Using Sub-word-level Information for Confidence Estimation with Conditional Random Field Models

M.S. Seigel, P.C. Woodland

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

Abstract

The task of word-level confidence estimation (CE) for automatic speech recognition (ASR) systems stands to benefit from the combination of suitably defined input features from multiple information sources. However, the information sources of interest may not necessarily operate at the same level of granularity as the underlying ASR system. The research described here builds on previous work on confidence estimation for ASR systems using features extracted from word-recognition lattices, by incorporating information at the sub-word level. Furthermore, the use of Conditional Random Fields (CRFs) with hidden states is investigated as a technique to combine information for word-level CE. Performance improvements are shown using the sub-word-level information in linear-chain CRFs with appropriately engineered feature functions, as well as when applying the hidden-state CRF model at the word level.

Index Terms: confidence estimation, hidden-state conditional random fields, speech recognition, sub-word-level information

1. Introduction

As shown in previous work [1], multiple word-level information sources can be combined to improve CE for ASR systems using linear-chain CRF models. CRFs were first proposed as a discriminative modelling framework for segmenting and labelling sequence data [2]. The sequential nature of the model contributes to improvements in word-level CE performance, and results in confidence estimates which are more accurate over utterances/word sequences. It was also shown that the largest performance gains are achieved when combining information from alternative systems. However, rich sources of information which are potentially useful for CE do not necessarily operate at the word level. This work therefore investigates techniques through which predictor features at both the word and sub-word level may be effectively combined within a single, structured framework. Examples of sub-word-level units typically used in ASR include phones, graphemes and morphemes.

The sub-word-level predictor features investigated originate from two sources. The first of these is the set of recognition lattices output by the ASR system, which have been marked with sub-word-level information. The second source is a separate sub-word recogniser, the aim of which is to produce competing predictor features to augment those available from the underlying ASR system. This approach is similar to that investigated in [3], in which a phone-loop recogniser was used for this purpose.

This work was in part supported by DARPA under the GALE and RATS programs via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

The linear-chain CRF framework detailed in [1] is extended and applied such that sub-word-level information sources may be incorporated effectively. This was achieved by engineering feature functions specifically for the sub-word-level task, which is a contribution of this work. The application of hidden-state CRF models to the CE task is another contribution. Such models are able to capture interesting characteristics of the relationship between individual words and sequences of words. Performance improvements are shown over comparable CE approaches based on the Cambridge Arabic ASR system, thus proving the utility of the contributions.

The first part of this paper will introduce the hidden-state CRF model investigated, and draw comparisons with related work using such models. Thereafter, the sub-word-level predictor features and feature functions will be detailed. Finally, the experimental setup and evaluation for both the linear-chain and hidden-state CRF approaches to CE with sub-word-level information will be presented.

2. Hidden-state CRF Model for CE

Including hidden variables in the CRF model makes it suitable to be applied in tasks where there is some hidden structure. A benefit of this approach is that it does not impose strong independence assumptions, as is the case with models such as HMMs. Arbitrary feature functions which act on the predictor features may also be defined. CRFs with hidden states have been applied to many tasks, such as phone classification [4], phone recognition [5] and gesture recognition [6] to name a few, in which they typically yield good performance. Such a model, which is particularly suited to capturing vital sub-structure information in word sequences is investigated.

In contrast to the hidden CRF (HCRF) models employed in [4], the model used does not represent the joint distribution of a single label $y$ and a hidden sequence $h$, but rather the joint distribution of an entire label sequence $y$ and hidden state sequence $h$, all conditioned on the sequence of observation vectors $X$. The resulting graphical model structure, which is similar to the Latent-Dynamic CRF (LD-CRF) [6], is shown in figure 1b. For the CE task, $x_j$ represents a sub-word-level/word-level vector of predictor features, $y_j$ represents the corresponding label (“Correct/Incorrect”) and $h_j$ is the hidden variable assigned to $x_j$ and $y_j$. Marginalising out over hidden state sequences $h$ in the set of possible sequences $\mathcal{H}$ yields the following:

$$p(y|X) \propto \sum_{h \in \mathcal{H}} \exp \left( \sum_k \lambda_k t_k(y, h) + \sum_l \mu_l g_l(y, h, X) \right)$$

where $t_k(y, h)$ are the transition feature functions and $g_l(y, h, X)$ are the observation feature functions, with param-
Figure 1: Graphical structure of the models used, where shaded vertices correspond to variables observed during training.

