CONFIDENCE MEASURES FOR DIALOGUE MANAGEMENT IN THE CU COMMUNICATOR SYSTEM

Rubén San-Segundo[1], *Bryan Pellom, Wayne Ward*

Center for Spoken Language Understanding. University of Colorado
Boulder, Colorado 80309-0594, USA , http://cslu.colorado.edu

José M. Pardo
Grupo de Tecnología del Habla. Universidad Politécnica de Madrid
Ciudad Universitaria s/n, 28040 Madrid, Spain, http://www-gth.die.upm.es

ABSTRACT

This paper provides improved confidence assessment for detection of word-level speech recognition errors and out-of-domain user requests using language model features. We consider a combined measure of confidence that utilizes the language model back-off sequence, language model score, and phonetic length of recognized words as indicators of speech recognition confidence. The paper investigates the ability of each feature to detect speech recognition errors and out-of-domain utterances as well as two methods for combining the features contextually: a multi-layer perceptron and a statistical decision tree. We illustrate the effectiveness of the algorithm by considering utterances from the ATIS airline information task as either in-domain and out-of-domain for the DARPA Communicator task. Using this hand-labeled data, it is shown that 27.9% of incorrectly recognized words and 36.4% of out-of-domain phrases are detected at a 2.5% false alarm rate.

1. INTRODUCTION

Detection and handling of ill-posed queries, speech recognition errors, or even out-of-domain user input are important issues in the design of any robust spoken dialogue system. Without confidence assessment of speech input, errors made during speech recognition and speech understanding can lead to human-computer dialogues that diverge from the user's intended goals. When conflicts occur between the user's goals and the system's response, the user is left confused, frustrated, and often dissatisfied with the interaction.

In April 1999, the University of Colorado (CU) speech group began development of the CU Communicator system [1], a Hub-compliant implementation of the DARPA Communicator task [2,3]. The CU Communicator is a telephone-based spoken language system for accessing up-to-date airline, hotel, and car rental information from the Internet. During conversational interaction, users call the system to make travel plans, inquire about flight times and availability, and select car rental and hotel information for each leg of a planned trip. The spoken language system represents the combination of continuous speech recognition, natural language parsing, speech understanding, and event-driven dialogue management.

In this paper, we investigate methods for using statistical language model for confidence assessment of user input within the context of the Communicator task. Our work furthers the ideas proposed by Uhrik and Ward [4] in which the language model back-off information from the speech recognizer was utilized to detect out-of-domain phrases for a medical transcription task. Here, it was suggested that utterances with low word-error rates often result from word sequences for which trigrams exist within the back-off language model. When speech recognition errors occur (e.g., from poor channel, speaker characteristics, or out-of-domain requests), the decoded utterances often result from backed-off units such as bigrams and unigrams [5]. We propose that the sequences generated from such back-off events can play an important part in detection of word-errors and out-of-domain utterances in a dialogue system.

2. MOTIVATION

Statistical language model can provide a powerful mechanism for detecting out-of-domain utterances in a speech understanding system. As an illustrative example, consider a simple scenario of confidence assessment using the CMU Sphinx-II speech recognizer, a Communicator task trigram language model, and the back-off based confidence measure proposed in [4]. To simulate an out-of-domain condition, a total of 410 sentences from the Wall Street Journal (WSJ) development test set were submitted through the decoder. For an in-domain case, 699 sentences from the ATIS airline information task deemed to be "in-domain" for the DARPA Communicator task were utilized. For each decoded sentence, the utterance-level confidence measure proposed in [4] was computed and binned to produce the histogram in Fig. 1. It can be seen for the case of two very different task domains that separation of in-domain from out-of-domain input is quite easy. In fact, 99.9% of the Communicator domain utterances were correctly accepted while 99.3% of the WSJ sentences were correctly rejected

[1] This work was performed during a visiting internship at the Center for Spoken Language Understanding and has been supported by Spanish Education Ministry grant AP97-20257252 and by DARPA through ONR grant #N00014-99-1-0418.

0-7803-6293-4/00/$10.00 ©2000 IEEE.

While WSJ and ATIS represent two substantially different task domains, similar situations can occur when users make inquiries related to services not provided by dialogue system. Consider, for example, the speech recognition output for the phrase, "*I wanna go from Denver to Istanbul Turkey on June first*", compared with the recognized output from the phrase, "*Is there any discount for student conference travel?*" shown in Fig. 2. Here we can see that the in-domain utterance results in a bigram back-off followed by sequences of trigrams while the out-of-domain query results in a sequence of unigram back-offs. The notion the decoder tends to back off to bigram and unigram sequences during out-of-vocabulary or out-of-domain events provides the basis for the confidence algorithm proposed earlier in [4]. In this paper, we further this idea by explicitly modeling the *sequence* of the back-off and language model score events to improve detection of out-of-domain utterances.

