Model Estimation in Machine Translation

John DeNero
Some slides borrowed from Dan Klein and David Chiang
Stitching Together Fragments

Machine translation system:

Yo lo haré después  →  Model of translation  ←  I will do it later
Stitching Together Fragments

Parallel corpus gives translation examples:

I will do it gladly  
Yo lo haré de muy buen grado

You will see later  
Después lo veras

Machine translation system:

Yo lo haré después  
Model of translation

I will do it later
Stitching Together Fragments

Parallel corpus gives translation examples:

I will do it gladly

Yo lo haré de muy buen grado

You will see later

Después lo verás

Machine translation system:

Yo lo haré después

Model of translation

I will do it later
Stitching Together Fragments

Parallel corpus gives translation examples:

I will do it gladly

Yo lo haré de muy buen grado

You will see later

Después lo veras

Machine translation system:

Yo lo haré
después

Model of translation

I will do it
later
Stitching Together Fragments

Parallel corpus gives translation examples:

I will do it gladly
Yo lo haré de muy buen grado

You will see later
Después lo verás

Machine translation system:

Yo lo haré
después

I will do it
later

Model of translation
Stitching Together Fragments

Parallel corpus gives translation examples:

I will do it gladly

Yo lo haré de muy buen grado

You will see later

Después lo verás

Machine translation system:

Yo lo haré
después

I will do it
later
An Example Syntax-Based Translation

Arabic source sentence:

ورفض الباز الأدلاء بآي تصريحات فور وصوله إلى المقاطعة

Wednesday, June 16, 2010
An Example Syntax-Based Translation

Arabic source sentence:

ورفض الباز الأدلاء بآي تصريحات فور وصوله إلى المقاطعة

Reference translation from a human translator:

Al-baz declined to make any statements upon his arrival in the province
An Example Syntax-Based Translation

Arabic source sentence:
ورفض البار الأدلاء بَيْن تصريحات فور وصوله إلى المقاطعة

Reference translation from a human translator:
Al-baz declined to make any statements upon his arrival in the province
An Example Syntax-Based Translation

Arabic source sentence:
ورفض البار الأدلاء بأي تصريحات فور وصوله إلى المقاطعة

Reference translation from a human translator:
Al-baz declined to make any statements upon his arrival in the province

Features:

<table>
<thead>
<tr>
<th>feature</th>
<th>weight</th>
<th>value</th>
<th>product</th>
</tr>
</thead>
<tbody>
<tr>
<td>derivation-size</td>
<td>0.41</td>
<td>8</td>
<td>3.30</td>
</tr>
<tr>
<td>glue-rule</td>
<td>3.89</td>
<td>2</td>
<td>7.78</td>
</tr>
<tr>
<td>green</td>
<td>-0.08</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gt_prob</td>
<td>0.40</td>
<td>36.18</td>
<td>14.43</td>
</tr>
<tr>
<td>identity</td>
<td>-9.97</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>is_lexicalized</td>
<td>-0.65</td>
<td>6</td>
<td>-3.91</td>
</tr>
<tr>
<td>lex_pfs</td>
<td>1.02</td>
<td>5.47</td>
<td>5.60</td>
</tr>
<tr>
<td>lex_pfe</td>
<td>0.31</td>
<td>4.44</td>
<td>1.39</td>
</tr>
<tr>
<td>lm1</td>
<td>1</td>
<td>22.76</td>
<td>22.76</td>
</tr>
<tr>
<td>lm1-unk</td>
<td>30.08</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lm2</td>
<td>0.74</td>
<td>26.66</td>
<td>19.79</td>
</tr>
<tr>
<td>lm2-unk</td>
<td>-39.18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>missingWord</td>
<td>-1.29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>model1inv</td>
<td>1.02</td>
<td>10.60</td>
<td>10.81</td>
</tr>
<tr>
<td>model1invm</td>
<td>1.35</td>
<td>11.29</td>
<td>15.22</td>
</tr>
<tr>
<td>nonmonotone</td>
<td>4.17</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>olive</td>
<td>1.95</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>psm1-unk</td>
<td>0.50</td>
<td>24.65</td>
<td>12.30</td>
</tr>
<tr>
<td>text-length</td>
<td>-3.87</td>
<td>15</td>
<td>-58.05</td>
</tr>
<tr>
<td>trivial_cond_prob</td>
<td>0.41</td>
<td>3.34</td>
<td>1.38</td>
</tr>
<tr>
<td>unk-rule</td>
<td>19.28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>reported totalcost</td>
<td>52.82</td>
<td>V · W</td>
<td>52.82</td>
</tr>
</tbody>
</table>
Thank you, I will do it gladly.