Parameters $\lambda_k$ and $\mu_k$ respectively. Each of the $k$ transition feature functions corresponds to a combination of previous/current label and hidden state values $(y', y, h', h)$. Each of the $q_k$ feature functions takes the current observation vector $x_i$ as an argument and corresponds to a current label and hidden state value $(y, h)$. These are enumerated over the set of discrete values a predictor feature can take, or a quantisation parameter in the continuous case. The feature functions are expressed as follows:

$$t_{y', y, h', h} = \delta(y_{i-1}, y') \delta(h_{i-1}, h') \delta(y_i, y) \delta(h_i, h)$$ (1)

$$g_{y, h}(x_i) = z(x_i) \delta(y_i, y) \delta(h_i, h)$$ (2)

where $i$ is an index within the observation/label sequence, $\delta$ is the Kronecker delta function, and $z(x_i)$ is a function applied to the current observation vector. This function is dependent on the specific type of observation feature function. The model parameters $\lambda_k$ and $\mu_k$ are estimated by optimising the conditional log-likelihood of the model using a gradient-based technique (Limited-Memory BFGS [7]).

Although the structure of the model described is similar to that described in [6], hidden variables are not constrained to take values in a disjoint set for each label, thus yielding a more general model. This model is capable of capturing the dynamic nature of the label sequence, as well as that of a hidden state space (which can represent additional structure). This structure could correspond to a soft segmentation of words into sequences of Correct/Incorrect words, which is in line with the concept of regions of confidence within recogniser output.

### 3. Sub-word-level Predictor Features and Feature Functions

The acoustic models of the ASR system may be used to acquire timing information for the individual sub-word units which comprise the word arcs in the recognition lattices. Given this information, posterior probabilities for each sub-word unit may be calculated directly from the lattices. The procedure for doing so is similar to that for the word-level Lattice Arc Posterior Ratio (LAPR) [1]. Arc posterior probabilities $p(a|O)$ are calculated for each of the word-level arcs $a$ in a lattice given the acoustic observation vectors $O$ (as described in [8]), and are replicated over all sub-word arcs $s$ comprising the word arc to yield $p(s|O)$. For a given sub-word unit along the 1-best path $r$ through a lattice, the posterior probabilities of sub-word arcs $s$ in the set of intersecting arcs $I$ with the same label $s_l$ as the reference sub-word $r_i$ are summed. This sum is normalised by the sum of the posterior probabilities for all sub-word arcs in $I$, yielding the following expression for the Lattice Sub-Arc Posterior Ratio (LSAPR):

$$LSAPR_{s} = \sum_{s_l} \delta(s_l, r_i) p(s|O) / \sum_{s_l} p(s|O)$$

An alternative, lightweight recogniser which is not constrained by the language model in the ASR system is a potentially useful source of information for CE. This is due to the fact that predictor features produced by such a source are more closely matched to the acoustic evidence. A Multi-Layer Perceptron (MLP)-based recogniser acting at the sub-word level was used as such an alternative source. This recogniser outputs posterior probabilities for each of the sub-word units at the frame level. For a given sub-word unit in the 1-best path through the underlying ASR system lattice, MLP posterior probabilities for the same sub-word unit on the corresponding frame interval are averaged. This yields the average MLP posterior predictor feature (AMP). The highest scoring sub-word unit hypothesised by the MLP on the same 1-best sub-word intervals is another useful predictor feature, as these may be significantly different from those hypothesised by the ASR system.

#### 3.1. Feature Functions

Continuous predictor features are represented using spline feature functions [9], as they have proven to yield good performance in previous work [1]. These are defined for each of the discrete “knot points” on the continuous interval for a particular predictor feature, at which a cubic polynomial is fitted in the spline approximation for the distribution. Supplementary feature functions were implemented in order to take advantage of some of the structural characteristics of the data, and the available sub-word-level predictor features.

