Large Margin Estimation of n-gram Language Models for Speech Recognition via Linear Programming

Vladimir Magdin, Hui Jiang (Department of Computer Science and Engineering, York University, Canada)

YORK UNIVERSITÉ UNIVERSITY

Abstract

- **Contribution: a novel discriminative training algorithm for back-off n-gram language models for use in LVCSR**
- The LME-based objective function uses a metric between correct transcriptions and word-graph-encoded competing hypotheses
- The nonlinear LME objective function is approximated by a linear function of LM parameters, which leads to a linear programming solution
- Experimental results on the SPINE1 speech recognition task show a relative reduction in word error rate of close to 2.5%

Language Models in ASR

Recognition is performed via the MAP decision rule:

$$\hat{W} = \mathop{\textbf{arg max}}_{W} \Pr(W|X) = \mathop{\textbf{arg max}}_{W} \Pr(X|W) \cdot \Pr(W)$$

W - sequence of word labels

X - sequence of acoustic observations

Language models in automatic speech recognition:

- LMs constrain the search space of hypotheses
- $\Pr(W)$ is modeled via n-gram LMs (e.g. Katz back-off LM)

Language model issues:

- ML criterion not directly related to recognition performance
- n-gram LMs not tailored for a particular application
- n-gram LM parameters are crudely approximated via smoothing

Discriminative Training of LMs via Soft-Margin LME

Based on the principle of *large margin estimation*, maximize the minimum margin between correct transcription and competing hypotheses:

$$d(W|\Lambda) = \ln\left[\Pr(W|\Lambda) \cdot \mathcal{A}(W)\right] - \max_{W' \in \mathcal{G} \setminus W} \ln\left[\Pr(W'|\Lambda) \cdot \mathcal{A}(W')\right]$$

Replacing the maximization with a log-summation (soft-max operation) for mathematical reasons:

$$d(W|\Lambda) = \ln\left[\Pr(W|\Lambda) \cdot \mathcal{A}(W)\right] - \ln \sum_{W' \in \mathcal{G} \setminus W} \ln\left[\Pr(W'|\Lambda) \cdot \mathcal{A}(W')\right]$$

Incorporating the minimum margin over the support set S and the error over the error set $\mathcal{E}$:

Original Soft-Margin LME Objective Function (log-domain)

$$\mathop{\textbf{arg max}}_{\Lambda} \left[\min_{W_n \in S} d(W_n|\Lambda) - \epsilon \cdot \frac{1}{|\mathcal{E}|} \sum_{W_i \in \mathcal{E}} d(W_i|\Lambda) \right]$$

Solving Soft-Margin LME via Linear Programming

The original LME objective function is computationally intractable, so we approximate it with a linear function of individual LM parameters:

$$\Lambda = \{\lambda_i, \eta_j, \mu_k, \phi_l, \psi_m | i \in P_3, j \in P_2, k \in P_1, l \in Q_2, m \in Q_1\}$$

(λ_i, η_j, μ_k - tri-gram, bi-gram, uni-gram log-conditional probs.
ϕ_l, ψ_m - bi-gram, uni-gram back-off weights in log-domain)

APPROXIMATION STEP: Approximate original objective function with a simpler linear function

$$d(W|\Lambda) = \ln\left[\Pr(W_n|\Lambda) \cdot \mathcal{A}(W_n)\right] - \ln \sum_{W' \in \mathcal{G}_n} \left[\Pr(W'|\Lambda) \cdot \mathcal{A}(W')\right]$$

Correct Transcription:

$$\ln[\Pr(W|\Lambda) \cdot \mathcal{A}(W)] = \ln \prod_{t=1}^{R_W} \Pr(w_t|w_{t-2}w_{t-1}) + B'$$
$$= \sum_{t=1}^{R_W} \ln \Pr(w_t|w_{t-2}w_{t-1}) + B'$$
$$= \sum_{i \in P_3} a'_i(W) \cdot \lambda_i + \cdots + \sum_{m \in Q_1} e'_m(W) \cdot \psi_m + B'$$

- B' is a constant related to acoustic scores
- a'_i through e'_m denote counts of individual LM parameters

Competing Hypotheses (Word-graph):

$$\ln \sum_{W' \in \mathcal{G}} [\Pr(W'|\Lambda) \cdot \mathcal{A}(W')] \approx \sum_{W' \in \mathcal{G}} \ln[\Pr(W'|\Lambda) \cdot \mathcal{A}(W')] \cdot \Pr(W'|\Lambda^{(n)}, \mathcal{G})$$
$$= \sum_{W' \in \mathcal{G}} \sum_{t=1}^{R_{W'}} \ln \Pr(w'_t|w'_{t-2}, w'_{t-1}) \cdot \gamma_{W'} + B''$$
$$= \sum_{i \in P_3} a''_i(\mathcal{G}) \cdot \lambda_i + \cdots + \sum_{m \in Q_1} e''_m(\mathcal{G}) \cdot \psi_m + B''$$

- B'' is a constant related to acoustic scores
- a''_i through e''_m are computed using posterior probabilities of sequences of arcs obtained from the forward-backward algorithm with history (Wessel et al., 2001)

OPTIMIZATION STEP: Maximize approximate objective function to obtain an improved LM

Approximate LME Objective Function

$$\tilde{d}(W|\Lambda) = \sum_{i \in P_3} a_i(W, \mathcal{G}) \cdot \lambda_i + \cdots + \sum_{m \in Q_1} e_m(W, \mathcal{G}) \cdot \psi_m$$

$a_i(W, \mathcal{G}) = a'_i(W) - a''_i(\mathcal{G})$, etc...

$$\mathop{\textbf{arg max}}_{\Lambda} \left[\min_{W_n \in S} \tilde{d}(W_n|\Lambda) - \epsilon \cdot \frac{1}{|\mathcal{E}|} \sum_{W_i \in \mathcal{E}} \tilde{d}(W_i|\Lambda) \right]$$

$\Rightarrow$

Approximate LME objective function as a standard LP

$$\mathop{\textbf{arg max}}_{\Lambda, \rho} \left[\rho - \frac{\epsilon}{|\mathcal{E}|} \cdot c^T \Lambda \right]$$
$$\forall\, W_n \in \mathcal{S} : \tilde{d}(W_n|\Lambda) \geq \rho$$
$$\forall\, \lambda_i \in \Lambda : \lambda_i^{(n)} - \tau \leq \lambda_i \leq \lambda_i^{(n)} + \tau$$
$$\rho \geq 0$$
$$c = \left[\sum_{W \in \mathcal{E}} a_1 \cdots \sum_{W \in \mathcal{E}} a_I \quad \ldots \quad \sum_{W \in \mathcal{E}} e_1 \cdots \sum_{W \in \mathcal{E}} e_M \right]$$

Experiments

- The discriminative training algorithm was evaluated using the (SPINE1) data set (Quiet subset): 5210 training, 2030 test utterances
- The LM was built using the CMU-Cambridge Statistical Language Modeling toolkit: 1210 unigrams, 12880 bigrams, and 27924 trigrams
- HMM model training, wordgraph generation, and recognition were done with HTK.
- GNU Linear Programming Kit was used to solve the linear programming problems.

Reduction in Word and Sentence Error Rates

	Baseline	MMIE	LME
WER (%)			
Training Set	11.97	5.49 (54.14)	5.29 (55.81)
Test Set	27.00	26.38 (2.30)	26.30 (2.59)
SER (%)			
Training Set	25.6	12.20 (52.34)	11.6 (54.69)
Test Set	43.15	42.46 (1.60)	42.36 (1.83)