Discriminative Training for Complementariness in System Combination

Daniel Willett1, Chuang (Mike) He2

Nuance Communications, Acoustic Modeling Research, 1Aachen, Germany; 2Burlington, MA
daniel.willett@nuance.com, mike.he@nuance.com

Abstract

In recent years, techniques of output combination from multiple speech recognizers for improved overall performance have gained popularity. Most commonly, the combined systems are established independently. This paper describes our attempt to directly target joint system performance in the discriminative training objective of acoustic model parameter estimation. It also states first promising results.

Index Terms: speech recognition, system combination, discriminative training, CNC, ROVER, MMI, MCE

1. Introduction

The success of system combination approaches such as ROVER [1] and CNC [2,3] depends on two aspects. The recognition systems being combined need to be sufficiently good while at the same time they need to be complementary. This implies that they need to make different errors. Usually, multiple systems targeted for system combination are established by means of modest variations in model configuration. Such variations comprise front-end, phone-set, language model, model topology, training data selection and weighting, training and decoding software among others. A first joint training objective was formulated in [4]. The approach yields a data weighting for a second system based on posteriors estimated on the reference system. It results in a stronger influence of utterances that are only poorly modeled by the reference system in the Maximum Likelihood parameter estimation of the second system. However, no indication is given on how to integrate the approach in a discriminative training (DT) framework which itself weights words according to posterior probabilities.

2. DT for Complementariness

Discriminative objectives usually target the optimization of a criterion in which the likelihood of the correct word sequences is set in relation to the most likely recognition alternatives. The objectives of Minimum Classification Error (MCE) training as well as that of Maximum Mutual Information Estimation (MMIE) can be formulated as in Eq.1.

$$\lambda_{\text{MMIE}} = \arg \max_{\lambda} \sum_{w} \left( \log p(W, p(X | W)) - \log \sum_{w'} p(W) p_{\lambda}(X | W) \right)$$  \hspace{1cm} (1)

U represents the set of training utterances, Ww the uttered word sequence and Xw the acoustic observation. The term within f() is often referred to as the misclassification measure. The loss function f() itself is usually assumed to be sigmoidal in MCE, i.e. f(x)=1/(1+e^-x), and identity f(x)=x in MMIE.

$$p_{\lambda}(X | W) = \prod_{w \in W} p_{\lambda}(Xw | w)$$  \hspace{1cm} (2)

For simplicity we neglect context-dependency across words such that the observation likelihood can be rewritten as a product over word likelihoods as in Eq.2. Our approach now starts off from this representation and replaces the single model word scores with a weighted sum over word scores accumulated over all models. Hence, we formulate the multi-model word sequence likelihoods according to Eq.3 with R being the set of models.

$$p_{\lambda}(X | W) = \prod_{w \in W} \sum_{r \in R} n_r p_{\lambda}(Xw | w)$$  \hspace{1cm} (3)

The scaling factors n_r normalize likelihood ranges and also account for potentially different model weights in system combination. The rationale for this kind of incorporation of word scores accumulated over all models is that combination approaches perform a per-word averaging over posteriors from each system. The posterior probabilities themselves are scaled (sums of) word likelihoods. We knowingly neglect that this scaling in fact varies per utterance.

3. Experimental evaluations

The parameter estimation according to DT objectives as in Eq.1 with the utterance likelihoods approximated over multiple models as in Eq.3 allows many flavors of implementation. The approach that we took so far is training two systems successively, the first according to single model MCE and the second according to the joint criterion with the first model remaining untouched. The proposed method showed 2-5% relative WER reduction as compared to system combination over two independently trained systems in initial LVCSR experiments. The amount of improvement varies with model configuration and recognizer setup.

4. Conclusions

We have sketched our approach for training a multitude of models aiming at performance in system combination. First results validate its soundness and demonstrate its potential for gradually stacking models, each of which targeting the errors that persist in combination with the other models. Future work will look into parallel discriminative training of multiple systems. With each system making use of different features (i.e. front-end) and/or model configurations, we expect the approach to allow each model to contribute discrimination on those parts of the data that it is particularly well suited for.

5. References


