DETECTING DELETIONS IN ASR OUTPUT

M.S. Seigel and P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
{mss46,pcw}@eng.cam.ac.uk

ABSTRACT

In this work, the novel task of detecting deletions within automatic speech recognition (ASR) system output is investigated. Deletion-informed confidence estimation is proposed as an approach which simultaneously yields a confidence score in a word being correct, as well as a deletion confidence score which indicates whether a deletion is likely to occur in the output. The sequential nature of conditional random field (CRF) models is exploited as a means through which this can be achieved. It is shown that this sequence structure is crucial in yielding useful deletion detection scores, with an equivalent non-sequential model proven to be unsuitable for the task. The deletion-informed confidence estimation approach is also shown to outperform one where deletion confidence scores are estimated as a classification task separate from that of overall confidence estimation.

***Index Terms*—** Deletion detection, conditional random fields, confidence estimation.

1. INTRODUCTION

In the literature on confidence estimation for ASR systems, many approaches make use of classifiers to estimate a confidence score by performing binary classification, e.g. [1, 2, 3, 4, 5, 6]. The premise in this approach being that words may be classified as either being correct or incorrect, and the probability of the word being correct may be taken as the confidence score. The definition of incorrect (or erroneous) words is limited to those for which the underlying ASR system has produced some output, as information or features can be obtained for such events. This implies that only substitution and insertion errors may be considered. In this work, the definition of errors within the confidence estimation task is extended to include that of deletions. By modelling this type of error in addition to those typically considered, it is possible to simultaneously estimate a confidence score in a word being correct, as well as a deletion confidence score indicating whether a deletion is likely to occur. This will be referred to as deletion-informed confidence estimation. To the authors' knowledge, this type of error has not yet been considered as part of the confidence estimation problem in the existing literature. Many downstream applications stand to benefit from the availability of deletion confidence scores. For example, this may be particularly useful in applications where words may be deleted from information-bearing "slots", such as in information extraction and dialogue systems.

This paper begins by defining the task and introducing an approach using conditional random fields (CRFs) [7] for deletion-informed confidence estimation. Thereafter, the predictor features used as input to this model are detailed. Results of applying this approach in a large-scale ASR system to estimate deletion confidence scores are then presented.

2. CRF MODELS FOR COMBINED CONFIDENCE AND DELETION MODELLING

In a direct classification-based approach to word-level confidence estimation, a set of predictor features is associated with each word in the ASR output. This information is used as input to a classifier, which estimates a measure of confidence in that word being correct. If a word is not hypothesised by the ASR system, there is no evidence for that word having occurred. Consequently, no predictor features can be associated with such a deleted word. In the proposed approach, this unseen event in the observed output is modelled through exploiting the transition (or sequence) structure of CRF models. Deletion regions are defined here as regions within the output in which one or more deletions may occur. The information encoded in the sequence of hypothesised words and their associated features is used to predict when transitioning from an existing word hypothesis, which includes both correct and incorrect words, into such a deletion region which may have a single or multiple consecutive deletions. Deletions do not compete with or replace words in the hypothesis, as may be the case with null words in confusion networks.

The framework for confidence estimation with CRF models described in previous work [8, 9] is adapted for this new task. The general expression for a CRF which models the sequence of discrete labels $\mathbf{Y}$ conditioned on a sequence of input feature vectors $\mathbf{X}$ is the following:

This work was in part supported by a grant from the Nuance Foundation. The paper does not necessarily reflect the position or the policy of the Nuance Foundation, and no official endorsement should be inferred.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

$$p(\mathbf{Y}|\mathbf{X}) \propto \exp\Big(\sum_k \lambda_k t_k(\mathbf{Y}) + \sum_l \mu_l g_l(\mathbf{Y},\mathbf{X})\Big)$$

where the k transition feature functions $t_k(\mathbf{Y})$ with parameters λ_k represent transitions between labels in $\mathbf{Y}$, and the l observation feature functions $g_l(\mathbf{Y},\mathbf{X})$ with parameters μ_l relate the input features $\mathbf{X}$ to the labels $\mathbf{Y}$. An illustration of how this model is applied to the task for a synthetic example is shown in Figure 1.

