Corpus-Based MT

Modeling correspondences between languages

Sentence-aligned parallel corpus:
- Yo lo haré mañana
  I will do it tomorrow
- Hasta pronto
  See you soon
- Hasta pronto
  See you around

Machine translation system:

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

Alignment Error Rate

- Alignment Error Rate
  - □ = Sure
  - ○ = Possible
  - ■ = Predicted

Model Parameters

- Emissions: \( P(F) = \text{Gracias \mid E} \times \text{Thank} \)
- Transitions: \( P(A_2 = 3) \)

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
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

Second model is basically a filter on the first

Precision jumps, recall drops

End up not guessing hard alignments

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 AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 E→F</td>
<td>82/58 30.6</td>
</tr>
<tr>
<td>Model 1 F→E</td>
<td>85/58 28.7</td>
</tr>
<tr>
<td>Model 1 AND</td>
<td>96/46 34.8</td>
</tr>
<tr>
<td>Model 1 INT</td>
<td>93/69 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)

\[ P(f, e) = \prod_j P(a_j = i f, j, I) P(f_j | e_i) \]

\[ P(\text{dist} = i - j) \frac{1}{Z} e^{-a(i-j)} \]

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

\[ P(f_j | e_i) \]

\[ P(a_j = i f, j, I) \]

- Start with \( P(f_j | e_i) \) uniform, including \( P(f_j | \text{null}) \)
- For each sentence:
  - Calculate posterior over English positions

\[ P(a_j = i f, e_j) = \frac{P(f_j | e_i) P(a_j = i f, j, I) P(f_j | e_i)}{\sum_e P(a_j = i f, l, J) P(f_j | e_j)} \]

- (or just use best single alignment)
- Increment count of word \( f \) with word \( e \), by those amounts
- Also re-estimate distortion probabilities for model 2
- Iterate until convergence
Example: Model 2 Helps

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

Phrase Movement

On Tuesday Nov. 4, earthquakes rocked Japan once again

The HMM Model

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)
  - 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>
IBM Models 3/4/5

Mary did not slap the green witch
Mary not slap slap NULL the green witch
Mary no daba una botefada a la bruja verde

[from Al-Onaizan and Knight, 1998]

Examples: Translation and Fertility

<table>
<thead>
<tr>
<th>the</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>t(f</td>
</tr>
<tr>
<td>le</td>
<td>0.497</td>
</tr>
<tr>
<td>la</td>
<td>0.207</td>
</tr>
<tr>
<td>pas</td>
<td>0.442</td>
</tr>
<tr>
<td>non</td>
<td>0.059</td>
</tr>
<tr>
<td>rien</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Mary did not slap slap the green witch
Mary not slap slip the green witch
Mary no daba una botefada a la bruja verde

Examples: Idioms

Example: Nodding

| f  | t(f | e) | φ  | m(φ | e) |
|----|-----|----|-----|
| signe | 0.164 | 4   | 0.542 |
| la  | 0.123 | 3   | 0.293 |
| tete | 0.097 | 2   | 0.167 |
| oui  | 0.096 | 1   | 0.163 |
| fait | 0.073 | 0   | 0.023 |
| que  | 0.073 | 0   | 0.023 |
| hoche | 0.054 |     |     |
| hoche | 0.046 |     |     |
| faire | 0.030 |     |     |
| me  | 0.024 |     |     |
| approuve | 0.019 |     |     |
| qui  | 0.019 |     |     |
| un  | 0.012 |     |     |
| fai | 0.011 |     |     |

Example: Morphology

Example: Some Results

Some Results

- [Och and Ney 03]

