CONFIDENCE METRICS BASED ON N-GRAM LANGUAGE MODEL BACKOFF BEHAVIORS

C. Uhrik
Berdy Medical Systems
4909 Pearl East Circle, Suite 202
Boulder, Colorado, USA 80301
Tel. 303-417-1603, FAX 303-417-1662, E-mail: uhrik@berdy.com

W. Ward
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
Tel. 303-442-8807, FAX 303-417-1662, E-mail: whw@cs.cmu.edu

ABSTRACT

We report results from using language model confidence measures based on the degree of backoff used in a trigram language model. Both utterance-level and word-level confidence metrics proved useful for a dialog manager to identify out-of-domain utterances. The metric assigns successively lower confidence as the language model estimate is backed off to a bigram or unigram. It also bases its estimates on sequences of backoff degree. Experimental results with utterances from the domain of medical records management showed that the distributions of the confidence metric for in-domain and out-of-domain utterances are separated. Use of the corresponding word-level confidence metric shows similar encouraging results.

1. INTRODUCTION

Speech recognition systems typically produce a rank ordered set of hypotheses using acoustic and language models. When designing a Spoken Language System, it is important to be able to estimate confidence that an utterance has been understood correctly by the system. Among other things, this estimate is needed by a Dialog Manager component to decide how to respond to the utterance. It can be used to identify problematic words or utterances which should initiate user interactions to verify, clarify or correct errors.

This work extends the work of [1] and [2], using the degree of backoff in the language model estimate for a word string as the basis for a confidence estimate. Although here we speak specifically of trigram language models, the techniques presented are easily extended to generalized N-gram models.

Our basic recognition system is a modified Sphinx II HMM-based system [3], using a backed-off trigram language model [4]. We propose an utterance-level confidence metric based on the backoff behavior of the trigram language model. The metric takes into account the observations that:

1) language model probabilities resulting from trigrams are associated with fewer errors than those resulting from bigram backoffs, and likewise bigram backoffs are more reliable than unigram backoffs [3];

2) language model probabilities resulting from runs of trigrams are more confident than those resulting from runs of bigram backoffs, and runs of bigram backoffs are more confident than runs of unigram backoffs.

Two measures are used. One provides confidence estimates for utterances as a whole, and the other for individual words within utterances.

2. A BACKOFF-BASED UTTERANCE-LEVEL CONFIDENCE MEASURE

First we consider assigning a confidence to an utterance as a whole. For an utterance consisting of a sequence of words $w_1\ w_2\ ...\ w_i\ ...\ w_n$, one assigns word local confidences relative to the backoff behavior of a backed-off trigram model,

conf(i) = 1.0 if $P(w_i)$ derives from a trigram,

conf(i) = 0.8 if $P(w_i)$ derives from a bigram-bigram backoff, both $p(w_{i-2},w_{i-1})$ and $p(w_{i-1},w_i)$ exist

conf(i) = 0.6 if $P(w_i)$ derives from a bigram, $p(w_{i-1},w_i)$ exists

conf(i) = 0.4 if $P(w_i)$ derives from a bigram-unigram backoff, $p(w_{i-2},w_{i-1})$ and $p(w_i)$ exist

conf(i) = 0.3 if $P(w_i)$ derives from a unigram-unigram backoff, both $p(w_{i-1})$ and $p(w_i)$ exist

conf(i) = 0.2 if $P(w_i)$ derives from a simple unigram, i.e., $p(w_i)$ exists but w_{i-1} not seen

conf(i) = 0.1 if w_i is completely UNKNOWN,

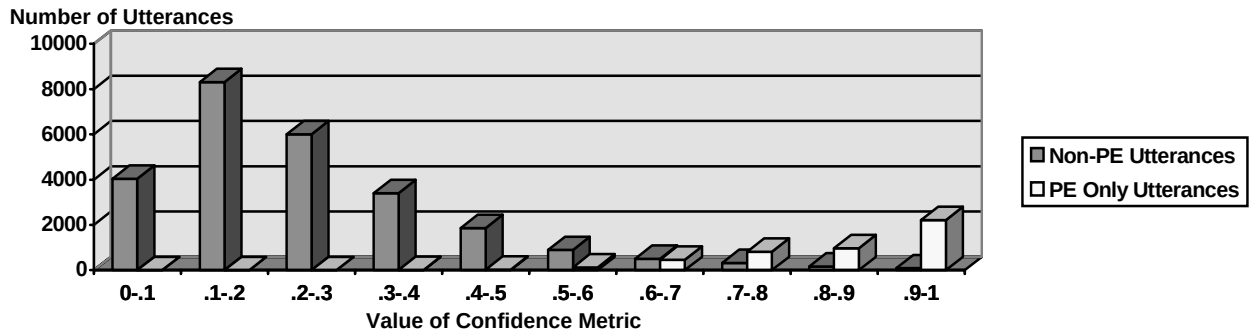

Figure 1. Distribution of Confidence Values for 25,585 Non-Physical-Exam Utterances and 4,599 Physical Exam Utterances