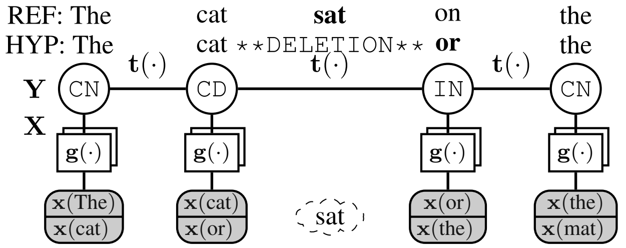

Fig. 1. Illustration of the CRF-based deletion-informed confidence estimation approach using a synthetic training example.

A deletion region may occur following any word. This is represented by augmenting the standard set of confidence estimation labels Y in the sequence $\mathbf{Y}$, indicating whether a word is correct (C) or incorrect (I), with an indicator of whether the current word occurs before a deletion (D) or not (N). For instance, in Figure 1, the label $Y =$ CD corresponds to a word which is correct ("cat"), and occurs before a deletion region (for the word "sat"). It should be noted that separate labels are used to indicate when a deletion occurs before the first word in a hypothesis. The transition feature functions, $\mathbf{t}(\cdot)$, allow the model to capture the desired characteristics of transitions into and out of deletion regions, as well as that of the label sequences (e.g. multiple consecutive correct/incorrect/deleted words). The observation feature functions, $\mathbf{g}(\cdot)$, relate the observed predictor features for each word ($\mathbf{x}$) to the current label Y. Spline feature functions [10] are used to represent the continuous features. The approach defined in this way implies the model makes a decision on whether a deletion should occur after the current time. The feature vector for each word is therefore extended to include the predictor features for the next word as well as the current word. This is intended to provide some context to the decision on the placement of deletion boundaries. It should also compensate for the asymmetric nature of the approach in predicting whether a deletion immediately follows a word, but not whether it could precede a word.

The labels encode information pertaining to both whether or not the word is correct, as well as whether a deletion region will occur. In this deletion-informed confidence estimation approach, standard confidence scores as well as deletion confidence scores may therefore be obtained by computing the marginal probability of the required labels, at each point in the word sequence.

2.1. Predictor features

Instead of an exhaustive set of predictor features, a selection of word and sub-word-level posterior probabilities that proved useful in previous work [9], are input to the CRF model to estimate standard confidence and deletion confidence scores. The first of these is the word-level posterior computed from the recognition lattice, similar to that in [11, 12]. This posterior probability is computed for a word W_i at position i within a sequence of words in an ASR hypothesis, and is referred to as the lattice arc posterior ratio (LAPR):

$$\text{LAPR}(W_i, \mathcal{I}) = \frac{\sum_{A\in\mathcal{I}} \delta(\texttt{word}(A), W_i) p(A|\mathbf{X})}{\sum_{A\in\mathcal{I}} p(A|\mathbf{X})}$$

where $\mathcal{I}$ is the set of arcs in the lattice which intersect with the word hypothesis for W_i and A is an arc within this set. The Kronecker delta function (δ) matches the word identities of the arc $\texttt{word}(A)$ with the word W_i, and $p(A|\mathbf{X})$ is the posterior probability computed for arc A during a forward-backward pass over the lattice. The word-level lattices may be marked up with sub-word timing (context-dependent graphemes in this work). Given the resulting set of sub-word arcs, a sub-word-level posterior referred to as the lattice sub-arc posterior ratio (LSAPR) can be computed for each sub-word U_i in the hypothesis:

$$\text{LSAPR}(U_i, \mathcal{I}) = \frac{\sum_{S\in\mathcal{I}} \delta(\texttt{sub-word}(S), U_i) p(S|\mathbf{O})}{\sum_{S\in\mathcal{I}} p(S|\mathbf{O})}.$$

where $\mathcal{I}$ is the set of sub-word arcs in the lattice which intersect with the sub-word hypothesis U_i and S is a sub-word arc within this set. The Kronecker delta function (δ) matches U_i with the context-independent sub-word $\texttt{sub-word}(S)$ for arc S, and $p(S|\mathbf{O})$ is taken as the posterior probability of the word arc to which the sub-word arc S belongs. This predictor feature is averaged over the sub-words in a word hypothesis to produce the average lattice sub-arc posterior ratio (ALSAPR). An additional frame-level sub-word classifier is also used to provide complementary information to the sub-word scores computed using the lattices produced by the underlying recogniser. The posterior probability of a sub-word U_i hypothesised by the ASR system on the interval from t_b to t_e, is computed using the alternative system to yield the alternative sub-word posterior (ASWP):

