CATEGORICAL UNDERSTANDING USING STATISTICAL NGRAM MODELS

Alexandros Potamianos[†]

Bell Labs, Lucent Tech., 600 Mountain Ave.
Murray Hill, NJ 07974-0636, U.S.A.
email: potam@research.bell-labs.com

Giuseppe Riccardi and Shrikanth Narayanan

AT&T Labs–Research, 180 Park Ave,
Florham Park, NJ 07932-0971, U.S.A.
email: {dsp3,shri}@research.att.com

ABSTRACT

In this paper, the speech understanding problem in the context of a spoken dialog system is formalized in a maximum likelihood framework. Word and dialog-state n-grams are used for building categorical understanding and dialog models, respectively. Acoustic confidence scores are incorporated in the understanding formulation. Problems due to data sparseness and out-of-vocabulary words are discussed. Incorporating dialog models reduces relative understanding error rate by 15-25%, while acoustic confidence scores achieve a further 10% error reduction for a computer gaming application.

1. INTRODUCTION

Recent efforts in natural language understanding have focused on statistically-based approaches. Research is motivated by the increasing complexity of spoken dialog systems, e.g., user-initiated dialog, multiple actions and attributes per dialog turn. A typical statistical understanding system decodes incoming utterance in three stages: acoustic decoding, semantic parsing of recognizer output (rule-based mapping from text to semantic labels), and context-dependent semantic decoding (statistical mapping from semantic labels to actions and attributes) [2, 3]. In this paper, *actions* are defined as application-specific operations that are independent of the human-computer interface, e.g., input and presentation modalities. *Attributes* are parameters associated with a specific action; (some of) these parameters need to be instantiated to perform the action.

Breaking down the understanding problem into semantic parsing and semantic decoding is appropriate for certain tasks (e.g., travel reservations) where there are few actions and many attributes expressed with short low-perplexity phrase fragments. However, there are many applications, e.g., gaming [4], call-routing [1], where there are many actions with few attributes, and actions are often expressed with high-perplexity phrase fragments. For these applications the emphasis of the understanding system lays on building statistical mapping from user-input to actions. The understanding problem thus becomes mostly a categorical classification problem where the classes are the application-dependent actions. Once the action is recognized, the attributes associated with the action can be identified through semantic parsing.

In this paper, we concentrate on the problem of categorical classification of actions from speech input. The organization of the paper is as follows: First, a finite state machine dialog flow model is defined. Next, the understanding from speech signal problem (called "speech understanding") is formalized as a maximum likelihood problem and is shown to be equivalent to a two-dimensional decoding problem. Then the speech understanding problem is decomposed into dialog-dependent acoustic decoding (see [5]) and understanding from transcription. Word and dialog state n-grams are used for building understanding models and dialog models respectively. Finally, results are presented for a computer gaming application.

2. DIALOG FLOW MODEL

In this section, a formal representation of the dialog flow of a general human-machine interaction with multimodal input and output is introduced. A user-initiated finite-state dialog structure is assumed which is typical for gaming applications. The central notion of the dialog flow model is the state S_t at turn t that is defined in terms of user input and system output. If W_t is the user input (e.g., speech transcription) to the application and P_t is the output in response to input W_t, then a typical transaction is

$$\ldots \underbrace{W_{t-1} \rightarrow P_{t-1}}_{S_{t-1}} \mapsto \underbrace{W_t \rightarrow P_t}_{S_t} \mapsto \underbrace{W_{t+1} \rightarrow P_{t+1}}_{S_{t+1}} \ldots \quad (1)$$

where $W \rightarrow P$ transitions are determined by the understanding system and dialog manager, $P \mapsto W$ transitions are determined by the user, and S_t is the *dialog state* at dialog turn t. A total of K dialog states are available $\{s_k,\ k = 1, \ldots, K\}$. In practice, a dialog state can be associated with no action, e.g., extraneous speech input, or with multiple actions. For simplicity we assume that only one action is allowed per dialog turn and thus dialog action and dialog state are used interchangeably in this paper (generalization of this framework to multiple actions per dialog state is straightforward).