When modelling transitions in the linear-chain CRF at the sub-word level, information regarding the word-level transition structure is lost. A long-range transition feature function, much like the skip-chain feature functions introduced in [10], was therefore implemented. This feature function is only active on word boundaries, which are marked explicitly within a predictor feature argument passed to the function. The word-level transitions are therefore modelled separately from the sub-word-level transitions. This feature function (WB) is expressed as follows:

$$W_{y', y, h', h}(x_i) = \delta(x_i, h) t_{y', y, h', h}$$

where $x_i$ is the predictor feature that indicates whether the observation is at a word boundary (8) or within a word (1), and $t_{y', y, h', h}$ is the transition feature function (equation 1).

Feature functions which make direct use of sub-word-unit identities and corresponding scores were implemented. The first of these is a pair of discrete “string match” feature functions (returning 1 or 0). These feature functions (referred to as SM), compare the string values of two elements in the predictor feature vector, and are active if the strings are matched/different. A similar pair of feature functions which return a continuous value was investigated. This feature function (referred to as SMV), performs the same string match function, and returns the value of an arbitrary continuous predictor feature if these strings are matched/different. These feature functions, which specify the function $z(x_i)$ in equation 2, are expressed as follows:

$$SM_{y, h}(x_i, x_{i+1}) = \delta(x_i, x_{i+1}) \delta(y_i, y) \delta(h_i, h)$$

$$SMV_{y, h}(x_i, x_{i+1}) = \delta(x_i, x_{i+1}) \delta(y_i, y) \delta(h_i, h)$$

where $x_i$ and $x_{i+1}$ are the strings to be compared, and $x_i$ is the continuous predictor feature to return.

The feature functions in this section have been expressed in their most general form and therefore include hidden variables. However, these definitions hold for the linear-chain CRF model, in which case the number of possible hidden state values is 1.
4. Experiments

Confidence estimation is carried out for a state-of-the-art recogniser which forms part of the 2010 Cambridge Arabic ASR system [11]. Multiple sub-word unit representations of Arabic are used within this system, requiring separate recognisers for each representation. For the sake of simplicity, a grapheme-based system is used for the experiments. 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 output lattices from the second decoding pass, which are typically dense lattices that represent a large hypothesis space.

The dev10d, dev10r and dev10c subsets of the 2010 GALE development data were used. The dev10d subset consists of difficult/high error rate portions of the data, while the dev10r data has a more typical error rate distribution. The dev10c dataset is one in which particular care was taken in producing the reference transcriptions. A subset of the 2009 GALE development data (dev09sub), and the non-sequestered portion of the 2009 GALE evaluation data (eval09ns) were also used. The dev10c, dev10r and dev09sub datasets comprise the training dataset (27.5 hours and 9700 utterances). The dev10d (18.5 hours and 7609 utterances) and eval09ns (6.5 hours and 1554 utterances) datasets were held out for evaluation. The word error rate on these datasets is 32.8% and 14.1% respectively.

The MLP-based grapheme recogniser was trained on 140 hours of Arabic broadcast news data (the GALE p4r3 dataset) using the ICSI QuickNet MLP neural network software [12]. The input vector consists of 14 perceptual linear prediction (PLP) features with deltas and double-deltas over a window of 9 frames. The MLP includes one hidden layer of 3500 nodes, and an output layer of 37 softmax nodes (one for each grapheme). The cross validation accuracy of this recogniser is 66.59%.

Experiments are evaluated using the word-level normalised cross entropy (NCE) metric. It is also interesting to assess the accuracy of the confidence measures assigned to word sequences. The utterance-level mean absolute deviation (UMAD) metric [1] is particularly suited to this, and is therefore used.

A summary of the predictor features and feature functions used in all experiments is provided in table 1.

<table>
<thead>
<tr>
<th>Feature/Function Description</th>
<th>dev10d</th>
<th>eval09ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>L(S)APR (W)AMP WB SM(V)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Predictor features and feature functions used.

