a Hierarchy of Abstraction Spaces". *Artificial Intelligence 5*, 2 (1974), 115-135.

15. Walker, D.E. SRI Research on Speech Understanding. In Lea, W.A., Ed., *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1980, pp. 294 - 315.

16. Wilensky, R.. *Planning and Understanding*. Addison Wesley, Reading, MA, 1983.

17. Woods, W.A., Bates, M., Brown, G., Bruce, B., Cook, C., Klovstad, J., Makhoul, J., Nash-Webber, B., Schwartz, R., Wolf, J., and Zue, V. Speech Understanding Systems - Final Technical Report. Tech. Rept. 3438, Bolt, Beranek, and Newman, Inc., Cambridge, MA, 1976.

18. Young, S. R., Hauptmann, A. G., Ward, W. H., Smith, E. T. and Werner, P. "High Level Knowledge Sources in Usable Speech Recognition Systems". *Communications of the ACM 32*, 2 (1989), .