Given the equivalence between action and dialog state, we define $\mathcal{I}_k$, $k = 1, \ldots, K$ to be the set of all user inputs that trigger state s_k. User input class $\mathcal{I}_k$ will be referred henceforth as a *dialog state class*. Note that $\mathcal{I}_k$ contain semantically equivalent utterances W since they all trigger the same action s_k.

The understanding problem is formulated here as determining the dialog state S_t given user input W_t. This is equivalent to the categorical classification of W_t in one of the K classes $\{\mathcal{I}_k\}$. For spoken dialog systems, at dialog turn t, only the sequence of acoustic vectors O_t is observable, while the input speech transcription W_t and the dialog state S_t are hidden variables[1]. Thus the understanding problem in the context of a spoken dialog system, involves a joint search over the W_t, S_t state space.

[†] This work was performed while A. Potamianos was with AT&T Labs–Research, Florham Park, NJ.

[1] For other input modalities, e.g., mouse and keyboard, user input W_t is directly observable.

3. THE CATEGORICAL UNDERSTANDING PROBLEM

As discussed in Section 2 the understanding problem is defined here as determining the dialog state[2] S_t given the speech input O_t. The maximum likelihood (ML) approach to this problem is based on maximizing the joint posterior probability

$$\max_{S_t, W_t} P(S_t, W_t | O_t, S_1..S_{t-1}) \quad (2)$$

where S_t is the dialog state, W_t is the user input (mapped to a sequence of words) and O_t is the acoustic observation sequence at dialog turn t. This ML problem is equivalent to

$$\max_{S_t, W_t} P(O_t|W_t) P(W_t|S_1..S_t) P(S_t|S_1..S_{t-1}) \quad (3)$$

under the simplistic[3] assumption that the acoustic observations are independent of the dialog state, i.e., $P(O_t|W_t S_1..S_t) \approx P(O_t|W_t)$. Eq. (3) suggests that acoustic decoding and understanding should be investigated as a single problem. Moreover, dialog state-dependent language models and understanding models could potentially be merged into a single model that computes $P(W_t|S_1..S_t)$. In practice, instead of jointly maximizing Eq. (3) with respect to W_t and S_t it is typical to first maximize the posterior probability with respect to W_t and then with respect to S_t, i.e.,

$$\hat{W}_t = \arg\max_{W_t} P(O_t|W_t) P(W_t|S_1..S_{t-1}) \quad (4)$$

$$\hat{S}_t = \arg\max_{S_t} P(\hat{W}_t|S_1..S_t) P(S_t|S_1..S_{t-1}) \quad (5)$$

where $P(W_t|S_1..S_t)$ was approximated by $P(W_t|S_1..S_{t-1})$ in the first equation since S_t is unknown at decoding time. The two step likelihood maximization although suboptimal is often used in practice because it decomposes the general understanding problem into two simpler, well-studied problems: standard acoustic decoding and understanding from transcription. Indeed, $\hat{W}_t$ is the solution to the decoding problem, i.e., maximizing the product of the acoustic and language likelihood scores[4], while $\hat{S}_t$ is the solution to the understanding problem given a transcription $\hat{W}_t$, i.e., maximizing the product of the understanding and dialog likelihood scores.

In practice, the probabilities in Eqs. (4),(5) are estimated based on (imperfect) acoustic λ_A, language λ_L, understanding λ_U and dialog λ_D models. Confidence scores can be attached to the various information streams (acoustic, language, understanding and dialog) based on the "quality" of the corresponding model. These confidences are typically used to compute exponential weights that adjust the dynamic range of the information sources. These weights γ_A, γ_L, γ_U, γ_D, are task-dependent and can be time-varying, e.g., acoustic confidence scores can be computed at the word level and used to weight the language and understanding model probabilities. Thus, we can rewrite the understanding problem as

$$\hat{W}_t = \arg\max_{W_t} P(O_t|W_t, \lambda_A)^{\gamma_A} P(W_t|S_{t-1}, \lambda_L)^{\gamma_L} \quad (6)$$

$$\hat{S}_t = \arg\max_{S_t} P(\hat{W}_t|S_t, \lambda_U)^{\gamma_U} P(S_t|S_{t-1}, \lambda_D)^{\gamma_D} \quad (7)$$

