**Statistical NLP**

**Spring 2010**

**University of California**

**CA**

**NLP**

**Berkeley**

Lecture 16: Word Alignment

Dan Klein – UC Berkeley

---

**Machine Translation: Examples**

**Atlanta, preso il killer del palazzo di Giustizia**

**ATLANTA** - Lo grande palazzo che per 30 anni ha attirato attori è finito. Brian McCall, l'uomo che aveva ucciso tre persone a giudizio di Giustizia e che non è stato mai condannato, è stato arrestato dalla polizia, dopo aver cercato rifugio nell'edificio che appartenne alla comunità. McCall è stato arrestato per proposta di omicidio, per aver dinamitato l'edificio di una è di uno di loro, è stato arrestato dalla polizia.

**Atlanta, taken the killer of the palace of Justice**

**ATLANTA** - The great building that for 30 years has attracted actors is gone. Brian McCall, the man who has killed three people in a trial of justice and who has never been convicted, has been arrested by the police, after seeking refuge in the building that belonged to the community. McCall has been arrested for attempted murder, for having dynamited the building of one of them, it was eventually extinguished.

---

**Corpus-Based MT**

**Model of translation**

- Sentence-aligned parallel corpus:
  - Yo lo haré mañana
  - Hasta pronto
  - Hasta pronto

- Machine translation system:
  - Yo lo haré pronto
  - I will do it soon

---

**Phrasal / Syntactic MT: Examples**

**Levels of Transfer**

- **English** to **Spanish**
  - I will do it tomorrow
  - Hasta pronto

**Phrases**

- **English** to **Spanish**
  - Yo lo haré mañana (I will do it tomorrow)
  - Hasta pronto (See you soon)

---

**Phrasal / Syntactic MT: Examples**

- **U.S. President Barack Obama to announce**
  - Tuesday, new measures to help Americans. General Motors and Chrysler have already received the first installment of federal bailout funds, and are expected to receive more. Obama is also expected to announce an extension of the federal tax credit for new car purchases.

- **General Motors and Chrysler are likely to receive**
  - Tuesday, new measures to help Americans. General Motors and Chrysler have already received the first installment of federal bailout funds, and are expected to receive more. Obama is also expected to announce an extension of the federal tax credit for new car purchases.

---

**Statistical NLP**

**Spring 2010**

**University of California**

**CA**

**NLP**

**Berkeley**

Lecture 16: Word Alignment

Dan Klein – UC Berkeley
MT: Evaluation

- Human evaluations: subject measures, fluency/adequacy
- Automatic measures: n-gram match to references
  - NIST measure: n-gram recall (worked poorly)
  - BLEU: n-gram precision (no one really likes it, but everyone uses it)
- BLEU:
  - P1 = unigram precision
  - P2, P3, P4 = bi-, tri-, 4-gram precision
  - Weighted geometric mean of P1-4
  - Somewhat hard to game...

Automatic Metrics Work (?)

Human evaluations: subject measures, fluency/adequacy

Automatic measures: n-gram match to references
- NIST measure: n-gram recall (worked poorly)
- BLEU: n-gram precision (no one really likes it, but everyone uses it)

BLEU:
- P1 = unigram precision
- P2, P3, P4 = bi-, tri-, 4-gram precision
- Weighted geometric mean of P1-4

Brevity penalty (why?)

Somewhat hard to game...

Today

- The components of a simple MT system
  - You already know about the LM
  - Word-alignment based TMs
    - IBM models 1 and 2, HMM model
    - A simple decoder
- Next few classes
  - More complex word-level and phrase-level TMs
  - Tree-to-tree and tree-to-string TMs
  - More sophisticated decoders

Word Alignment

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. Align words with a probabilistic model
2. Infer presence of larger structures from this alignment
3. Translate with the larger structures
1-to-Many Alignments

**IBM Model 1 (Brown 93)**
- Alignments: a hidden vector called an alignment specifies which English source is responsible for each French target word.

\[ a = a_1, \ldots, a_j \]

\[ P(f|u|c) = 1 \prod_{i=1}^{j} P(a_j = i) P(f_j|a_i) = \prod_{i=1}^{j} \frac{1}{i+1} P(f_j|a_i) \]

\[ P(f|c) = \sum_{a} P(f|a|c) \]

**IBM Models 1/2**

<table>
<thead>
<tr>
<th>E:</th>
<th>Thank you , I shall do so gladly .</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:</td>
<td>Gracias , lo haré de muy buen grado .</td>
</tr>
</tbody>
</table>

**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**

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

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

Joint Training?

- Overall:
  - Similar high precision to post-intersection
  - But recall is much higher
  - More confident about posting 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>

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>

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)
  \[ P(f, a|v) = \prod_j P(a_j = t(j, t, r) P(f_j|v_j)) \]
  \[ P(dist = i - \frac{1}{2}) \]
  \[ \frac{1}{2^d - a(v - \frac{1}{2})} \]
- 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)
- 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
    - Increment count of word $f_j$ with word $e_i$ by these amounts
    - Also re-estimate distortion probabilities for model 2
- Iterate until convergence

Phrase Movement

- 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

- Model 2 preferred global monotonicity
- We want local monotonicity:
  - Most jumps are small
- HMM model (Vogel 96)
  \[ P(f, a | e) = \prod_j P(a_{j+1} \mid a_j) P(f_j \mid e_i) \]
  \[ P(a_{j+1} = a_j - 1) \]
- Re-estimate using the forward-backward algorithm
- Handling nulls requires some care
- What are we still missing?
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>

