Machine Translation: Examples

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

**ATLANTA** - La grande paura che per 20 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia a che ha poi abbandonato un agente di droga, è stato consegnato alla polizia, dopo aver cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, oltre di una popolare area metropolitana, era rimasto paralizzato.

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

**ATLANTA** - The great fear that for 20 hours gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to the palaces of justice and that an addict has then killed, is delivered to the police, after a time to have gathered shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For the day, the center of the city, center of the Coca Cola and of the 1996 Olympics, is a popular area of metropolitan area, remained paralyzed.
Machine Translation

Madame la présidente, votre présidence de cette institution a été marquante.  
Mrs Fontaine, your presidency of this institution has been outstanding.  
Madam President, president of this house has been discoveries.  
Madam President, your presidency of this institution has been impressive.

Je vais maintenant m'exprimer brièvement en irlandais. 
I shall now speak briefly in Irish . 
I will now speak briefly in Ireland . 
I will now speak briefly in Irish .

Nous trouvons en vous un président tel que nous le souhaitions.  
We think that you are the type of president that we want.  
We are in you a president as the wanted.  
We are in you a president as we the wanted.

Levels of Transfer

![Levels of Transfer Diagram]


(Vauquois triangle)
General Approaches

- Rule-based approaches
  - Expert system-like rewrite systems
  - Interlingua methods (analyze and generate)
  - Lexicons come from humans
  - Can be very fast, and can accumulate a lot of knowledge over time (e.g. Systran)

- Statistical approaches
  - Word-to-word translation
  - Phrase-based translation
  - Syntax-based translation (tree-to-tree, tree-to-string)
  - Trained on parallel corpora
  - Usually noisy-channel (at least in spirit)

MT System Components

Source language model $P(e)$ gives $e$;
Decoder $P(f|e)$ gives $f$.

Best $e$ is the argmax of $P(e|f) = \arg\max_e P(f|e)P(e)$.
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-to-Many Alignments

<table>
<thead>
<tr>
<th>And₁</th>
<th>the₂</th>
<th>programme₃</th>
<th>has₄</th>
<th>been₅</th>
<th>implemented₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le₁</td>
<td>programme₂</td>
<td>a₃</td>
<td>été₄</td>
<td>miss₅</td>
<td>en₆</td>
</tr>
</tbody>
</table>
Many-to-1 Alignments

Many-to-Many Alignments
A Word-Level TM?

- What might a model of \( P(f|e) \) look like?

\[
e = e_1 \ldots e_I
\]
\[
f = f_1 \ldots f_J
\]

\[
P(f|e) = \prod_j P(f_j|e_1 \ldots e_I)
\]

What can go wrong here?

How to estimate this?

---

IBM Model 1 (Brown 93)

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

\[
a = a_1 \ldots a_J
\]

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

\[
= \prod_j \frac{1}{I+1} P(f_j|e_i)
\]

\[
P(f|e) = \sum_a P(f, a|e)
\]
IBM Model 1

- Obvious first stab: greedy matchings
- Better approach: re-estimated generative models

\[
P(f|e) = \sum_a P(f, a|e)
\]

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

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

- Basic idea: pick a source for each word, update co-occurrence statistics, repeat

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
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)

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

\[
P(dist = i - j \frac{I}{J})
\]

\[
\frac{1}{Z}e^{-\alpha(i-j)}
\]

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

---

Example

<table>
<thead>
<tr>
<th>the branches they intend to close</th>
</tr>
</thead>
<tbody>
<tr>
<td>les embranchements que ils songeaient à fermer</td>
</tr>
</tbody>
</table>
EM for Models 1/2

- Model 1 Parameters:
  - Translation probabilities (1+2) \( P(f_j|e_i) \)
  - Distortion parameters (2 only) \( P(a_j = i|j, I, J) \)
- 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 \( f_j \) 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.
Phrase Movement

The HMM Model

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

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

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

Distortion $\theta_d$

- $p(\text{the} \rightarrow \text{le}) = 0.6$
- $p(\text{railroad} \rightarrow \text{ferroviaire}) = 0.2$
- $p(\text{NULL} \rightarrow \text{le}) = 0.1$

Translation $\theta_t$

- $p(\text{le} \rightarrow \text{the}) = 0.53$
- $p(\text{la} \rightarrow \text{the}) = 0.24$
- $p(\text{ferroviaire} \rightarrow \text{railroad}) = 0.19$
- $p(\text{le} \rightarrow \text{NULL}) = 0.12$

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 não slap slash slash NULL the green witch
Mary no daba una botefada a la bruja verde

(from Al-Onaizan and Knight, 1998)
### Examples: Translation and Fertility

#### the

| f  | t(f | e) | φ  | n(φ | e) |
|----|-------|----|--------|
| le | 0.497 | 1  | 0.746  |
| la | 0.207 | 0  | 0.254  |
| les| 0.155 | 1  | 0.260  |
| l' | 0.086 | 0  | 0.254  |
| ce | 0.018 | 1  | 0.260  |
| cette | 0.011 | 1 | 0.260 |