[2]Recognizing the attributes associated with a dialog state is an equally important part of the understanding problem, however, it is often trivially solved by rule-based parsing of the recognizer output.

[3]Acoustics and specifically prosody carry significant semantic information [8].

[4]Note that the language score is conditioned on the dialog state history, i.e., the language models used are dialog-state dependent [5].

provided that[5]: $P(W_t|S_1..S_{t-1}, \lambda_L) = P(W_t|S_{t-1}, \lambda_L)$, $P(W_t|S_1..S_t, \lambda_U) = P(W_t|S_t, \lambda_U)$, $P(S_t|S_1..S_{t-1}, \lambda_D) = P(S_t|S_{t-1}, \lambda_D)$.

The decoding problem of Eq. (6) is not addressed in this paper. In [5], we have presented results for different ways of computing $P(W_t|S_{t-1}, \lambda_L)$, i.e., training state-dependent and state-adapted language models, and quoted relative understanding error rate reduction up to 30% over the state-independent language models $P(W_t|\lambda_L)$. In the next two sections we propose simple markovian models for the understanding λ_U and dialog λ_D models. Furthermore, the incorporation of acoustic confidence scores in the exponential stream weights γ_U, γ_D is discussed.

4. UNDERSTANDING MODEL

A typical statistical approach to the problem of estimating $P(W_t|S_t, \lambda_U)$ involves constructing a model L_k from each one of the state classes $\mathcal{I}_k$, $k = 1, \ldots, K$. The understanding model is then defined as $\lambda_U = \{L_k, k = 1, \ldots, K\}$ and

$$P(W_t|S_t = s_k, \lambda_U) = P(W_t|L_k). \quad (8)$$

For spoken dialog systems, user input W_t is a text string and $\mathcal{I}_k$ is a set of transcribed sentences. A markovian model for $\mathcal{I}_k$ is the variable n-gram stochastic automaton [6]. Recall that n-grams have been used extensively for language modeling and well-established learning algorithms exist in the literature. If L_k is the n-gram statistical model trained from $\mathcal{I}_k$ and the input utterance $W_t = w_1 w_2..w_N$, $W_t \in \mathcal{I}_k$ then the computation of the word sequence probability is done as follows:

$$P(W_t|L_k) \approx \prod_n P(w_n|w_1..w_{n-1}, L_k). \quad (9)$$

In many cases, the training corpus has small amounts of data which leads to poor estimates of $P(W_t|L_k)$. Techniques for dealing with sparse training data can be borrowed from the language modeling literature, e.g., introduction of semantic classes (such as cities, dates, digits) back-off techniques etc. A formal evaluation of smoothing techniques in λ_U training in terms of understanding performance remains to be done. Another issue that stems from data sparseness is the high confusability of utterances that lay on understanding class boundaries. In such cases, discriminative approaches could be used to train the understanding models.

4.1. Out-of-vocabulary Words

Out-of-vocabulary (OOV) words in the transcribed input string W_t is a common problem for large vocabulary systems. Moreover, OOV words might appear even when W_t is the output of an automatic speech recognizer because the vocabulary V_k of understanding model L_k is a subset of the vocabulary used for recognition. To deal with OOV words a simple garbage node is introduced in the understanding model finite state machine with insertion penalty c_{oov}. Specifically, if the input utterance $W_t = w_1 w_2..w_N$ is represented as $\bigoplus_n w_n$ then

$$P(W_t|L_k) \approx P(\bigoplus_{n: w_n \in V_k} w_n | L_k) \; [(c_{oov})^{\sum_n \delta(w_n \notin V_k)}] \quad (10)$$