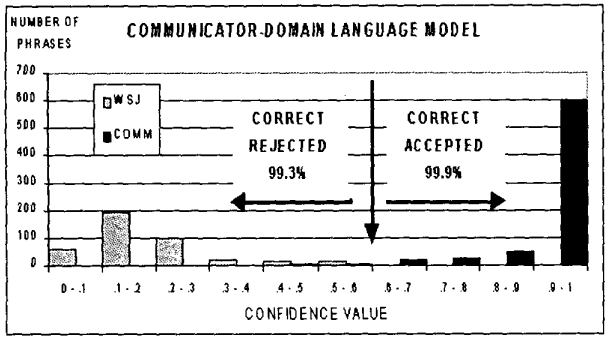

Figure 1. Distribution of confidence values for 410 WSJ utterances and 699 Communicator domain ATIS utterances decoded using the CMU Sphinx-II speech recognizer with a Communicator task trigram language model.

IN DOMAIN	
UTTERANCE:	I WANNA GO FROM DENVER TO ISTANBUL TURKEY ON JUNE FIRST
RECOGNIZED:	<s> I WANNA GO FROM DENVER TO ISTANBUL TURKEY ON JUNE FIRST </s>
	<s> I WANNA GO FROM [city] TO [city_state] ON [month] [month_number] </s>
LM BACK-OFF:	B T T T T T T T T T
OUT OF DOMAIN	
UTTERANCE:	IS THERE ANY DISCOUNT FOR STUDENT CONFERENCE TRAVEL
RECOGNIZED:	<s> EASTERN ANY DISCOUNT THIRFTY CAR TO TRAVEL </s>
	<s> EASTERN ANY DISCOUNT [car_rental_company] FOR TO TRAVEL </s>
LM BACK-OFF:	U U U U U U B

Figure 2. Example of in-domain and out-of-domain utterances and resulting LM back off for (T) trigram, (B) bigram and (U) unigram sequences.

3. CONFIDENCE FEATURES

When assigning confidence to an utterance, we first calculate a confidence measure for each word. The word-level values are then accumulated over the entire utterance to obtain a phrase-level confidence. For low confidence phrases, the dialogue manager must decide to re-prompt the user for specific information, make a clarification, or take other actions based on the current system state. In this work, the following language model features were considered:

- Language Model Back-Off: The back-off behavior of a N-gram language model can be utilized to assess confidence of the speech recognition output. Higher confidence is assumed to derive from sequences of trigrams that appear within the language model. Lower confidence utterances are assumed to derive from lower-order back-off sequences.
- Language Model Score: The second feature considered is the log-probability for each word in a sequence as computed from a back-off language model. This feature provides additional information because two words with the same back-off sequence will often have different language model probabilities. High probability word sequences are assumed to indicate high confidence while low probability word sequences are assumed to indicate low confidence.
- Phonetic Length of Word: It is often the case that when out-of-vocabulary words are spoken, the speech recognizer will tend to make sequences of substitutions and insertions involving short, monosyllabic words.

Next, we consider several methods for assessing the contributions of each feature both individually and as combined estimators of confidence.

4. FEATURE ANALYSIS

4.1 Individual Feature Assessment

Our first experiments investigated the ability of each feature to detect in-domain versus out-of-domain phrases in the Communicator domain. For the LM score and phonetic length features, a linear mapping was utilized to constrain the possible range of values to the [0,1] interval. For the LM back-off feature, each back-off type (e.g., trigram, bigram-bigram, etc.) was mapped to a unique value on the [0,1] interval using the method described in [4]. The sequences of back-off confidence values are smoothed over a 5-word context window and accumulated over an entire utterance to produce a phrase-level confidence value [4].

4.2 Multi-layer Perceptron Context Model

We have considered two methods for incorporating context information within the decision process. First, we combined word-level features over a 5-word context window using a multi-layer perceptron (MLP). In this study, the features were quantized to 50 binary inputs (i.e., 10 quantized inputs per word of context). The hidden layer consisted of 75 units and one output node was used to model the word-level confidence. During weight estimation, a target value of 1 is assigned when the decoder correctly recognizes the word and a value of 0 is assigned during incorrect recognition (e.g., substitutions and deletions). By accumulating these word-level confidence values across all or part of an utterance, we can assign phrase-level confidence to the system input.

