Machine Translation: Examples

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

**ATLANTA** - La grande polizia che per 20 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che
ha rinunciato a una festa di corteo, ha colpito per 20 ore 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 giunta e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasta paralizzata.

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

**ATLANTA** - The great fear that for 20 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of justice and that
had abandoned the party of a party, has been delivered to the police, after to have tried to find shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Senate and of Giochi 1996, heart of one popolosa metropolitan area, was 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.

History

- 1950’s: Intensive research activity in MT
- 1960’s: Direct word-for-word replacement
- 1966 (ALPAC): NRC Report on MT
  - Conclusion: MT no longer worthy of serious scientific investigation.
- 1966-1975: `Recovery period’
- 1975-1985: Resurgence (Europe, Japan)
- 1985-present: Gradual Resurgence (US)

Levels of Transfer

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)
The Coding View

- “One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’”

- Warren Weaver (1955:18, quoting a letter he wrote in 1947)

MT System Components

Language Model

\[ P(e) \]

Translation Model

\[ P(f|e) \]

oberved

\[ f \]

decoder

\[ \text{argmax} \ P(e|f) = \text{argmax} \ P(f|e)P(e) \]

\[ e \]

Finds an English translation which is both fluent and semantically faithful to the French source.
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
  - nous . . . acceptons
  - acceptons . . .
  - votre . . .
  - opinion . .
  - we . .
  - accept . .
  - your . .
  - view .

1-to-Many Alignments

- Diagram showing 1-to-Many alignments with words and positions.
Many-to-1 Alignments

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

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

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

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

How to estimate this?

What can go wrong here?

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

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

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

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

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

---

IBM Model 1 [Brown et al, 93]

- Alignments: a hidden vector called an alignment specifies which English source is responsible for each French target word. \( a = a_1 \ldots a_J \)

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

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

Evaluating TMs

- How do we measure TM quality?
  - Method 1: use in an end-to-end translation system
    - Hard to measure translation quality
    - Option: human judges
    - Option: reference translations (NIST, BLEU scores)
  - Method 2: measure quality of the alignments produced
    - Easy to measure
    - Hard to know what the gold alignments should be
    - May not correlate with translation quality (like perplexity in LMs)
Alignment Error Rate

- Alignment Error Rate

- = Sure
  O = Possible
  ■ = Predicted

\[
AER(A, S, P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)
= \left(1 - \frac{3 + 3}{3 + 4}\right) = \frac{1}{7}
\]

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)

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

\[ P(i - j^{\frac{1}{2}}) \]

\[ \frac{1}{Z} e^{-\alpha(i-j^{\frac{1}{2}})} \]

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

- les embranchements
- que
- ils songeaient
- à
- fermer

the branches they intend to close

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 $P(f_j|null)$
- 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 $e_i$ 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

- 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

\[ p(e) \]
\[ p(a_j | a_{j-1}; \theta_d) \]
\[ p(f_j | e_i; \theta_t) \]

**Distortion** $\theta_d$

- $p(\rightarrow \text{le}) = 0.6$
- $p(\rightarrow \text{la}) = 0.2$
- $p(\rightarrow \text{...}) = 0.1$

**Translation** $\theta_t$

- $p(\text{the} \rightarrow \text{le}) = 0.53$
- $p(\text{the} \rightarrow \text{la}) = 0.24$
- $p(\text{railroad} \rightarrow \text{ferroviaire}) = 0.19$
- $p(\text{NULL} \rightarrow \text{le}) = 0.12$
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 not slap slap slap the green witch

Mary not slap slap NULL the green witch

Mary no daba una botefada a la verde bruja

Mary no daba una botefada a la bruja verde

[Al-Onaizan and Knight, 1998]

