WHEELS: A CONVERSATIONAL SYSTEM IN THE AUTOMOBILE CLASSIFIEDS DOMAIN

Helen Meng, Senis Busayapongchai†, James Glass, Dave Goddeau, Lee Hetherington, Ed Hurley, Christine Pao, Joe Polifroni, Stephanie Seneff, and Victor Zue

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, MA 02139 USA
BellSouth Information Systems †
Tucker, GA 30084 USA
http://www.sls.lcs.mit.edu

ABSTRACT

WHEELS is a conversational system which provides access to a database of eletronic automobile classified advertisements. It leverages off the existing spoken language technologies from our GALAXY system, and enables users to search through a database of 5,000 automobile classifieds. The current end-to-end system can respond to spoken or typed inputs, and produces a short list of entries meeting the constraints specified by the user. The system operates in mixed-initiative mode, asking for specific information but not requiring compliance. The output information is conveyed to the user with visual tables and synthesized speech. This system incorporates a new type of category bigram, created with the innovative use of the natural language component. Future plans to extend the system include operating in a displayless mode, and porting the system to Spanish.

1. INTRODUCTION

Over the past year we have been developing a conversational system named WHEELS, interfacing to a database of electronic automobile classified ads. The ultimate goal of this work, done in conjunction with BellSouth Intelliventures, is to develop a displayless, bilingual (English and Spanish) conversational system that can be deployed over the telephone network. We plan to utilize the context of this application to experiment with new speech technologies. The intermediate goals of this project, from a system development perspective, are (1) to port our GALAXY framework to the WHEELS domain, using telephone speech input, (2) to develop a displayless version of the conversational system, including an interface with the narrow-band recognizer for telephone-based displayless deployment, and (3) to extend the system to Spanish. We have completed the first stage, and are beginning development of the displayless system. At present, our end-to-end system is capable of responding to speech or typed inputs, and produces a short list of entries meeting user-specified constraints (e.g. in make, model, price, year, mileage, etc.). The output information is conveyed to the user via a visual table display and synthesized speech.

The following section describes the various components of our system. Section 3 describes our experiments in generating a category bigram from a context free grammar. In Section 4 we discuss the data collection efforts done in conjunction with BellSouth Intelliventures. A results section reports on recognition performance and parse coverage. We conclude with our future plans.

[1]This research was supported by a research contract from BellSouth Intelliventures.

2. SYSTEM COMPONENTS

Figure 1 shows a block diagram of WHEELS. This system preserves the GALAXY client-server architecture [3]. The GALAXY client interfaces between the user and the system. It also contains TINA [6] and GENESIS [1] libraries which are used for natural language understanding and language generation respectively. GENESIS is utilized to paraphrase the input sentence, so as to reflect the interpretation of the input by the system, serving to alert the user to any recognition errors. The WHEELS server is responsible for discourse and dialogue modeling as well as database retrieval. It makes use of GENESIS for its response generation. A detailed description of the dialogue component can be found in [4]. The system utilizes the FastFind database, provided to us by BellSouth Intelliventures. It contains about 5,000 entries of automobile classified advertisements that appeared in the *Atlanta Journal and Constitution*.

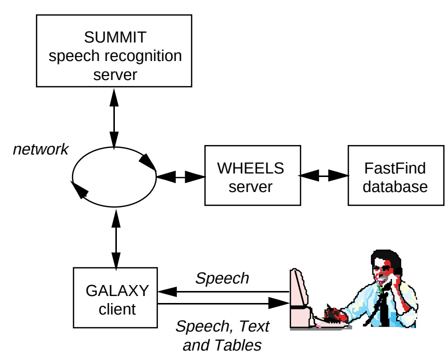

Figure 1: System Architecture of WHEELS.

