Spoken Language Variation over Time and State in a Natural Spoken Dialog System

Allen L. Gorin and Giuseppe Riccardi

AT&T Labs, Speech Research
180 Park Avenue
Florham Park, N.J. 07932
{algor,dsp3}@research.att.com

ABSTRACT

We are interested in adaptive spoken dialog systems for automated services. Peoples' spoken language usage varies over time for a fixed task, and furthermore varies depending on the state of the dialog. We will characterize and quantify this variation based on a database of 20K user-transactions with AT&T's experimental '*How May I Help You?*' spoken dialog system. We then report on a language adaptation algorithm which was used to train state-dependent ASR language models, experimentally evaluating their improved performance with respect to word accuracy and perplexity.

1. INTRODUCTION

There exist a variety of interactive speech systems in laboratories around the world, some even in actual service. There are, however, many open issues concerning how to provide robustness for large populations of non-expert users. We address the problem of creating natural spoken dialog systems for automated services. By *natural*, we mean that the machine recognizes and understands what people actually say, in contrast to what a system designer hoped they would say. This approach is in contrast with menu-driven or strongly-prompted systems, where many users are unable or unwilling to navigate such highly structured interactions. This research targets shifting the burden from human to machine, wherein the system adapts to peoples' language, as contrasted with forcing users to learn the machine's jargon.

We have been investigating methods for spoken language understanding from fluent speech [G95][G97]. Recognizing fluent speech over the telephone is a difficult task, at best. Similarly, a complete linguistic analysis of people's natural language is also not in hand. For any particular task, however, one observes that there are some events which are crucial to detect and analyze correctly, others not so. We have quantified this notion via information theory, defining *salience* as the mutual information of a linguistic event for the random variable representing machine actions [G95]. We have reported on algorithms which automatically acquire and exploit salient words [G95], phrases [R97] and grammar fragments [W97][A98]. We have embedded these methods within an experimental spoken dialog system [A97][B96], which has been evaluated on 20K user-transactions.

Peoples' spoken natural language is highly variable. A first and well-studied dimension of variation is over a large *user population*. Different people use different words and sentence structure to convey the same meaning. The second variation is over *time*. The ensemble user-behavior changes as does the world (e.g., ten years ago nobody asked for internet access.) Furthermore, there are shifts in language usage as people adapt to speaking with machines. The third variation is over *dialog state*. Depending on the dialog history, people will of course respond differently.

The current incarnation of AT&T's "*How May I Help You?*" prototype system was trained in three major stages. The *first* step was to collect data on what people say to human agents, then to automatically train language models for both recognition and understanding. This enabled us to automatically map what people say to what they want [G97]. The *second* step was to embed this recognition/understanding mechanism in a dialog system [A97], with place-holder grammars for those stages where there was no training data. We then collected data on 8K transactions of this system with live traffic. The *third* step was to exploit that dialog data to adapt the original human/human language models, compensating for language variations over time and dialog-state. This adapted spoken dialog system was then evaluated on 12K transactions.

In Section 2 we describe the language variability over these three databases. The language model adaptation algorithm is described in Section 3. This algorithm is experimentally evaluated in Section 4, measuring improvements in word accuracy and perplexity resulting from the adapted ASR language model.

2. Measuring Language Variability

2.1 Databases

The first database was generated from recordings of users talking with human agents, responding to the prompt "*AT&T. How may I help you?*' The characteristics of this data and early experiments were detailed in [G97]. We denote this set of 10K human/human interactions by HH.

These models were embedded in a spoken dialog system, using place-holder models at those states where we had no data. This dialog system was then run on 8K user-transactions. The resulting data on these human/machine interactions was split into training and test, denoted by HM_1^{train} and HM_1^{test} respectively. The HM_1 data was further partitioned based on a coarse dialog state. There are many notions of dialog state in the literature. In fact, the dialog manager in this system [A97] has *no* explicit representation of state. But, in these experiments we model

users' response to various equivalence classes of prompts. This is a first-order approximation to dialog history. Examples from these various classes are shown in Table 1.

Prompt Class	Example
GREETING	AT&T, How May I Help You?
BILLING_METHOD	How would you like to bill this call?
CARD_NUMBER	May I have your card number, please?
CONFIRMATION	Do you need me to give you credit?
PHONE_NUMBER	What number would you like to call?
REPROMPT	Sorry. Please briefly tell me how may I help you?

Table 1. Dialog State as Prompt-Equivalence Classes

After adaptively training the state-conditional models by exploiting the data HM_1 , the system was then run for an additional 12K user-transactions. This data is denoted HM_2, and is used in this paper only for testing. It is similarly split with respect to dialog-state, as above.

2.2 Utterance Length

As was observed in [G97], the number of words per utterance in HH is unimodal and highly skewed with a long tail. In Figure 1, we compare that to the length distribution for responses to the greeting prompt in HM_1 . First, observe that the HM_1 histogram is bimodal. One mode corresponds to *menu-speak*: when people are aware that they're talking with a machine, then they sometimes speak in short fragments. Interestingly, while some of the menu-speak corresponds to keywords on deployed menus, many do not. Instead, these short phrases often correspond to the salient fragments which were derived from the HH natural language database. Observe also that the second mode of HM_1 is almost identical to the single mode of the HH responses. Thus, we can view the HM greeting-responses as a mixture of menu-speak and natural spoken language, with the second component similar to the natural language in HH.

Also in Figure 1, we observe that the HM_1 distribution tail falls off much faster than for HH. Upon inspection, we observe that the very long utterances in HH are accompanied by an agent's back-channel utterances such as '*uh-huh*', encouraging the customer to continue talking. In the case of HM, there is no such back-channel encouragement from the machine, so people don't tell long stories as often. Finally, also in Figure 1, we plot the length distribution for responses to a reprompt in HM_1, observing that it is also unimodal and similar to the HH distribution of natural language responses to a human agent.

We then measure the length distribution for responses to confirmation prompts, as shown in Figure 2. The responses are divided into three categories: *explicit affirmations*, *explicit denials*, and *other*. Explicit affirmation/denials are sentences which contain the words *yes* or *no* or some variant thereof. These are sometimes spoken in isolation, or as a prepend to a natural language utterance to provide further task information. For example, responding to the prompt '*Do you want to make a credit card call?*', as user might respond '*Yes, the card number is xxxxxxx.*' The *other* category occurs during context-switching, error recovery or user-confusion.

Observe that the affirmation-length distribution is unimodal and tends to comprise shorter utterances than the denials. The explicit denials are a bimodal mixture of short responses plus a second mode at the same position as for the greeting prompts. These modes correspond to people answering '*no*' or some variant (short utterances) or to people using natural language, often with '*no*' prepended.. Thus, it is more likely for *no* to be followed by additional spoken information than it is for *yes*. Finally, the '*other*' responses also have their mode at that same position, corresponding to the natural language distribution.

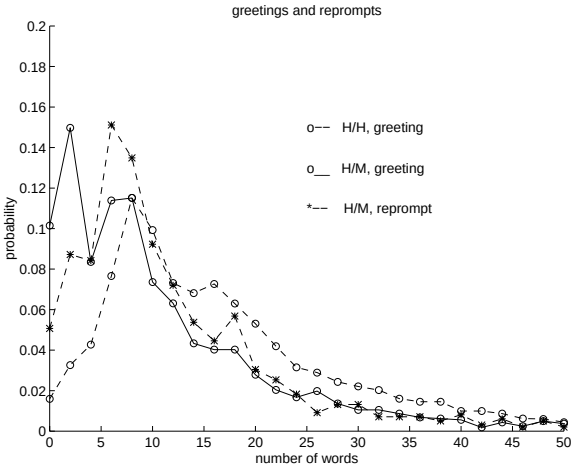

Figure 1. Utterance Length Distribution for Responses to Greetings and Reprompts.

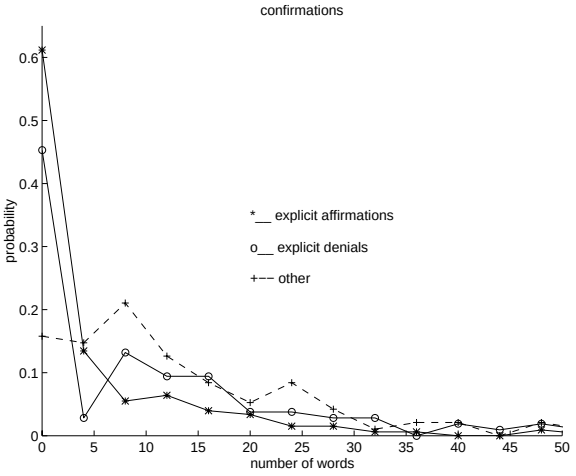

Figure 2. Utterance Length Distribution for Responses to Confirmation Prompts.

3. Language Adaptation for Spoken Dialog

There is a large literature on training and adapting language models for ASR. Typically, one assumes that the training data is drawn from the same source as the test set. In a spoken dialog system, however, a user's response depends strongly on the history of the dialog to that point. Thus, rather than merely modeling the probability of an utterance, we instead model the *conditional* probability given some dialog state and history. Let $w_1, w_2, ..., w_N$ denote the word sequence in an utterance of length N. Denote by $\mathbf{s_i}$ a particular state, where $0 < i < K$ and K is the number of states. The state-conditional probability can be expressed as follows.

$$(1) \quad P(w_1, w_2, ..., w_N \mid s_i) = \prod_j P(w_j \mid w_1, w_2, ..., w_{j-1}; s_i)$$

The state conditioning partitions the database, providing insufficient training data for each state in isolation. This is one motivation for adaptive training. A second motivation derives from the goal of naturalness: people should be able to say anything at anytime and have the machine respond appropriately. This user-behavior occurs during on-line error recovery and context-switching in natural spoken dialog. An illustrative example of a dialog with context-switching is as follows.

M: How may I help you?
U: I want to put this on my charge card.
M: What is your card number?
U: Uh, I can't find it. Can I make this a collect call?
M: What number would you like to call?
U: Good question. I need John Smith's number in Newark.
M: Please hold on for directory assistance.

In other works [P97] [S96], researchers have similarly partitioned data sets, then created separate, often disjoint models for each dialog state. That approach restricts a user, at a particular point in the dialog, to only saying what has been previously encountered there.

The stochastic language model in these experiments is the Variable Length N-gram Stochastic Automaton (VNSA) as described in [R96]. This is an automatically-trained non-deterministic stochastic finite state machine, which efficiently approximates n-gram, phrase-based and class-based models. First, a VNSA-model is trained from the HH data, denoted λ^T [G97]. The HH database comprised only responses to the greeting prompt, so that λ^T is derived from that initial dialog state alone. The set of all user-responses in HM_1^{train} is partitioned into training T_i, development B_i and test sets E_i, for each dialog state $\mathbf{s_i}$. The HH greeting model λ^T is adapted for each state $\mathbf{s_i}$ using the data T_i, via maximizing the log likelihood

$$(2) \quad \lambda_i^* = \arg\max_{\lambda_i^A} \log P(T_i \mid \lambda_i^A) \ .$$

The adapted model λ_i^* is constrained to be a linear interpolation λ_i^A of λ^T with some state-dependent model λ_i. Starting from λ^T, for each subset T_i viterbi training is run to obtain a state-dependent model λ_i. For any states t_i and t_j in the VNSA language model, the transition probabilities in the interpolated models λ_i^A $(0<A<1)$ are computed via

$$(3) \quad P_i^A(t_j \mid t_{j-1}) = \alpha_i \ P^T(t_j \mid t_{j-1}) + (1 - \alpha_i) P_i(t_j \mid t_{j-1})$$

There is no closed-form solution for the parameters α_i in equation 2 constrained by equation 3. Hence, development set B_i is used to iteratively discover the local maximum over a finite number of α_i values. This adaptation for each dialog state s_i is illustrated in Figure 3.

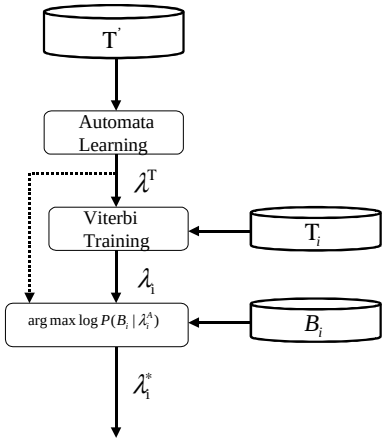

Figure 3. ASR Language Model Adaptation

4. Experimental Evaluation

Recall that λ^T is a language model trained from HH: peoples' responses to a human agent's greeting. The state-conditional model λ_i^* for state s_i was obtained by adapting λ^T with the data T_i and B_i from HM_1^{train}. One method to evaluate the utility of this adaptation is to compute their test-set perplexities on the partitioned test sets E_i in HM_1^{test}, as shown in Table 2. Also shown is the perplexity on HM_2.

Recall from [G97] that the test-set entropy of HH was 18.2. The responses to the greeting prompt in HM_1 occurred later in time, with a modified prompt to 'tip our hand' that people were talking with a machine [B96]. The language variation in both time and state is illustrated by each row of Table 2. For example, the model of greeting-responses λ^T models the utterances in HH

significantly better than any utterances in the E_i. Furthermore, the adapted language model λ_i^* provides a significantly lower perplexity for the human/machine data than the human/human data. Observe also that the adapted model does a better job of modeling the greeting-responses in HM_1, as compared to HM_2. This confirms our intuition that people's responses are 'simpler' in HM than HH, as discussed also in our earlier analysis of utterance-length.