$$\text{ASWP}(U_i, t_b, t_e) = \frac{\sum\limits_{t\in[t_b,t_e]} \sum\limits_{S\in\mathcal{S}} \delta(\texttt{sub-word}(S), U_i) p(S|t)}{\sum\limits_{t\in[t_b,t_e]} \sum\limits_{S\in\mathcal{S}} p(S|t)}$$

where t is a frame, S is a sub-word targets in the complete set of sub-words $\mathcal{S}$, and $p(S|t)$ is the posterior probability output by the alternative recogniser at time t for the sub-word S. This predictor feature is averaged over the sub-words comprising a word-level hypothesis to yield the average alternative sub-word posterior (AASWP) predictor feature.

3. EXPERIMENTS

A number of experiments were carried out in which deletion-informed confidence estimation models are developed and evaluated. This is performed for the 1-Best Viterbi output from a recogniser which is part of the 2010 Cambridge Arabic ASR system [13]. A system which uses a graphemic representation at the sub-word-level is used. The acoustic training data used consists of 1538 hours of audio, comprising broadcast news and conversation data. The language model is trained from 1.2G words, with a vocabulary of 350k words. The decoding structure includes multiple passes, the first two of which are the main lattice generation phase, with adaptation applied in subsequent passes. Experiments are based on the output lattices from the second decoding pass, which are typically dense and therefore informative.

Subsets of the 2010 GALE development data, a subset of the 2009 GALE development data, and a subset from the 2009 GALE evaluation data were used. The training dataset consists of three of these datasets and includes 9700 utterances. The dev10d and eval09ns datasets were held out for evaluation, and have error rates of is 32.8% and 14.1% respectively. The alternative recogniser used for additional sub-word-level predictor features is an MLP-based grapheme classifier, with 37 context-independent grapheme targets. This was trained on 140 hours of Arabic broadcast news data (the GALE p4r3 dataset) using the ICSI QuickNet MLP software [14]. The cross validation accuracy of this classifier is 66.59%.

The training data is scored to produce labels indicating whether each word is correct or incorrect. Where deletions occur, the label of the preceding word is augmented with the deletion marker. A baseline similar to that described in [15], which uses a decision tree with a piecewise linear mapping of the baseline predictor feature (LAPR) was developed.

3.1. Evaluation

A metric commonly used in confidence estimation evaluation is the normalised cross entropy (NCE) metric [16]. This metric is a normalised representation of the information gain in assigning confidence scores to the set of correct and incorrect words, over assuming the empirical accuracy of the system is the likelihood of each word being correct. The deletion-informed confidence estimation approach developed to model deletion regions is capable of estimating "classical" confidence scores, which will be evaluated using the NCE metric.

As the task of having a confidence measure in deletions has not previously been studied (to the authors' knowledge), there is no standard metric for evaluating performance. It is assumed that there is the potential for a deletion region to exist in a slot after every word hypothesised by an ASR system, as well as before the first word in every utterance. The goal in detecting deletions is therefore formulated as being that of estimating a probability with which a deletion region is likely to occur in each such slot. This definition is analogous to that in the "classical" confidence estimation task concerned with word correctness. A modified version of the NCE metric is therefore used for this purpose. This modified metric represents the information gain in assigning scores to each of the potential deletion regions in the slots where these may occur, over assuming these scores are equal to the empirical deletion rate on the test data for the ASR system. This results in the definition of the DNCE metric, defined as follows:

$$\text{DNCE} = \frac{H_m + \sum\limits_{w \in \mathcal{D}} \log_2(\hat{p}_d(w)) + \sum\limits_{w \in \mathcal{N}} \log_2(1 - \hat{p}_d(w))}{H_m}$$

where $\hat{p}_d(w)$ is the confidence in a deletion occurring in a particular slot w, $\mathcal{D}$ is the set of slots in which deletions occur and $\mathcal{N}$ is the set of slots in which deletions do not occur. Given d slots in which words are deleted from the ASR hypotheses, out of N total slots where deletions could possibly occur, the empirical average probability of a deletion region being present is $P_d = \frac{d}{N}$. The empirical entropy is:

$$H_m = -d \log_2(P_d) - (N - d) \log_2(1 - P_d). \qquad (1)$$

A DNCE score of around zero indicates that a system performs similarly to one which has knowledge of the empirical deletion probability on the test data, and uses this probability as the deletion confidence score. Ideal performance corresponds to a DNCE score of 1.