2.1. Speech Recognition

The speech recognition component utilizes a stochastic segment-based recognizer developed in our group called SUMMIT [5]. Context independent phone models were initially trained on the ATIS sentence subset of the MACROPHONE corpus obtained from the Linguistic Data Consortium. They were later retrained on a large corpus

Figure 2: Parse tree for the sentence "SHOW ME BLACK CHEVROLET CORVETTE CONVERTIBLES."

```
[ clause: display
   topic: [named_auto
        number: pl
        color: black
        make: chevrolet
        model: corvette
        car_type: convertible ] ]
```

Figure 3: Semantic frame for the sentence "SHOW ME BLACK CHEVROLET CORVETTE CONVERTIBLES."

sentence	→	[word*1] np_auto [word*2] np_auto [word*3]
np_auto	→	make | model | make_model | auto
make_model	→	*honda accord*
	→	*ford ranger*
auto	→	*sports utility* (*van* | *vehicle...*)

Figure 4: Representative entries from the category bigram rules file. "|" indicates alternatives; [] indicates optional.

2.3. Dialogue Modeling

The WHEELS domain server has two main responsibilities: to manage the dialogue interaction, and to retrieve the appropriate items from the database. The user can specify cars by make, model, year, price, mileage, and other features such as color, place of origin, car-type (e.g., sport utility vehicle), etc. The semantic frame generated from the user's spoken input is received by the WHEELS domain server, and cast into the current dialogue context. This is accomplished via a "form-filling" mechanism, where the information from the semantic frame is entered into the appropriate "slot" of an electronic form (or "E-form"), which is a structure for maintaining coherence in each dialogue [4]. All the information present in the E-form is used to generate an SQL query to access the database.

In addition, the system prompts the user for further constraints to narrow down the search (e.g., "A price range would be helpful."), based on information *absent* from the E-form. It can subsequently accomodate both a cooperative response (which answers the prompt directly) and a non-cooperative response (which ignores the prompt). Such flexibility is inherent in the design of the discourse model. Further details regarding dialogue modeling in WHEELS can be found in a companion paper [4].

3. DERIVED CATEGORY BIGRAM

Whenever a new domain is first developed, there is always the problem of insufficient training data for the recognizer's language model. This problem can potentially be overcome by deriving a category bigram from the grammar rules that are being developed in parallel for the natural language component. Such a bigram has the advantage that it can be well matched to the constraints of the natural language component.

The basic principle, as applied here, is to use a simple context-free grammar to make explicit any obvious patterns in the domain, and cover the rest of the words through a flexible all-word rule. To this end, we have developed the capability of expressing a flexible grammar that intermixes a *word** rule with a set of recursive context-free rules that cover the prominent word patterns in the domain. The recursive rules specify, for example, various ways of naming cars, as in "Chevrolet Corvette convertible."

Representative rules for this grammar are given in Figure 4. The rules show three unique *word** labels, tagged positionally relative to an *np_auto* category. Although all three instances have the same all-word model,[2] their probability distributions will differ significantly, due to their differing positions in the sentence.[3] These differences yield significant improvements in the perplexity of the result-

[2] Augmented with obvious categories such as *color* and *digit*.

[3] For example, we expect that the category *word*2* would consists mostly of conjunctions, e.g., "I'D LIKE HONDA CIVICS *AND* PRELUDES."

ing model, over what would be realized with a single shared *word** model.

In order to produce a preliminary category bigram from such a grammar, whatever small amount of training utterances available initially are parsed using this covering grammar. Rule counts are tabulated from parsed sentences and converted to conditional sibling-sibling probabilities, which can then be multiplied out to the preterminal level, yielding a standard category bigram. For high perplexity points where sparse data are likely to be problematic, the frequency counts can be augmented with estimates obtained from other sources, such as the frequency of occurrence in the database. Such database-driven frequency counts are spliced into the rules for make and model, giving strong preference to the Fords and Hondas, for example.

This strategy has been found to be useful in the WHEELS domain. It continued to be beneficial even after we had obtained a large amount of training data (see Section 6). We believe that the main advantage of the derived bigram technique in WHEELS is its ability to generalize across all references to make-model pairs in cars, with respect to the external context in the sentence, while at the same time preserving the strong ties between particular makes and their associated models. This leads directly to better recognition of make-model pairs. A standard category bigram, with the obvious MAKE and MODEL categories, would not be able to enforce the constraint that a "Ford Taurus" is far preferred over a "Ford Civic." An undesirable alternative would be to explode the vocabulary with tied words enumerating all legitimate make-model pairs, but this would lead to far greater inefficiencies in the recognizer.

