OPEN MICROPHONE SPEECH UNDERSTANDING: CORRECT DISCRIMINATION OF IN DOMAIN SPEECH

James Hieronymus, Greg Aist, and John Dowding***

Code TI, RIACS, and University of California, Santa Cruz
NASA Ames Research Center
Moffett Field, CA 94305

ABSTRACT

An ideal spoken dialogue system listens continually and determines which utterances were spoken to it, understands them and responds appropriately while ignoring the rest. This paper outlines a simple method for achieving this goal which involves trading a slightly higher false rejection rate of in domain utterances for a higher correct rejection rate of Out of Domain (OOD) utterances. The system recognizes semantic entities specified by a unification grammar which is specialized by Explanation Based Learning (EBL), so that it only uses rules which are seen in the training data. The resulting grammar has probabilities assigned to each construct so that overgeneralizations are not a problem. The resulting system only recognizes utterances which reduce to a valid logical form which has meaning for the system and rejects the rest. A class N-gram grammar has been trained on the same training data. This system gives good recognition performance and offers good out of domain discrimination when combined with the semantic analysis. The resulting systems were tested on a Space Station Robot Dialogue Speech Database and a subset of the OGI conversational speech database. Both systems run in real time on a PC laptop and the present performance allows continuous listening with an acceptably low false acceptance rate. This type of open microphone system has been used in the Clarissa procedure reading and navigation spoken dialogue system which is being tested on the International Space Station.

1. INTRODUCTION

Deciding when the system is being spoken to represents a continuing problem for spoken dialogue systems, especially when other people are in the environment where the system is being used. Early spoken dialogue systems used "push-to-talk" as a way to indicate start and end of user speech intended for the system This method was used at MIT and at SRI [1], [2]

A later development was the attention phrase or name which was used by several systems. The Bell Labs system used "Watson" as the name of the assistant. [3] Prefixing any command to the system with the name allowed the system to ignore any other speech. This unfortunately ignores the talker if he or she forgets to use the word of address.

A more recent development has been open-microphone speech recognition, where the end pointing is based on the presence of sound (assumed to be speech) instead of silence for a sufficiently long time. Systems using such an approach include the many telephone based systems such as HMIHY [4], spoken translation systems such as the Japanese-English system by Karaorman et al. [5] and a Spanish-English system by Roe et al [6].

The fundamental assumption is made that the user is always talking to the system, or in a speech translation system is always talking to another person in another language via the system. Therefore the system should always attempt to interpret everything that the user says, no matter how unlikely the interpretation, and respond to that in a reasonable way.

Systems also implemented barge in, which is the ability of the user to answer and give information as soon as the system has said enough. Once again this could be based on just hearing a speech like sound or actually doing recognition with a grammar. [7] Unfortunately a back channel would trigger the barge in and cancel the system output for both varieties of system. Incorrect barge in is annoying for users, but can be simply repaired by saying "Say that again."

Over the past few years, the RIALIST group at NASA Ames has developed systems for tasks in which the usual assumptions do not hold. The user may be speaking to the dialogue system or

1) speaking to other people in the area or remotely
2) speaking to other dialogue agents over a communications link

This means that the system must discriminate against speech which is not directed to it. This paper presents the first part of the solution to this problem, the development of a speech understanding system which can discriminate between speech which is semantically meaningful to it and speech which has no meaning within the system. This still leaves speech which has meaning within the system, but which is spoken to another person or agent. It is possible to use context and the particular response expected in this particular turn of the dialogue, as a mechanism to curtail false positives. For example a "yes" utterance would be accepted only if a yes/no question had been asked. This method is presently used in the NASA systems and decreases the false positives for these short utterances.

Discriminating between in domain and conversational speech is more than a two choice problem of deciding the closest distance to the representative set. The problem with conversational speech is that the topics, vocabularies, and expressions change with the situation. So while it might be tempting to make a language model for the conversational database used in testing the OOD discrimination and use utterance verification techniques, this would not generalize to a real world application. Rather we have chosen to concentrate on improving the in domain discrimination of the dialogue system and depend on that to eliminate OOD utterances. In addition to the method used here, binary decision trees and SVM's based on various phrase properties may increase performance. This will be the subject of further research.

2. RULE BASED LANGUAGE MODELS

Most of the present dialogue systems use either n-gram grammars or hand built finite state grammars for speech recognition. N-gram grammars require large amounts of transcribed speech data in order to train accurate models. Because these grammars never have enough training data to cover all possible word sequences which may potentially be said to the system, the grammar is backed off to allow previously unseen word sequences to be recognized. This has the unfortunate side effect of allowing a large number of false accepts, for data which is out of domain or ungrammatical. The system in effect coerces any input speech to be the chain of the most probable n-gram sequences recognizable by the grammar. Thus conversations with coworkers or other communication channels become sources of speech recognition errors for an n-gram system. Thus recognition performance for word and class n-gram systems has been very good, but discrimination of out of domain utterances has not been good. Tools for constructing n-gram grammars have been made available to the research community, so it is relatively easy to construct an n-gram grammar system. [8] [9] [10]

Hand built finite state grammars require large amounts of human effort to develop and require extensive rewriting when switching to a new domain. These grammars also tend to be fragile in that only a few ways of expressing an action or request are allowed or designed in. Many of the commercial telephone dialogue systems use this style of grammar and provide tools for constructing and compiling them. Over a long period of use, grammatical constructions which are in common use but are not in the original grammar are added and these systems become more natural to use.

