Machine Translation: Examples

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

Atlanta, firebombed 25th street, has apparently been burned. An arson attack by a group of individuals, who claimed they were in protest against the police and the courts, has caused extensive damage to the building.

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

ATLANTA -- The man who has been identified as the person responsible for the arson attack on the city police station yesterday was arrested by the police today.

---

**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)


---

**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)
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

MT System Components

Language Model

source

\[ P(e) \]

\[ \rightarrow \]

e

channel

\[ P(f|e) \]

\[ \rightarrow \]

f

best

decoder

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

\[ e \]

\[ e \]

Finds an English translation which is both fluent and semantically faithful to the French source

A Word-Level TM?

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

\[ f = f_1 \ldots f_j \]

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

\[ a = a_1 \ldots a_j \]

\[ \begin{align*}
P(f|a) &= \prod_j P(f_j|a_1 \ldots a_j) \\
&= \prod_j \frac{1}{I+1} P(f_j|a_i)
\end{align*} \]

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

How to estimate this?

What can go wrong here?

IBM Model 1 (Brown 93)

1-to-Many Alignments

Many-to-1 Alignments
Many-to-Many Alignments

Monotonic Translation

Le Japon secoué par deux nouveaux séismes

Japan shaken by two new quakes

Local Order Change

Japan is at the junction of four tectonic plates

Le Japon est au confluent de quatre plaques tectoniques

IBM Model 2

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

EM for Models 1/2

Model 1 Parameters:
- Translation probabilities (1-2)
- Distortion parameters (2 only)

Start with $P(f_i | e_i)$ uniform, including $P(f_{j \in \text{null}})$

For each sentence:
- For each French position $j$
  - Calculate posteriors over English positions
    - $P(a_j = i | f, e) = \frac{P(a_j = i | f, e) P(f_j | e_i)}{\sum_k P(a_j = k | f, e) P(f_j | e_i)}$

- (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

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
  - ○ = Possible
  - ■ = Predicted

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

\[
= \left(1 - \frac{3 + 3}{3 + 4}\right) = \frac{2}{7}
\]

### 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):
  
  \[
P(f, o|e) = \prod_i P(o_i|o_{i-1})P(f_i|e_i)\prod_i P(a_i - a_{i-1})
  \]

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

### 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>30.9</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>(y^1)</td>
<td>40.6</td>
<td>35.6</td>
<td>29.6</td>
<td>25.9</td>
</tr>
<tr>
<td>Model 2</td>
<td>(y^2)</td>
<td>46.7</td>
<td>39.3</td>
<td>22.0</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM</td>
<td>(y^2)</td>
<td>26.3</td>
<td>23.3</td>
<td>15.0</td>
<td>10.8</td>
</tr>
<tr>
<td>Model 3</td>
<td>(y^2)</td>
<td>43.6</td>
<td>27.5</td>
<td>20.5</td>
<td>18.0</td>
</tr>
<tr>
<td>Model 4</td>
<td>(y^2)</td>
<td>27.5</td>
<td>22.5</td>
<td>16.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Model 5</td>
<td>(y^2)</td>
<td>41.7</td>
<td>25.1</td>
<td>17.3</td>
<td>14.1</td>
</tr>
<tr>
<td>Model 6</td>
<td>(y^2)</td>
<td>26.1</td>
<td>20.2</td>
<td>13.1</td>
<td>9.4</td>
</tr>
<tr>
<td>Model 7</td>
<td>(y^2)</td>
<td>26.3</td>
<td>21.8</td>
<td>13.3</td>
<td>9.3</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.
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 is 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!

Decoding, Anyway

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

- Greedy decoding:
  - Assign each French word its most likely English translation
  - Operations:
    - 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)

- You should be able to build a model 1/2 translator now
- More on word alignment, decoding next class

WSD?

- Remember when we discussed WSD?
  - Word-based MT systems rarely have a WSD step
  - Why not?