Constrained Discriminative Training of N-gram Language Models

Ariya Rastrow[1], Abhinav Sethy[2] and Bhuvana Ramabhadran[2]

[1]Human Language Technology Center of Excellence,
and Center for Language and Speech Processing, Johns Hopkins University, MD, USA
[2]IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
ariya@jhu.edu {asethy,bhuvana}@us.ibm.com

Motivation

- Language Model plays a crucial role in identifying the correct hypothesis in many natural language processing (NLP) systems, such as Automated Speech Recognition (ASR) and Machine Translation (MT)
- Statistical Language Models (SLMs) are conventionally trained using *Maximum Likelihood* (ML) estimation on large quantities of *text*.
- Given their role in selection of the correct hypothesis from the output space of NLP systems, it is expected that language models can benefit from discriminative training.
- Among discriminative approaches proposed in the literature are:
  - Conditional Random Fields (CRF) based technique
  - Perceptron based algorithm
  - Minimum Classification Error (MCE) based discriminative training

Our Approach:

We propose a two-step procedure:
1. Generating discriminative updates based on MCE criterion
2. Applying N-gram updates such that the following issues are addressed:
   - **Updating Back-off Probabilities for N-grams** for which there is no explicit parameter value in the initial LM
   - **Normalization** to ensure that the updated N-grams conform to a valid probability distribution
   - **Global Constraint** to ensure that the trained models do not deviate too much from the initial maximum likelihood trained models

Minimum Classification Error Discriminative Training

Decode Speech Data with the Initial Model

Truth: BOB FRANKEN REPORTING LIVE TO US THIS MORNING FROM LITTLE ROCK ARKANSAS

W_i: BOB FRANKEN REPORTING LIVE TO ASSIST MORNING FROM LITTLE ROCK ARKANSAS

- An Automated Speech Recognition (ASR) system is used to decode the training data and generate multiple hypothesis (N-best). The N-gram updates are obtained by comparing the correct word sequence and the corresponding N-best list generated by the ASR system. The likelihood of any sequence of words is:

$$g(X_i, W; \Lambda, \Gamma) = \alpha \log P(X_i|W, \Lambda) + \log P(W|\Gamma)$$

- MCE Objective Function is defined as:

$$d(X_i; \Lambda, \Gamma) = -g(X_i, W_0; \Lambda, \Gamma) + G(X_i, W_1, \cdots, W_N; \Lambda, \Gamma)$$

  - N-best average score is

$$G(X_i, W_1, \cdots, W_N; \Lambda, \Gamma) = \log\left(\frac{1}{N}\sum_{r=1}^{N} \exp[g(X_i, W_r; \Lambda, \Gamma)\eta]\right)^{\frac{1}{\eta}}.$$

  - $d > 0$ implies an error in the recognition of the utterance
  - η controls the weighting of the different hypotheses in the N-best list.
- Generalized Probabilistic Descent (GPD) algorithm is used to determine the N-gram updates in log-space:

$$\frac{\partial d(X_i; \Lambda, \Gamma)}{\partial \log P_{w_y|w_x}} = -I(W_0, w_x, w_y) + \sum_{r=1}^{N} C_r I(W_r, w_x, w_y)$$

Applying N-gram Updates

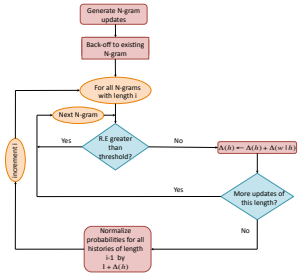


FIG. 1. *Algorithm for applying N-gram updates generated based on minimum-classification-error criterion.*

- Updating Missing N-grams
  - The back-off probability which is used to obtain the probability of the new N-gram is updated
  - This does not result in an increase in the LM size
  - Updating the back-off probabilities affects a larger number of N-grams
  - There is a need to constrain updates
- Normalization
  - The objective function does not impose any constraint on the updates which ensures that the updated LM conforms a probability distribution
  - The fact that normalization needs to be done for for all histories *h*, for which there is at least one update, makes it computationally expensive

$$\sum_{w \in S(h)} p(w|h) + \beta(h)\left(1 - \sum_{w \in S(h)} p(w|h')\right) = 1$$

  - The normalization factor for all explicit probabilities (*p(w|h)*) and back-off weights is found to be

$$1 + \sum_{w \in S(h)} \Delta(w|h)$$

  - Procedure of updating N-grams should begin with the lower order N-grams and expand to the higher order N-grams
- Relative Entropy (R.E) based Constraint
  - The N-gram updates are obtained based on local regions of mismatches of ASR output w.r.t the reference
  - The global effect of the updated N-grams on the language model needs to be considered to ensure they do not deviate too much from maximum likelihood estimation
  - Assuming N-grams affect the state of LM roughly independently, *Kullback-Leibler* (KL) distance at each step can be calculated as

$$D(q_j||q_{j+1}) = q_j(h)\left[\log(1 + \Delta(w_x|h)) - q_j(w_x|h).\delta(w_x|h)\right]$$

