Multi-Task Learning for End-to-End ASR Word and Utterance Confidence with Deletion Prediction

David Qiu¹, Yanzhang He¹, Qiujia Li², Yu Zhang¹, Liangliang Cao¹, Ian McGraw¹

¹Google, LLC, USA, ²University of Cambridge, UK
¹{q david, yanzhanghe, ngyuzh, llcao, imcgraw}@google.com, ²ql264@cam.ac.uk

Abstract

Confidence scores are very useful for downstream applications of automatic speech recognition (ASR) systems. Recent works have proposed using neural networks to learn word or utterance confidence scores for end-to-end ASR. In those studies, word confidence by itself does not model deletions, and utterance confidence does not take advantage of word-level training signals. This paper proposes to jointly learn word confidence, word deletion, and utterance confidence. Empirical results show that multi-task learning with all three objectives improves confidence metrics (NCE, AUC, RMSE) without the need for increasing the model size of the confidence estimation module. Using the utterance-level confidence for resoring also decreases the word error rates on Google’s Voice Search and Long-tail Maps datasets by 3-5% relative, without needing a dedicated neural rescoring.

Index Terms: automatic speech recognition, confidence estimation, deletion prediction, multi-task learning, Transformer

1. Introduction

Confidence scores are often used to assess the reliability of speech recognizers [12]. They have been widely used for various downstream tasks, including semi-supervised and active learning [4, 8, 9, 20, 21, 22], system combination [9, 10, 11], dialog systems [12, 13, 14] and keyword spotting [15, 16, 17]. For traditional hidden Markov model (HMM)-based automatic speech recognition (ASR) systems, the simplest approach is to use the posterior probability for each hypothesized word. More advanced methods have been proposed to better estimate word-level confidence, including linear models [11, 18], conditional random fields [19], and neural networks [20, 21, 22, 23].

However, for end-to-end (E2E) ASR models such as recurrent neural network transducers (RNN-T) and attention-based sequence-to-sequence models, word posterior cannot be approximated well from the tree-like “lattice” where the prediction of each token conditions on the full history of previous tokens. Autoregressive decoders also tend to be overconfident [24]. To solve this challenge, several model-based methods have been proposed to estimate word and utterance-level confidence for E2E models. For example, [25] proposed to train a token-level (e.g. graphemes or word-pieces [26]) confidence estimation module (CEM) on top of a given E2E model and the word-level confidence can be simply obtained by averaging the token-level scores. Since word-level confidence is more commonly used for applications such as keyword spotting [15, 16, 17] and systems combination [9, 10, 27] proposed to train CEMs for word-level confidence directly, where issues such as multiple possible word-piece tokenizations are addressed. Utterance-level confidence scores are also important for applications such as data selection [4, 6, 7]. One can directly average the word-level confidence scores for utterance-level confidence, but averaging is sub-optimal since deletion errors are not accounted for. [28, 29] proposed to directly learn utterance-level confidence scores for E2E systems and [29] showed that rescoring n-best hypotheses based on utterance-level confidence estimator can also reduce word error rates (WER). Previous works generally model confidence scores at a single level (i.e. token, word or utterance). This paper explores joint optimization of word-level and utterance-level confidence estimation in a single model with multi-task learning.

Model-based approaches typically formulate the confidence estimation problem as a binary classification task [20, 21, 22, 23, 24, 25, 27] where correct tokens, words, or utterances should have confidence scores close to 1, and 0 otherwise. For word-level confidence estimation, scores between 1 and 0 are only assigned to words that appear in the hypotheses. This covers substitution and insertion errors for ASR. However, deletion errors are either ignored or not explicitly addressed. In other words, averaging the confidence scores yields an estimate for the word correct ratio (WCR), but deletion errors need to be considered for estimating WER. As shown in [20, 22], deletion estimation is also important for downstream tasks. For example, even with a good confidence estimator, utterances with high confidence may have high WER because of deletion errors. Tasks such as semi-supervised training can be adversely affected without explicit modeling of deletion errors [24]. In this work, we explore estimating deletion errors along with word-level and utterance-level confidence scores in a single model.

