Discriminative Training II: MERT

April 3, 2014
The Noisy Channel

\[-\log p(g|e)\]

\[-\log p(e)\]
As a Linear Model

\[-\log p(g|e)\]

\[-\log p(e)\]
As a Linear Model

\[-\log p(g|e)\]

\[-\log p(e)\]
As a Linear Model
As a Linear Model

\[-\log p(g|e)\]

Improvement 1:

change \( \tilde{w} \) to find better translations
As a Linear Model

\[-\log p(\mathbf{g}|\mathbf{e})\]

\[-\log p(\mathbf{e})\]

\[\mathbf{w}\]
As a Linear Model

\[-\log p(g | e)\]

\[-\log p(e)\]
As a Linear Model

\[-\log p(g|e)\]

\[-\log p(e)\]
As a Linear Model

Improvement 2:

Add dimensions to make points separable
Parameter Learning: Review
K-Best List Example

$h_1$ vs $h_2$

- $0.8 \leq \ell < 1.0$
- $0.6 \leq \ell < 0.8$
- $0.4 \leq \ell < 0.6$
- $0.2 \leq \ell < 0.4$
- $0.0 \leq \ell < 0.2$
Fit a linear model
Fit a linear model
K-Best List Example

- 0.8 ≤ ℓ < 1.0
- 0.6 ≤ ℓ < 0.8
- 0.4 ≤ ℓ < 0.6
- 0.2 ≤ ℓ < 0.4
- 0.0 ≤ ℓ < 0.2
Limitations

• We can’t optimize corpus-level metrics, like BLEU, on a test set
• These don’t decompose by sentence!
• We turn now to a kind of “direct cost minimization”
MERT

- **Minimum Error Rate Training**
- Directly target an automatic evaluation metric
  - BLEU is defined at the corpus level
  - MERT optimizes at the corpus level
- **Downsides**
  - Does not deal well with > ~20 features
MERT

Given weight vector $w$, any hypothesis $\langle e, a \rangle$ will have a (scalar) score $m = w^\top h(g, e, a)$

Now pick a search vector $v$, and consider how the score of this hypothesis will change:

$$w_{\text{new}} = w + \gamma v$$
MERT

Given weight vector \( w \), any hypothesis \( \langle e, a \rangle \) will have a (scalar) score \( m = w^\top h(g, e, a) \)

Now pick a search vector \( v \), and consider how the score of this hypothesis will change:

\[
\begin{align*}
  w_{new} &= w + \gamma v \\
  m &= (w + \gamma v)^\top h(g, e, a)
\end{align*}
\]
MERT

Given weight vector $w$, any hypothesis $\langle e, a \rangle$ will have a (scalar) score $m = w^\top h(g, e, a)$

Now pick a search vector $v$, and consider how the score of this hypothesis will change:

$$ w_{\text{new}} = w + \gamma v $$

$$ m = (w + \gamma v)^\top h(g, e, a) $$

$$ = w^\top h(g, e, a) + \gamma v^\top h(g, e, a) $$
MERT

Given weight vector \( \mathbf{w} \), any hypothesis \( \langle e, a \rangle \) will have a (scalar) score \( m = \mathbf{w}^\top \mathbf{h}(g, e, a) \)

Now pick a search vector \( \mathbf{v} \), and consider how the score of this hypothesis will change:

\[
\mathbf{w}_{\text{new}} = \mathbf{w} + \gamma \mathbf{v} \\
m = (\mathbf{w} + \gamma \mathbf{v})^\top \mathbf{h}(g, e, a) \\
= \mathbf{w}^\top \mathbf{h}(g, e, a) + \gamma \mathbf{v}^\top \mathbf{h}(g, e, a) \\
= \mathbf{w}^\top \mathbf{h}(g, e, a) + \gamma \mathbf{v}^\top \mathbf{h}(g, e, a) \\
= a \gamma + b
\]
Given weight vector $w$, any hypothesis $\langle e, a \rangle$ will have a (scalar) score $m = w^\top h(g, e, a)$

Now pick a search vector $v$, and consider how the score of this hypothesis will change:

$$w_{\text{new}} = w + \gamma v$$

$$m = (w + \gamma v)^\top h(g, e, a)$$

$$= w^\top h(g, e, a) + \gamma v^\top h(g, e, a)$$

$$m = a\gamma + b$$

*Linear function in 2D!*
MERT

\[ m \]

\[ \gamma \]
Recall our k-best set \( \{ \langle e_i^*, a_i^* \rangle \}_{i=1}^K \)
Recall our k-best set \( \{ \langle e_i^*, a_i^* \rangle \}_{i=1}^K \)
\[ \langle e_{162}^*, a_{162}^* \rangle \quad \text{and} \quad \langle e_{28}^*, a_{28}^* \rangle \]

\[ \langle e_{73}^*, a_{73}^* \rangle \]
MERT

\[ \langle e_{162}^*, a_{162}^* \rangle \]

\[ \langle e_{28}^*, a_{28}^* \rangle \]

\[ \langle e_{73}^*, a_{73}^* \rangle \]

events

\( m \)