4. DATA COLLECTION

During the fall of 1995 BellSouth has used the WHEELS system to collect domain-specific data from two cities in Florida – Jacksonville and Melbourne. The data collection system was set up in two shopping malls, and therefore our subjects involve real users interacting with our system in a realistic environment. We have obtained over 14,000 utterances from about 600 speakers (male-to-female ratio was approximately 2:1). A wizard typed user queries directly into the WHEELS system, and speech was recorded from the telephone handset. Subjects were requested to continue the dialogue until the session was complete. In many cases, the subject chose to provide us with multiple dialogue sessions. The domain-specific data have been used for training our speech recognizer, broadening the coverage of the TINA grammar, refining the WHEELS domain server, and offering guidance for the design of the dialogue component.

5. EXAMPLE DIALOGUES

Figures 5 and 6 illustrate the capabilities of WHEELS. Figure 5 shows an interaction between the system and a cooperative speaker, who always responds directly to the system's prompts. Figure 6 is an exchange between the system and a non-cooperative speaker, who often ignores the system's prompts. A dialogue session is considered complete when the user succesfully narrows the search to five or fewer cars.

S: *Welcome to the Electronic Automobile Classifieds. What car would you like?*

U: SHOW ME LISTINGS OF HONDA CIVICS

S: *There are 49 Honda Civics. A price range would be helpful.*

U: HOW ABOUT LESS THAN EIGHT THOUSAND DOLLARS

S: *There are 37 Honda Civics costing less than 8,000 dollars. What year or range of years are you interested in?*

U: BETWEEN NINETEEN EIGHTY SIX AND NINETEEN NINETY TWO

S: *There are 29 Honda Civics costing less than 8,000 dollars between 1986 and 1992. What mileage range are you interested in?*

U: UNDER ONE HUNDRED THOUSAND

S: *There are 9 Honda Civics costing less than 8,000 dollars with mileage less than 100,000 miles between 1986 and 1992. Can you be more specific?*

U: SHOW ME THE CHEAPEST ONES

S: *Here are the 5 cheapest Honda Civics with mileage less than 100,000 miles between 1986 and 1992. Can I help you with something else?*

U: YES PLEASE

Figure 5: Sample dialogue between a cooperative user and the WHEELS system.

6. EXPERIMENTAL RESULTS

In this section we report on the overall natural language coverage of the collected data, as well as the measured perplexities and recognition performances for our test data.

In terms of grammar coverage, we found that about 12% of the collected sentences did not parse. Representative examples of the non-parsable sentences are shown in Figure 7. Some of these are reasonable sentences, and our rules can be expanded to accomodate them. In other cases, users are clearly playing with the system, or reaching outside of its domain of expertise.

Recognition results are summarized in Table 1. Our acoustic and language models were trained on over 10,000 WHEELS utterances, and recognition was measured on a disjoint test set of about 1,200 utterances. The simple bigram gave a perplexity of 9.1, which *increased* to 9.9 with the derived category grammar. However, the derived language model, even though it had a higher measured perplexity, resulted in a 2% improvement in word recognition accuracy.

Data set	# utts	Bigram	Perplexity	Accuracy
Test	1200	standard	9.1	67.4%
Test	1200	category	9.9	69.3%
IVR (make-model)	383	standard	6.1	74.8%
IVR (make-model)	383	category	5.9	76.3%

Table 1: Perplexity and recognition results for the regular and category bigram

BellSouth has been involved in the development of an Interactive Voice Response (IVR) application thus far restricted to requests for make and model, intended to replace a touch-tone input. In order to

S: *Welcome to the Electronic Automobile Classifieds. What car would you like?*

U: CAN YOU SHOW ME SOME CONVERTIBLES

S: *There are 208 convertible cars. Can you provide a make or a make and model?*

U: RED ONES

S: *There are 39 red convertible cars. Which particular model are you interested in?*

U: HOW ABOUT BLACK ONES

S: *There are 41 black convertible cars. Can you provide a make or a make and model?*

U: ANY COLOR

S: *There are 208 convertible cars. Can you provide a make or a make and model?*

U: A CORVETTE

S: *There are 8 corvette convertibles. A price range would be helpful.*

U: BETWEEN TWENTY AND THIRTY THOUSAND

S: *There are 4 Corvette convertibles costing between 20,000 dollars and 30,000 dollars. Can I help you with something else?*

U: NO THANKS

S: *Have a nice day!*

Figure 6: Sample dialogue between a non-cooperative user and the WHEELS system.