2.1 Typed Unification Grammar

bold

Typed Unification grammars can potentially overcome the lack of discrimination in n-gram grammars by recognizing semantic meanings of the input speech. This allows many different ways of giving the same command, while discriminating against OOD utterances. Semantic grammars have traditionally required trained linguists to write them and each new domain needed many iterations with real data to insure coverage. For real tasks broad coverage rule based grammars allow phrases and sentences which are grammatical, but only a few of the constructions are actually used by humans within a particular domain. A way is needed to specialize the grammar and its underlying rule set to a new domain. The Explanation Based Learning (EBL) machine learning method developed for language by Manny Rayner [11] allows a general rule based grammar to be specialized to a new sub-domain automatically, given a corpus of training data. It is only necessary to make sure that all of the sentences in the new domain are parsible by the unification grammar, before applying EBL to it. EBL then prunes down the number of rules to those seen in the training corpus. A further specialization by training probabilities for the elements in the grammar, gives very good recognition performance. We also provide results from a non-probabilistic grammar to show how important this step is. In our system the result of utterance recognition is a logical form obtained from a second parsing step using the Gemini system [12]. Thus utterances which result in a valid logical form are "within domain" and those which are not are rejected.

Another possibility is to use a class n-gram grammar for recognition, and then use the unification grammar to determine if the utterance has meaning within the system. This method was also tried and worked well within this domain.

2.2 Unification Grammar for PSA

bold

The Personal Satellite Assistant (PSA) is a robotic assistant which is designed to navigate around in the International Space Station (ISS) propelled by fans. It is capable of making measurements and examining the status of various components visually. [13] The language consists of commands to navigate to various parts of the spacecraft and perform measurements. The commands can be elliptic and contain pronominal references. Further details of the task and language can be found in the following reference [14]. The grammar is a large coverage grammar which was constructed by hand to cover a much larger domain than the PSA domain. This grammar and grammar compilation tools are publicly available in the open source Regulus project [16], which began as a joint effort between NASA Ames and Fluency Ltd. EBL was used to prune down the number of rules to those necessary to parse the training corpus. This makes a more compact grammar which runs faster in the Nuance Communication speech recognition engine [17], than a non-specialized grammar. This grammar still lacks probabilities, so a further step of using a training corpus to compute probabilities results in a probabilistic grammar.

In our system the result of utterance understanding is a logical form obtained from the parsing step. Thus utterances which result in valid logical forms are "within domain" and those which are not are rejected. This allows the system to discriminate between in domain and out of domain utterances in a principled way.

3. CLASS N-GRAM LANGUAGE MODELS

A class trigram grammar for the PSA domain was constructed with 5394 training utterances containing approximately 22,000 words. The classes can either be constructed by hand, using knowledge of the domain or categories from a unification grammar or completely automatically [15]. The automatically generated classes have the problem that they often provide classes of inhomogeneous words. Our simple class n-gram grammar uses noun compounds as destination classes, time classes and number classes as recommended by Andreas Stolcke. These classes were constructed by hand. The SRI language modeling toolkit [9] was used to construct the class n-gram grammar. The resulting grammar was then compiled into a Nuance grammar.

4. PERFORMANCE MEASUREMENT

In order to test the claim that the EBL Unification grammar based system is able to discriminate between speech in domain and speech out of domain, a series of experiments was performed. The first experiment was to test the recognition of in domain speech by the systems on PSA dialogues. These dialogues command the robot to go to locations and to measure gases, pressure and radiation. The commands are complete sentences, but tend to be short. However the grammar allowed the cascading of requests, so that the utterance "Measure the temperature at the pilot's seat and the crew hatch and the pressure at the lockers." is a legal utterance. This means that a strict word limit on the length of the utterances would not be effective in discriminating between the in domain and out of domain utterances. The PSA data was segmented into a training and test portion, with 2211 utterances in the training set and 3888 in the test set.

The second test set consisted of short utterances from the conversational OGI 11 language corpus, cut from the "stories-at" section of the corpus. There were 117 utterances in this test set and each utterance consisted of a single sentence or phrase. This makes this data comparable in length to the PSA in domain utterances, to eliminate any length effects.

The third test set consisted of OOD utterances recorded during the PSA data collection. These 25 utterances serve as a verification that the error rates are similar between the OGI data and OOD PSA data.

Recognizer	AER	Reject	F Accept
PSA Class	6.86 %	6.35 %	0 %
PSA	9.39 %	4.75 %	0 %
PSA Pcfg	6.57 %	4.35 %	0 %

Table 1: In-domain Performance with Minimum WER

The three grammars were compiled into Nuance grammars, the PSA EBL grammar, the PSA EBL probabilistic grammar and the PSA class n-gram grammar. These were tested on all of the data sets, and the results shown for the systems trained to minimize the word error rate in Table 1 and 2. The categories in the tables and plots are Accepted Error Rate (AER) which is the WER on non-rejected utterances, Rejection (which can be false or correct), False Accept (F Accept). The performance for the Class n-gram and the PCFG grammar were comparable for the in domain data, except that the Class n-gram has a higher rejection rate by approximately 2 %. It can be seen that optimizing for WER results in systems which have poor rejection of out of domain utterances.