[5]The markovian assumption for the dialog state sequence is supported from the data for the "Carmen Sandiego" task (see Section 5). However, one can argue that in practice $P(.|S_1..S_t) \approx P(.|S_t)$ for most dialog interactions.

where V_k is the vocabulary drawn from the $\mathcal{I}_k$ training set, $w_n \in V_k$ signifies that word w_n is in V_k, $\delta(w_n \notin V_k)=1$ for out of vocabulary (OOV) word (else 0) and c_{oov} is a task dependent constant penalty for deletion of OOV words from input W_t.

To avoid deletions from the input utterances that could have deleterious effects when computing n-gram probabilities, OOV words can be modeled explicitly in L_k by labeling a subset of the words in the training set of each class as "OOV". For example a round-robin (a.k.a. hold-one-out) technique can be used for the sentences in each training set $\mathcal{I}_k$ and words in the held-out utterances that don't belong in the state's training dictionary are labeled as OOV. In practice, we have observed improved performance when OOV labels are explicitly modeled as in Eq. (10) rather than included in $\mathcal{I}_k$ training. Thus for our experiments a positive OOV insertion penalty c_{oov} in computed from held-out data. Understanding accuracy as a function of c_{oov} is shown in Fig. 1 for two understanding tasks (see also Results section).

4.2. Incorporating Acoustic Confidence Scores

According to [7], an acoustic confidence score can be generated for each decoded word in an utterance as the ratio of the likelihood of the "foreground" and "background" (utterance verification) model. The acoustic confidence scores AC are normalized, i.e., $AC \in [0,1]$, as discussed in [7].

The acoustic confidence scores can be used to scale the dynamic range of the understanding model probabilities for each word in a sentence or equivalently scale the log probabilities. The argument is that low confidence words should be weighted less in the understanding decision. Specifically, assuming for simplicity that there are no OOV words, the understanding model probabilities can be expressed as

$$P(W_t|L_k)^{\gamma_U} \approx \prod_n P(w_n|w_1..w_{n-1},L_k)^{\frac{c+1}{c+AC(w_n)}} \quad (11)$$

where c is a smoothing constant experimentally determined from held out data. Note that acoustic confidence scores can also be incorporated in the language model as stream weights γ_L or explicitly as word tags (see [7]).

5. DIALOG MODEL

In this section, a statistical dialog model for computing the dialog state transition probability, i.e., $P(S_t|S_1..S_{t-1})$ is defined. A simple state n-gram model is used for this purpose. Note that according to the definition of a dialog state given in Eq. (1) (user-initiated dialog) a model of the sequence of states $S_1..S_t$ is actually a *user model*, since the user input fully determines dialog state transitions. In practice, we have found that for user-initiated dialogs a state bigram $P(S_t|S_{t-1})$ models well the short-time dialog state dependencies. For example, for the "Carmen Sandiego" task (see next section) the n-gram perplexities of the finite state dialog model were: state unigram 12.2, state bigram 4.0, state trigram 3.9, state fourgram 4.0 (total of 15 dialog states, 6039 dialog turns for training and 2050 for testing).

6. TASK DESCRIPTION

The algorithms proposed above have been applied to a gaming application, the "Carmen Sandiego" task. In [4], data have been collected and analyzed from 160 children ages 8-14 using voice to interact with the popular computer game "Where in the U.S.A. is Carmen Sandiego?" by Brøderbund Software. To successfully complete the game, i.e., arrest the appropriate suspect, two subtasks had to be completed: (i) to determine the physical characteristics of the suspect and issue an arrest warrant, and (ii) to track the suspect's whereabouts in one of fifty U.S. states. The game is rich in dialog subtasks including: navigation and multiple queries (talking to cartoon characters on the game screen), database entry (filling the suspect's profile), and database search (look up clues in a geographical database). Using the dialog flow notation introduced in Section 2 we have defined *15 dialog states* for this application. For a better understanding of the semantic description of the dialog states see [4]. All collected utterances W_t have been manually assigned to the correct state s_k that they trigger according to the definition of $\mathcal{I}_k$. The training set consists of 6039 utterances collected from 51 speakers and the test set consists of 2050 utterances from 20 speakers. A typical dialog between the user and the system is shown next. Dialog state labels are shown on the right and attributes are underlined.