3.2. Results

Experimental results for CRF-based deletion-informed confidence estimation are shown in Table 1. Here, the decision tree baseline is made up of two systems, one trained specifically for the classical confidence estimation task and one for the deletion detection task. The NCE scores for standard confidence estimation are seen to be comparable to other results in the table. However, the performance in terms of the deletion metric (DNCE) is poor, with the scores being negative. This indicates that the baseline system in fact performs worse than a naïve system which assumes the empirical deletion rate of the system is equal to the actual average probability (on the test data) of a deletion occurring after each word. This is to be expected, as there is not necessarily a high degree of correlation between the current word-level posterior used by this model (LAPR), and the likelihood of a deletion following the current word. Moreover, this model is unable to capture the sequence information which is vital in this task, and is therefore an unsuitable modelling approach. This result is verified by making use of a non-sequential maximum entropy model in which the LAPR feature is represented using spline feature functions. This system yields similar performance to the decision tree baseline, proving that the sequence information is crucial. Approaches which attempt to classify the onset of deletion regions based purely on the observations for a current word are therefore clearly not able to perform this task.

System	dev10d		eval09ns	
	DNCE	NCE	DNCE	NCE
Baseline LAPR	-0.047	0.314	-0.019	0.358
MaxEnt LAPR	-0.044	0.312	0.010	0.351
1: ALSAPR	0.098	0.248	0.093	0.232
2: AASWP	0.034	0.047	0.042	0.056
3: LAPR	0.117	0.334	0.128	0.356
4: AASWP+LAPR	0.127	0.347	0.133	0.361
5: ALSAPR+AASWP	0.101	0.250	0.096	0.232
6: $P(t)$	0.130	0.347	0.137	0.361
7: +$P(t+1)$	0.134	0.347	0.143	0.362
8: MaxEnt $P(t) + P(t+1)$	0.084	0.337	0.128	0.360
9: DELONLY $P(t) + P(t+1)$	0.109	–	0.132	–

Table 1. Results for deletion detection systems, evaluated in terms of the deletion metric (DNCE), and the standard confidence metric (NCE). Results are shown for dev10d and eval09ns. LAPR: lattice arc posterior ratio, ALSAPR: average lattice sub-arc posterior ratio, AASWP: average alternative sub-word posterior. $P(t)$=LAPR(t)+ALSAPR(t)+AASWP(t)). DELONLY is used to depict the system for deletion detection only.

Comparing the CRF-based system (3) which uses the word-level posterior (LAPR) only, with the baseline or the maximum entropy model using the same information, it is clear that the sequential nature of the model contributes to improve performance for this task. This sequential structure is exploited such that deletion regions can be modelled and thereby detected. The results in terms of the NCE metric show that the inclusion of the deletion modelling aspect does not impact negatively on the model's capability to assign confidence scores on word correctness. The CRF-based system outperforms the decision tree baseline by a considerable margin on both the dev10d and eval09ns datasets.

Results for a CRF-based system which makes use of the full set of posterior predictor features P, which includes the lattice arc posterior ratio (LAPR), the lattice sub-arc posterior ratio (ALSAPR) and the alternative recogniser sub-word posterior (AASWP) are shown (system 7). Large improvements over the system making use of LAPR in isolation (system 3) are observed, both in terms of evaluation for the classical confidence estimation task (NCE) and in the deletion task (DNCE). This improvement is 11.1% and 7% relative on dev10d and eval09ns respectively on DNCE, and 3.9% and 1.4% relative on dev10d and eval09ns respectively in NCE. These results highlight the utility of these additional sub-word-level predictor features for both tasks.

The predictor features for the next hypothesised word in the sequence, referred to as the right-context features, are denoted by $P(t+1)$. It is useful for the system to be able to make local decisions on whether there is a deletion following the current word, using information pertaining to both the current word, and the right-context word. The predictor feature vector for each hypothesised word is therefore augmented with the right-context predictor features, yielding CRF systems 7 and 9 in Table 1. The use of these features results in relative improvements in DNCE over a system without them (6) of 3.1% and 4.4% on dev10d and eval09ns respectively. When applying a decision threshold equal to the test set deletion rate to the output, this system is shown to yield an absolute reduction in the error rate (falsely detected and missed deletions) of 15% over the baseline system. Furthermore, it is seen that this CRF system outperforms an equivalent, non-sequential maximum entropy model (8) in terms of DNCE, showing relative improvements of 21.4% and 10.5%.