4.1. Sub-word-level Linear-chain CRF

The use of sub-word-level predictor features in a linear-chain CRF model, which itself operates at this level was investigated. The sub-word units used are graphemes, as this is the representation employed by the underlying recogniser. As a baseline grapheme-level predictor feature, the word-level posterior (LAPR) is replicated over all constituent graphemes of that word. To successfully apply the linear-chain CRF model to the word-level CE task using the grapheme-level predictor features, an expanded label set is used. These labels indicate whether the overall word is “Correct/Incorrect”, as well as whether the current grapheme observation is at a word boundary or within a word. Grapheme-level confidence scores are obtained by computing the marginal probability of both the word-start and within-word “Correct” labels for a grapheme observation. The resulting scores are averaged over the constituent graphemes of a word to yield the final word-level confidence score.

The WB feature functions are used to model within-word grapheme transitions separately from between-word transitions, by including the mapping of graphemes to word boundaries as a predictor feature. The SM feature functions match the identity of the 1-best grapheme hypothesised by the ASR system, with the highest scoring grapheme output on the same interval by the MLP recogniser. The SMV feature functions match the aforementioned grapheme identities and return the value of AMP, the posterior probability assigned by the MLP recogniser.

4.1.1. Results

Experimental results are shown in Table 2, in which the baseline corresponds to a decision tree-based system similar to that described in [8], which makes use of the expanded word-level LAPR predictor feature. Continuous predictor features are represented using spline feature functions in the CRF models, with the same number of knot points as quantisation intervals in the baseline system. Performance improvements correspond to increases in the NCE metric and decreases in the UMAD metric.

<table>
<thead>
<tr>
<th>System</th>
<th>dev10d</th>
<th>eval09ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.325 12.64 0.356 6.98</td>
<td></td>
</tr>
<tr>
<td>LAPR</td>
<td>0.300 10.96 0.313 6.68</td>
<td></td>
</tr>
<tr>
<td>LAPR+WB</td>
<td>0.338 11.35 0.360 6.64</td>
<td></td>
</tr>
<tr>
<td>LAPR+LSAPR+WB</td>
<td>0.344 11.08 0.362 6.56</td>
<td></td>
</tr>
<tr>
<td>LAPR+WB+SM</td>
<td>0.348 11.04 0.361 6.60</td>
<td></td>
</tr>
<tr>
<td>LAPR+WB+SMV</td>
<td>0.352 10.92 0.364 6.57</td>
<td></td>
</tr>
<tr>
<td>LAPR+AMP+WB</td>
<td>0.353 10.87 0.365 6.59</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Sub-word level linear-chain CRF results.

The CRF system using only the LAPR predictor feature has no explicit knowledge of word boundaries, and therefore results in word-level NCE performance inferior to that of the baseline. However, including word boundary information through the WB feature functions yields improvements in NCE and UMAD. Interestingly, incorporating word boundary information results in a decrease in UMAD performance with an increase in NCE performance. This suggests that without word boundary information, the confidence scores are quite accurate on average, but do not necessarily align with the words. With the inclusion of a true grapheme-level predictor feature (LSAPR), gains are seen in both metrics over the LAPR+WB system. Using the SM feature functions yields slightly larger improvements, suggesting the alternative information from the MLP is useful. Including the AMP predictor feature through the SMV feature functions builds on these gains. Representing the AMP predictor feature with a spline yields the best system with relative improvements over the baseline on dev10d of 8.6% and 14% in NCE and UMAD respectively. These results show that sub-word-level information is indeed useful for word-level CE.

4.2. Word-level Linear-chain CRF

The use of sub-word-level predictor features in a linear-chain CRF model which itself operates at the word level was inves-
tigated. The sub-word-level AMP predictor feature is represented at the word level by averaging these scores over each word (referred to as WAMP). The SM feature function is used to match the word-level sequence of graphemes assigned by the underlying system and the MLP recogniser. The SMV feature functions are used to match the afore-mentioned grapheme sequences, and return the value of the WAMP predictor feature. This experimental setup was chosen to match that investigated in Section 4.1 as closely as possible.

