Lattice Segmentation and Minimum Bayes Risk Discriminative Training

Vlasios Doumpiotis, Stavros Tsakalidis, William Byrne

Center for Language and Speech Processing
Department of Electrical and Computer Engineering
The Johns Hopkins University
Baltimore, MD  U.S.A.
Minimum Bayes-Risk Speech Recognition-Decoding Under a Loss Function

Goal: Decoders that minimize the Expected Loss \( E_{P(W,A)} [l(W, \delta(A))] \)

1. Evaluate the expected loss of each hypothesis \( W' \)
\[
E(W') = \sum_{W \in W} l(W, W') P(W|A)
\]

2. Select the hypothesis with least expected loss:
\[
\delta_{MBR}(A) = \arg\min_{W' \in W} \sum_{W \in W} l(W, W') P(W|A)
\]

For efficiency these operations are carried out over lattices

MBR often improves MAP hypotheses :
e.g. : ROVER, Information Retrieval, Statistical Machine Translation
Particularly useful for system combination

Given a loss function and a set of models, MBR is all about search
Segmental Minimum Bayes Risk Decoding

- Lattice MBR can be expensive - suppose we chop up the lattice?

- Segmentation induces a loss function: \( l_t(W, W') = l(W_1, W_1') + l(W_2, W_2') + l(W_3, W_3') \)

- Lattice MBR becomes a sequence of smaller, simpler MBR operations

\[
\delta(A) = \arg\min_{W' \in W_1} \sum_{W \in W_1} l(W, W') P_1(W|A) \cdot \arg\min_{W' \in W_2} \sum_{W \in W_2} l(W, W') P_2(W|A) \cdot \arg\min_{W' \in W_3} \sum_{W \in W_3} l(W, W') P_3(W|A)
\]

- Problem: Any non-ideal segmentation restricts string alignments.

- Trade-off between search and approximation errors:
  - Fewer search errors in the sub-problems
  - Possibly poor approximation to the original search problem
**Discriminative Training for MBR**

- MBR decoding can improve over MAP for a given set of models
- If the goal is to apply MBR (or ROVER or system combination), why not train the models with that in mind?

  Suppose we have labelled training data \((\tilde{W}, A)\)

  One possibility is to minimize the *Expected Risk*

  \[
  \theta^* = \arg\min_{\theta} \sum_{W' \in \mathcal{W}} l(\tilde{W}, W') P(W'|A; \theta)
  \]

  - Variations of this have been studied for ASR

    Kaiser, Horvat, & Kacic, 2000; Povey & Woodland, 2002

    - good idea
    - difficult to implement
    - What happens if we use the loss function induced by lattice pinching?
Pinched Lattice MMI

- If restricted to loss functions induced lattice pinching, minimum risk estimation becomes

\[ \theta^* = \arg\min_\theta \sum_{i=1}^{N} \sum_{W' \in \tilde{W}_i} l(\tilde{W}_i, W') P_i(W'|A, \tilde{W}; \theta) \]

- Suppose the segments are pruned wrt \( P_i(W'|A, \tilde{W}; \theta) \) leaving at most binary choices

For many segments \( l(\tilde{W}_i, W') = 0 \ \forall \ W' \in \tilde{W}_i \)

- Let \( C \) be the indices of the sets \( \tilde{W}_i \) that are identified to contain errors

\[ \theta^* = \arg\max_\theta \sum_{i \in C} P_i(\tilde{W}_i|A, \tilde{W}; \theta). \]

- Implemented via MMI over pinched lattices

  - Pinched Lattice MMI improves the probability of the truth in the low-confidence regions
  - By contrast, MMI improves the posterior probability of the entire sentence
SMBR - Discriminative Training

Training criterion captures global risk

\[ \theta^* = \arg\max_{\theta} \sum_{i \in C} P_i(\tilde{W}_i|A, \tilde{W}; \theta). \]

But many different types of errors can be found in the segments in \( C \)

Why not train individual models for different error types?

Partition the error segments into different types: \( C = \{C_1, \ldots, C_K\} \)

Introduce different models for each error type and optimize each separately:

\[ \theta_k^* = \arg\max \sum_{i \in C_k} P_i(\tilde{W}_i|A; \tilde{W}; \theta_k) \quad k = 1, \ldots, K \]
Lattice Cutting for Estimation and Search

Simple task - OGI Alphadigits

Generate training set lattices:

- Segment lattices (PLC-1)
- Focus on low confidence confusion sets and select frequently observed confusion pairs
  - Train specialized sets of models to resolve each distinct confusion

On the test set:

- Segment lattices with respect to the MAP hypothesis and pinch
- Rescore the pinched lattices using the specialized model sets
SMBR-DT Performance

- OGI AlphaDigits - 37 word vocabulary, no language model
- Attempt to correct 50 binary error types identified by Period 1 lattice cutting

![Graph showing SMBR-DT Performance]

- WER vs. Iteration
- Lines represent different methods: MMI, MMIpl, SMBR-DT
Within-Class Error Analysis

- SMBR-DT avoids MMI’s tendency to fix some error types at the expense of other types.
- Arguably due to a better match between training and evaluation criteria.

Eurospeech, 2003  Lattice Segmentation and Minimum Bayes Risk Discriminative Training
**Acoustic Codebreaking**

An unsupervised, divide-and-conquer approach to ASR:

- Perform a first decoding pass to generate lattices
- Perform lattice segmentation to locate and characterize recognition errors

Unsupervised identification of recognition errors

- Search training data to create training sets for all these errors
- Train specialized models for each type of error

Can use different types of models for different types of errors:

- SVMs, language models, pronunciation models, ...

- Decode pinched lattice using the specialized models
- *Repeat until all errors are resolved*

Currently being applied to LVCSR
Conclusions

• Discriminative Training for Segmental Minimum Bayes Risk rescoring
  • Better match between the criteria that determine training and testing algorithms
  • Implemented by MMI with pinched lattices and extended model sets
• SMBR - Acoustic Codebreaking
  • identifies regions likely to contain errors
  • models are trained to resolve these errors
  • models are incorporated in subsequent SMBR decoding passes
• Goal is to apply these techniques to large vocabulary ASR
  • Major issue is how to identify consistent and manageable confusion sets at a level of detail consistent with acoustic modeling