The results reported for this task show that the performance of the confidence estimation task is not affected negatively by the additional capability of the model to detect deletions, which is useful as a single model can be used to estimate scores for both types of events. To investigate whether the converse is true, models were trained with labels indicating the presence of deletions only, and not word correctness. These are therefore purely deletion detection systems. Results for a system of this type (8) are shown in Table 1. This system yields considerably inferior DNCE performance to that of the combined model. This implies that the information encoded in whether or not a word is correct is useful in predicting deletions, and the combined approach is therefore justified.

4. CONCLUSION

This paper addressed the novel task of detecting deletions in ASR output, and proposed a solution in terms of a CRF-based confidence estimation model. Detecting the presence of deletions was cast as an extension of the confidence estimation problem, as deletion-informed confidence estimation. The proposed approach exploits the sequential nature of the CRF model, to predict when transitioning into deletion regions of one or more deleted words. This approach results in a system which is capable of simultaneously performing confidence estimation in a general sense, whilst estimating a measure indicating whether a deletion region is likely to occur following a word hypothesised by the ASR system. The results showed that the key aspect of this approach is indeed its sequential nature. Non-sequential approaches such as the decision tree and maximum entropy model were shown to be unable to perform this task successfully. The word level posterior (LAPR) proved useful in predicting deletions, with the sub-word-level posterior predictor features contributing to yield further improvements in the accuracy of deletion confidence scores. The ability of the deletion-informed confidence estimation system to estimate accurate "standard" confidence scores does not suffer with the extension to the deletion task. This approach outperforms one which detects deletions without information on word-correctness, showing that the two tasks are closely related and worth modelling simultaneously.

5. REFERENCES

[1] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Proceedings of Eurospeech*, 1997.

[2] Mitch Weintraub, Francoise Beaufays, Ze'ev Rivlin, Yochai Konig, and Andreas Stolcke, "Neural-network based measures of confidence for word recognition," in *Proceedings of ICASSP*, 1997.

[3] L. Gillick, Y. Ito, and J. Young, "A probabilistic approach to confidence estimation and evaluation," in *Proceedings of ICASSP*, 1997.

[4] Daniel Willett, Andreas Worm, Christoph Neukirchen, and Gerhard Rigoll, "Confidence measures for HMM-based speech recognition," in *Proceedings of ICSLP*, 1998.

[5] Alberto Sanchis, Alfons Juan, and Enrique Vidal, "A word-based naıve bayes classifier for confidence estimation in speech recognition," *IEEE transactions on Audio, Speech and Language Processing.*, vol. 20, no. 2, pp. 565–574, 2012.

[6] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Proceedings of ICASSP*, 2007.

[7] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML*, 2001.

[8] Matthew S. Seigel and Philip C. Woodland, "Combining information sources for confidence estimation with crf models," in *Proceedings of Interspeech*, 2011.

[9] Matthew S. Seigel and Philip C. Woodland, "Using sub-word-level information for confidence estimation with conditional random field models," in *Proceedings of Interspeech*, 2012.

[10] Dong Yu, Li Deng, and Alex Acero, "Using continuous features in the maximum entropy model," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1295–1300, 2009.

[11] Gunnar Evermann and Philip C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proceedings of ICASSP*, 2000.

[12] Frank Wessel, Ralf Schlter, Klaus Macherey, and Hermann Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288–298, 2001.

[13] F. Diehl, M. J. F. Gales, X. Liu, M. Tomalin, and P. C. Woodland, "Word boundary modelling and full covariance Gaussians for Arabic speech-to-text systems," in *Proceedings of Interspeech*, 2011.

[14] David Johnson, "ICSI QuickNet software package," 2004.

[15] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proceedings of NIST Speech Transcription Workshop*, 2000.

[16] M. Siu, H. Gish, and F. Richardson, "Improved estimation, evaluation and applications of confidence measures for speech recognition," in *Proceedings of Eurospeech*, 1997.