User input/System output	Dialog State
W_{t-3}: Tell me about the suspect?	S_{t-3}: TellmeAbout
P_{t-3}: She is neither long- nor short-legged	
W_{t-2}: Her height is average	S_{t-2}: EnterFeature
P_{t-2}: ... [updating suspect's drawing]	
W_{t-1}: Where did the suspect go?	S_{t-1}: WhereDid
P_{t-1}: She is picking peonies in Bloomington	
W_t: Go to Indiana	S_t: GoToState
P_t: ... [travel theme]	

Four of the fifteen states (or actions) have attributes, e.g., "GoToState" in the above example has "Indiana" as an attribute. The understanding problem is defined here as determining the dialog state label and attribute(s), e.g., "GoToState" and "Indiana", given the recognized user input $\hat{W}_t$.

7. EXPERIMENTAL RESULTS

Context independent hidden Markov Models (HMMs) using three states and sixteen Gaussians to model each phone were trained from the acoustic data. State-dependent word trigram language models were used for recognition (see [5]). The word accuracy (WACC) on the 2050 sentence recognition task was 78% with this acoustic and language model configuration. Dialog state attribute recognition was at 73% which is comparable to the WACC for this application. This is expected because most attributes are one or two words long (the average word length of an attribute is 1.4).

Word unigram, bigram and trigram models were trained for each dialog state class $\mathcal{I}_k$ and used as understanding models L_k. The test set perplexity was very different for each of the understanding models L_k, e.g., for word trigram models the perplexity ranged from 1.4 to 12.6. State unigram and bigrams were used for dialog modeling. The OOV word insertion penalty in Eq. 10 was set to $c_{oov} = 10$. In Fig. 1, understanding accuracy from correct and recognized transcription is shown as a function of c_{oov} for two tasks (see [1] for HMIHY task description and state of the art performance). Understanding accuracy is computed as the number of correctly classified state labels over the total number of state labels. Note the large error rates for small OOV penalty values.

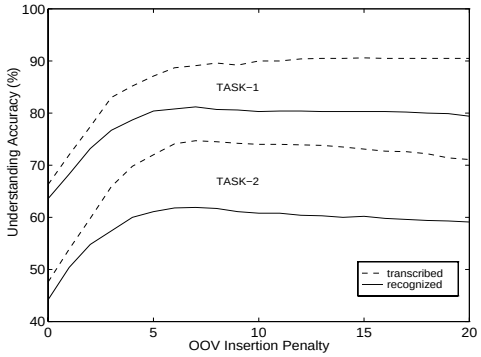

Figure 1: Understanding accuracy as a function of the OOV insertion penalty for two classification tasks (TASK1 = 'Carmen Sandiego', TASK2 = 'HMIHY') for transcribed and recognized utterances.

	Dialog Model		
Understanding Model	none	unigram	bigram
unigram	91.8%	93.2%	94.1%
bigram	92.6%	93.2%	94.4%
trigram	92.4%	93.6%	94.3%

Table 1: Understanding accuracy (%) from correct transcriptions.

Understanding accuracy from text and recognized transcriptions (WACC: 78%) is shown in Table 1 and Table 2, respectively, for different understanding and dialog model order. The more complex understanding models (bigram or trigram) perform significantly better (about 10% relative error rate reduction) than the unigram model in the presence of recognition errors. However, when the correct transcriptions are used for understanding the simple unigram model performs almost as well. In both cases, the difference in performance between bigram and trigram understanding models is negligible for this task. By incorporating the dialog model in the understanding process performance improves significantly. Overall, an additional 15% relative error rate reduction is achieved by incorporating a dialog model (25% for correct transcriptions). The understanding accuracy improvements are significant both when adding a state unigram model and when upgrading to a state bigram model.