Examples: Translation and Fertility

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>not</th>
</tr>
</thead>
<tbody>
<tr>
<td>le</td>
<td>0.497</td>
<td>1</td>
</tr>
<tr>
<td>la</td>
<td>0.207</td>
<td>0</td>
</tr>
<tr>
<td>les</td>
<td>0.155</td>
<td></td>
</tr>
<tr>
<td>l’</td>
<td>0.086</td>
<td></td>
</tr>
<tr>
<td>ce</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>cette</td>
<td>0.011</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>n(φ</th>
<th>e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>me</td>
<td>0.497</td>
<td>2</td>
</tr>
<tr>
<td>pas</td>
<td>0.442</td>
<td>0</td>
</tr>
<tr>
<td>non</td>
<td>0.029</td>
<td>1</td>
</tr>
<tr>
<td>rien</td>
<td>0.011</td>
<td></td>
</tr>
</tbody>
</table>

Example: Idioms

<table>
<thead>
<tr>
<th></th>
<th>n(φ</th>
<th>e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>signe</td>
<td>0.164</td>
<td>4</td>
</tr>
<tr>
<td>la</td>
<td>0.123</td>
<td>3</td>
</tr>
<tr>
<td>tête</td>
<td>0.007</td>
<td>2</td>
</tr>
<tr>
<td>oui</td>
<td>0.086</td>
<td>1</td>
</tr>
<tr>
<td>fait</td>
<td>0.073</td>
<td>0</td>
</tr>
<tr>
<td>que</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>hoche</td>
<td>0.054</td>
<td></td>
</tr>
<tr>
<td>faire</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>me</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>approuver</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td>qui</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>un</td>
<td>0.011</td>
<td></td>
</tr>
</tbody>
</table>

Example: Morphology

<table>
<thead>
<tr>
<th></th>
<th>n(φ</th>
<th>e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>devrait</td>
<td>0.330</td>
<td>1</td>
</tr>
<tr>
<td>devraient</td>
<td>0.123</td>
<td>0</td>
</tr>
<tr>
<td>devoirs</td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td>faut</td>
<td>0.086</td>
<td></td>
</tr>
<tr>
<td>doit</td>
<td>0.014</td>
<td></td>
</tr>
<tr>
<td>doiveint</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td>devons</td>
<td>0.024</td>
<td></td>
</tr>
<tr>
<td>devoirs</td>
<td>0.023</td>
<td></td>
</tr>
</tbody>
</table>

Example: Morphology

<table>
<thead>
<tr>
<th></th>
<th>n(φ</th>
<th>e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>model 1 INT</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>HMM E→F</td>
<td>11.4</td>
<td></td>
</tr>
<tr>
<td>HMM F→E</td>
<td>10.8</td>
<td></td>
</tr>
<tr>
<td>HMM AND</td>
<td>7.1</td>
<td></td>
</tr>
<tr>
<td>HMM INT</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>GIZA M4 AND</td>
<td>6.9</td>
<td></td>
</tr>
</tbody>
</table>

Some Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Training scheme</th>
<th>0.8K</th>
<th>8K</th>
<th>128K</th>
<th>1.47M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td></td>
<td>50.9</td>
<td>34.3</td>
<td>39.6</td>
<td>38.9</td>
</tr>
<tr>
<td>Dice+C</td>
<td></td>
<td>46.3</td>
<td>39.6</td>
<td>35.0</td>
<td>34.0</td>
</tr>
<tr>
<td>Model 1</td>
<td>1'5'</td>
<td>40.6</td>
<td>33.6</td>
<td>26.8</td>
<td>25.9</td>
</tr>
<tr>
<td>Model 2</td>
<td>1'2'</td>
<td>46.7</td>
<td>29.3</td>
<td>22.0</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM</td>
<td>1'</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>17.5</td>
<td>14.1</td>
</tr>
<tr>
<td>Model 6</td>
<td>1'2'3'4'5'6'</td>
<td>26.1</td>
<td>20.2</td>
<td>13.1</td>
<td>9.4</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)
  -> Now let me mention some of the disadvantages.
  -> 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

Decoding, Anyway

- Simplest possible decoder:
  - Enumerate sentences, score each with TM and LM

- Greedy decoding:
  - Assign each French word it’s most likely English translation
  - Operators:
    - Change a translation
    - Insert a word into the English (zero-fertile French)
    - Remove a word from the English (null-generated French)
    - Swap two adjacent English words
  - Do hill-climbing (or annealing)
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

### Dynamic Programming Table

<table>
<thead>
<tr>
<th>Length</th>
<th>Tokens/Word</th>
<th>Error/Word</th>
<th>Insertion/Word</th>
<th>Deletion/Word</th>
<th>NE</th>
<th>PSE</th>
<th>CSE</th>
<th>PSE</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>17</td>
<td>57.76</td>
<td>16</td>
<td>98</td>
<td>44</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>0</td>
<td>stacks</td>
<td>0.07</td>
<td>18</td>
<td>98</td>
<td>38</td>
<td>45</td>
<td>5</td>
<td>2</td>
<td>10.0</td>
</tr>
<tr>
<td>0</td>
<td>17</td>
<td>407.04</td>
<td>25</td>
<td>90</td>
<td>24</td>
<td>57</td>
<td>1</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>0</td>
<td>tokens</td>
<td>5.94</td>
<td>36</td>
<td>79</td>
<td>20</td>
<td>58</td>
<td>4</td>
<td>9</td>
<td>10.0</td>
</tr>
<tr>
<td>0</td>
<td>grammars</td>
<td>2.46</td>
<td>43</td>
<td>79</td>
<td>20</td>
<td>58</td>
<td>4</td>
<td>9</td>
<td>10.0</td>
</tr>
</tbody>
</table>

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