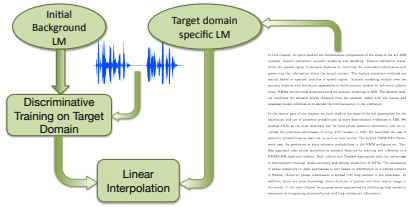
  - Using a threshold (tuned on held-out set) for above KL distance, the final set of N-grams is selected

Language Model Adaptation

- Language Model Adaptation is crucial when the training data does not match the test data being decoded
  - Adapting to new domain/genre
- Linear Interpolation based methods are most commonly used to adapt LMs to new domain

$$p(w|h) = \lambda p_B(w|h) + (1 - \lambda)p_A(w|h)$$

- Discriminative Training followed by adaptation
  - Background model is first discriminatively trained using speech data from target domain
  - In the second step, the discriminatively trained LM is interpolated with target specific LM



Experimental Setup

- The LVCSR system is based on the *2007 IBM Speech recognition system* for GALE Distillation Go/No-go Evaluation. The acoustic models used in this system are state-of-the-art discriminatively trained models and are the same ones used for all experiments presented in this work.
- For LM adaptation experiments, the background LM (p_B) is the Broadcast News LM which is built on following training text:
  - 1996 CSR Hub4 Language Model data
  - EARS BN03 closed captions
  - GALE Phase 2 Distillation GNG Evaluation Supplemental Multilingual data
  - Hub4 acoustic model training transcripts
  - TDT4 closed captions
  - TDT4 newswire
  - GALE Broadcast Conversations
  - GALE Broadcast News
- The MIT lectures data set (176K words, 21 hours, 20 lectures given by two speakers) serves as the target domain set (p_A) for language model adaptation experiments
  - 16 hours for building target specific LM and use as speech data for discriminative training
  - 2.5 hours for evaluation
  - 2.5 hours as development set (for tuning *R.E* threshold)
- The OOV rate on the target (MIT) domain using the source (BN) domain lexicon is high (1.65%). However, acoustic scores obtained from reference alignments are needed for discriminative training
  - The reference (truth) is substituted with the oracle path of the lattice and serves as a sloppy reference

Result

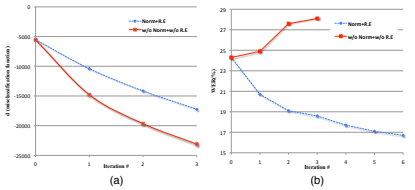


FIG. 2. *Result on the training: (a) sum of d (misclassification function) over all utterances (b) WER(%) for different iterations using both methods.*

Language Model	Training	Dev	Eval
BN-LM (%)	24.3	24.7	24.7
Interp.-LM(%)	-	17.9	18.5
Disc. Train-First Pass (%)	18.6	22.4	22.7
Disc. Train-Second Pass (%)	16.7	22.1	22.6
Disc.Train-First Pass+16 hour LM(%)	-	**17.5**	**18.2**
Disc.Train-Second Pass+16 hour LM(%)	-	17.5	18.2

TABLE I. *WER(%) for discriminatively trained and interpolated LMs on the training, development and evaluation test set. The best performance is achieved when discriminatively-trained LM is interpolated with the MIT-LM*

LM	$WER(\%)$
BN-LM	14.1
Disc. Train-First Pass	12.5
Disc. Train-Second Pass	12.4

TABLE II. *Oracle WER(%) of the baseline and discriminatively-trained LMs on the training data*

N-gram Updates	Overlap w/ Eval N-grams(%)
Regular Updates	6.23
Back-off Updates	30.45

TABLE III. Overlap between evaluation N-grams and discriminative updates

Conclusion

We have introduced a framework for discriminative training of language models. The following key points summarize this work:

- ✓ Relative entropy based constraint and normalization allow for regularization of N-gram updates
- ✓ Discriminative training on out-of-domain data serves as an adaptation method. The best performance is achieved when discriminatively trained LM is interpolated with an LM built on the out-of-domain text
- ✓ The overall performance improvements for adaptation are modest and additive to standard linear-interpolation method

References:

[1] H.-K.J.Kuo, E.Fosler-Lussier, H.Jiang, and C.-H.Lee, "Discriminative training of language models for speech recognition," in Proc. ICASSP, vol. 1, 2002, pp. 325–328.
[2] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in ACL, 2004, pp. 47–54.
[3] A. Stolcke, "Entropy-based pruning of back-off language models," in Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.
[4] M.Bacchiani, B.Roark, and M.Saraclar, "Unsupervised language model adaptation," in Proc. ICASSP, 2003, pp. 224–227.