This paper makes the following contributions: 1) Analyzes the word-level CEM in [27] to show that it mitigates the over-confidence problem when subword deletions are present. 2) Proposes using neural networks and Poisson regression to learn arbitrary deletion length, and integrate it with the word-level CEM to estimate WER. 3) Demonstrates that jointly learning the word confidence, utterance confidence, and deletion length prediction tasks improves all reported confidence metrics over single-task word-level or utterance-level CEMs. 4) Shows that using the utterance-level confidence to resore RNN-T hypotheses reduces WERs on industry-scale datasets.

2. Model

2.1. Notation

In general, an E2E ASR model processes a sequence of acoustic features a, and outputs a token sequence [y₁, ..., y_M] as its hypothesis. Popular choices for the set of tokens include graphemes and word-pieces (WP). In this paper, we focus on WP. The goal of the confidence model is to use both a, and the embedding of the WP sequence, b, as inputs, and output a score.
for every word and every utterance. To distinguish between WP and words, we use $M$ and $L$ to refer to the number of WPs and words in a hypothesis, respectively. We use $y_i$ and $w_j$ to denote the $i$-th WP and $j$-th word, respectively.

We use the Transformer $[31, 32]$ as the base feature extractor for all of the proposed tasks:

$$[f(y_1), \ldots, f(y_M)] = \text{Transformer}(\text{CA}(a), \text{SA}(b)), \quad (1)$$

where the cross-attention block $\text{CA}$ attends to $a$, and the self-attention block $\text{SA}$ attends to $b$. We refer to the Transformer’s output at the $i$-th token as $f(y_i)$.

### 2.2. Baseline Word-Level Confidence Estimation

For the baseline CEM, we use a multi-layer perception (MLP) to process the Transformer’s output into 3-D outputs:

$$[c(y_i), i(y_i), s(y_i)] = \text{Softmax}(\text{MLP}(f(y_i))). \quad (2)$$

For training, we adapt the word-level training loss introduced in $[22]$. To recap, the loss gathers the output of $[2]$, but only at the indices corresponding to the last WP of each word (see Table 1). The gathered output can be re-indexed as $[c_w(w_1), \ldots, c_w(w_L)], [i_w(w_1), \ldots, i_w(w_L)],$ and $[s_w(w_1), \ldots, s_w(w_L)]$, to denote the probabilities that the words are tagged by Levenshtein edit distance $[33]$ as “correct”, “insertion”, and “substitution”, respectively. Then, over every word, we compute the cross-entropy loss $\mathcal{L}_{\text{word}}$ between the ground truth tags and the outputs:

$$- \sum_{j=1}^{L} \left[ y_{j}^{\text{cor}} \log c_w(w_j) + y_{j}^{\text{ins}} \log i_w(w_j) + y_{j}^{\sub} \log s_w(w_j) \right].$$

(3)

This also yields a simple estimate for the WCR:

$$\hat{\mu}_{\text{w}} = \frac{\sum_{j=1}^{L} c_w(w_j)}{L}. \quad (4)$$

Although word-level deletions are not modeled, the subword/WP deletion problem is implicitly addressed by the loss. Table 1 shows an example where the first word in the hypothesis contains a WP deletion. WP edit distance completely ignores any WP deletions because deletions are not part of the hypothesis. On the other hand, because the word-level CEM is trained with the word-level edit distance serving as the ground truth, hypothesized words that contain WP deletions are guaranteed to be labeled as word-level substitutions. This resolves an overconfidence problem that is common in token-level CEMs trained to model WP edit distance, especially for a smaller set of tokens such as graphemes.

