DIRECT SUB-WORD CONFIDENCE ESTIMATION WITH HIDDEN-STATE CONDITIONAL RANDOM FIELDS

M.S. Seigel and P.C. Woodland

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

ABSTRACT

The estimation of accurate confidence scores for sub-word-level units within automatic speech recognition (ASR) system transcriptions is investigated in this work. This is achieved through the application of linear-chain and hidden-state conditional random field (CRF) models to the task. A method for evaluating the significance of results quoted in terms of the normalised cross entropy (NCE) is also introduced. Instead of using sub-word-level information to improve word-level confidence scores, sub-word and word-level predictor features are combined to improve the accuracy of confidence scores in each sub-word being correct. The use of CRFs to model transitions between consecutive correct/incorrect sub-words yields large performance improvements. The scale of these gains is shown to increase further with the application of hidden-state CRFs. This is attributed to the fact that the hidden states make it possible for longer-span runs of consecutive correct/incorrect sub-words to be modelled, with these runs also not being constrained by word-level boundaries.

Index Terms— Hidden-state conditional random fields, confidence estimation, sub-words.

1. INTRODUCTION

Confidence estimation is often defined on the word-level, such that the desired confidence scores indicate whether an entire word within ASR output is likely to be correct or not. This is partly due to the fact that word-level transcriptions are commonly employed by downstream systems which make use of confidence scores. Modern large vocabulary speech recognisers use sub-word level representations for acoustic modelling. Information defined at the sub-word level is therefore readily available for use in confidence estimation. Confidence scores defined on this level may prove useful in applications where sub-word sequences are important, such as in OOV detection, pronunciation modelling, sub-word-level keyterm detection and unsupervised adaptation.

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.

One approach to estimating sub-word confidence scores is to compute sub-word-level posterior probabilities to use as confidence scores directly (e.g. [1, 2, 3, 4]). Classification approaches are often used to estimate word-level confidence scores (e.g. [5, 6, 7, 8, 9, 10]). In this paper, the models operate at the sub-word-level. Useful features are extracted for sub-words and combined to estimate confidence scores for these sub-word units. Examples of related work using this classification-based approach on the sub-word level make use of neural networks [11] and support vector machines [12].

In this work, linear-chain CRFs [13] and hidden-state CRFs are explored for this task. CRF models are sequential in nature, and are therefore suited to capturing the structure in runs of consecutive errors. Hidden-state CRF models are an extension of linear-chain CRFs to include hidden state variables. These models are similar to latent-dynamic CRF models [14], but impose fewer constraints on the hidden-state structure. The premise for the use of a hidden-state model for this task is that these states can capture the dynamics of longer-range hidden structure in the sequence of correct/incorrect sub-words. This builds on previous work [15] which did not make use of a hidden-state CRFs on the sub-word-level, and was concerned with computing word-level scores rather than sub-word-level scores directly.

Statistical significance tests have been developed to asses the significance of word error rate results for ASR systems [16]. These are based on the difference between systems in terms of the number of correct/erroneous words in the output. A metric often used in the evaluation of confidence estimation systems is the NCE metric. While this metric is useful, differences in this measure are difficult to interpret. In this work, a novel technique for evaluating the significance of confidence estimation results in terms of NCE scores is proposed.

The word and sub-word-level predictor features utilised for the confidence estimation task will be described in the first part of this paper. Thereafter, the hidden-state CRF-based framework for this task is presented. The proposed approach for testing the significance of NCE results is then presented. Results of performing direct sub-word-level confidence estimation are finally given and discussed.

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

2. PREDICTOR FEATURES

In a direct classification-based approach to sub-word-level confidence estimation, a set of predictor features is associated with each context-independent sub-word in the 1-Best transcription. Word and sub-word-level posterior probabilities are used as predictor features in this work. The baseline feature is the word-level lattice arc posterior ratio (LAPR). This feature is computed from the ASR lattice for a word W_i at position i within the 1-Best hypothesis, over a set of arcs $\mathcal{I}$ which intersect with that hypothesis and is defined as:

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

where A is an arc within $\mathcal{I}$. The Kronecker delta function (δ) is used to match the word W_i with the word identity of an arc, $\texttt{word}(A)$, and $P(A|\mathbf{X})$ is the posterior probability for arc A, computed during a forward-backward pass. To obtain a predictor feature defined on the sub-word level, word-level recognition lattices are marked up with temporal information at the sub-word-level. This effectively yields a set of context-dependent sub-word arcs (for graphemes in this work). The lattice sub-arc posterior ratio (LSAPR) is computed for each sub-word U_i in the 1-Best hypothesis, over a set of sub-word arcs $\mathcal{I}$ which intersect with this grapheme in the lattice:

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

where S is a sub-word arc within $\mathcal{I}$. The Kronecker delta function (δ) matches the sub-word U_i with the identity of the central sub-word, $\texttt{sub-word}(S)$, within the context-dependent sub-word arc S, and $P(S|\mathbf{O})$ is assumed to be equal to the posterior probability of the parent word arc for the sub-word S. An alternative frame-level sub-word classifier is used to incorporate competing information into the confidence estimation process. For a sub-word U_i in the 1-Best hypothesis on the interval from t_b to t_e, the alternative sub-word posterior (ASWP) is computed as:

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

where t is a frame within the audio, S is a sub-word in the complete set of sub-words $\mathcal{S}$. The posterior probability $P(S|t)$ is the posterior at time t for the sub-word S in the sub-word classifier's output. Substituting the best scoring sub-word from the grapheme classifier for U_i in 2 yields the best alternative sub-word posterior (BASWP).

3. HIDDEN-STATE CRFS FOR DIRECT SUB-WORD CONFIDENCE ESTIMATION

The framework for confidence estimation with CRF models presented in previous work [17, 15] is extended for this task.

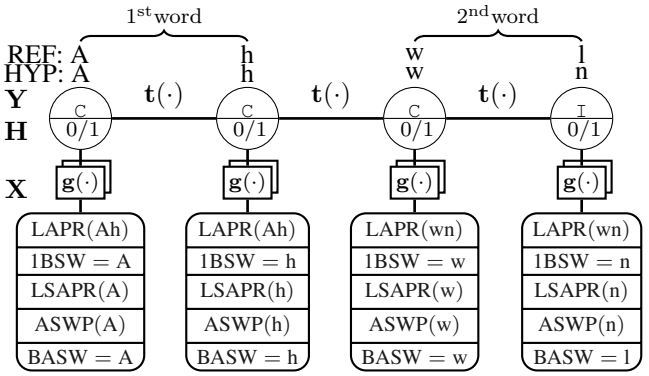

Fig. 1. Figure illustrating direct grapheme-level confidence estimation using CRF models for an Arabic training example. REF=reference, HYP=hypothesis. $\mathbf{Y}$ is the ideal sequence of correct/incorrect labels, hidden state variables in $\mathbf{H}$ may take values of 0/1 and $\mathbf{X}$ are input predictor features.

The general expression for a hidden-state CRF which models the sequence of discrete labels $\mathbf{Y}$ conditioned on a sequence of input feature vectors $\mathbf{X}$, marginalising over hidden-state sequences $\mathbf{H}$ is the following:

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

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

The example illustrated in Figure 1 is that of directly estimating confidence scores for the sequence of graphemes `A-h w-n` in the 1-Best hypothesis, where the reference transcription is `A-h w-l`. This implies there is one misrecognised grapheme in the second word. The sequence of labels indicating whether the sub-word is correct (`C`) or incorrect (`I`), is depicted as $\mathbf{Y}$. A sequence of hidden states $\mathbf{H}$ has the same length as $\mathbf{Y}$, with hidden states able to take the value 0 or 1. The transitions between the label/state pairs are captured with the transition feature functions $\mathbf{t}(\cdot)$. The feature vector of predictor features is $\mathbf{X}$, which are related to the labels and hidden states through the observation feature functions $\mathbf{g}(\cdot)$. Spline feature functions [18] are used to represent the continuous-valued predictor features, using 8 evenly-spaced knot points in the approximation, as proved useful in [17].

The confidence score for an individual sub-word is taken as the marginal probability of the label `C` for that sub-word within the sequence. This marginal is computed using the forward-backward algorithm during test.

4. EVALUATION AND SIGNIFICANCE TESTING

A statistical significance test determines the probability with which a null hypothesis (H_0) may be rejected. Here, the null hypothesis is that the mean of the differences between the NCE scores of two competing systems is zero. If this null hypothesis is proven to be true, it implies that the performance of the two systems is not significantly different. This test is formulated in a manner similar to that of the matched pair test commonly used to test significance of word error rate results for ASR transcriptions [16]. A segmentation of the ASR output which produces K segments for which the NCE scores are statistically independent is assumed. The NCE score for a segment i is computed as follows:

$$\text{NCE}(i) = \frac{H_i + \sum_{S \in \mathcal{C}_i} \log_2(\hat{P}(S)) + \sum_{S \in \mathcal{I}_i} \log_2(1 - \hat{P}(S))}{H_i} \quad (1)$$

where $\hat{P}(S)$ is the confidence score for an output symbol (e.g. word/sub-word) S, $\mathcal{C}$ is the set of all correct symbols in the evaluation data and $\mathcal{I}$ is the set of all incorrect symbols. Given n correct symbols out of a total of N, the empirical accuracy is $P_c = \frac{n}{N}$, and the maximum empirical entropy is:

$$H = -n \log_2(P_c) - (N - n) \log_2(1 - P_c). \quad (2)$$

The normalised sub-word cross entropy (NSCE) is used for evaluation in this work, where the symbols in Equations 1 and 2 over which the metric is computed are sub-words.

