IMPROVED SPONTANEOUS DIALOGUE RECOGNITION USING DIALOGUE AND UTTERANCE TRIGGERS BY ADAPTIVE PROBABILITY BOOSTING

Ramesh R. Sarukkai & Dana H. Ballard

Dept. of Computer Science
University of Rochester
Rochester, NY-14627
e-mail: { sarukkai , dana }@cs.rochester.edu

ABSTRACT

Based on the observation that the unpredictable nature of conversational speech makes it almost impossible to reliably model sequential word constraints, the notion of *word set error criteria* is proposed for improved recognition of spontaneous dialogues. The basic idea in the TAB algorithm is to predict a *set* of words based on some *a priori* information, and perform a re-scoring pass wherein the probabilities of the words in the predicted word set are amplified or *boosted* in some manner. An adaptive gradient descent procedure for tuning the *word boosting* factor has been formulated. Two novel models which predict the required word sets have been presented: *utterance triggers* which capture within-utterance long-distance word inter-dependencies, and *dialogue triggers* which capture local temporal dialogue-oriented word relations. The proposed Trigger and Adaptive Boosting (TAB) algorithm have been experimentally tested on a subset of the TRAINS-93 spontaneous dialogues and the TRAINS-95 semispontaneous corpus, and have resulted in improved performances.

1. Motivation

Spontaneous speech exhibits very different properties than continuous read speech. This is especially true for conversational dialogues, where multiple speakers interact simultaneously, to perhaps, achieve a common goal. Spontaneous speech is not well-structured, and contains unpredictable features like speech restarts, mid-utterance corrections, and other speech repairs. Such attributes increase perplexity, and make it extremely difficult to build grammars or statistical language models which capture the "inherent stochasticity" in human dialogues [13].

Based on the observation that the unpredictable nature of spontaneous conversational speech makes it almost impossible to reliably model sequential word constraints, the notion of *word sets* is proposed in order to alleviate some of the aforementioned problems. The key idea is to define the word set error criteria, which measures the mismatch between the set of words from the speech recognition output, and the required correct word set. It is argued that minimization of the word set error criteria is a loose form of imposing the sequential word matching constraint. The integration of this constraint into the sequential search algorithm is performed by *boosting* the language model weights in a word-dependent fashion, the boosting factor being adapted so as to minimize the word set mismatch.

2. Word Sets

Bayes rule is applied in most continuous speech recognition systems to find the most likely word sequence. If $\mathcal{W}$ is a word sequence, given the acoustic observation $\mathcal{X}$, then a combined estimate is

$$Pr(\mathcal{W})Pr(\mathcal{X}|\mathcal{W})/Pr(\mathcal{X}) \quad (1)$$

It is important to note that the $Pr(\mathcal{X})$ term is often unknown, and left out from the Bayes expansion in the above equation: thus, the language model and acoustic model probability products should be viewed as *scores* rather than probabilities. In practice, however, since the sources of information are very different, and the true probability distributions cannot be accurately estimated, the straightforward application of Bayes rule will not lead to a satisfactory recognition performance. Therefore, it is common to weight the acoustic and language model scores separately so as to optimize performance on some held-out training data. The language weights may also be tuned using actual acoustic and language model scores in a unified stochastic model as was demonstrated by [5]. [2] have explored the idea of estimating Hidden Markov Model parameters so as to maximize speech recognition accuracy using an acoustically confusable list of words. The primary difference of this work is that interpreting the language model weights as "boosting" values enables the formulation of utterance specific triggering effects in dialogues so as to improve speech recognition accuracy.

Thus, the general form of the balanced score is:

$$Pr(\mathcal{W})^{\alpha}Pr(\mathcal{X}|\mathcal{W}) \quad (2)$$

One common method is to assign a fixed weight for each language model, and the parameter α is usually determined using a EM or gradient descent method to optimize performance on a held-out train data. A fixed value for α is

commonly used for most HMM-based speech recognition systems. Since $Pr(\mathcal{W})$ is between zero and one, the smaller the α, the more $Pr(\mathcal{W})$ is "boosted".

The *Word Set Boosting* framework extends the above concepts in the following manner. First, let us assume there is some *a priori* information which enables us to predict a set of words Ψ for the particular speech utterance $\mathcal{X}$ in consideration. Additionally α is modified into two terms as $\alpha_{LM}\alpha_{WS}$, where α_{LM} is the usual language model weight, and α_{WS} is a word dependent language model weight.

For any word w, α_{WS} is defined as:

$$\begin{aligned} \alpha_{WS}(w) &= \Omega_w \; if \; w \in \Psi \\ &= 1 \quad otherwise \end{aligned} \tag{3}$$

where Ω_w is a word-dependent factor which is adapted to minimize speech recognition system word error rate. The vector Ω is termed as the *word boosting vector*. Thus, the language model scores of *every* word present in the predicted word set, Ψ are boosted by the corresponding Ω_w factor during the search process.