Examples: Translation and Fertility

|      | f  | ∀(f) | n(∀|e) |
|------|----|------|--------|
| the  | le | 0.497| 0.746  |
|      | la | 0.207| 0.254  |
|      | les| 0.155|        |
|      | l’ | 0.086|        |
|      | ce | 0.018|        |
|      | cett | 0.011|        |

|      | f  | ∀(f) | n(∀|e) |
|------|----|------|--------|
| not  | ne | 0.497| 0.735  |
|      | pas| 0.442| 0.154  |
|      | non| 0.029| 0.107  |
|      | rien| 0.011|        |

farmers

|      | f  | ∀(f) | n(∀|e) |
|------|----|------|--------|
|      | agriculteurs | 0.442| 2 0.731|
|      | les | 0.418| 1 0.228|
|      | cultivateurs | 0.046| 0 0.039|
|      | producteurs | 0.021|        |
### Example: Idioms

**nodding**

<table>
<thead>
<tr>
<th>f</th>
<th>$t(f \mid e)$</th>
<th>$\phi$</th>
<th>$n(\phi \mid e)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>signe</td>
<td>0.164</td>
<td>4</td>
<td>0.342</td>
</tr>
<tr>
<td>la</td>
<td>0.123</td>
<td>3</td>
<td>0.293</td>
</tr>
<tr>
<td>tête</td>
<td>0.097</td>
<td>2</td>
<td>0.167</td>
</tr>
<tr>
<td>oui</td>
<td>0.086</td>
<td>1</td>
<td>0.163</td>
</tr>
<tr>
<td>fait</td>
<td>0.073</td>
<td>0</td>
<td>0.023</td>
</tr>
<tr>
<td>que</td>
<td>0.073</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hoche</td>
<td>0.054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hocher</td>
<td>0.048</td>
<td></td>
<td></td>
</tr>
<tr>
<td>faire</td>
<td>0.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>me</td>
<td>0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>approuve</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>qui</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>un</td>
<td>0.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>faites</td>
<td>0.011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Example: Morphology

**should**

<table>
<thead>
<tr>
<th>f</th>
<th>$t(f \mid e)$</th>
<th>$\phi$</th>
<th>$n(\phi \mid e)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>devrait</td>
<td>0.330</td>
<td>1</td>
<td>0.649</td>
</tr>
<tr>
<td>devraient</td>
<td>0.123</td>
<td>0</td>
<td>0.336</td>
</tr>
<tr>
<td>devrions</td>
<td>0.109</td>
<td>2</td>
<td>0.014</td>
</tr>
<tr>
<td>faudrait</td>
<td>0.073</td>
<td></td>
<td></td>
</tr>
<tr>
<td>faut</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doit</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aurait</td>
<td>0.041</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doivent</td>
<td>0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>devons</td>
<td>0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>devrais</td>
<td>0.013</td>
<td></td>
<td></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>$1^52^5$</td>
<td>46.7</td>
<td>29.3</td>
<td>22.0</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM</td>
<td>$1^2H^5$</td>
<td>26.3</td>
<td>23.3</td>
<td>15.0</td>
<td>10.8</td>
</tr>
<tr>
<td>Model 3</td>
<td>$1^22^33^3$</td>
<td>43.6</td>
<td>27.5</td>
<td>20.5</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>$1^2H^53^3$</td>
<td>27.5</td>
<td>22.5</td>
<td>16.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Model 4</td>
<td>$1^22^33^34^3$</td>
<td>41.7</td>
<td>25.1</td>
<td>17.3</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>$1^2H^33^34^3$</td>
<td>26.1</td>
<td>20.2</td>
<td>13.1</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>$1^2H^34^3$</td>
<td>26.3</td>
<td>21.8</td>
<td>13.3</td>
<td>9.3</td>
</tr>
<tr>
<td>Model 5</td>
<td>$1^2H^34^35^3$</td>
<td>26.5</td>
<td>21.5</td>
<td>13.7</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>$1^2H^33^34^35^3$</td>
<td>26.5</td>
<td>20.4</td>
<td>13.4</td>
<td>9.4</td>
</tr>
<tr>
<td>Model 6</td>
<td>$1^2H^34^36^3$</td>
<td>26.0</td>
<td>21.6</td>
<td>12.8</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>$1^2H^33^34^36^3$</td>
<td>25.9</td>
<td>20.3</td>
<td>12.5</td>
<td>8.7</td>
</tr>
</tbody>
</table>