Decoding Revisited: Easy-Part-First & MERT

February 26, 2015
Translating the Easy Part First?

The tourism initiative addresses this for the first time.

Both hypotheses translate 3 words. The worse hypothesis has a better score.
Estimating Future Cost

- Future cost estimate: how expensive is translation of rest of sentence?
- Optimistic: choose cheapest translation options
- Cost for each translation option
  - translation model: cost known
  - language model: output words known, but not context
    → estimate without context
  - reordering model: unknown, ignored for future cost estimation
Cost Estimates from Translation Options

The tourism initiative addresses this for the first time

-1.0  -2.0  -1.5  -2.4  -1.4  -1.0  -1.0  -1.9  -1.6

-4.0  -2.5  -2.2

-1.3  -2.4

-2.7

-2.3

-2.3

-2.3

Cost of cheapest translation options for each input span (log-probabilities)
## Cost Estimates for all Spans

- Compute cost estimate for all contiguous spans by combining cheapest options

<table>
<thead>
<tr>
<th>first word</th>
<th>future cost estimate for $n$ words (from first)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>-1.0</td>
</tr>
<tr>
<td>tourism</td>
<td>-2.0</td>
</tr>
<tr>
<td>initiative</td>
<td>-1.5</td>
</tr>
<tr>
<td>addresses</td>
<td>-2.4</td>
</tr>
<tr>
<td>this</td>
<td>-1.4</td>
</tr>
<tr>
<td>for</td>
<td>-1.0</td>
</tr>
<tr>
<td>the</td>
<td>-1.0</td>
</tr>
<tr>
<td>first</td>
<td>-1.9</td>
</tr>
<tr>
<td>time</td>
<td>-1.6</td>
</tr>
</tbody>
</table>

- Function words cheaper (*the*: -1.0) than content words (*tourism*: -2.0)
- Common phrases cheaper (*for the first time*: -2.3) than unusual ones (*tourism initiative addresses*: -5.9)
Combining Score and Future Cost

- Hypothesis score and future cost estimate are combined for pruning
  - left hypothesis starts with hard part: the tourism initiative
    score: -5.88, future cost: -6.1 → total cost -11.98
  - middle hypothesis starts with easiest part: the first time
    score: -4.11, future cost: -9.3 → total cost -13.41
  - right hypothesis picks easy parts: this for ... time
    score: -4.86, future cost: -9.1 → total cost -13.96
f: Maria no dio una bofetada a la bruja verde

Future costs make these hypotheses comparable.
Other Decoding Algorithms

- A* search
- Greedy hill-climbing
- Using finite state transducers (standard toolkits)
A* Search

- Uses *admissible* future cost heuristic: never overestimates cost
- Translation agenda: create hypothesis with lowest score + heuristic cost
- Done, when complete hypothesis created
Greedy Hill-Climbing

• Create one complete hypothesis with depth-first search (or other means)

• Search for better hypotheses by applying change operators
  – change the translation of a word or phrase
  – combine the translation of two words into a phrase
  – split up the translation of a phrase into two smaller phrase translations
  – move parts of the output into a different position
  – swap parts of the output with the output at a different part of the sentence

• Terminates if no operator application produces a better translation
Decoding algorithm

• Translation as a search problem

• Partial hypothesis keeps track of
  • which source words have been translated (coverage vector)
  • $n-1$ most recent words of English (for LM!)
  • a back pointer list to the previous hypothesis + (e,f) phrase pair used
  • the (partial) translation probability
  • the estimated probability of translating the remaining words (precomputed, a function of the coverage vector)

• Start state: no translated words, E=<s>, bp=nil

• Goal state: all translated words
Decoding algorithm

• \( Q[0] \leftarrow \text{Start state} \)
• for \( i = 0 \) to \(|f|-1\)
  • Keep \( b \) best hypotheses at \( Q[i] \)
  • for each hypothesis \( h \) in \( Q[i] \)
    • for each untranslated span in \( h.c \) for which there is a translation \(<e,f>\) in the phrase table
      • \( h' = h \) extend by \(<e,f>\)
      • Is there an item in \( Q[|h'.c|] \) with \( = \text{LM state} \)?
        • yes: update the item bp list and probability
        • no: \( Q[|h'.c|] \leftarrow h' \)
  • Find the best hypothesis in \( Q[|f|] \), reconstruction translation by following back pointers
Parameter Learning: Review
K-Best List Example

\[
\begin{align*}
\mathbf{w} & \quad \mathbf{h}_1 \\
\mathbf{h}_2 & \\
\end{align*}
\]

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

\[ \begin{align*}
\mathbf{w} & \quad \text{weights}\n\end{align*} \]

- \(0.8 \leq \ell < 1.0\)
- \(0.6 \leq \ell < 0.8\)
- \(0.4 \leq \ell < 0.6\)
- \(0.2 \leq \ell < 0.4\)
- \(0.0 \leq \ell < 0.2\)
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
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:

$$w_{\text{new}} = w + \gamma v$$
Now pick a search vector \( \mathbf{v} \), and consider how the score of this hypothesis will change:

\[
\mathbf{w}_{\text{new}} = \mathbf{w} + \gamma \mathbf{v}
\]

\[
m = (\mathbf{w} + \gamma \mathbf{v})^\top \mathbf{h}(\mathbf{g}, \mathbf{e}, \mathbf{a})
\]
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 $w$, 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:

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

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

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

$$m = a \gamma + b$$
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:

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

Linear function in 2D!
MERT

\[ \gamma \]

\[ m \]
Recall our k-best set $\{\langle e_i^*, a_i^* \rangle \}_{i=1}^K$
Recall our k-best set \( \{ \langle e_i^*, a_i^* \rangle \}_{i=1}^{K} \)
\[ \langle e_{162}^*, a_{162}^* \rangle \]
\[ \langle e_{28}^*, a_{28}^* \rangle \]
\[ \langle e_{73}^*, a_{73}^* \rangle \]
\[ \langle e_{162}^*, a_{162}^* \rangle \]

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

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

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

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

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

errors
MERT

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

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

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

\[ \text{errors} \]
MERT

\[ m \]

\[ \gamma \]

errors

\[ \gamma \]
Let $w_{\text{new}} = \gamma^* v + w$
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
<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 \text{\mathcal{G} is an acyclic hypergraph and K is a semiring}
2: for q in topological order in \mathcal{G} do
3: \quad if B(q) = \emptyset then
4: \quad \quad \alpha(q) \leftarrow 1 \quad \triangleright \text{assume states with no in-edges are axioms}
5: \quad else
6: \quad \quad \alpha(q) \leftarrow 0
7: \quad \quad for all e \in B(q) do
8: \quad \quad \quad k \leftarrow w(e) \quad \triangleright \text{all in-coming edges to node q}
9: \quad \quad \quad for all r \in t(e) do
10: \quad \quad \quad \quad k \leftarrow k \otimes \alpha(r) \quad \triangleright \text{all tail (previous) nodes of edge e}
11: \quad \quad \quad \alpha(q) \leftarrow \alpha(q) \oplus k
12: 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 \((\mathbb{K}, \oplus, \otimes, 0, 1)\) be defined as follows:

<table>
<thead>
<tr>
<th>(\mathbb{K})</th>
<th>A set of points in the plane that are the extreme points of a convex hull.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A \oplus B)</td>
<td>(\text{conv}[A \cup B])</td>
</tr>
<tr>
<td>(A \otimes B)</td>
<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>
<tr>
<td>(\overline{0})</td>
<td>(\emptyset)</td>
</tr>
<tr>
<td>(\overline{1})</td>
<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!