Task-Specific Minimum Bayes-Risk Decoding using Learned Edit Distance

Izhak Shafran
The Johns Hopkins University

Collaborator: William Byrne
Motivation

We know that:

- ASR lattice contains a large number of hypotheses.
- Most decoders use MAP as their best guess.
- WER of MAP $\gg$ WER of the best hypothesis (Oracle).

Can we pick a hypothesis better than MAP? How?
MAP is Not Optimal for WER

Maximum *Aposteriori* (MAP) Estimate:

\[ \hat{W} = \arg \max_W P(W|O) = \arg \max_W P(O|W)P(W) \]

**MAP** can be derived from, the more general, Minimum Bayes-Risk (MBR) decision rule, which is given by:

\[ \hat{W} = \arg \min_W R(W) = \arg \min_W \sum_{W'} C(W, W')P(W'|O) \]

**MAP ≡ MBR** if \( C(W, W') = 1 - \delta(W, W') \).

So, **MAP optimizes Sentence Error Rate, not WER!**
Outline

- Previous Work
- Task-Specific Minimum Bayes-Risk
- Experiments on a Large Vocabulary Task
- Conclusions
Mangu et al 2001:

- Collapse lattices into sausages using heuristics.
- Instead of picking a path in the lattice, the problem is simplified to picking a word at each position.
- Words are selected using a set of rules.
- Rules learned automatically from a training corpus.
Error Correcting LM

Roark et al 2004, Murat et al 2004:

- Re-score lattice w/ an error corrective language model.
- Pick the best scoring hypothesis.
- N-gram weights of the error correcting LM is learned using a perceptron algorithm or conditional random field.
Direct Use of Minimum Bayes-Risk

\[ \hat{W} = \arg \min_W R(W) = \arg \min_W \sum_{W'} C(W, W') P(W'|O) \]

When \( C(W, W') \) is the edit distance, it optimizes WER!

For each \( W \), \( R(W) \) is the sum of weighted edit distance from other \( W \)'s. Typically, the distances are computed using fixed elementary costs (e.g. \( \text{ins}=3 \), \( \text{del}=3 \), \( \text{sub}=4 \)).
Consider three hypotheses:

1. Look who’s here
2. Luke is here
3. Yeah right here

Acoustically, the (1) and (2) are very different from (3). But, all substitutions costs the same in standard edit distance. So,

\[ C(W_1, W_2) = C(W_2, W_3) = C(W_3, W_1) = 2 \times 4 \]

These distances do not reflect the acoustic similarities!
Need a Learned Edit Distance

Replace fixed elementary costs with costs that reflect acoustic similarities and the confusabilities inherent to the task.

\[ c(\text{\textit{who's}}, \text{\textit{is}}) < c(\text{\textit{who's}}, \text{\textit{yeah}}) \]

Learn elementary edit distance from data.
Consider the following model:

\[ \hat{W} = \arg \max_W \sum_{W'} P(W|W')P(W'|O) \]

Under this model, the best hypothesis is:
The decision rule for the stochastic model is:

\[ \hat{W} = \arg \max_W \sum_{W'} P(W|W') P(W'|O) \]

Note, the similarity with MBR decoding:

\[ \hat{W} = \arg \min_W \sum_{W'} C(W, W') P(W'|O) \]

For a given alignment, \( a \), using simple memoryless model:

\[ P(W|W') = \prod_i P(W_i = w | W_{a(i)} = w') \]

So, we choose elementary edit cost:

\[ c(w, w') = - \log P(w|w') \quad (\text{Ristad et al 1999}) \]
MALACH Task

- Corpus: Testimonies of Czech Holocaust survivors.
- 42k Vocab, contains large number of colloquial words.
- 84 hrs of training data, and 2 hrs of test.
- AM: 3-state HMM, 4k states, 16-mix GMMs, trained using MMI criterion.
Parameter Estimation

\[ P(w|w') \] was estimated using a Viterbi approx.

1. Decode the training data using the recognition model.
2. Compute the best alignment with transcripts using current \( P(w|w') \).
3. Re-estimate new \( P(w|w') \).
4. Iterate 2-3.

Note:

- To replicate test conditions, it is important to exclude the transcript of the utterance from the LM used for decoding.
- \( P(w|w') \) was smoothed using a simple additive constant.
Parameter Estimation

Costs of edit operations learned from the corpus:
ASR Results

<table>
<thead>
<tr>
<th>Decoder</th>
<th>Word Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MAP</td>
<td>45.4</td>
</tr>
<tr>
<td></td>
<td>N=50</td>
</tr>
<tr>
<td></td>
<td>N=200</td>
</tr>
<tr>
<td>2. MBR with untrained costs</td>
<td>45.2</td>
</tr>
<tr>
<td>3. MBR with learned edit distance</td>
<td>44.5</td>
</tr>
<tr>
<td></td>
<td>44.4</td>
</tr>
</tbody>
</table>

Advantages of MBR with learned edit distance:

1. Better than MAP (45.4 ↔ 44.4)
2. Better than MBR with untrained costs (45.0 ↔ 44.4)
3. Provides most of the gain with smaller set of hypotheses, thus makes the MBR possible at low computational cost
Conclusions

1. This work extends the minimum Bayes-risk framework to capture some of the systematic errors from ASR.

2. To learn the systematic errors we have introduced a stochastic transduction model, whose parameters were learned from data.

3. Our experiments demonstrate significant gains over MAP (1%) and MBR with untrained edit distance (0.6%).

4. Further, most of the gain in performance is obtained by considering only as few as 50 N-best hypotheses. This significantly reduces the computational cost of MBR.

5. The technique is general enough to be applicable in other task such as MT and bio-informatics.
Questions ...