Dialog State	**Baseline λ^T on HM_1**	**Adapted λ_i^* on HM_1**	**Adapted λ_i^* on HM_2**
GREETING	17.3	12.8	13.8
REPROMPT	15.1	13.2	13.8
BILLING METHOD	17.0	6.4	8.4
PHONE NUMBER	19.8	12.8	15.7
CARD NUMBER	21.2	15.0	16.1
CONFIRMATION	27.8	11.2	31.6

Table 2. Perplexity Reduction via Adaptation to Dialog State

In Table 3, we provide corresponding measurements of word accuracy at each dialog state for these adapted models. An additional column is provided, giving the word-accuracy using the place-holder language models used during the HM_1 trial and data collection. The word accuracy is improved over the baseline system across all dialog states. We remark that for the card and phone number responses, this is the accuracy for all words, not just the digits. A detailed discussion of the language distribution and baseline performance for utterances containing embedded digit sequences is in [R98]. We also remark that task accuracy is >> word accuracy, as detailed in [G97]. The latest reported result is 91% correct call-classification on the HH greeting-responses [W97].

Prompt-Class	**HM_1 trial**	**Baseline (λ_T)**	**Adapted (λ_i^*)**
GREETING	52.4	52.4	56.2
REPROMPT	56.7	56.7	57.4
BILLING_METHOD	60.0	62.9	64.0
PHONE_NUMBER	70.0	79.2	82.1
CARD_NUMBER	72.5	84.5	87.0
CONFIRMATION	39.6	54.4	58.3

Table 3. Percent Word Accuracy of State-Adapted Models

For the number queries (PHONE and CARD), the place-holder grammars in the HM_1 trial were merely digit loops with appropriate constraints and garbage models at each end. Although most of the tokens in those utterances were indeed digits, there were still 15% non-digit tokens. Thus, adapting a large vocabulary grammar improves word accuracy over the digit-only grammars.

5. CONCLUSIONS

We have presented a language adaptation algorithm for training state-conditional models in a natural spoken dialog system. These models allow users to say anything at anytime in the dialog. This algorithm was evaluated with respect to perplexity and word accuracy on a database of 20K human-machine transactions. A next step is to evaluate its impact on natural language understanding rate. A further step is to refine and extend the notion of dialog state in these experiments.

6. REFERENCES

[A97] A. Abella and A.L. Gorin, "*Generating Semantically Consistent Inputs to a Dialog Manager*," Proc. Eurospeech, Greece, pp. 1879-1882, Sept. 1997.

[A98] A. Arai, J. Wright, G. Riccardi and A. Gorin, "*Grammar fragment acquisition using syntactic and semantic clustering*," to appear in Proc. ICSLP Sydney, 1998.

[B96] S.J. Boyce and A.L. Gorin, "*Designing User Interfaces for Spoken Dialog Systems,*" Proc. Intl. Symp. On Spoken Dialog (ISDD), pp. 65-68, Philadelphia, Oct. 1996.

[G95] A.L. Gorin, "*On Automated Language Acquisition*," 97(6), pp. 3441-3461, Journal of the Acoustical Society of America (JASA) (June 1995).

[G97] A.L. Gorin, G. Riccardi and J.H. Wright, "*How may I Help You?*", Speech Communication 23 (1997) pp. 113-127.

[P97] C. Popovici and P. Baggia "*Specialized Language Models using Dialog Predictions*", Proc. ICASSP'97 , pp. 815-818.

[R96] G. Riccardi, R. Pieraccini and E. Bocchieri, " Stochastic Automata for Language Modeling ", Computer Speech and Language, vol. 10(4), pp. 265-293, 1996.

[R97] G. Riccardi, A.L. Gorin, A. Ljolje and M. Riley, "*Spoken Language Understanding for Automated Call-Routing*," Proc. ICASSP, pp. 1143-1146, Munich 1997.

[R98] M. Rahim, "*Connected Digit Recognition in Natural Spoken Dialog*," submitted to ICASSP '99.

[S96] H. Sakamoto and S. Matsunaga, " *Continuos Speech Recognition using Dialog-Conditioned Stochastic Language Model*", Proc. ICSLP, pp. 841-844, Yokohama, 1994.

[W97] J.H. Wright, A.L. Gorin and G. Riccardi, "*Automatic Acquisition of Salient Grammar Fragments for Call-Type Classification*", Proc. Eurospeech, Greece, pp. 1419-1422, Sept. 1997.