### 2.3. Deletion Length Prediction

To estimate the WER, the model needs to estimate the number of deletions as a secondary task. Here, we define a sequence of deletion length random variables for $1 \leq j \leq L + 1$ as:

$$E_j = \begin{cases} 
\#(\text{del}) \text{ before the first word}, & \text{if } j = 1 \\
\#(\text{del}) \text{ after the last word}, & \text{if } j = L + 1 \\
\#(\text{del}) \text{ between } j - 1 \text{ and } j\text{-th word}, & \text{otherwise}.
\end{cases}$$

(5)

The number of deletions can be any non-negative integer. Existing works such as $[22]$ converts the problem into binary prediction by clipping $E_j$ at 1. In contrast, we predict $E_j$ directly by modeling $P(E_j | f(w_j); \theta)$ to follow the Poisson distribution, which is well-suited for count data. We use a MLP to parameterize it and predict the mean of $E_j$ as below:

$$E_j = \sum_{j=1}^{L} c_w(w_j) L.$$  

$$r_w(w_j) = \text{MLP} (f(w_j); \theta) \quad (6)$$

Let $e_j$ be the ground truth realization of $E_j$. With maximum likelihood estimation for Poisson regression, the training loss

![Diagram](image_url)
for deletions is given by:
\[ \mathcal{L}_{\text{deletion}} = - \sum_{j=1}^{L+1} \left[ e_j r_w(w_j) - \exp(r_w(w_j)) \right]. \] (7)

During inference, the WER estimate can be computed as
\[ \mu_{\text{total}} = \frac{\hat{\mathcal{D}} + \hat{T} + \hat{S}}{L + \hat{D} - \hat{T}}, \]
where \( \hat{D} = \sum_{j=1}^{L+1} \exp(r_w(w_j)), \hat{T} = \sum_{j=1}^{L} \nu(w_j), \) and \( \hat{S} = \sum_{j=1}^{L} s_w(w_j). \)

### 2.4. Utterance-Level Confidence Estimation

A coarse way of including deletion information is to define a tertiary utterance-level prediction task, with the ground truth
\[ e_{\text{att}} = \begin{cases} 1, & \text{if utterance WER} = 0 \\ 0, & \text{otherwise.} \end{cases} \]

The presence of any deletions causes the value of \( e \) to be 0, and this signal is backpropagated to the internal features in the Transformer feature extractor to help the word-level confidence task. This objective was proposed in [28], albeit in a single-task setup that ignores word confidences. Intuitively, an utterance feature is a summary of the sequence of WP features. Thus, to extract the utterance features and make the prediction, we use the hierarchical attention proposed in [34]:
\[ u_i = \tanh(W_1 f(y_i) + b) \]
\[ \alpha_i = \frac{\exp(u_i^T u_m)}{\sum_{m=1}^{M} \exp(u_i^T u_m)} \]
\[ \hat{\mu}_{\text{att}} = \sigma \left( \text{MLP} \left( \sum_{i=1}^{M} \alpha_i f(y_i) \right) \right). \] (10)

The parameters \( W_1, b, \) and \( u_2 \) that generate \( \hat{\mu}_{\text{att}} \) can be trained with the binary cross-entropy loss:
\[ \mathcal{L}_{\text{att}} = - [e_{\text{att}} \log \hat{\mu}_{\text{att}} + (1 - e_{\text{att}}) \log(1 - \hat{\mu}_{\text{att}})]. \] (11)

\( \hat{\mu}_{\text{att}} \) provides an estimate of the probability that the entire utterance is recognized with zero WER, which is useful for ranking utterances in the order of their confidences.

### 2.5. Multi-task Objective

The model can be trained by combining [3], [7], and [11]:
\[ \mathcal{L}_{\text{total}} = \frac{1}{L} \mathcal{L}_{\text{word}} + \frac{\lambda_{\text{deletion}}}{L+1} \mathcal{L}_{\text{deletion}} + \lambda_{\text{att}} \mathcal{L}_{\text{att}}. \]

In this paper, \( \lambda_{\text{deletion}} = 0.5, \lambda_{\text{att}} = 1 \), when the respective losses are active, and \( \mathcal{L}_{\text{total}} \) is further normalized over the mini-batch.

### 3. Experimental Setup

#### 3.1. Model Architecture

The ASR’s architecture is a RNN-T that follows [35]. It has 8 LSTM layers in the encoder and 2 LSTM layers in the prediction network. Each LSTM layer is unidirectional, with 2,048 units and a projection layer with 640 units. A time-reduction layer is added between 2nd and 3rd encoder layers, which halves the sequence length by concatenating every two input frames. The RNN-T is frozen during CEM training.

