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 Automatic 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 SLM.
   - 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 ideal maximum likelihood trained models.

Minimum Classification Error (MCE) based discriminative training

An Automated Speech Recognition (ASR) system is used to decode the training data and generate multiple n-best hypotheses. 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:

\[
g(X_n; W,T,A,\Gamma) = \alpha \log P(X_n; W,T) + \log P(W|T)\]

MCE Objective Function is defined as:

\[
d(X_n; A,\Gamma) = -(\alpha g(X_n; W,T,A,\Gamma) + \beta g(X_n; W,T,A,\Gamma))\]

where:
- \(X_n\) is n-gram update
- \(X_n = x_{n-1}x_n\) is n-gram
- \(W\) is training set
- \(T\) is target specific

N-best average score is:

\[
g(X_n; W,T,A,\Gamma) = \sum_{i=1}^{N} \log P(x_{n-1}x_n|W,T,A,\Gamma)\]

\(g(x_{n-1}x_n)\) is an error in the recognition of the utterance.

The global update function for all histories \(\hat{X}\) is obtained by taking the logarithm of the product of the N-best average score of each history:

\[
\hat{g}(X_n; W,T,A,\Gamma) = \log \sum_{i=1}^{N} \log P(x_{n-1}x_n|W,T,A,\Gamma)\]

The normalized back-off weight is found to be:

\[
\hat{g}(X_n; W,T,A,\Gamma) = \frac{\sum_{i=1}^{N} \log P(x_{n-1}x_n|W,T,A,\Gamma)}{\sum_{i=1}^{N} \log P(x_{n-1}x_n|W,T,A,\Gamma)}\]

Procedure of updating n-grams should begin with the lower order n-grams and expand to the higher order n-grams.

Relative Entropy (RE) based Constraint

The n-gram updates are obtained based on local regions of n-grams and the reference.

The global update of the n-grams on the language model needs to be constrained to ensure they do not deviate too much from maximum likelihood estimation.

Assuming N-grams affect the state of LM roughly independently, Back-off-kullback (KL) distance at each step can be calculated as:

\[
\hat{D}_{\text{KL}}(\hat{p}(w|\hat{X}) || p(w|x_{n-1}x_n|X_n; W,T,A,\Gamma)) = \sum_{w \in x_{n-1}x_n} \hat{p}(w) \log \frac{\hat{p}(w)}{p(w|x_{n-1}x_n|X_n; W,T,A,\Gamma)}\]

Using a threshold (based on hold-out set) for above KL distance, the final set of n-grams is selected.

Applying N-gram Updates

- 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 large 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 to a probability distribution.
- The fact that normalization needs to be done for all histories \(\hat{X}\) for which there is at least one update, makes it computationally expensive.

\[
\sum_{x_{n-1}x_n} \hat{p}(w|x_{n-1}x_n|X_n; W,T,A,\Gamma) = 1\]

The normalization factor for all explicit probabilities \(p(w|x_{n-1}x_n|X_n; W,T,A,\Gamma)\) and back-off weights is found to be:

\[
1 + \sum_{x_{n-1}x_n} \hat{p}(w|x_{n-1}x_n|X_n; W,T,A,\Gamma)\]

Procedure of updating n-grams should begin with the lower order n-grams and expand to the higher order n-grams.

Relative Entropy (RE) based Constraint

The n-gram updates are obtained based on local regions of n-grams and the reference.

The global update of the n-grams on the language model needs to be constrained to ensure they do not deviate too much from maximum likelihood estimation.

Assuming N-grams affect the state of LM roughly independently, Back-off-kullback (KL) distance at each step can be calculated as:

\[
\hat{D}_{\text{KL}}(\hat{p}(w|\hat{X}) || p(w|x_{n-1}x_n|X_n; W,T,A,\Gamma)) = \sum_{w \in x_{n-1}x_n} \hat{p}(w) \log \frac{\hat{p}(w)}{p(w|x_{n-1}x_n|X_n; W,T,A,\Gamma)}\]

Using a threshold (based on hold-out set) for above KL distance, the final set of n-grams is selected.

Experimental Setup

- The LVCSP system is based on the 2007 IBM Speech recognition system for DARPA Distribution 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:
  - TITUS Spanish Language Multisite
  - DARPA BMDL cover corpus
  - MUSC2 Storage
  - Saudi-Arabia
  - TID 07 spoken text
  - TID 07 phone
  - GRN-2006 Broadcast News
  - GRN-2006 Broadcast News

- The MT adaptation data set is (176K words, 21 hours, 30 lectures given by two speakers) serves as the target domain set (\(T_D\)) for language model adaptation experiments.
  - 14 hours for building specific LM and use as speech data for discriminative learning.
  - 2.5 hours for models.
  - 2.5 hours for development set (for tuning RL threshold).

- The OCR data on the target (MT) domain using the source (BN)-domain lexicon is high (97.60%). However, acoustic scores obtained from reference alignments are needed for discriminative training.
  - The reference (hyp) is substituted with the oracle path of the lattice and serves as a noisy reference.

Language Model Adaptation

- Language Model Adaptation is crucial when the training data does not match the training data set being decoded.
  - Adapting to new domain/gene
  - Linear interpolation methods are most commonly used to adapt LMs to new domains:
  \[
p(w|h) = \lambda p_{\text{adm}}(w|h) + (1 - \lambda)p_{\text{adm}}(w|h)\]
  - 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.

Table I: MCE% of discriminatively trained LMs in the training data.

<table>
<thead>
<tr>
<th>LM Type</th>
<th>MCE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>10</td>
</tr>
<tr>
<td>Norm</td>
<td>5</td>
</tr>
<tr>
<td>Rx%</td>
<td>3</td>
</tr>
</tbody>
</table>

Table II: Overview of the adaptation and discriminative training.

<table>
<thead>
<tr>
<th>Condition</th>
<th>MCE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>+iscx Train</td>
<td>5</td>
</tr>
<tr>
<td>+iscx Train</td>
<td>10</td>
</tr>
<tr>
<td>+iscx Train</td>
<td>15</td>
</tr>
</tbody>
</table>

Table III: Overview of the discriminative training and adaptation.

<table>
<thead>
<tr>
<th>Condition</th>
<th>MCE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>+iscx Train</td>
<td>5</td>
</tr>
<tr>
<td>+iscx Train</td>
<td>10</td>
</tr>
<tr>
<td>+iscx Train</td>
<td>15</td>
</tr>
</tbody>
</table>

Table IV: Overview of the discriminative training and adaptation.

<table>
<thead>
<tr>
<th>Condition</th>
<th>MCE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>+iscx Train</td>
<td>5</td>
</tr>
<tr>
<td>+iscx Train</td>
<td>10</td>
</tr>
<tr>
<td>+iscx Train</td>
<td>15</td>
</tr>
</tbody>
</table>

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 ML is interpolated with an LM built on the out-of-domain text.
- Overall performance improvements for adaptation are modest and additive to standard error/interpolation method.

References:
