IBM Translation Models

Instructor: Yoav Artzi

Slides adapted from Michael Collins
The Noisy Channel Model

- **Goal:** translate from French to English
- Have a model $p(e|f)$ to estimate the probability of an English sentence $e$ given a French sentence $f$
- Estimate the parameters from training corpus
- A noisy channel model has two components:
  - $p(e)$: the language model
  - $p(f|e)$: the translation model
- Giving:
  
  $$p(e|f) = \frac{p(e, f)}{p(f)} = \frac{p(e)p(f|e)}{\sum_e p(e)p(f|e)}$$

  and

  $$\arg \max_e p(e|f) = \arg \max_e p(e)p(f|e)$$
Overview

- IBM Model 1
- IBM Model 2
- EM Training of Models 1 and 2
IBM Model 1: Alignments

- How do we model $p(f|e)$?
- English sentence $e$ has $l$ words $e^1 \ldots e^l$
  French sentence $f$ has $m$ words $f^1 \ldots f^m$
- An **alignment** $a$ identifies which English word each French word originated from
- Formally, an alignment $a$ is:
  \[ \{a_1, \ldots, a_m\} \text{ where } a_j \in 0\ldots l \]
- There are $(l + 1)^m$ possible alignments
IBM Model 1: Alignments

\[ l = 6, \ m = 7 \]

\[ e = \text{And the program has been implemented} \]

\[ f = \text{Le programme a ete mis en application} \]
IBM Model 1: Alignments

\[ l = 6, \ m = 7 \]

\[ e = \text{And the program has been implemented} \]

\[ f = \text{Le programme a été mis en application} \]

- One alignment is

\[ \{2, 3, 4, 5, 6, 6, 6\} \]
IBM Model 1: Alignments

\[ l = 6, \ m = 7 \]
\[ e = \text{And the program has been implemented} \]

\[ f = \text{Le programme a ete mis en application} \]

• Another (bad!) alignment is

\[ \{1, 1, 1, 1, 1, 1, 1, 1\} \]
IBM Model 1: Alignments

\[ l = 6, \ m = 7 \]

\[ e = \text{And the program has been implemented} \]

\[ f = \text{Le programme a été mis en application} \]

- Another (bad!) alignment is

\[ \{1, 1, 1, 1, 1, 1, 1, 1\} \]
Alignments in the IBM Models

• We define two models:

\[ p(a|e, m) \quad p(f|a, e, m) \]

• Giving:

\[ p(f, a|e, m) = p(a|e, m)p(f|a, e, m) \]

• Also:

\[ p(f|e, m) = \sum_{a \in A} p(a|e, m)p(f|a, e, m) \]

where \( A \) is a set of all possible alignments
Most Likely Alignments

\[ p(f, a|e, m) = p(a|e, m)p(f|a, e, m) \]

- We can also calculate:

\[ p(a|f, e, m) = \frac{p(f, a|e, m)}{\sum_{a \in A} p(f, a|e, m)} \]

for any alignment \( a \)

- For a given \( f, e \) pair, can also compute the most likely alignment (details in notes)

- The original IBM models are rarely used for translation, but still key for recovering alignments
Example Alignment

• French:
  le conseil a rendu son avis, et nous devons à présent adopter un nouvel avis sur la base de la première position.

• English:
  the council has stated its position, and now, on the basis of the first position, we again have to give our opinion.

• Alignment:
  the/le council/conseil has/à stated/rendu its/son position/avis,, and/et now/présent,NULL on/sur the/le basis/base of/de the/la first/première position/position,, NULL we/nous again/NULL have/devons to/a give/adopter our/nouvel opinion/avis,. 
IBM Model 1: Alignments

• In IBM Model 1 all alignments $a$ are equally likely:

$$p(a|e, m) = \frac{1}{(1 + l)^m}$$

• Reasonable assumption?
  – Simplifying assumption, but it gets things started …
IBM Model 1: Translation Probabilities

• Next step: come up with an estimate for 
  \[ p(f|a, e, m) \]

• In Model 1, this is:
  \[
  p(f|a, e, m) = \prod_{j=1}^{m} t(f_j|e_{a_j})
  \]
IBM Model 1: Example

