Robust Parsing for Spoken Language Systems[1]

Stephanie Seneff

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 U.S.A.

ABSTRACT

This paper describes a recent extension to the MIT ATIS (Air Travel Information Service) system, which allows it to answer a question when a full linguistic analysis fails. This "robust" parsing capability was achieved through minor extensions of pre-existing components already in place for the full linguistic analysis component. Robust parsing is applied only after a full analysis has failed, and it involves the two stages of 1) parsing a set of phrases and clauses, and 2) gluing them together to obtain a single semantic frame encoding the full meaning of the sentence. In a recent evaluation on text input collected at multiple sites within the DARPA community, less than two thirds of the sentences yielded a full parse, but the overwhelming majority of the remaining sentences were analyzed correctly by the robust parsing scheme. We also analyzed the results when text input was replaced by recognizer outputs [3]. Even though the recognizer produced greater than 50% sentence error rate, the drop in score (%correct - %incorrect) was only 10 percentage points. This result leads to the conclusion that most of the recognizer errors are harmless in terms of meaning analysis, as long as a robust mechanism for accounting for the parsable phrases is in place.

INTRODUCTION

Current approaches to the language understanding aspect of spoken language systems tend to fall into two categories. In syntax-driven formulations [1,4,10], a complete syntactic analysis is performed which attempts to account for *all* words in an utterance. While providing strong linguistic constraints to the speech recognition component and a useful structure for further linguistic analysis, such an approach can break down in the presence of unknown words, novel linguistic constructs, recognition errors, and some spontaneous speech events (such as restarts at the word or phrase level). In contrast, semantic-driven approaches [2,5,9] tend to derive their understanding by spotting key words and phrases in the utterance. While this approach can potentially provide better coverage and deal with ill-formed sentences, it provides less constraint for the speech recognizer, and may not be able to adequately interpret complex linguistic constructs.

[1]This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

This paper describes our efforts to develop a language understanding component that combines the advantages of both of these approaches. Our strategy has been to gradually relax the constraint that the syntactic analysis must account for all of the words in an utterance. Our current implementation takes the form of a two stage process. In the first step, our parser [7] searches for a complete linguistic analysis. Failing that, constraints of the parser are relaxed to permit the recovery of parsable phrases and clauses within the sentence. These fragments are fused together using a mechanism that closely resembles our discourse history mechanism [8]. Thus the *robust* parser was integrated into the overall system with minimal changes to existing components.

OVERVIEW OF THE MIT ATIS SYSTEM

ATIS, or Air Travel Information Service, is the designated common task of the DARPA Spoken Language Systems (SLS) Program [6]. It is an air travel information system that is designed to provide travel assistance using spoken input. It currently knows about only 11 cities (9 airports) in the U.S., and 8 airlines serving these cities. The system can answer questions about such topics as departure and arrival times of each flight, the type of aircraft used, and meals served. In addition, the MIT version can guide the user through making a flight reservation.

The natural language and response generation components of the system make use of a semantic frame representation of the meaning which serves as the input for database access, spoken response generation, and history management. The frame design is flexible enough to be readily extended to other domains. Domain-dependent aspects of the system are entered mainly through table-driven mechanisms, with very little explicit programming required.

Processing of a sentence involves several steps. The first step is to provide a parse tree for the input word stream. If a full linguistic analysis fails, then a *set* of parse trees accounting for key phrases is recovered. A single semantic frame is derived from the parse tree(s), and is then integrated

0-7803-0532-9/92 $3.00 © 1992 IEEE

with available frames from the history. Both an SQL query and a text response are generated from the completed frame. The verbal response is spoken to the subject and a table is retrieved from the database through the ORACLE database management system. Finally, the system examines the goal plan and optionally initiates an additional response, based on its assessment of a likely follow-up dialogue. Detailed descriptions of the MIT ATIS system and the discourse/dialogue model can be found in [8,11].

ROBUST PARSING

This paper will focus on those aspects of the system that are most relevant to the robust parsing scheme. Since we already had in place an inheritance mechanism which could respond appropriately in context to cryptic phrases such as "aircraft" or "first class," we surmised that the same mechanism could be utilized effectively to fuse together parsed fragments *within* a single sentence. The only important distinction between such a sentence-internal history mechanism and the existing sentence-*external* history mechanism is that nothing from the internal history can be overwritten.