Recognizer	AER	Reject	F Accept
PSA Class	95.7 %	42.6 %	57.5 %
PSA	98.2 %	31 %	69 %
PSA Pcfg	100 %	33 %	67 %

Table 2: Out of domain performance with Minimum WER

Next the systems are tuned to optimize out of domain utterances by turning up the rejection threshold, with the goal of balancing false rejections in the in domain utterances with false acceptances in the out of domain utterances. This is done by increasing the weighting of the grammar and by increasing the rejection threshold. The results are shown in Table 3 and 4 for rejection thresholds of 50 and grammar weights of 6 for PSA_PCFG and 7 for PSA Class n-gram.

Recognizer	AER	Reject	F Accept
PSA Class	6.86 %	6.35 %	0 %
PSA	9.39 %	4.75 %	0 %
PSA Pcfg	7.00 %	4.6 %	0 %

Table 3: In-domain Performance with Minimum FAcc

The AER performance is only slightly worse on the in domain recognition and the false rejection also increases slightly. However this makes a huge difference in the false accept rate for out of domain utterances.

Recognizer	AER	Reject	F Accept
PSA Class	85.3 %	91.49 %	8. 51%
PSA	97.3 %	92.59 %	7.41 %
PSA Pcfg	81.5 %	95.74%	4.26 %

Table 4: OOD performance with Minimum F Accept

The Pcfg system provides superior performance for rejecting out of domain utterances while preserving in domain performance.

8. CONCLUSIONS

Spoken dialogue systems benefit greatly from being able to determine whether the user is talking to the system or to another person. By using a unification based grammar with probabilities compiled into a Nuance recognition grammar we are able to discriminate in domain from out of domain utterances with acceptable accuracy. .

11. REFERENCES

[1] J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken language understanding in the MIT VOYAGER system," Speech Communication 17:1-18, 1995.

[2] A. Stent, J. Dowding, J. Gawron, E. Bratt, and R. Moore, "The CommandTalk spoken dialogue system", Proc. ACL, 1999.

[3] R. D. Sharp, E. Bocchieri, C. Castillo, S. Parthasarathy, C. Rath, M. Riley and J. Rowland, "The Watson speech Speech Recognition Engine," Proc. ICASSP97, vol. 05, no. 5, pp. 4065-4068, 1997.

[4] A. L. Gorin, B.A. Parker, R. M. Sachs, and J. G. Wilpon, "How May I Help You?" Proc. IVTTA, Basking Ridge, NJ, 1996.

[5] M. Karaorman, T. H. Applebaum, T. Itoh, M. Endo, Y. Ohno, M. Hoshimi, T, Kamai, K. Matsui, K. Hata, S. Perarson, J-C. Junqua, "An Experimental Japanese English Interpreting Video Phone System," Proc. ICSLP96, 1996.

[6] D. B. Roe, F.C.N. Pereira, R.W. Sproat, M.D. Riley, P. J. Moreno and A. Macarron, "Efficient grammar processing for a spoken language translation system." Proc. ICASSP-92, Vol.1, pp. 213-216, 1992.

[7] N. Stroem and S Seneff, "Intelligent Barge-in in Conversational Systems," Proc. ICSLP 2000, 2000.

[8] P. Clarkson and R. Rosenfeld, "Statistical Language Modeling using the CMU-Cambridge Toolkit," Eurospeech 97, 1997.

[9] A. Stolcke, "SRILM-An Extensive Language Modeling Toolkit," Proc. ICSLP 2002, Denver, CO, 2002.

[10] M. K. Brown and B. M. Buntschuh, "A context free grammar compiler for speech understanding systems," Proc. ICSLP94, pp. 21-24, 1994.

[11] M. Rayner, "Applying explanation-based generalization to natural language processing, Proc. Int. Conf. on Fifth Generation Computer Systems," pp. 1267-1274, 1988.

[12] J. Dowding, JM Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "Gemini: A Natural Language System for Spoken-Language Understanding," Proc. ACL 1993, pp. 54-61, 1993.

[13] B.A. Hockey, J. Dowding, G. Aist, J. Hieronymus, "Targeted Help and Dialogue about Plan," Proc. ACL-02 Companion Volume , 2002.

[14] M. Rayner, B. A. Hockey and F. James, "A Compact Architecture for Dialogue Management Based on Scripts and Meta-Outputs," Proc. Applied Natural Language Processing 2000, pp. 54-60, 2000.

[15] S. Martin, J. Liermann, H. Ney, "Algorithms for Bigram and Trigram Word Clustering." Proc. EUROSPEECH-95, Madrid, 1995, pp.1253-1256.

[16] Regulus, http://sourceforge.net/projects/regulus, 2005

[17] Nuance Communications, http://www.nuance.com