\( \gamma \)
MERT

\[ \langle e_{162}^*, a_{162}^* \rangle \]

\[ \langle e_{28}^*, a_{28}^* \rangle \]

\[ \langle e_{73}^*, a_{73}^* \rangle \]

errors
MERT

\[ m \]

\[ \gamma \]

errors

---

---

---
MERT
Let $w_{new} = \gamma^* v + w$
MERT

• In practice “errors” are sufficient statistics for evaluation metrics (e.g., BLEU)
• Can maximize or minimize
• How do you pick the search direction?
Dynamic Programming
MERT
Other Algorithms

- Given a hypergraph translation space
- In the Viterbi (Inside) algorithm, there are two operations
  - **Multiplication** (extend path)
  - **Maximization** (choose between paths)
- **Semirings** generalize these to compute other quantities
# Semirings

<table>
<thead>
<tr>
<th>semiring</th>
<th>$\mathbb{K}$</th>
<th>$\oplus$</th>
<th>$\otimes$</th>
<th>$\bar{0}$</th>
<th>$\bar{1}$</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean</td>
<td>${0,1}$</td>
<td>$\lor$</td>
<td>$\land$</td>
<td>$0$</td>
<td>$1$</td>
<td>idempotent</td>
</tr>
<tr>
<td>count</td>
<td>$\mathbb{N}_0 \cup {\infty}$</td>
<td>$+$</td>
<td>$\times$</td>
<td>$0$</td>
<td>$1$</td>
<td></td>
</tr>
<tr>
<td>probability</td>
<td>$\mathbb{R}_+ \cup {\infty}$</td>
<td>$+$</td>
<td>$\times$</td>
<td>$0$</td>
<td>$1$</td>
<td></td>
</tr>
<tr>
<td>tropical</td>
<td>$\mathbb{R} \cup {-\infty, \infty}$</td>
<td>max</td>
<td>$+$</td>
<td>$-\infty$</td>
<td>$0$</td>
<td>idempotent</td>
</tr>
<tr>
<td>log</td>
<td>$\mathbb{R} \cup {-\infty, \infty}$</td>
<td>$\oplus_{\log}$</td>
<td>$+$</td>
<td>$-\infty$</td>
<td>$0$</td>
<td></td>
</tr>
</tbody>
</table>
Inside Algorithm

\[ \alpha(q_{\text{goal}}) = \bigoplus_{d \in \mathcal{G}} \bigotimes_{e \in d} w(e) \]

1: function INSIDE(\mathcal{G}, K)  \quad \triangleright \ \mathcal{G} \text{ is an acyclic hypergraph and } K \text{ is a semiring}
2: for \ q \text{ in topological order in } \mathcal{G} \text{ do}
3: \quad \text{if } B(q) = \emptyset \text{ then}
4: \quad \quad \alpha(q) \leftarrow 1 \quad \quad \quad \triangleright \ \text{assume states with no in-edges are axioms}
5: \quad \text{else}
6: \quad \quad \alpha(q) \leftarrow 0
7: \quad \quad \text{for all } e \in B(q) \text{ do}
8: \quad \quad \quad k \leftarrow w(e)
9: \quad \quad \quad \text{for all } r \in t(e) \text{ do}
10: \quad \quad \quad \quad k \leftarrow k \otimes \alpha(r) \quad \quad \triangleright \ \text{all in-coming edges to node } q
11: \quad \quad \alpha(q) \leftarrow \alpha(q) \oplus k \quad \quad \triangleright \ \text{all tail (previous) nodes of edge } e
12: \text{return } \alpha
Point-Line Duality

- Represent a set of lines as a set of points (and vice-versa)

- \( y = mx + b \Rightarrow (m, -b) \)

- The slope between dual points is the intersection x-axis of the pair of lines

- An upper envelope is dual to a lower convex hull
Convex Hull Semiring

**Definition 2. The Convex Hull Semiring.**

Let \((K, \oplus, \otimes, \overline{0}, \overline{1})\) be defined as follows:

<table>
<thead>
<tr>
<th>(K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A set of points in the plane that are the extreme points of a convex hull.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(A \oplus B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{conv} \ [A \cup B])</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(A \otimes B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>convex hull of the Minkowski sum, i.e., (\text{conv}{(a_1 + b_1, a_2 + b_2) \mid (a_1, a_2) \in A \land (b_1, b_2) \in B})</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(\overline{0})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\emptyset)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(\overline{1})</th>
</tr>
</thead>
<tbody>
<tr>
<td>{(0,0)}</td>
</tr>
</tbody>
</table>

**Theorem 1.** The Convex Hull Semiring fulfills the semiring axioms and is commutative and idempotent.
Theorem 2

- The Inside algorithm with the computes the convex hull dual to the MERT upper envelope generated from the \( \infty \)-best list of derivations
Summary

• Evaluation metrics
  • Figure out how well we’re doing
  • Figure out if a feature helps
  • Train your system
• What’s a great way to improve translation?
  • Improve evaluation!