<table>
<thead>
<tr>
<th>Model</th>
<th>Training scheme</th>
<th>0.5K</th>
<th>8K</th>
<th>128K</th>
<th>1.47M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td>C</td>
<td>36.9</td>
<td>43.4</td>
<td>39.6</td>
<td>38.9</td>
</tr>
<tr>
<td>Dice-C</td>
<td></td>
<td>46.3</td>
<td>37.6</td>
<td>35.0</td>
<td>34.0</td>
</tr>
<tr>
<td>Model 1</td>
<td>1+2</td>
<td>40.6</td>
<td>33.6</td>
<td>28.6</td>
<td>25.9</td>
</tr>
<tr>
<td>Model 2</td>
<td>1+2+3</td>
<td>46.7</td>
<td>29.3</td>
<td>22.0</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM1</td>
<td>1+3+3</td>
<td>26.3</td>
<td>23.3</td>
<td>15.0</td>
<td>10.8</td>
</tr>
<tr>
<td>Model 3</td>
<td>1+2+3</td>
<td>43.6</td>
<td>27.5</td>
<td>20.5</td>
<td>18.0</td>
</tr>
<tr>
<td>Model 4</td>
<td>1+2+3+4</td>
<td>27.5</td>
<td>22.5</td>
<td>16.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Model 5</td>
<td>1+2+3+4+5</td>
<td>41.7</td>
<td>25.1</td>
<td>17.3</td>
<td>14.1</td>
</tr>
<tr>
<td>Model 6</td>
<td>1+2+3+4+5+6</td>
<td>21.8</td>
<td>13.3</td>
<td>9.3</td>
<td>9.3</td>
</tr>
<tr>
<td>Model 7</td>
<td>1+2+3+4+5+6+7</td>
<td>25.9</td>
<td>20.3</td>
<td>12.5</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Decoding

- In these word-to-word models
- Finding best alignments is easy
- Finding translations is hard (why?)

it is not clear.

CE NE EST PAS CLAIR.
Bag “Generation” (Decoding)

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)

⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)

⇒ In our organization research has two missions.
⇒ In our missions research organization has two.

Bag Generation as a TSP

- Imagine bag generation with a bigram LM
- Words are nodes
- Edge weights are P(w|w')
- Valid sentences are Hamiltonian paths
- Not the best news for word-based MT!

IBM Decoding as a TSP

Greedy Decoding

Stack Decoding

- Stack decoding:
  - Beam search
  - Usually A* estimates for completion cost
  - One stack per candidate sentence length

- Other methods:
  - Dynamic programming decoders possible if we make assumptions about the set of allowable permutations

Stack Decoding

- Stack decoding:
  - Beam search
  - Usually A* estimates for completion cost
  - One stack per candidate sentence length

- Other methods:
  - Dynamic programming decoders possible if we make assumptions about the set of allowable permutations
Phrase-Based Systems

Phrase table (translation model)

Sentence-aligned corpus

Word alignments

Phrase-Based Decoding

Decoder design is important: [Koehn et al. 03]

The Pharaoh “Model”

Segmentation
Translation
Distortion

Where do we get these counts?

The Pharaoh “Model”

P(f|e) = P((f_i)|e) \prod \phi(f_i|e) d(a_i - b_{i-1})

\frac{1}{K} \frac{\text{count}(f_i,e)}{\text{count}(e)} \alpha^{a_i-b_{i-1}}

Phrase Weights

How the MT community estimates P(f|e)

Parallel training sentences provide phrase pair counts.

Gracias, lo haré de muy buen grado.
Thank you, I shall do so gladly.

All phrase pairs are counted, and counts are normalized.

P(f|e) = \frac{\text{count}(f,e)}{\text{count}(e)}

Counting Phrase Pairs

Input:
Gracias, lo haré de muy buen grado.
Thank you, I shall do so gladly.

Gloss
Thanks that of very good degree
Phrase Scoring

- Learning weights has been tried, several times:
  - [Marcu and Wong, 02]
  - [DeNero et al, 06]
  - ... and others
- Seems not to work well, for a variety of partially understood reasons
- Main issue: big chunks get all the weight, obvious priors don’t help
  - Though, [DeNero et al 08]

Phrase Size

- Phrases do help
  - But they don’t need to be long
  - Why should this be?

Lexical Weighting

\[
\phi(F_i|\tilde{e}_i) = \frac{\text{count}(F_i|\tilde{e}_i)}{\text{count}(\tilde{e}_i)} \times \text{p_w}(F_i|\tilde{e}_i)
\]

The Pharaoh Decoder

- Probabilities at each step include LM and TM

Hypothesis Lattices

- Problem: easy partial analyses are cheaper
  - Solution 1: use beams per foreign subset
  - Solution 2: estimate forward costs (A*-like)
Remember when we discussed WSD?
- Word-based MT systems rarely have a WSD step
- Why not?