The other functionality that was needed was a mechanism to bracket a sentence into a set of parsed phrases/clauses representing the most extensive analysis possible, given the limitations of the grammar. This was done by modifying the parser and the grammar in minor ways. The grammar is written as a set of context free rewrite rules with constraints, and is converted automatically to a network form, where each node in the network represents a particular category (which might be a semantic name such as *a-place* or a syntactic one such as *predicate*). In full-sentence analysis mode, only the *sentence* category is allowed to terminate, and only at the *end* of the sentence. In the relaxed mode, on the other hand, a set of categories representing important clauses and phrases are allowed to terminate, and such termination can occur anywhere in the sentence.

When operating in robust mode, the parser proceeds left-to-right, as usual, but begins by producing an exhaustive set of possible parses beginning at the first word of the sentence. The parse that consumes the most words is then selected[2]. The parser begins again at the first subsequent word, repeating the procedure. Whenever no parses are returned, the parser advances by one word and tries again. Eventually a set of parsed phrases are returned.

To produce a single semantic frame for the parsed phrases, the system invokes a modified discourse history mechanism. The parsed phrases are first individually converted to semantic frames, which are delivered in sequence to a sentence-internal inheritance mechanism. In the standard history mechanism, the presence of certain attributes in the new frame masks inheritance of certain other attributes from the history. Furthermore, whenever a value for a given attribute occurs in the current frame and also in the history frame, the value of that attribute from the history is overwritten. The sentence-internal history mechanism remembers everything, however, and may occasionally have to return multiple clauses, whenever the collection of frames are judged to be too disjoint. This would be the case, for example, for the input: "I'll take flight twelve oh nine. What ground transportation is available in Denver?"

[2] In a more sophisticated form, the score may take into account N-best outputs and/or parse probabilities.

```
SENTENCE:
"(WHAT ARE THE MEALS) AND (AIRCRAFT FOR FLIGHT
 TWO EIGHTY ONE) AND ALSO FOR (FLIGHT TWO OH ONE)"

FRAME:
[Existential clause
 Topic: [(aircraft meal) for: [(flight) number: (281, 201)]]]

RESPONSE:
Here are meals for flights 281 and 201 from boston to denver.
 AIRLINE  NUMBER  FROM  TO    MEALS SERVED
 UNITED    201    BOS   DEN    BREAKFAST
 UNITED    281    BOS   DEN    DINNER

Here are the aircraft for flights 281 and 201 from boston
to denver.
 AIRLINE  NUMBER  FROM  TO   AIRCRAFT    COMPANY
 UNITED   201     BOS   DEN  DC8       MCDONNELL DOUGLAS
 UNITED   281     BOS   DEN  DC8       MCDONNELL DOUGLAS
```

Figure 1: Example sentence to illustrate robust-parsing mechanism. Parentheses in the sentence indicate parsed phrases.

An example, shown in Figure 1, will help to explain the difference between the two history mechanisms. The sentence, "What are the meals and aircraft for flight two eighty one and also for flight two oh one," is treated by the parser as three sequential entries: "What are the meals," "aircraft for flight 281," and "flight 201." If this sequence were delivered to the sentence-external history mechanism, the last phrase would be interpreted as "aircraft for flight 201." Sentence internally, however, the result would become "meals and aircraft for flights 281 and 201." Once the sentence is fully fused, the external history is brought in, and the sentence may inherit further constraints from the dialogue context, as shown in the figure, where it picks up a source and destination from the discourse context.

Further examples of robust parsing on sentences spoken by actual users are shown in Figure 2. In all three cases, we believe the system produced reasonable answers to the questions. The tables are omitted due to space limitations, but the verbal response gives a clear indication of the system's interpretation.

EVALUATION

In order to evaluate the degree of success in interpreting questions through the robust parsing scheme, we needed a measure of correctness for a set of unseen sentences. Fortunately, the DARPA community has been developing an evaluation scheme over the past year and a half, based on a comparison between the table generated by the system and a

```
TEXT:
"(WHAT ARE THE CHEAPEST FLIGHTS FROM BOSTON TO ATLANTA)
 WITH (CONNECTING FLIGHTS IN PHILADELPHIA)"

RESPONSE:
"Here are the cheapest connecting flights from Boston
 to Atlanta connecting in Philadelphia."
 <shows table>

TEXT:
"(I WOULD LIKE TO FLY FROM BOSTON TO ATLANTA AND STOP
 IN PHILADELPHIA) (ARE THERE ANY PLANES) (THAT) DO (THAT)"

RESPONSE:
"Here are the aircraft for the flights with stops from
 Boston to Atlanta connecting in Philadelphia."
 <shows table>

TEXT:
"(HOW LONG) DOES (FLIGHT TWO EIGHTY ONE) REMAIN (IN DENVER)

RESPONSE:
"Here are the connection times for flight 281 from Boston
 to San Francisco connecting in Denver.
 <shows table>
```

Figure 2: Three examples of robustly parsed sentences, taken from the training data. Tables are omitted in the interest of conserving space.

