Phrase-Based Systems

Sentence-aligned corpus

Word alignments

Phrase table (translation model)

Many slides and examples from Philipp Koehn or John DeNero
The Pharaoh “Model”

\[ P(e|g) = P(\{\tilde{g}_i\}|g) \prod_{i} \phi(\tilde{e}_i|\tilde{g}_i) d(a_i - b_{i-1}) \]

Segmentation  Translation  Distortion

Phrase-Based Decoding

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dic</th>
<th>uns</th>
<th>bofedas</th>
<th>a</th>
<th>las</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
</tr>
<tr>
<td>did not</td>
<td>slap</td>
<td>by</td>
<td>the</td>
<td>green</td>
<td>witch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did not give</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
If monotonic, almost an HMM; technically a semi-HMM

for (fPosition in 1...|f|)
for (lastPosition < fPosition)
for (eContext in eContexts)
for (eOption in translations[fPosition])
  ... combine hypothesis for (lastPosition ending in eContext) with eOption

If distortion... now what?
Pruning: Beams + Forward Costs

Maria no dio una bofetada a la bruja verde

- Problem: easy partial analyses are cheaper
  - Solution 1: use beams per foreign subset
  - Solution 2: estimate forward costs (A*-like)

The Pharaoh Decoder

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio una bofetada</th>
<th>a la bruja verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>did not give a slap to the green witch</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>did not slap to the the witch</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>slap to the the witch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio una bofetada</th>
<th>a la bruja verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>did not slap the green witch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hypothesis Lattices

Word Alignment

1. Align words with a probabilistic model
2. Infer presence of larger structures from this alignment
3. Translate with the larger structures
What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?

Unsupervised Word Alignment

- Input: a *bitext*: pairs of translated sentences

  *nous acceptons votre opinion* .
  *we accept your view* .

- Output: *alignments*: pairs of translated words

  - When words have unique sources, can represent as a (forward) alignment function \( a \) from French to English positions
1-to-Many Alignments

Alignments: a hidden vector called an alignment specifies which English source is responsible for each French target word.

\[ a = a_1 \ldots a_j \]

\[
P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i) = \prod_j \frac{1}{i+1} P(f_j|e_i)
\]

\[
P(f|e) = \sum_a P(f, a|e)
\]
IBM Models 1/2

E: Thank you, I shall do so gladly.

A: 1 2 3 4 5 6 7 8 9

F: Gracias, lo haré de muy buen grado.

Model Parameters

Emissions: \( P(\text{F} = \text{Gracias} | \text{E} = \text{Thank}) \)

Transitions: \( P(\text{A}_2 = 3) \)

Evaluating TMs

- How do we measure quality of a word-to-word model?
  - Method 1: use in an end-to-end translation system
    - Hard to measure translation quality
    - Option: human judges
    - Option: reference translations (NIST, BLEU)
    - Option: combinations (HTER)
    - Actually, no one uses word-to-word models alone as TMs
  - Method 2: measure quality of the alignments produced
    - Easy to measure
    - Hard to know what the gold alignments should be
    - Often does not correlate well with translation quality (like perplexity in LMs)
Alignment Error Rate

- **Alignment Error Rate**

  - ☐ = Sure
  - ☐ = Possible
  - ■ = Predicted

  \[
  AER(A, S, P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)
  \]

  \[
  = \left(1 - \frac{3 + 3}{3 + 4}\right) = \frac{1}{7}
  \]

Problems with Model 1

- There's a reason they designed models 2-5!
- Problems: alignments jump around, align everything to rare words
- Experimental setup:
  - Training data: 1.1M sentences of French-English text, Canadian Hansards
  - Evaluation metric: alignment error Rate (AER)
  - Evaluation data: 447 hand-aligned sentences

- on a enregistré 1,122,000 divorces sur le continent
- le terme ferroviaire est << chargement sur demande >>
- the railroad is demand loading
Intersected Model 1

- Post-intersection: standard practice to train models in each direction then intersect their predictions [Och and Ney, 03]

- Second model is basically a filter on the first
  - Precision jumps, recall drops
  - End up not guessing hard alignments

<table>
<thead>
<tr>
<th>Model</th>
<th>P/R</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 E→F</td>
<td>82/58</td>
<td>30.6</td>
</tr>
<tr>
<td>Model 1 F→E</td>
<td>85/58</td>
<td>28.7</td>
</tr>
<tr>
<td>Model 1 AND</td>
<td>96/46</td>
<td>34.8</td>
</tr>
</tbody>
</table>

Joint Training?