One accounts for the reduced reliability of *runs* of bigram and unigram backoffs by taking the confidence of an N-gram's environment,

$$CONF(i) = conf(i-2)*conf(i-1)*conf(i) .$$

The confidence of an utterance is then the average of word confidences, CONF(i),

$$CONF(utt=\{w_1w_2...w_n\}) = \Sigma\, CONF(w_i) / n .$$

3. EXPERIMENTS WITH THE UTTERANCE-LEVEL CONFIDENCE METRIC

We used the utterance-level confidence metric in an experiment to differentiate in-domain utterances from out-of-domain utterances. The test set data consisted of transcripts of patient reports dictated by physicians. A subset of the utterances in the reports comes from physical examinations, which is effectively a separable subdomain of the larger medical records domain. The physical exam utterances were labeled as in-domain utterances, and all utterances from other subsections of the reports (history of present illness, chief complaint, medications, problem list, etc.) as out-of-domain utterances. A trigram language model was trained from the physical exam utterances, and the confidence metric was computed for all utterances based on this language model.

Experimental results (Figure 1) show that the distributions of the confidence metric for in-domain and out-of-domain utterances are indeed well separated. A threshold value of 0.55 for the confidence metric can be used to discriminate in-domain utterances from out-of-domain utterances with a total error of 7.5% for false positives and false negatives combined.

4. A BACKOFF-BASED WORD-LEVEL CONFIDENCE MEASURE

Using the utterance-level metric allows hypothetical utterances coming out of a speech recognizer to be highlighted as *suspicious* utterances relative to the language model, giving a system the ability to semiautomatically detect where corrections are needed. However it is even more useful to know exactly which words are in error. Towards this end, a closely related word-level-confidence metric is adapted from the utterance-level metric.

Again, one accounts for runs of bigram and unigram backoffs by considering a three word window around a particular word, w_i, and assigning the confidence of that word as follows:

$$CONF(i) = conf(i-1)*conf(i)*conf(i+1) ,$$

Note, the difference from the previous definition of CONF(i) is the *centering* of the window. This allows the words before or after the word in question to have equal influence on the confidence.

5. EXPERIMENTS WITH THE UTTERANCE-LEVEL CONFIDENCE METRIC

Preliminary results using 850 Physical Exam utterances run through our modified SPHINX II recognizer indicate that while the confidence metric is not perfect at detecting misrecognized words, it does have some discriminating ability. Considering incorrectly recognized words with high confidence as false positives and correctly recognized words with low confidence as false negatives, the false positive and false negative rates bottom out at 20% for a threshold of 0.55 for the confidence metric, assuming equal scoring weight for false positives and false negatives. I.e., in 80% of the cases where the recognizer introduced an error, the condition CONF<0.55 identified the error word, and in 20% of the cases where no error occurred, the metric falsely signaled an error.

Errors naturally affect not just one word but a series of words. Consider the following example,

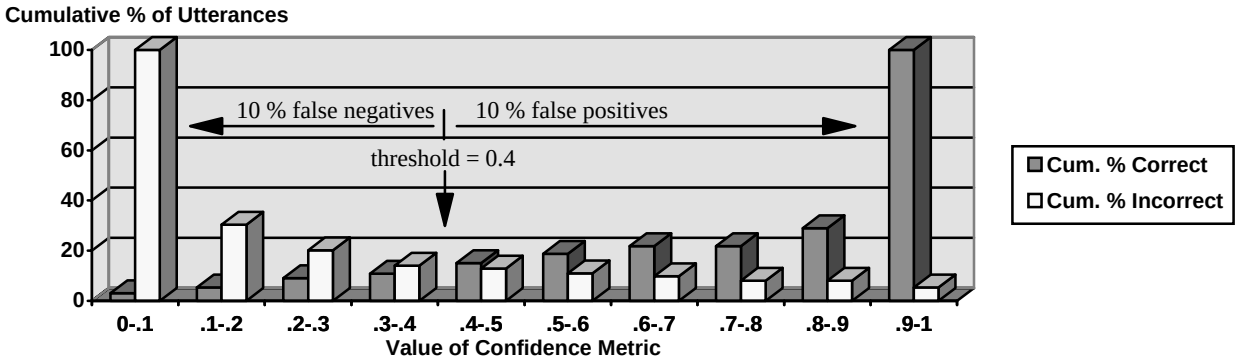

Figure 2. False Positives and False Negatives for 885 Physical Exam Utterances for the Word-Level Confidence Metric

Given:
the patient is awake alert and oriented times three

Recognizer:
the patient _ awake alert worrying times three

For the sequence *the patient awake* and similarly for *alert worrying times*, the confidence metric will signal an aberrant situation. However, the causes for the aberration are quite different - in one case, a deletion error and in the other case, a substitution error involving more than one word. Thus, using the confidence metric, one can mark a phrase sequence that is odd, but one can not reliably say a-priori exactly which word is wrong. Because the confidence metric only indicates that a word stands a low chance of following other words, the process of marking low confidence words as errors is easily confused by deletion or insertion errors and multiword substitutions on the part of the recognizer, disabling it from indicating the exact spot where an error occurs.