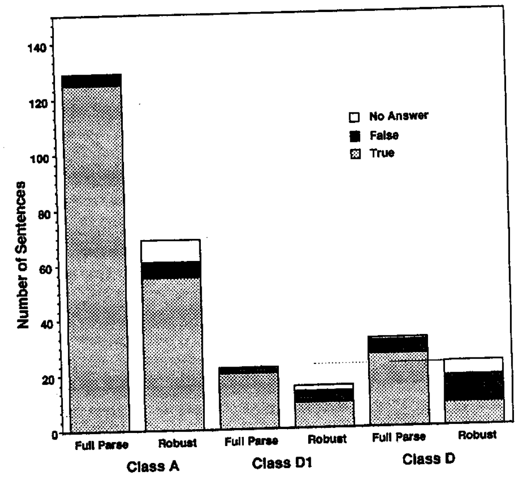

Figure 3: Results for the DARPA Dry Run test material, text input, broken down by sentence type. Class A: Context Independent; Class D1: Context Dependent on a Single Query; Class D: Context Dependent on Multiple Queries. The robust parser is used only when a full parse fails.

set of two "min/max" tables provided by trained annotators. These two tables specify the minimum and maximum requirement for expected entries from the database, where the maximum table, a recent addition to the evaluation schema, addresses the overgeneration issue, so that a system cannot indiscriminately answer all columns from the table in the hopes of accidentally providing the requested information.

The DARPA community recently performed a "dry run" evaluation, in which a set of sentences that had been collected at four sites (BBN, MIT, CMU, and SRI) were made available for testing, along with min/max annotations. The sentences had been collected under varying conditions, sometimes including a speech recognizer, and sometimes including a "wizard" who could rephrase the question before submitting it to the site-dependent data collection system. Thus they represent a fairly diverse set of collection conditions. The sentences were labelled according to four distinct categories: Class A (no context required), Class D1 (context-dependent on only one preceding sentence), Class D (context dependent on multiple sentences), and Class X (unanswerable). Out of a total of 362 sentences, 290 were "answerable," (Class A, D, or D1). The sentences for a given dialogue are presented in order to the system being tested, and it must deal with the sentence in context to come up with an appropriate answer[3]. No partial credit is given for a "nearly correct" answer, and systems are penalized for wrong answers, so that the score is defined as the difference between percent correct and percent incorrect.

Because the evaluation mechanism penalizes systems for incorrect answers, we augmented the robust parser with a capability for detecting certain key words, such as "between," which, if not properly understood, would most likely lead to an incorrect answer. Another heuristic, most relevant when a speech recognizer is included, was to refuse to answer if an unknown flight number was detected in the sentence. We used these sentences to update the discourse context, but gave a NO ANSWER response for evaluation.

Evaluating Text Input: A breakdown of the results for our system on text input, with robust parsing included, is given in Figure 3. All of the columns under "robust" mode would have given a NO ANSWER response without the robust parser. Over half of the answers must be correct in order to yield a net gain in score. For the Class A and Class D1 sentences, this requirement was met with a comfortable margin. Although the Class D, robustly parsed sentences yielded a greater number of incorrect answers than correct ones, this result is misleading, because the majority of the errors were not due to failures in the robust parsing algorithm. For instance, five sentences concerned a fare "less than one thousand dollars." A minor bug in the number interpretation routine led to an incorrect answer to all of these questions. An additional four sentences failed due to a minor problem in the the external history mechanism. Overall, we were quite encouraged by the result of this evaluation, and it leads us to believe that the robust parsing mechanism provides a powerful enhancement of the system's capabilities.

Evaluating Recognizer Hypotheses: In addition to testing the robust parsing mechanism on the correct orthographic transcription, we were interested in seeing how much the performance breaks down when the robust parser is pro-

[3]Sentence categories are *not* provided to the system being tested.

vided with errorful recognition outputs. The recognizer component of the system at MIT is still under development; in particular, we have not yet incorporated a robust mechanism for dealing with filled pauses and other disruptive events, which are fairly prevalent in the collected data. A recognizer has been developed at SRI International which does have a mechanism to account for some spontaneous speech events. Partly as a consequence, it gave a good recognition performance (11.9% word error rate, 52.2% sentence error rate) on the October dry run data [3]. The SRI researchers agreed to provide us with their recognizer's outputs. In spite of the fact that over half of the sentences contained incorrect words, the score for overall *understanding* (as measured by the min/max comparison) dropped by only 10 points[4]. We are encouraged by this result, because it indicates that many of the recognizer errors are harmless to understanding, particularly when a robust parsing mechanism can disregard misrecognized words.

