Machine Translation: Examples

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

**ATLANTA** - La grande oscurità che per 20 ore ha attanagliato l'Atlantico è finita. Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha minacciato di attentato a se stesso, 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 cattedrale e dei Giocchi 1996, ha subito 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 in the Palace of Justice and that had threatened self-murder, is delivered to the police, after to have tried 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 Cathedral and of the Games 1996, heart of one popolosa metropolitan area, was remained paralyzed.
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:

- Yo lo haré pronto
  - I will do it soon
- I will do it around
- See you tomorrow

Levels of Transfer

- Interlingua
- Semantics
- Syntax
- Phrases
- Words

P(English (E) | Yo lo haré mañana)

<table>
<thead>
<tr>
<th>Action</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>will do it</td>
<td>0.8</td>
</tr>
<tr>
<td>will do so</td>
<td>0.2</td>
</tr>
</tbody>
</table>

P(English (E) | mañana)

<table>
<thead>
<tr>
<th>Action</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>tomorrow</td>
<td>0.7</td>
</tr>
<tr>
<td>morning</td>
<td>0.3</td>
</tr>
</tbody>
</table>
### World-Level MT: Examples

<table>
<thead>
<tr>
<th>Foreign Original</th>
<th>Reference Translation</th>
<th>IBM4+N-grams+Stack</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>la politique de la haine.</em></td>
<td>politics of hate.</td>
<td>the policy of the hatred.</td>
</tr>
<tr>
<td><em>nous avons signé le protocole.</em></td>
<td>we did sign the memorandum of agreement.</td>
<td>we have signed the protocol.</td>
</tr>
<tr>
<td><em>où était le plan solide ?</em></td>
<td>but where was the solid plan ?</td>
<td>where was the economic base ?</td>
</tr>
</tbody>
</table>

### Phrasal / Syntactic MT: Examples

- Le président américain Barack Obama doit annoncer lundi de nouvelles mesures en faveur des constructeurs automobile. General motors et Chrysler avaient déjà bénéficié fin 2008 d'un prêt d'urgence cumulé de 17,4 milliards de dollars, et ont soumis en février au Trésor un plan de restructuration basé sur un total de 22 milliards de dollars d'aides publiques supplémentaires.

- Interrogé sur la chaîne CBS dimanche, le président a toutefois clairement précisé que le gouvernement ne préterait pas d'argent sans de fortes contreparties. "Il faudra faire des sacrifices à tous les niveaux", a-t-il prévenu. "Tout le monde devra se réunir autour de la table et se mettre d'accord sur une restructuration en profondeur".

- General Motors et Chrysler sont engagés dans des négociations avec le principal syndicat de l'automobile. Les constructeurs souhaitaient diminuer leurs cotisations aux caisses de retraites, et accorder en échange des actions aux syndicats. Ils souhaiteraient également négocier des baisses des salaires.

- U.S. President Barack Obama to announce Monday new measures to help automakers. General Motors and Chrysler had already received late in 2008 a cumulative emergency loan of 17.4 billion dollars, and submitted to the Treasury in February a restructuring plan based on a total of 22 billion dollars in additional aid.

- Interviewed on CBS Sunday, the president has clearly stated that the government does not lend money without strong counterparts. "We must make sacrifices at all levels," he warned. "Everyone should gather around the table and agree on a profound restructuring."

- General Motors and Chrysler are engaged in negotiations with the major union of the car. Manufacturers wishing to reduce their contributions to pension funds, and give in exchange for the shares to trade unions. They would also negotiate lower wages.
MT: Evaluation

- Human evaluations: subject measures, fluency/adequacy

- Automatic measures: n-gram match to references
  - NIST measure: n-gram precision (worked poorly)
  - BLEU: n-gram recall (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...

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:
The American [?] international airport, not the office as everyone calls self the sand Arab rich business [?] and so on electronic mail, which sends out, The three will be able after public place and on the airport to start the biochemical attack [?] highly alerts after the maintenance.

Automatic Metrics Work (?)
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

X

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?

Z

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

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

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

many-to-1 Alignments

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.

\[ a = a_1 \ldots a_J \]

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

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

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

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

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

- 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

Problems: alignments jump around, align everything to rare words
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)

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

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

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

- 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 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 \( f \) with word \( e_i \) by these amounts
- Also re-estimate distortion probabilities for model 2

- Iterate until convergence
Example

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

On Tuesday Nov. 4, earthquakes rocked Japan once again

Phrase Movement

Des tremblements de terre ont à nouveau touché le Japon jeudi 4 novembre.
IBM Models 1/2

The HMM Model

Model Parameters

Emissions: \( P(F_1 = \text{Gracias} \mid E_{A_1} = \text{Thank}) \)
Transitions: \( P(A_2 = 3) \)
The HMM Model

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

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

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

| f   | f(f|e) |
|-----|-------|
| nationale | 0.469 |
| national  | 0.418 |
| nationaux | 0.054 |
| nationales | 0.029 |

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 slap NULL the green witch

Mary no daba una botefada a la verde bruja

Mary no daba una botefada a la bruja verde

[n(3|slap), P(NULL), t(la|the), d(jji)]

[from Al-Onaizan and Knight, 1998]
Examples: Translation and Fertility

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

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

Example: Idioms

|      | $t(f | e)$ | $\phi$ | $n(\phi | e)$ |
|------|-----------|--------|---------------|
| agriculteurs | 0.442 | 2      | 0.731         |
| les   | 0.418     | 1      | 0.228         |
| cultivateurs | 0.046 | 0      | 0.039         |
| producteurs | 0.021 |        |               |

Example: Idioms

nodding

|      | $t(f | e)$ | $\phi$ | $n(\phi | e)$ |
|------|-----------|--------|---------------|
| 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     |        |               |
| hoche | 0.054     |        |               |
| hocher| 0.048     |        |               |
| faire | 0.030     |        |               |
| me    | 0.024     |        |               |
| approuve | 0.019 |        |               |
| qui   | 0.019     |        |               |
| un    | 0.012     |        |               |
| faites| 0.011     |        |               |
Example: Morphology

\[ \text{should} \]

<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>devraisent</td>
<td>0.123</td>
<td>0</td>
<td>0.336</td>
</tr>
<tr>
<td>devriez</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>1^5^2^3</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>1^5^2^3^3^3</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^5^3^3</td>
<td>27.5</td>
<td>22.5</td>
<td>16.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Model 4</td>
<td>1^5^2^3^4^3^3</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^5^3^4^3</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^5^4^3</td>
<td>26.3</td>
<td>21.8</td>
<td>13.3</td>
<td>9.3</td>
</tr>
<tr>
<td>Model 5</td>
<td>1^5^H^5^3^4^5^3</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^5^3^4^5^5</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^5^4^6^3</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^5^3^4^6^3</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.
⇒ 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)

---

**Greedy Decoding**

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

- `NULL well understood, it talks about a great victory`, `translateOneWord(3, he)`
  - `bien entendu, il parle de une belle victoire`

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

- `NULL quite naturally, he talks about a great victory`, `translateTwoWords(1, quite, 2, naturally)`
  - `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>sentence 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</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>2</td>
<td>1</td>
<td>10</td>
<td></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</td>
<td>5.67</td>
<td>20</td>
<td>75</td>
<td>24</td>
<td>57</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>greedy</td>
<td>2.66</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>