This observation suggests a better use of the metric. When the metric is low, one should consider the cause. It could be that the word before or the word after is really the source of an error. Accordingly, the worst case within a sliding window of three words was considered in using the word-level confidence metric:

$$\begin{aligned} CONF'(i) = \text{Min} (& conf(i-2)*conf(i-1)*conf(i) , \\ & conf(i-1)*conf(i)*conf(i+1) , \\ & conf(i)*conf(i+1)*conf(i+2)) , \end{aligned}$$

Furthermore, one gets the additional information of which of the 3-word sequences generated the deflated confidence value.

Modifying the experiment to implement this style of error identification, both false positives and false negatives fall to 10% at a threshold value of about 0.4 (see Figure 2).

6. USING THE CONFIDENCE METRICS IN A SPEECH RECOGNITION SYSTEM

Since the experiments are somewhat complicated by the windowing procedure, it serves to clarify them by considering how they are actually used in a number of examples in the context of an integrated speech recognition system.

We adapted a color coding scheme to alert the user to particular words or whole utterances that are questionable. Low confidence words (CONF´(i)<0.4) are highlighted RED (here presented as shaded text), and utterances that have a low utterance-level confidence (CONF(utt) <0.55) are in ITALICS. In this way the user is immediately alerted to potentially odd utterances. But what does a user do with this feedback?

One scenario is that the user has uttered an out-of-domain utterance. E.g.,

the patient has a history of hypertension and diabetes

In this case, so many runs of low-confidence word-neighborhoods leave little doubt that the whole utterance is worth questioning. The user has attempted to enter an utterance which properly belongs to the medical history subdomain rather than to physical exams.

Consider another example where so many of the subphrases in the utterance have never been encountered in training the language model that the utterance as a whole is questioned:

the patient walks with a very pronounced limp

This happens to be a slightly atypical, but still acceptable instance of an observation falling under the heading of *general patient description* within a physical exam. Naturally, this category of utterances is the most widely variable of any others, and thus it benefits from augmenting the language model training set with exemplars of such atypical or not-often-seen utterances.

Another scenario is that subsets of words are highlighted in red, indicating some sequences of words have low confidence, e.g., in the two following misrecognitions:

midline lumbar scar secondary to is previous surgery

tonsils were essentially normal for her hate

In these cases, the user must make corrections, but having problem areas already highlighted makes it considerably easier than a thorough proofreading. Again, the language model can benefit from retraining. A log file of corrected utterances is automatically generated during interactive edits and used for off-line retraining of the model, so that when such utterances are re-encountered in the future by the recognizer, its language model will have higher weight for utterances previously missed.

7. CONCLUSION

Two confidence metrics based on the backoff behaviors of trigram language models are introduced here. The one metric identifies neighborhoods of words that are likely to have been misrecognized; the other identifies utterances that are likely to be out-of-domain utterances. The metrics take into account that trigrams are more reliable than bigrams, bigrams are more reliable than unigrams, runs of bigrams are even less reliable, and runs of unigrams are least reliable of all.

Experiments show that the utterance-level confidence metric performed well, with only a 7.5% total error for false positives and false negatives. The word-level confidence metric does not perform as well partly because of the confusion that arises regarding exactly which word is in error when the recognizer makes insertions and deletions. Thus, we adapt it with a windowing strategy in order to highlight neighborhoods of words where errors are likely to have occurred. In this way, the number of total errors (false positives + negatives) is cut from 20% to 10%.

The language model is only one component of a larger picture. Additional confidence sources should be exploited - e.g., lower level acoustic confidence and higher level dialog or semantic-context based confidences.

8. ACKNOWLEDGMENT

This project is supported in part by an ATP Cooperative Agreement, Number 70NANB5H1184, from the National Institute of Standards and Technology.

9. REFERENCES

[1] Chase, L., Rosenfeld, R., and Ward, W., *"Error-Responsive Modifications to Speech Recognizers: Negative N-grams"*, ICSLP 1994.

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

[3] Ravishankar, M.K., *"Efficient Algorithms for Speech Recognition"* Unpublished PhD Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, May 1996.

[4] Katz, S., *"Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer"*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-35 pp.400-401, 1987.