4.3 Decision-Tree Context Model

Alternatively, we have considered using a statistical decision tree to model the impact that context has on assigning word-level confidence. The decision tree is implemented using the raw scores of each feature (i.e., the language model score itself, the actual phonetic length of the word, as well as the back-off type). During training, questions are asked about the feature sequence and nodes are split to maximize detection of incorrectly recognized words. For example, "Q: Did the previous word result from a bigram back-off?" The splitting questions are designed around a 5-word context window as in the case of the MLP and the best-question used to split the tth node is computed based on the node's impurity, I(t),

$$I(t) = 2\, p(CORRECT/t)\, p(INCORRECT/t) \qquad (1)$$

where, p(CORRECT/t) is the probability of CORRECT words in the node t and p(INCORRECT/t) is the probability of INCORRECT words in the node t. The question that results in the lowest impurity is used to split the node. The splitting is stopped when one of two conditions is met: (1) when the number of training vectors in a given node is less than 10, and (2) when all the decisions on a node produce a new node without vectors.

After the tree is constructed T_0, we prune it for the optimum sub-tree. In our experiments we use *Minimal Cost-Complexity Pruning*. We must calculate the sequence of sub-tree T_1, $T_2,\ldots,T_n$, that minimize the cost-complexity until the root node is reached. Then we evaluate all of these sub-trees, using the evaluating set, and select the best one. More details can be seen in [6].

5. EXPERIMENTAL EVALUATION

5.1 Experimental Configuration

We have taken 7388 phrases from ATIS [7] travel information task and tagged each by hand as either "in" or "out" of the Communicator task domain. The resulting partition contained 5110 in-domain and 2288 out-of-domain phrases respectively. From this set of phrases, a held-out set of 1000 random sentences were used for final algorithm test, leaving the remaining for training and development test. Using Round-Robin methods, the experimental results were repeated 4 times using different held-out sets of data in order to minimize the impact of data selection. The language model used in the experiments has been generated with the in-domain ATIS phrases of the training set in addition to approximately 25,000 phrases collected during a pilot study of the Carnegie Mellon Communicator system [8]. We point out that discrimination of in-domain versus out-of-domain phrases is quite difficult for the ATIS data since all phrases relate to airline information queries. Therefore, phrases labeled as out-of-domain for this study often represent user queries for services that would generally not be provided by a Communicator-like dialogue system.

5.2 Experimental Results

5.2.1 Word-Level Confidence Results

Table 1 summarizes the correct detection rates for word-level recognition errors at false alarm rates of 2.5% and 5.0%. It can be seen that language model (LM) score provides the best indicators for word-level confidence. For example, using LM score alone, 39% of mis-recognized words were detected at a false alarm rate of 5%. Furthermore, we found that phonetic-length of the word provides very little relationship to word-level confidence. When the features are combined contextually using a 5-word window, we see further improvements. For example, 43.2% of the mis-recognized words are correctly detected at a false alarm rate of 5.0%. Overall, the MLP provided superior classification results compared with the decision tree combination method.

(A) Individual Feature Assessment			
	2.5% FA	**5.0% FA**	**Min. Error**
Length	0.0%	6.5%	14.3%
LM Back off	17.2%	22.0%	13.4%
LM score	22.2%	39.0%	12.6%
(B) Combining all Features			
Decision Tree	26.4%	40.3%	12.1%
MLP	**27.9%**	**43.2%**	**12.1%**

Table 1. Correct detection of mis-recognized words at a 2.5% and 5.0% false alarm (FA) rate. Minimum classification error is also shown. Simulation results are shown for (A) classification using individual features and (B) classification using contextual information and combined feature sets.

Considering that the recognition rate was 85.7% (ER 14.3%) the LM score reduce the incertitude of word confidence in 2.2%

5.2.2 Utterance-Level Confidence Results

Table 2 summarizes results obtained for utilizing each language model feature to detect out-of-domain utterances. In Table 2(A), we see once again that phonetic word-length is a very poor indicator of confidence and that LM score alone provided the best overall measure (i.e., 38.3% correct detection at a 5% false alarm rate).

For contextually combined features, shown in Table 2(B), the MLP provided superior accuracy compared with the decision tree method. Here, the method combining all the features using an MLP resulted in 44.8% correct detection at a 5% false alarm rate. The best overall results for out-of-domain utterance detection was obtained by removing the phonetic-length feature from the training vectors of the MLP. In this case, 46.7% correct detection of out-of-domain utterances was obtained at an overall 5% false alarm rate. We point out that these results are

promising considering the fact that separation of in-domain from out-of-domain ATIS data for the Communicator task is quite difficult since all inquiries are related to airline information services.

Our language model used in the above experiments consisted of a combination of "in-domain" ATIS phrases augmented with approximately 25,000 utterances from the CMU Communicator pilot data [8]. It is interesting to consider the impact that the additional data has on the classification problem. Therefore, we conducted one additional experiment in which the system language model was constructed only from in-domain utterances labeled from the ATIS data set. The resulting classification rate is summarized in Table 3. Here we see that the resulting minimum error rate is only 0.2% lower for the ATIS only language model compared to that constructed from ATIS and Communicator data. Since the "in-domain" ATIS phrases represent a subset of the entire Communicator task (airline, hotel, and car rental reservations), we see that the similarity in results is to be expected.

(A) Individual Feature Assessment			
	2.5% FA	**5.0% FA**	**Min. Error**
Length	2.7	9.3	29.2
LM Backoff	23.2	35.8	19.7
LM score	**24.6**	**38.3**	**19.3**
(B) Combining all Features			
Decision Tree	22.9	34.5	21.4
MLP	34.3	44.8	18.0
MLP (without length)	**36.4**	**46.7**	**18.0**

Table 2. Correct detection of out-of-domain phrases at a 2.5% and 5.0% false alarm (FA) rate. Minimum classification error is also shown. Simulation results are shown for (A) classification using individual features and (B) classification using contextual information and combined feature sets.

Combining LM Score and LM Back-off			
	2.5% FA	**5.0% FA**	**Min. Error**
(A) ATIS+COMM	36.4	46.7	18.0
(B) ATIS ONLY	31.6	44.5	17.8

Table 3. Correct detection of out-of-domain phrases combining LM score and LM Back-off sequence for a 2.5% and 5.0% false alarm rate. Results are shown for (A) the MLP feature combination with language model from ATIS and Communicator pilot data, and (B) the language model constructed from in-domain ATIS utterances.

6. CONCLUSIONS

In this paper we have considered an improved mechanism for combining contextual information in order to detect speech recognition errors and out-of-domain phrases within the context of the CU Communicator system. Specifically, it was shown that the context sequence of LM scores and LM back-off sequences over a 5-word window can be combined using a decision tree or multi-layer perceptron to provide improved discrimination compared with the method proposed in [4]. The efficacy of the approach was demonstrated by hand-labeling utterances from the ATIS task as either in-domain or out-of-domain for the Communicator task. Given this partitioned data of airline information queries, the proposed algorithm was shown to detect 27.9% of mis-recognized words and 36.4% of out-of-domain utterances at a 2.5% false alarm rate. Using this feature context modeling, we have increased correct detection by more than 10% from the baseline experiments for word confidence (17.2% Table 1) and utterance confidence (23.2% Table 2) reducing the classification error 1.3% and 1.7% respectively. From experimental observations, word phonetic length is not a significant indicator of word-level or phrase-level confidence.

Another important conclusion is that the multi-layer perceptron and the decision tree are more powerful mechanisms to combine the 5-word context information. In each experiment we found that the multi-layer perceptron provides improved classification compared with the decision tree method. However, we point out that the questions used in node splitting for the decision tree method offers more insight into the relative importance of each feature considered in the study.

7. REFERENCES

[1] Ward W., Pellom B. "The CU Communicator System," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Keystone Colorado, 1999.

[2] Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., Zue, V., "Galaxy-II: A Reference Architecture for Conversational System Development," *Proc. ICSLP-98*, Sydney Australia, Vol. 3, pp. 931-934, 1998.

[3] http://fofoca.mitre.org

[4] Uhrik C., Ward W. "Confidence metrics based on n-gram language model backoff behaviors", *Proc. Eurospeech-97*, Rhodes, Greece, Sep 1997.

[5] Chase L. *Error-Responsive Feedback Mechanisms for Speech Recognizers.* Unpublished PhD Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, April 1997.

[6] Breiman L., Friedman J.H., Olshen R. A., Stone C.J. *Classification and Regression Trees* Ed. Wadsworth & Brooks/Cole advanced books & software. 1984.

[7] Hemphill, C., Godfrey, J., Doddington, G., "The ATIS Spoken Language Systems Pilot Corpus", *DARPA Speech and Natural Language Workshop*, June 1990..

[8] Eskenazi M., Rudnicky A., Gregory K., Constantinides P., Brennan R., Bennett C., Allen J., "Data collection and processing in the Carnegie Mellon Communicator," *Proc. Eurospeech-99*, Budapest, Hungary, Sep 1999.