3. Adaptive Word Probability Boosting

Different words need different boosting factors. An adaptive procedure is required to learn the α_{WS} parameters based on some criteria. Such a learning procedure is given below:

Definition 1 *Φ_l is the set of words in the correct word sequence for the l^{th} utterance.*

Definition 2 *Γ_l is the word set corresponding to the best path word sequence generated by the speech recognition system for the l^{th} utterance.*

Definition 3 *Ψ_l is the predicted word set to be boosted for the l^{th} utterance.*

Definition 4 *The error function for the utterance l is defined to be*

$$e_l = \frac{1}{1 + e^{-\mathcal{S}_l}} \tag{4}$$

where

$$\mathcal{S}_l = - \sum_{\omega \in \{\Psi_l \cap \Phi_l\}} [\tau - \Omega_\omega] + \sum_{\varpi \in \{\Psi_l \cap \overline{\Phi}_l \cap \Gamma_l\}} [\tau - \Omega_\varpi] \tag{5}$$

Thus, two sets of words are in consideration in the above equation. The first is the set of words present in the correct word set Φ_l, and the predicted word set Ψ_l : this acts like the biasing term of the intermediate variable. Note that this set is constant for a particular utterance during the adaptation process. τ is a constant defined such that all elements of Ω are less than τ.

The second set is the set of words present in the predicted word set Ψ_l, and the best path word sequence Γ_l, but absent in the correct word set Φ_l. Thus, this set corresponds to the wrong words that have been added to the best path word sequence.

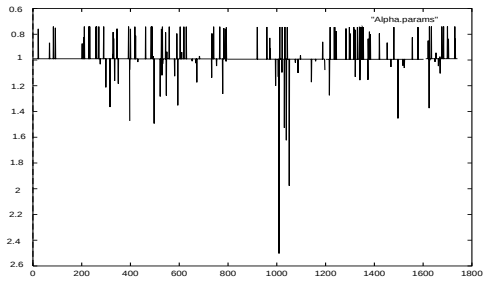

Figure 1: Adapted Boosting Vector Ω after applying TAB1.1 on TRAINS-95. x-axis : word index ; y-axis: boost factor;

The gradient adaptation derived in a previous paper [9] is given by

$$\Delta\Omega_\omega = \eta \sum_l e_l \, (1 - e_l) \, \mu_l^\omega \tag{6}$$

where η is the learning step-size, and

$$\begin{aligned} \mu_l^\omega &= -\lambda_1 \; if \quad \omega \in \{\Psi_l \cap \Phi_l \cap \Gamma_l\} \\ &= -\lambda_2 \; if \quad \omega \in \{\Psi_l \cap \Phi_l \cap \overline{\Gamma}_l\} \\ &= \lambda_3 \; if \quad \omega \in \{\Psi_l \cap \overline{\Phi}_l \cap \Gamma_l\} \\ &= 0 \quad otherwise \end{aligned} \tag{7}$$

The above equation can be easily interpreted intuitively. If the words are in the word set corresponding to the intersection of the predicted and correct word sets (i.e. the correctly predicted words), then decrease the boosting parameter by some value, thus in effect increasing the probability of the word.

On the other hand, if a word is not in the correct word set but is present in the best path word sequence produced by the speech recognizer, then increase the boosting parameter, so as to diminish its probability during the re-scoring pass.

Figure 1 shows an example of the boost factors for different words after 14 iterations (see later sections). The parameters were all initialized at 1.00. It is clear that the iterative process selectively biases some words by having lower boost factors than others.

4. Trigger Models

[3] discuss the concept of word associations, and the application of information theoretic criteria to choose word pairs. [8] also applies a modified version of the mutual information criteria to choose word triggers. In our work, the mutual information criteria is used to automatically select a subset of features relevant to the TRAINS domain. The features refer to *utterance* and *dialogue* triggers, which are described below:

4.1. Utterance Triggers

Traditional triggers specify long distance relations in a document. While there are such long distance effects present in conversational dialogues, their characteristics are quite different: local context is a very important part of dialogues[7]. Furthermore, the concept of word sets enables word-level features to bootstrap each other merely by their presence rather than their exact position in the sequence of words corresponding to the dialogues.

If the trigger word pairs that are chosen are restricted to be in the same *utterance*, then the trigger pair is termed as an *utterance trigger*. Note that, by very definition, utterance triggers are symmetric, while traditional triggers need not be. Utterance triggers try to capture long-distance word relations within an utterance.

The "significant" utterance triggers can be chosen using information theoretic measures such as *mutual information*. Experiments were performed on a subset of TRAINS-93 dialogues, and when the words in the word lattice were used to trigger a predicted word set using utterance triggers, the number of correct words present increased 68.40% to 89.51%, while the ratio of the new word set to the original lattice word set was 2.33. These results clearly indicate that the prediction method does introduce undetected correct words into the word set to be boosted during the re-scoring pass. For the TRAINS-95 corpus, all the triggers were used.

4.2. Dialogue Triggers

Conversational dialogues are complex dynamical phenomena. Intuitively it would seem that the incremental nature of multi-agent dialogue would enable the prediction of possible responses. For instance, [12] have used the *attentional structure* of sub-dialogues to predict the meaning structures the user is likely to communicate in an input. The set of utterances have two important roles. Firstly, a strong guidance can be provided for the speech recognition system so that error correction can be enhanced. Ambiguities can be resolved by biasing the recognition towards meaningful statements in the current context. Furthermore, expectations[7] can track the conversation as it jumps from one sub-dialog to another.

The concept of *dialogue triggers* is presented so that the temporal nature of conversational speech can be exploited without any higher level semantic processing. Word sets are created using the last few utterances in a dialogue. The triggered words correspond to the set of words present in the current utterance. Using a mutual information criteria, informative word pairs can be extracted. Thus, when the trigger word occurs, the triggered words can be expected to occur in the future utterance. Only a few preceding utterances (3 previous utterances in the experiments) in a dialogue are used along with the present utterance to generate possible word trigger pairs. Again mutual information measures enabled thresholding trigger pairs (threshold 0.1 millibit; word pair should have occurred at least 3 times) to give 6541 word pairs for TRAINS-93 task.

5. The Trigger and Adaptive Boosting (TAB) Algorithm

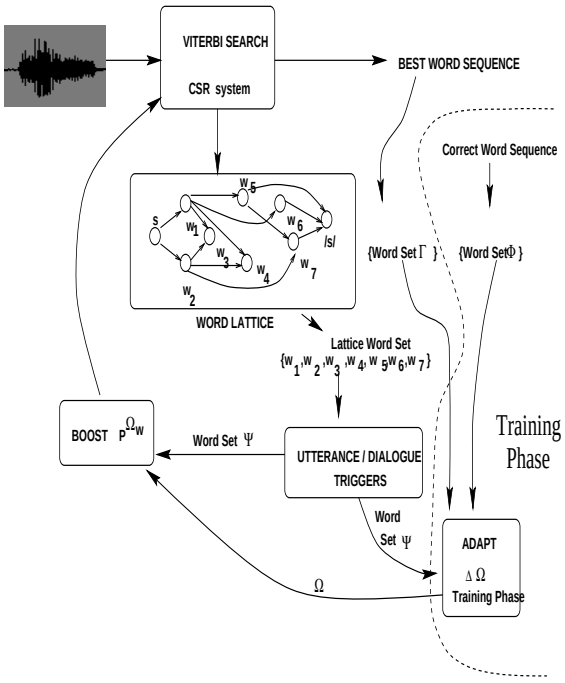

Figure 2: Overview of the Trigger and Adaptive Boosting (TAB) procedure

5.1. Two-Pass Boosting Phase

The *TAB* algorithm is summarized in figure 2. The actual implementation involves two passes. In the first pass, the conventional speech recognition search method is implemented, to give a pruned word lattice. The words in this lattice make up the lattice word set. The trigger tables are checked, and all the words that are triggered by the set of words in the lattice determines the predicted word set. These predicted word sets are boosted by the boosting factors when a second pass (re-scoring) is done, to give the re-scored best word sequence. In the case of *dialogue triggers*, the lattice word set corresponds to the words present in the lattices for the last three utterances in the dialogue order.

5.2. Adaptive Phase

During the adaptation phase, the two-pass boosting steps detailed earlier are applied. However, after each iteration, the word sets made from the best word sequence and the correct transcription are used along with the triggered word set to adapt the *boosting factors* as described in the earlier algorithm. Then the word boost factor is corrected in the batch learning process, and the next iteration uses this corrected version of the boost vector.

6. Experimental Results

The *TAB* algorithm was tested on the Trains-93 and Trains-95 domains[9]. Details of the databases and experimental

parameters can be found in [9]. Sphinx-ii[6] was the underlying speech recognition system used. Results are summarized in the following two tables[1].

Method	Bigram	Train	Test1	Test2
Utt	A+T93	2.71 %	5.16 %	2.49 %
Dia	A+T93	5.16 %	4.94 %	3.53 %
B2-LM	T93	7.17 %	6.66 %	3.11 %
Utt	T93	9.27 %	7.09 %	2.91 %
Dia	T93	7.96 %	9.24 %	4.15 %

Table 1: % decrease in absolute error for the TAB1.1 algorithm:TRAINS-93 corpus.

Method	Bigram	Train	Test1	Test2
Utt	A+T93	5.97 %	6.10 %	4.37 %
B2-LM	A+T93+T95	5.54 %	4.52 %	5.43 %
Utt	A+T93+T95	8.60 %	8.10 %	6.90 %

Table 2: % decrease in absolute error for the TAB1.1 algorithm:TRAINS-95 corpus.

7. Conclusions

The goal of this paper was to propose and evaluate word sets as a means of improving spontaneous dialogue recognition by machine. Word sets are predicted for a particular utterance based on some *a priori* information, and the probabilities of these words boosted by a word-dependent factor. In addition, a gradient descent based adaptation scheme was also derived. Furthermore, two sources of *a priori* information was proposed and studied: utterance triggers, and dialogue triggers. Utterance triggers model within-utterance long-distance inter-word relations, whereas dialogue triggers capture correlations between recent history of utterances and the future utterance. Improvements in accuracy were obtained on the TRAINS-93 and TRAINS-95 spontaneous speech dialogues. Furthermore, the word boosting approach suggests that a common language model can be tuned using the boosting parameters, and the derived word set error minimizing criteria, to bias the general-purpose language models towards topic-specific information, with only small additional memory and computational requirements.

8. References

[1] "Spoken Dialogue and Interactive Planning", James F. Allen, George Ferguson, Brad Miller, and Eric Ringger, in *Proc. of ARPA Spoken Language Technology Workshop*, Austin, TX, 1995.

[2] "Estimating Hidden Markov Model Parameters So As To Maximize Speech Recognition Accuracy", Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer, *IEEE Trans. on Speech and Audio Processing*, **vol. 1**, **no. 1**, pp:77-83, Jan. 1993.

[3] "Word Association Norms, Mutual Information, and Lexicography", Ken Church, and Patrick Hanks, COMPUTATIONAL LINGUISTICS, vol. 16, no. 1, 1990.

[4] "The TRAINS 93 Dialogues", Peter Heeman, and James F. Allen, *TRAINS Technical Note 94-2*, Dept. of Computer Science, Univ. of Rochester, March 1995.

[5] "Unified Stochastic Engine (USE) for Speech Recognition", Huang, Belin, Alleva, and Hwang, *Proc. of ICASSP'93*, pp:636-639.

[6] "An Overview of the SPHINX-II Speech Recognition System", Xuedong Huang, Fileno Alleva, Mei-Yuh Hwang, and Ronald Rosenfeld, *ARPA '93*.

[7] "The Effect of Context on the Intelligibility of Dialogue", David Novick, Karen Ward, and Benjamin Corliss, *Proc. of EUROSPEECH'95*, pp:1235-1238.

[8] "Adaptive Statistical Language Modeling: A Maximum Entropy Approach", Ronald Rosenfeld, CMU-CS-94-138.

[9] "Word Set Probability Boosting using Utterance and Dialogue Triggers for Improved Spontaneous Dialogue Recognition: The AB/TAB Algorithms", Ramesh R. Sarukkai, and Dana H. Ballard, URCSTR **601**, Dept. of Computer Science, Univ. of Rochester, Dec. 1995.

[10] "A Novel Word Pre-selection Method Based on Phonetic Set Indexing", Ramesh R. Sarukkai, and Dana H. Ballard, in the *Proc. of IEEE Intl. Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*.

[11] "The Distance Set Representation of Speech Segments", Ramesh R. Sarukkai, and Dana H. Ballard, in *Proc. of EuroSpeech'95*.

[12] "An Architecture for Voice Dialog Systems Based on Prolog-Style Theorem Proving", Ronnie Smith, Richard Hipp, and Alan Biermann, COMPUTATIONAL LINGUISTICS, 21:3, 1995.

[13] "Discourse Structure for Multi-Speaker Spontaneous Spoken Dialogs: Incorporating Heuristics into Stochastic RTNs", Sheryl Young, in *Proc. of IEEE ICASSP* , 1995.

Acknowledgements: Thanks to Prof. James Allen for his useful comments on this work. We are also grateful to our colleagues in the TRAINS group, particularly to Eric for building the ATIS+TRAINS93 bigram, and George for his help with the Sphinx-II code. Thanks to Alex Rudnicky and CMU for providing us with the Sphinx-II system.

[1] In the tables, A referes to $ATIS$, $B2-LM$ refers to a topic-specific benchmark bigram, $T93$ refers to $TRAINS93$ corpus, and $T95$ refers to $TRAINS95$ corpus