\( l = 6, \ m = 7 \)
\( e = \) And the program has been implemented

\( f = \) Le programme a ete mis en application

\( a = \{2, 3, 4, 5, 6, 6, 6\} \)
IBM Model 1: Example

<table>
<thead>
<tr>
<th>p(fle)</th>
<th>And</th>
<th>the</th>
<th>program</th>
<th>has</th>
<th>been</th>
<th>implemented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le</td>
<td>0.2</td>
<td>0.6</td>
<td>0.1</td>
<td>0.025</td>
<td>0.05</td>
<td>0.025</td>
</tr>
<tr>
<td>programme</td>
<td>0.05</td>
<td>0.2</td>
<td>0.45</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>a</td>
<td>0.1</td>
<td>0.1</td>
<td>0.15</td>
<td>0.2</td>
<td>0.15</td>
<td>0.3</td>
</tr>
<tr>
<td>ete</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td>mis</td>
<td>0.2</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.25</td>
<td>0.4</td>
</tr>
<tr>
<td>en</td>
<td>0.25</td>
<td>0.1</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1</td>
<td>0.05</td>
</tr>
<tr>
<td>application</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.9</td>
</tr>
</tbody>
</table>
IBM Model 1: Example

\[ l = 6, \ m = 7 \]
\[ e = \text{And the program has been implemented} \]
\[ f = \text{Le programme a ete mis en application} \]
\[ a = \{2, 3, 4, 5, 6, 6, 6\} \]

\[ p(f|a, e) = t(\text{Le|the}) \times t(\text{programme|program}) \]
\[ \times t(\text{a|has}) \times t(\text{ete|been}) \]
\[ \times t(\text{mis|implemented}) \times t(\text{en|implemented}) \]
\[ \times t(\text{application|implemented}) = 0.0006804 \]

\[ p(f, a \mid e, 7) = 8.26186E - 10 \]
IBM Model 1: The Generative Process

To generate a French string $f$ from an English string $e$:

- Step 1: Pick an alignment $a$ with probability $\frac{1}{(l+1)^m}$
- Step 2: Pick the French words with probability

$$p(f|a, e, m) = \prod_{j=1}^{m} t(f_j|e_{a_j})$$

The final result:

$$p(f, a|e, m) = p(a|e, m) \times p(f|a, e, m) = \frac{1}{(1 + l)^m} \prod_{j=1}^{m} t(f_j|e_{a_j})$$
... de la situation au niveau des négociations de l’ompi ...
... of the current position in the wipo negotiations ...

nous ne sommes pas en mesure de décider, ...
we are not in position to decide ...

... Le point de vue de la commission face à ce problème complexe .
... the commission ‘s position on this complex problem .
Overview

• IBM Model 1
• IBM Model 2
• EM Training of Models 1 and 2
IBM Model 2

• Only difference: we now introduce **alignment distortion parameters**
  \[ q(i|j, l, m) \]

• Probability that \( j \)'th French word is connected to \( i \)'th English word, given sentence length of \( e \) and \( f \) are \( l \) and \( m \)

• Define
  \[ p(a|e, m) = \prod_{j=1}^{m} q(a_j|j, l, m) \]
  where \( a = \{ a_1, \ldots, a_m \} \)

• Gives
  \[ p(f, a|e, m) = \prod_{j=1}^{m} q(a_j|j, l, m)t(f_j|e_{a_j}) \]
Example

\[ l = 6 \]
\[ m = 7 \]
\[ e = \text{And the program has been implemented} \]
\[ f = \text{Le programme a ete mis en application} \]
\[ a = \{2, 3, 4, 5, 6, 6, 6\} \]

\[ p(a \mid e, 7) = q(2 \mid 1, 6, 7) \times q(3 \mid 2, 6, 7) \times q(4 \mid 3, 6, 7) \times q(5 \mid 4, 6, 7) \times q(6 \mid 5, 6, 7) \times q(6 \mid 6, 6, 7) \times q(6 \mid 7, 6, 7) \]
\[ l = 6 \]
\[ m = 7 \]
\[ e = \text{And the program has been implemented} \]
\[ f = \text{Le programme a été mis en application} \]
\[ a = \{2,3,4,5,6,6,6\} \]

\[ p(f \mid a,e,7) = t(Le \mid \text{the}) \times \]
\[ t(\text{programme} \mid \text{program}) \times \]
\[ t(a \mid \text{has}) \times \]
\[ t(\text{ete} \mid \text{been}) \times \]
\[ t(\text{mis} \mid \text{implemented}) \times \]
\[ t(\text{en} \mid \text{implemented}) \times \]
\[ t(\text{application} \mid \text{implemented}) \]