The quantity required in order to test the null hypothesis is the difference in the NSCE values for two confidence estimation systems (A and B) on a segment i:

$$Z_i = \text{NSCE}_A(i) - \text{NSCE}_B(i)$$

The mean difference in the NSCE scores for the systems, μ_Z, the variance estimate of the Z_i values, σ_Z^2, and the test statistic W are defined as:

$$\hat{\mu}_Z = \sum_{i=1}^{K} \frac{Z_i}{K} \qquad \sigma_Z^2 = \frac{1}{K-1} \sum_{i=1}^{K} (Z_i - \hat{\mu}_Z)^2 \qquad W = \frac{\hat{\mu}_Z}{\frac{\sigma_Z}{\sqrt{K}}}.$$

For a sufficiently large number of segments (K), the distribution of W can be approximated by a zero-mean normal distribution with unit standard deviation. For the two-tailed significance test, the p-value is defined as:

$$p = 2p(X \geq |W|)$$

where X is a random variable which has the form of the standard normal distribution $\mathcal{N}(0, 1)$. The differences in the output of two systems may be considered statistically significant if this p-value is less than a desired level of confidence α. This signifies that the null hypothesis H_0 of there being no difference between the systems in question may be rejected. The value of α used here is 0.001 (the 99.9% confidence level).

5. EXPERIMENTS

Experiments in direct sub-word confidence estimation are carried out for the 1-Best Viterbi output of a recogniser which forms part of the 2010 Cambridge Arabic ASR system [19]. This recogniser employs a graphemic representation at the sub-word-level. The acoustic training data used consists of 1538 hours of audio, and comprises broadcast news (BN) and broadcast conversation (BC) data. The language model is trained on 1.2G words, with a vocabulary of 350k words being used. The decoding structure for this recogniser consists of multiple passes, the first two of which constitute the main lattice generation phase, with adaptation being applied in subsequent passes. All experiments are based on the large output lattices from the second decoding pass, which are typically information-rich and represent a large hypothesis space.

Three subsets of the 2010 GALE development data were used, as well as a subset of the 2009 GALE development data, and the non-sequestered portion of the 2009 GALE evaluation data. The training dataset comprises three of these datasets and contains 9700 utterances. The dev10d (7609 utterances) and eval09ns (1554 utterances) datasets were held out for evaluation. The word error rate on these datasets is 32.8% and 14.1% respectively. The alternative recogniser used for additional sub-word-level predictor features is an MLP-based frame-level classifier, which has 37 context-independent graphemes as targets. This was trained on 140 hours of Arabic broadcast news data using the ICSI QuickNet MLP neural network software [20]. 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 (an insertion or substitution). For substituted words, the hypothesised grapheme sequence is aligned with the reference grapheme sequence. In this way, graphemes which were recognised correctly within words which were incorrect overall are labelled as being correct. The baseline system uses a decision tree trained on the LAPR predictor feature, the leaf nodes of which yield optimal quantisation intervals and an associated mapped confidence score. A piecewise-linear mapping is applied over these intervals to yield smoothed scores, similar to that in [21].

5.1. Results

The results of experiments in which linear-chain and hidden-state CRFs are used to perform direct sub-word-level confidence estimation are presented in Table 1. A number is assigned to each system to aid comparisons. Results of statistical significance tests for systems of interest are also shown.

A sub-word-level CRF system which uses the LAPR predictor feature only (2) achieves improved NSCE performance over the baseline on dev10d, but yields worse performance on eval09ns. The significance test on this system against the

System		NSCE	
Num	Description	dev10d	eval09ns
1	Baseline : LAPR	0.208	0.266
2	CRF: LAPR	0.215	0.256
3	CRF: LAPR+LSAPR	0.251	0.305
4	CRF: LAPR+ASWP	0.254	0.284
5	CRF: LAPR+BASWP+ASWP	0.254	0.285
6	CRF: LAPR+ASWP+LSAPR	0.263	0.310
7	HCRF: LAPR+ASWP+LSAPR	0.309	0.343
8	CRF: 6+WLEV	0.274	0.320

Systems		dev10d+eval09ns	
A	B	ΔNCE	p($<$)
System 2	Baseline	0.002	1
System 6	Baseline	0.049	0.001
System 8	Baseline	0.060	0.001
System 7	Baseline	0.089	0.001
System 7	System 6	0.040	0.001

Table 1. Results for direct grapheme-level confidence estimation, evaluated in terms of NSCE (over all segments) on dev10d and eval09ns. LAPR=lattice arc posterior ratio, LSAPR=lattice sub-arc posterior ratio, ASWP=alternative sub-word posterior, BASWP=best alternative sub-word posterior, WLEV=word-level features. Significance test results on the combined datasets are shown. ΔNSCE = avg. difference in NSCE scores of systems, p($<$) = p-value for test.

baseline shows that the results are not statistically significant, with the p-value only being less than 1. The degraded performance on eval09ns may be because the word-level LAPR feature is used in isolation, which is static as it is repeated over sub-words within a word. The CRF is therefore unable to leverage the sequence information in consecutive sub-words effectively, and the word posterior may be inconsistent for correct sub-words within incorrect words and vice-versa. This also results in the CRF system being similar to the non-sequential baseline, and therefore not significantly different.