CONCLUSIONS

Through examining a large body of speech material collected from a general population of naive users, we have reached the conclusion that it is not feasible to design a grammar that can always achieve a complete linguistic analysis of every input sentence. We have simultaneously become aware that a system that could recover a partial analysis would also be valuable for overcoming some recognition errors. We have described in this paper our initial attempts to realize a partial analysis whenever a full parse fails, and have reported substantial performance improvements on test material as a direct consequence of this robust mechanism. We were able to leverage off of existing system components to a large extent, leading to a rapid development of the new robust parsing mechanism. This capability allowed the system to answer many more sentences than had previously been possible. Furthermore, the score dropped by only 10 points when a recognizer's outputs were substituted for the correct orthography, in spite of the fact that over half of the recognizer's orthographic transcriptions contained word errors.

We have begun to explore some possibilities for making use of a set of N-best recognizer outputs, by parsing a *network* of paths generated through an intelligent join of the top-N candidates. We can use the frequency of occurrence of a word in the top-N candidates as a measure of its robustness, and then select a path through the network that maximizes the selection of linguistically meaningful phrases that recurred among the top-N sentences. Another possibility we have been exploring in parallel is to generalize a single recognized sentence using "homonyms" representing commonly confused recognizer pairs such as "for"/"four" or "leaving"/"leave in." A network allowing these alternates can be inserted wherever one of the pair occurs. Parse probabilities can then become important for selecting the more likely candidate when both candidates provide a parse. Preliminary results for both of these experiments look encouraging, but further work needs to be done.

[4]with 21 fewer correct answers and 8 more wrong answers out of 290.

We are just beginning to incorporate robust parsing into our data collection procedure. It should be interesting to see whether the type of material collected changes dramatically as a consequence of the fact that the system can answer a much larger percentage of the questions.

ACKNOWLEDGEMENTS

I would like to thank Hy Murveit and John Butzberger of SRI for providing recognizer outputs for some training data and the dry run test material. Discussions with Jim Glass on robust parsing algorithms have been fruitful, and he has also played a major role in the network parsing experiments.

REFERENCES

[1] Bobrow, R., R. Ingria, and D. Stallard, "Syntactic and Semantic Knowledge in the DELPHI Unification Grammar," Proceedings DARPA Speech and Natural Language Workshop, pp. 230–236, June 1990.

[2] Jackson, E., D. Appelt, J. Bear, R. Moore, and A. Podlozny, "A Template Matcher for Robust NL Interpretation,", Proceedings DARPA Speech and Natural Language Workshop, pp. 190–194, Feb. 1991.

[3] Murveit, Hy, "Speech Recognitin in SRI's Resource Management and ATIS Systems," Proceedings DARPA Speech and Natural Language Workshop, pp. 94–100, Feb. 1991.

[4] Norton, L., M. Linebarger, D. Dahl, and N. Nguyen, "Augmented Role Filling Capabilities for Semantic Interpretation of Spoken Language," Proceedings DARPA Speech and Natural Language Workshop, pp. 125–133, Feb. 1991.

[5] Pieraccini, R., E. Levin, and C.H. Lee, "Stochastic Representation of Conceptual Structure in the ATIS Task," Proceedings DARPA Speech and Natural Language Workshop, pp. 121–124, Feb. 1991.

[6] Price P., "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. Third Darpa Speech and Natural Language Workshop*, pp. 91-95, Hidden Valley, PA , June 1990.

[7] S. Seneff, "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," Proceedings ICASSP-89, pp. 711–714, May 1989.

[8] Seneff, S., L. Hirschman, and V. Zue, "Interactive Problem Solving and Dialogue in the ATIS Domain," *Proc. Fourth Darpa Speech and Natural Language Workshop*, Pacific Grove, CA, February 1991.

[9] Ward, W., "The CMU Air Travel Information Service: Understanding Spontaneous Speech,", Proceedings DARPA Speech and Natural Language Workshop, pp. 127–129, June 1990.

[10] Zue, V., J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff, "Integration of Speech Recognition and Natural Language Processing in the MIT VOYAGER System," Proceedings ICASSP-91, pp. 713–716, May 1991.

[11] Zue, V., J. Glass, D. Goodine, L. Hirschman, H. Leung, M. Phillips, J. Polifroni, and S. Seneff, "The MIT ATIS System: Preliminary Development, Spontaneous Speech Data Collection, and Performance Evaluation," EUROSPEECH-91, pp. 537-540, Sept. 1991.