Finally, in Fig 2 results are shown when incorporating the acoustic confidence scores in the understanding model. Specifically, the understanding accuracy (UACC) is shown as a function of the smoothing parameter c in Eq. (11). Unigram understanding and dialog models were used for understanding and a (dialog state-independent) bigram language model was used for recognition. About a 10% relative understanding error reduction is achieved when incorporating acoustic confidence scores (UACC: 82.1% for $c = 0.2$ vs UACC: 80.3% baseline performance for $c = \infty$). Overall, the results are comparable with those

	Dialog Model		
Understanding Model	none	unigram	bigram
unigram	81.5%	82.6%	84.8%
bigram	84.3%	84.0%	86.3%
trigram	84.4%	84.7%	86.3%

Table 2: Understanding accuracy (%) from recognized transcriptions (78% word accuracy, λ_L is a state-dependent trigram language model).

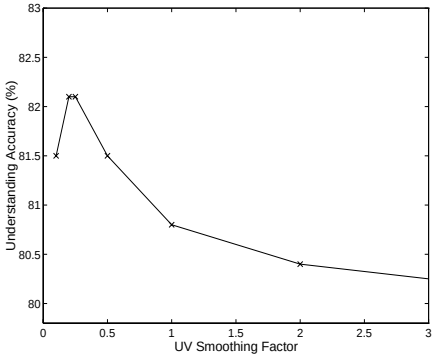

Figure 2: Understanding accuracy as a function of the acoustic confidence score smoothing factor c.

obtained with the understanding algorithms described in [7] for this task.

8. CONCLUSIONS

A categorical classification approach is proposed for speech understanding for applications where dialog actions are expressed with long high-perplexity speech fragments. The maximum-likelihood formulation of this problem in Eq. (3) suggests a unifying approach to language and understanding modeling. Language modeling techniques are successfully applied to the problem of training categorical understanding models and shown to provide results similar to fragment-based understanding models for certain tasks. Significant improvement in understanding accuracy is achieved by incorporating dialog models and acoustic confidence scores in the statistical formulation of the understanding problem.

Acknowledgments: The authors would like to express their sincere appreciation to Jerry Wright for providing baseline results of the "Carmen Sandiego" task using the HMIHY understanding system and to Rick Rose for his help with the utterance verification algorithms.

9. REFERENCES

[1] A. L. Gorin, G. Riccardi and J. H Wright, "How May I Help You?", Speech Communication, vol. 23, pp. 113-127, 1997.

[2] W. Minker, "Stochastically-Based Natural Language Understanding Across Tasks and Languages," EUROSPEECH, Rhodes, Greece, Sep. 1997.

[3] R. Pieraccini and E. Levin, "A spontaneous-speech understanding system for database query applications ," in *ESCA Workshop on Spoken Dialogue Systems - Theories and Applications*, 1995.

[4] A. Potamianos and S. Narayanan, "Spoken dialog Systems for Children", Proc. ICASSP, pp. 197-201, Seattle, 1998.

[5] G. Riccardi, A. Potamianos and S. Narayanan, "Language Model Adaptation for Spoken Dialog Systems", Proc. ICSLP, Sydney, Australia, Nov. 1998.

[6] G. Riccardi, R. Pieraccini and E. Bocchieri, "Stochastic Automata for Language Modeling", Computer Speech and Language, vol. 10(4), pp. 265-293, 1996.

[7] R. Rose, H. Yao, G. Riccardi and J. Wright, "Integration of utterance verification with statistical language modeling and spoken language understanding," Proc. ICASSP, Seattle, 1998.

[8] A. Stolcke et al, "Dialog Act Modeling for Conversational Speech," AAAI Spring Symposium, Stanford, California, Mar. 1998.