Including a "true" sub-word-level feature (LSAPR) in system 3 results in improvements in NSCE over the baseline of 21% and 15% relative on dev10d and eval09ns respectively. Combing the LAPR feature with the posterior from the alternative recogniser (ASWP) in system 4 yields slightly larger gains in NSCE on dev10d, but is not as beneficial on eval09ns. Including the BASWP feature in this system (5) yields a negligible improvement on eval09ns. This may be a result of the fact that this posterior is identical to ASWP when the systems have the same hypothesis, and when they disagree the posterior for a different sub-word is not informative.

The system which yielded the best performance using primarily sub-word-level predictor features is therefore one in which the word-level LAPR predictor feature is combined with the sub-word-level LSAPR feature and the ASWP feature (system 6). The results of this system are also proven to be significant, with a p-value of less than 0.001. Relative improvements over the LAPR-only sub-word-level configuration (system 2) of 22.3% and 21.1% relative in NSCE are achieved by this system on dev10d and eval09ns. This result shows that for sub-word-level confidence estimation, predictor features defined at the same level are most useful.

A hidden-state CRF model is applied to the feature combination of LAPR, ASWP and LSAPR (system 7). This system yields considerable performance improvements over the equivalent linear-chain CRF (system 6) of 26% and 10% relative for dev10d and eval09ns respectively. The results of significance testing carried out for this system against the baseline (which it outperforms by 0.089 absolute in NSCE), show that the results are significant, with a p-value below 0.001. Similarly, these results are seen to be significant in comparison with those for the equivalent linear-chain CRF (system 6), over which an improvement in NSCE of 0.04 is achieved.

Word-level predictor features (WLEV) used in previous work were included in the sub-word level system. This resulted in some improvements in NSCE performance, which are overshadowed by those obtained using the hidden-state CRF with sub-word-level features. The relative improvement in NSCE score over the baseline with this additional information is 31.7% and 20.3% on dev10d and eval09ns respectively. This result is statistically significant (with a p-value less than 0.001). Compared with system 6 which does not use these features, relative improvements of 4.2% and 3.2% in NSCE are seen on dev10d and eval09ns respectively. When applying a decision threshold of 0.5 to the combined dataset, the hidden-state CRF yields a relative reduction in error rate (words incorrectly classified as correct or incorrect) of 8.2% over the baseline (which has an error rate of 12.2%). This is 1.6% larger than for the equivalent linear-chain CRF.

6. CONCLUSION

This paper considered the task of estimating confidence scores directly on the sub-word-level, through the novel application of CRF and hidden-state CRF models. The sequential nature of the linear-chain CRF is exploited to model consecutive errors in ASR output and yield performance improvements. These gains are seen to increase dramatically when a hidden-state CRF model is used to capture more information on longer-spans of correct/incorrect sub-words. The combination of useful sub-word features (extracted from the underlying system and an external source), with word-level features are seen to improve performance. A statistical significance testing technique which addresses issues in interpreting results quoted in terms of the NCE metric was presented. This technique was used in the evaluation of systems in this work to prove the significance of the results reported for the direct sub-word confidence estimation task.

7. REFERENCES

[1] Z. Rivlin, M. Cohen, V. Abrash, and T. Chung, "A phone-dependent confidence measure for utterance rejection," in *Proceedings of ICASSP*, 1996.

[2] G. Williams and S. Renals, "Confidence measures for hybrid HMM/ANN speech recognition," in *Proceedings of Eurospeech*, 1997.

[3] 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.

[4] Wai Kit Lo, F. K. Soong, and S. Nakamura, "Generalized posterior probability for minimizing verification errors at subword, word and sentence levels," in *Proceedings of ISCSLP*, 2004.

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

[6] 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.

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

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

[9] 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.

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

[11] Lin Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition.," in *Proceedings of Eurospeech*, 1997.

[12] Shilei Huang, Xiang Xie, and Jingming Kuang, "Novel method to combine phone-level confidence scores using support vector machines," in *Proceedings of ICSP*, 2006.

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

[14] L. P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proceedings of CVPR*, 2007.

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

[16] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceddings of ICASSP*, 1989.

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

[18] 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.

[19] 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.

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

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