LATE EIGHTIES CAMARO Z28
RANGER XLT EXTENDED CAB
A 1992 AT POSSIBLY SEVEN THOUSAND
THREE OR FOUR PENNIES
CONVERTIBLE AUTOMATIC AIR AMFM RADIO ETCETERA
1959 DELOREAN V8 ENGINE STANDARD TRANSMISSION

Figure 7: Examples of user queries that the system did not understand.

assess the feasibility of such a system, we decided to measure the recognition performance using another disjoint data set of over 380 sentences which are limited to make and/or model requests. For this limited subdomain it is possible to measure "task completion" success rate. Our criteria for success are tabulated in Table 2.

As shown in Table 1, perplexity measured using the IVR make-model test set was 6.1 for the standard bigram and 5.9 for the derived bigram. A comparison between the perplexity measurements of the two test sets may suggest that the derived category bigram can better capture the constraints between makes and models, but this effect may be outweighed by constraints lost in the all-word modeling between categories. Recognition accuracies show consistent improvement between the two test sets as we move from the standard bigram to the derived bigram. The 74.8% word accuracy using the standard bigram translates to a task completion accuracy of 79.1%, while the 76.3% word accuracy using the derived bigram translates to a task completion accuracy of 79.4%. In the actual application, it is intended that the IVR system will verify the top few hypotheses through user feedback. If there is a positive verification, the task is considered complete. Otherwise, the user is asked to spell out the make and/or model of the car, and only the top-scoring spelling hypothesis is considered. Preliminary experiments indicate that augmentation with the spelling confirmation procedure would significantly boost the task completion accuracy to over 90%.

Ref Sentence	Hyp Sentence	Requirement for Success
make & model	make & model	makes & models agree
make & model	model only	models agree
make only	make only	makes agree
model only	make & model model only	models agree models agree

Table 2: Task completion success criteria for IVR make-model application

7. FUTURE WORK

The near-term goals with regards to the WHEELS project involve moving the development towards a displayless mode of operation, and porting the system to Spanish. In addition to general telephony issues that need to be addressed for a displayless application, it will present interesting challenges in *both* language generation and discourse management. The content of the classified ads will have to be delivered succintly to the user; there will also be additional subdialogs required for clarification, requesting information to be read or repeated, keeping track of what fraction of a search list has been read, referring to and inquiring about specific items in the list, as well as discourse-dependent help instructions. Porting WHEELS to the Spanish language will uncover its similarities and differences compared to English for parsing and generation purposes, and promote portability across languages.

8. ACKNOWLEDGEMENT

We would like to acknowledge the assistance and expertise of our collaborators at BellSouth Intelliventures, especially Dewey Anderson and Dave Anderson, for their help in coordinating the project and collecting the large body of field data in Florida.

9. REFERENCES

1. J. Glass, J. Polifroni and S. Seneff, "Multilingual Language Generation Across Multiple Domains," *Proceedings, International Conference on Spoken Language Processing*, pp. 983-986, Yokohama, Japan, Sept., 1994.
2. J. Glass, J. Chang, and M. McCandless, "A Probabilistic Framework for Feature-based Speech Recognition," *These Proceedings*, 1996.
3. D. Goddeau, E. Brill, J. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff, and V. Zue, "GALAXY: A Human-language Interface to On-line Travel Information," *Proceedings, ICSLP*, pp. 707-710, Yokohama, Japan, Sept., 1994.
4. D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai, "A Form-based Dialogue Manager for Spoken Language Applications," *These Proceedings.*
5. M. Phillips and D. Goddeau, "Fast Match for Segment-based Large Vocabulary Continuous Speech Recognition," *Proceedings, International Conference on Spoken Language Processing*, pp. 1359-1362, Yokohama, Japan, Sept., 1994.
6. S. Seneff, "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, Vol. 18, No. 1, pp. 61–86, 1992.