#### not

| f  | t(f | e) | φ  | n(φ | e) |
|----|-------|----|--------|
| ne | 0.497 | 2  | 0.735  |
| pas| 0.442 | 0  | 0.154  |
| non| 0.029 | 1  | 0.107  |
| rien| 0.011 | 1 | 0.107 |

### Example: Idioms

#### nodding

| f         | t(f | e) | φ  | n(φ | e) |
|-----------|-------|----|--------|
| signe     | 0.164 | 4  | 0.342  |
| la        | 0.123 | 3  | 0.293  |
| tête      | 0.097 | 2  | 0.167  |
| oui       | 0.086 | 1  | 0.163  |
| fait      | 0.073 | 0  | 0.023  |
| que       | 0.073 | 0  | 0.023  |
| hoche     | 0.054 | 0  | 0.023  |
| hocher    | 0.048 | 0  | 0.023  |
| faire     | 0.030 | 0  | 0.023  |
| me        | 0.024 | 0  | 0.023  |
| approuve  | 0.019 | 0  | 0.023  |
| qui       | 0.019 | 0  | 0.023  |
| un        | 0.012 | 0  | 0.023  |
| faites    | 0.011 | 0  | 0.023  |
Example: Morphology

<table>
<thead>
<tr>
<th>should</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
</tr>
<tr>
<td>devrait</td>
</tr>
<tr>
<td>devraiens</td>
</tr>
<tr>
<td>devriens</td>
</tr>
<tr>
<td>faudrait</td>
</tr>
<tr>
<td>faut</td>
</tr>
<tr>
<td>doit</td>
</tr>
<tr>
<td>aurait</td>
</tr>
<tr>
<td>doivent</td>
</tr>
<tr>
<td>devons</td>
</tr>
<tr>
<td>devrais</td>
</tr>
</tbody>
</table>

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></td>
<td>50.9</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>1^5</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^{3}\times 2^5</td>
<td>46.7</td>
<td>29.3</td>
<td>22.0</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM</td>
<td>1^{5}H^5</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}\times 3^3</td>
<td>43.6</td>
<td>27.5</td>
<td>20.5</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>1^{5}H^5\times 3^3</td>
<td>27.5</td>
<td>22.5</td>
<td>16.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Model 4</td>
<td>1^{5}\times 3^3\times 4^3</td>
<td>41.7</td>
<td>25.1</td>
<td>17.3</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>1^{5}H^5\times 3^3\times 4^3</td>
<td>26.1</td>
<td>20.2</td>
<td>13.1</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>1^{5}H^5\times 4^5</td>
<td>26.3</td>
<td>21.8</td>
<td>13.3</td>
<td>9.3</td>
</tr>
<tr>
<td>Model 5</td>
<td>1^{5}H^5\times 3^3\times 5^3</td>
<td>26.5</td>
<td>21.5</td>
<td>13.7</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>1^{5}H^5\times 3^4\times 5^3</td>
<td>26.5</td>
<td>20.4</td>
<td>13.4</td>
<td>9.4</td>
</tr>
<tr>
<td>Model 6</td>
<td>1^{5}H^5\times 4^6^3</td>
<td>26.0</td>
<td>21.6</td>
<td>12.8</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>1^{5}H^5\times 3^4\times 6^3</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.
  \[\Rightarrow\] Please give me your response as soon as possible.

*Reconstruction preserving meaning (8 of 38)*

- Now let me mention some of the disadvantages.
  \[\Rightarrow\] Let me mention some of the disadvantages now.

*Garbage reconstruction (6 of 38)*

- In our organization research has two missions.
  \[\Rightarrow\] 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)

---

**Greedy Decoding**

- **NULL well heard, it talks about a great victory.**
  - bien entendu, il parle de une belle victoire
  - translateTwoWords(2, understood, 0, about)

- **NULL well understood, he talks about a great victory.**
  - bien entendu, il parle de une belle victoire
  - translateOneWord(1, he)

- **NULL well understood, he talks about a great victory.**
  - bien entendu, il parle de une belle victoire
  - translateTwoWords(1, quite, 2, naturally)

- **NULL quite naturally, he talks about a great victory.**
  - bien entendu, il parle de une belle victoire
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

<table>
<thead>
<tr>
<th>sent length</th>
<th>decoder type</th>
<th>time (sec/sent)</th>
<th>search errors</th>
<th>translation errors (semantic and/or syntactic)</th>
<th>NE</th>
<th>PME</th>
<th>DSE</th>
<th>FSE</th>
<th>HSE</th>
<th>CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>IP</td>
<td>47.50</td>
<td>0</td>
<td>57</td>
<td>44</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>stack greedy</td>
<td>0.79</td>
<td>5</td>
<td>58</td>
<td>43</td>
<td>53</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>greedy</td>
<td>0.07</td>
<td>18</td>
<td>60</td>
<td>38</td>
<td>45</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>IP</td>
<td>499.00</td>
<td>0</td>
<td>76</td>
<td>27</td>
<td>74</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>stack greedy</td>
<td>5.67</td>
<td>43</td>
<td>75</td>
<td>20</td>
<td>38</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>33</td>
</tr>
</tbody>
</table>