Discriminative Language Modeling

Andreas Stolcke and Mitch Weintraub

SRI International

Presented at: 9th Hub-5 Conversational Speech Recognition Workshop, Linthicum Heights, MD

September 1998

Discriminative Language Modeling

Andreas Stolcke
Mitch Weintraub
Speech Technology and Research Laboratory
SRI International
Menlo Park, California

Overview

- Motivation
- Objective Functions
- Estimation of Discriminative N-gram LMs
- Experiments
- Issues for future work

Motivation

- Current LM training approaches try to minimize (unconditional) entropy of test data (= perplexity)
- If target data does not conform to model class (Gaussians, N-grams) then better classification can be expected from optimizing a *discriminative* objective function, e.g., LM entropy conditioned on acoustic data
- Discriminative training explicitly penalizes incorrect hypotheses at the expense of (more) correct ones.
- Discriminative training (potentially) allows LM to compensate for acoustic model errors, e.g., acoustic confusibility of words.
- Discriminative training has been tried variously for acoustic models (Maximum Mutual Information estimation): Bahl et al. (1983, 1986), Normandin (1991), Beaufays et al. (1998)

Discriminative Objective Functions

Define the *N-best posterior* p_k of the k-th N-best hypothesis W_k:

$$p_k = \frac{P_\theta(W_k)P(X|W_k)}{\sum_{j=1}^{N} P_\theta(W_j)P(X|W_j)}$$

$P_\theta(\cdot)$ is the language model with parameters θ
$P(X|\cdot)$ is the (fixed) acoustic model

Posterior of correct hypothesis Maximize log probability of correct (or least errorful) hypothesis W_{k^*}

$$R(\theta) = \log p_{k^*}$$

Expected Word Error Minimize average error of N-best hyps:

$$R(\theta) = -\sum_k p_k e_k$$

where e_k is error count for hypothesis W_k.

Estimation Algorithm

1. Initialize LM with smoothed maximum likelihood estimates
2. Reestimate LM parameters from a *training set* ("batch mode" parameter updates)
3. Evaluate objective function and/or word error on a held-out *cross-validation* set
4. Goto 2 while objective function or error improves

Estimation Approach 1

Perform gradient ascent on $\nabla_\theta R(\theta)$ while keeping parameters normalized (Gopalakrishnan et al. 1989):

$$\theta'_i = \frac{\frac{\partial R(\theta)}{\partial \log \theta_i} + D\theta_i}{\sum_{j=1}^{n}\left[\frac{\partial R(\theta)}{\partial \log \theta_j} + D\theta_j\right]}$$

where D is a 'suitably large' constant (in practice chosen to keep all parameters positive)

Applied to N-gram LMs:

- Jointly reestimate all N-grams with the same history (probabilities stay normalized)

How to handle back-off?

- Keep back-off mass constant, only reestimate explicit N-grams.
  Disadvantage: some N-grams never change in training.
- Or: expand all backed-off N-grams occurring in training to explicit higher-order N-grams.
  Disadvantage: creates many new parameters.

Sanity Check 1: Optimizing Unigram LM

Data

CallHome/CallFriend Spanish
44k training waveform segments
20k cross-validation waveform segments
(training + cross-validation set comprise available Spanish LVCSR training corpus)
100-best lists

Experiment

Use estimation approach 1 on a unigram LM. While unigram is a bad LM, discriminative reestimation should improve over ML unigram estimates. Note: no issue with handling back-off estimates here.

Result

NO improvement on cross-validation set, with either objective function.

Estimation Approach 2

Perform gradient ascent on $\nabla_{\log\theta} R(\theta)$ *without* normalizing: probabilities stay positive but don't sum to one.

$$\log\theta'_i = \log\theta_i + \epsilon \frac{\partial R(\theta)}{\partial \log\theta_i}$$

ϵ is step-size parameter controlling convergence speed/stability tradeoff

No problem handling backoff:

- Log backoff weights can be updated same as log probabilities
- Gradient can be propagated through backoff to lower-order N-grams, updates use cumulative gradient

Sanity Check 2

Data as in Sanity Check 1

Experiment

Use estimation approach 2 on a unigram LM

Objective function: average N-best error

Results

Iteration	Train errors	X-val errors
0	171887 (50.44)	83564 (53.46)
47	166298 (48.80)	81572 (52.19)
Lower bound	95969 (28.16)	48350 (30.93)

Non-normalized gradient ascent seems to be more effective!

Bigram Experiment

Data as before

use 1997 Spanish LVCSR eval set for testing

Experiment

Estimation approach 2 on a bigram LM

Back-off weights are reestimated, but not unigrams

Objective function: average N-best error

Results

Iter.	Train errs (%)	X-val errs (%)	Test WER
0	147417 (43.26)	72396 (46.32)	62.7
76	140073 (41.11)	71048 (45.45)	63.0
L.B.	95969 (28.16)	48350 (30.93)	

Cross-validation improvement doesn't carry over to independent test set (yet)

Issues for Future Work

Fundamental Problem

Cross-validation performance is biased because both AM and LM were trained on it

Things to try

- Parameter tying (e.g., word-dependent LM weight)
- Update lower-order N-grams as well
- Combine standard and discriminative models in test