4.2. Results

As is seen in the results presented in Table 3, improvements in UMAD and NCE are achieved on dev10d when including the SM feature functions in a system which uses the LAPR predictor feature. The SMV feature functions yield further gains in both metrics, which result from the inclusion of the WAMP predictor feature. Incorporating the WAMP predictor feature using a spline results in the best performing system of this type.

These results show that performance improvements are possible using sub-word-level information in a word-level model. However, representing sub-word-level information at the word level diminishes the impact of such predictor features.

<table>
<thead>
<tr>
<th>System</th>
<th>dev10d</th>
<th>eval09ns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NCE UMAD</td>
<td>NCE UMAD</td>
</tr>
<tr>
<td>LAPR</td>
<td>0.339 11.33</td>
<td>0.359 6.64</td>
</tr>
<tr>
<td>LAPR+SM</td>
<td>0.342 11.25</td>
<td>0.359 6.64</td>
</tr>
<tr>
<td>LAPR+SMV</td>
<td>0.346 11.09</td>
<td>0.360 6.60</td>
</tr>
<tr>
<td>LAPR+WAMP</td>
<td>0.346 11.06</td>
<td>0.361 6.58</td>
</tr>
</tbody>
</table>

Table 3: Results for the word-level linear-chain CRF with sub-word-level predictor features.

4.3. Word-level Hidden-state CRF

The effect of adding hidden variables to word-level models which incorporate sub-word-level predictor features was investigated. In determining the impact of the number of hidden state values, the most informative word-level predictor feature (LAPR) is used. The configuration which yielded the best linear-chain CRF system is also applied to a hidden-state CRF.

4.3.1. Results

The results presented in Table 4 show that further improvements over the CRF baseline are achieved in both NCE and UMAD when using a hidden-state CRF with two hidden state values. The improvements in UMAD suggest that regions of confidence within utterances are modelled more adequately, by capturing information about sequences of Correct/Incorrect words through the hidden structure. Increasing the number of hidden state values to 3 results in modest gains for the associated increase in model complexity. A system using two hidden state values which incorporates sub-word-level information through the WAMP predictor feature yields the best performance, with relative improvements over the CRF baseline on dev10d of 6.5% and 4% in NCE and UMAD. This result highlights the gains possible when using a hidden-state CRF model, and further supports the use of sub-word-level information.

<table>
<thead>
<tr>
<th>System</th>
<th>dev10d</th>
<th>eval09ns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NCE UMAD</td>
<td>NCE UMAD</td>
</tr>
<tr>
<td>CRF LAPR</td>
<td>0.339 11.33</td>
<td>0.359 6.64</td>
</tr>
<tr>
<td>HCRF H=2 LAPR</td>
<td>0.355 11.03</td>
<td>0.367 6.60</td>
</tr>
<tr>
<td>HCRF H=3 LAPR</td>
<td>0.357 11.09</td>
<td>0.371 6.60</td>
</tr>
<tr>
<td>HCRF H=2 LAPR+WAMP</td>
<td>0.363 10.81</td>
<td>0.368 6.55</td>
</tr>
</tbody>
</table>

Table 4: Word-level hidden-state CRF results, where “H” indicates the number of hidden state values.

in this work. It is shown that using sub-word-level information from the underlying ASR system and an alternative MLP-based recogniser yields improvements in CE performance. The most effective sub-word predictor feature proved to be the sub-word-level posterior probabilities output by an MLP-based recogniser. Information at the sub-word level is shown to be more useful when applied in a linear-chain CRF model operating at the sub-word level than at the word level, provided care is taken to capture aspects of the word level structure (through word boundary “skip-edge” feature functions). Applying the hidden-state CRF model at the word level yields further improvements in CE performance. This is attributed to the ability of the model to represent regions of confidence through the hidden state structure. Future work will investigate the application of the hidden-state CRF at the sub-word level.

5. Conclusions

Incorporating sub-word-level information in a CRF-based approach to CE for a state-of-the-art ASR system is investigated

6. References