- Overall:
  - Similar high precision to post-intersection
  - But recall is much higher
  - More confident about positing non-null alignments

<table>
<thead>
<tr>
<th>Model</th>
<th>P/R</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 E→F</td>
<td>82/58</td>
<td>30.6</td>
</tr>
<tr>
<td>Model 1 F→E</td>
<td>85/58</td>
<td>28.7</td>
</tr>
<tr>
<td>Model 1 AND</td>
<td>96/46</td>
<td>34.8</td>
</tr>
<tr>
<td>Model 1 INT</td>
<td>93/69</td>
<td>19.5</td>
</tr>
</tbody>
</table>
Monotonic Translation

Japan shaken by two new quakes

Le Japon secoué par deux nouveaux séismes

Local Order Change

Japan is at the junction of four tectonic plates

Le Japon est au confluent de quatre plaques tectoniques
IBM Model 2

- Alignments tend to the diagonal (broadly at least)

\[ P(f, a|e) = \prod_j P(a_j = i|j, I, J) P(f_j|e_i) \]
\[ P(\text{dist} = i - \frac{I}{J}) \]
\[ \frac{1}{Z} e^{-\alpha(i - \frac{J}{J})} \]

- Other schemes for biasing alignments towards the diagonal:
  - Relative vs absolute alignment
  - Asymmetric distances
  - Learning a full multinomial over distances

EM for Models 1/2

- Model 1 Parameters:
  - Translation probabilities (1+2)
  - Distortion parameters (2 only)

\[ P(f_j|e_i) \]
\[ P(a_j = i|j, I, J) \]

- Start with \( P(f_j|e_i) \) uniform, including \( P(f_j|\text{null}) \)

- For each sentence:
  - For each French position \( j \)
    - Calculate posterior over English positions

\[ P(a_j = i|f, e) = \frac{P(a_j = i|j, I, J) P(f_j|e_i)}{\sum_{i'} P(a_j = i'|j, I, J) P(f_j|e_{i'})} \]

- (or just use best single alignment)
- Increment count of word \( f \) with word \( e \) by these amounts
- Also re-estimate distortion probabilities for model 2

- Iterate until convergence
On Tuesday Nov. 4, earthquakes rocked Japan once again.

Des tremblements de terre ont à nouveau touché le Japon jeudi 4 novembre.
The HMM Model

E:
1  2  3  4  5  6  7  8  9
Thank you , I shall do so gladly.

A:

F:
Gracias , lo haré de muy buen grado.

Model Parameters
Emissions: P(F1 = Gracias | E1 = Thank)  Transitions: P(A2 = 3 | A1 = 1)

The HMM Model

- Model 2 preferred global monotonicity
- We want local monotonicity:
  - Most jumps are small
- HMM model (Vogel 96)

\[ P(f, a|e) = \prod_j P(a_j|a_{j-1})P(f_j|e_j) \]

- Re-estimate using the forward-backward algorithm
- Handling nulls requires some care
- What are we still missing?
HMM Examples

AER for HMMs

<table>
<thead>
<tr>
<th>Model</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 INT</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM E→F</td>
<td>11.4</td>
</tr>
<tr>
<td>HMM F→E</td>
<td>10.8</td>
</tr>
<tr>
<td>HMM AND</td>
<td>7.1</td>
</tr>
<tr>
<td>HMM INT</td>
<td>4.7</td>
</tr>
<tr>
<td>GIZA M4 AND</td>
<td>6.9</td>
</tr>
</tbody>
</table>