#### 3.2. Dataset

We report the performance on the five combinations of confidence training tasks shown in Table 2. The feature extractor in [1] consists of two Transformer decoder blocks with input embedding and internal dimensions of 640. For \( \mathcal{L}_{\text{deletion}} \), the MLP in [6] has three layers, with hidden and output dimensions of 320, 160, and 1. For \( \mathcal{L}_{\text{att}} \), we use the hierarchical attention mechanism in [8] and [9] to generate the utterance-level feature. The MLP in [10] has two layers, with hidden and output dimensions of 320 and 1, respectively. All training is performed in TensorFlow with the Lingvo [36] toolkit, using four hypotheses and random data downsampling to 8kHz [40] are applied to increase training data diversity.

The main test set includes ~14K Voice Search (VS) utterances extracted from Google traffic, which is anonymized and hand-transcribed. To test the generalization of the CEM, we create a list of proper nouns identified by an in-house named entity tagger that are common in the LM training data for conventional ASR but rare in the audio-text paired training data for E2E ASR models. We select 10K sentences from the LM test data for the Maps domain, such that each sentence contains at least one of these proper nouns, then synthesize the TTS audio (as in [41]) for them to create the Long-tails Maps test set.

#### 3.3. Evaluation Metrics

All evaluation metrics are reported on the top hypothesis. For word confidence, we adopt the same metrics in [27]: normalized cross-entropy (NCE) [42], area under the receiver operating characteristic curve (AUC-ROC) and the negative class precision recall curve (AUC-PR). Higher is better for these metrics.

To measure the ranking quality of the utterance confidence, we use the utterance-level AUC-ROC and AUC-PR (higher is better), where the positive / negative classes are the sets of utterances with zero / non-zero WER, respectively. To measure the accuracy of estimating WERs, we use the root mean squared error (RMSE, lower is better) between the utterance confidence and the ground truth (1 – WER).

Table 3 shows the availability of utterance-level confidence scores, and which one is preferred when multiple are available. Recall that \( \hat{\mu}_{\text{att}} \) models the probability that the utterance is recognized with zero WER. Thus, we prioritize using it for the utterance ranking tasks (AUC-ROC, AUC-PR). However, \( \hat{\mu}_{\text{att}} \) does not attempt to estimate the WER, and we de-prioritize it for the WER accuracy task.
4. Results

4.1. Joint Training Confidence Metrics

This section discusses the performance of all models on the confidence metrics. In Table 3, comparing Models $\hat{W}$, $\hat{U}$, $\hat{W}D$, and $\hat{W}U$ vs Model $\hat{W}$ shows that any joint training (especially with the utterance-level loss) helps improve word-level metrics. Comparing Models $\hat{W}U$ and $\hat{W}D$ vs Model $\hat{U}$ shows that joint training also helps improve the performance of $\hat{W}_u$ on utterance-level AUC-ROC. Comparing Models $\hat{W}D$ and $\hat{W}U$ vs Model $\hat{W}$ shows that modeling deletion and insertion rates helps achieve the most accurate WER estimate in terms of RMSE.

Since all five models have similar number of parameters (see Table 2), it is clear that the three tasks do not significantly increase the required model capacity, and their training signals complement each other to improve overall confidence accuracy.

4.2. Qualitative Example

To visualize the benefit of adding $L_{deletion}$ and $L_{edit}$, Table 5 shows an example utterance that contains a deletion. Model $\hat{W}UD$, through learning the utterance correctness and deletion lengths at every position, is able to use that knowledge to aid word confidence predictions. That is seen from the lower confidence score at the word “absolut”, which follows a deletion.

4.3. Rescoring Using Utterance Confidence

Figure 2 shows the improved utterance-level ROC and PR curves for the multi-task Model $\hat{W}U$ over single-task Models $\hat{W}$ and $\hat{U}$.

5. Conclusions

In this paper, we propose to extend the word-level CEM to jointly learn the length of deletions between every pair of words and the probability that the entire utterance is recognized correctly. Experimental results show that the proposed multi-task learning setup improves word and utterance-level confidence metrics without incurring significant model size increase. The utterance confidence can also be used to rescore first-pass RNN-T hypotheses and reduce WER by 3-5% relative.