Gracias,
lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias,
lo haré de muy buen grado.
The Alignment Problem in Translation

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
The Alignment Problem in Translation

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
In 1993, we translated words

Yo lo haré mañana
I will do it tomorrow
In 1993, we translated words

<table>
<thead>
<tr>
<th>English (E)</th>
<th>$P(E \mid \text{mañana})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>tomorrow</td>
<td>0.7</td>
</tr>
<tr>
<td>morning</td>
<td>0.3</td>
</tr>
</tbody>
</table>
In 1993, we translated words

| English (E) | P(E | mañana) |
|-------------|------------|
| tomorrow    | 0.7        |
| morning     | 0.3        |

In 1999, we translated phrases
In 1993, we translated words

<table>
<thead>
<tr>
<th>English (E)</th>
<th>$P(\ E \mid \text{mañana} \ )$</th>
</tr>
</thead>
<tbody>
<tr>
<td>tomorrow</td>
<td>0.7</td>
</tr>
<tr>
<td>morning</td>
<td>0.3</td>
</tr>
</tbody>
</table>

In 1999, we translated phrases

<table>
<thead>
<tr>
<th>English (E)</th>
<th>$P(\ E \mid \text{lo haré} \ )$</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>
In 1993, we translated words

<table>
<thead>
<tr>
<th>Spanish</th>
<th>English</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yo lo haré mañana</td>
<td>tomorrow</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>morning</td>
<td>0.3</td>
</tr>
</tbody>
</table>

In 1999, we translated phrases

<table>
<thead>
<tr>
<th>Spanish</th>
<th>English</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yo lo haré mañana</td>
<td>will do it</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>will do so</td>
<td>0.2</td>
</tr>
</tbody>
</table>

In 2004, we translated fragments

<table>
<thead>
<tr>
<th>Spanish</th>
<th>English</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yo lo haré mañana</td>
<td>I will do it tomorrow</td>
<td></td>
</tr>
</tbody>
</table>
In 1993, we translated words

\[ \text{Yo lo haré mañana} \]
\[ \text{I will do it tomorrow} \]

<table>
<thead>
<tr>
<th>English (E)</th>
<th>( P( E \mid \text{mañana} ) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>tomorrow</td>
<td>0.7</td>
</tr>
<tr>
<td>morning</td>
<td>0.3</td>
</tr>
</tbody>
</table>

In 1999, we translated phrases

\[ \text{Yo lo haré mañana} \]
\[ \text{I will do it tomorrow} \]

<table>
<thead>
<tr>
<th>English (E)</th>
<th>( P( E \mid \text{lo haré} ) )</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>

In 2004, we translated fragments

\[ \text{Yo lo haré mañana} \]
\[ \text{I will do it tomorrow} \]
**Features Match Model Structure**

**In 1993, we translated words**

<table>
<thead>
<tr>
<th>English (E)</th>
<th>$P(E \mid \text{mañana})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>tomorrow</td>
<td>0.7</td>
</tr>
<tr>
<td>morning</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**In 1999, we translated phrases**

<table>
<thead>
<tr>
<th>English (E)</th>
<th>$P(E \mid \text{lo haré})$</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>

**In 2004, we translated fragments**
In 1993, we translated words

| English (E) | \( P(\ E | \text{mañana} ) \) |
|------------|-----------------|
| tomorrow   | 0.7             |
| morning    | 0.3             |

In 1999, we translated phrases

| English (E) | \( P(\ E | \text{lo haré} ) \) |
|------------|-----------------|
| will do it | 0.8             |
| will do so | 0.2             |

In 2004, we translated fragments

\[
P(\text{will do it tomorrow}) = 0.8
\]
Aligning Structural Components

Today, we actually still align words
Aligning Structural Components

Today, we actually still align words

1. Learn to align words to words
Aligning Structural Components

Today, we actually still align words

1. **Learn** to align words to words

Yo lo haré mañana
I will do it tomorrow
Aligning Structural Components

Today, we actually still align words

1. **Learn** to align words to words

2. **Enumerate** larger structures given the alignment

Yo lo haré mañana

I will do it tomorrow
Aligning Structural Components

Today, we actually still align words

1. **Learn** to align words to words

2. **Enumerate** larger structures given the alignment

Yo lo haré mañana
I will do it tomorrow
Aligning Structural Components

Today, we actually still align words

1. **Learn** to align words to words

2. **Enumerate** larger structures given the alignment

3. **Translate with the larger structures**

*Yo lo haré mañana*
*I will do it tomorrow*
Thank you, I will do it gladly.

Gracias,
lo
haré
de
muy
buen
grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.

Frequency statistics on these rules guide translation.
Unsupervised Word Alignment

- **Input**: A large *bitext* of sentences and their translations
- **Learn**: Alignments and a probabilistic lexicon
Unsupervised Word Alignment

- **Input**: A large *bitext* of sentences and their translations
- **Learn**: Alignments and a probabilistic lexicon
Unsupervised Word Alignment

- **Input**: A large *bitext* of sentences and their translations
- **Learn**: Alignments and a probabilistic lexicon

<table>
<thead>
<tr>
<th>French (F)</th>
<th>$P(F \mid \text{your})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>votre</td>
<td>0.9</td>
</tr>
<tr>
<td>vous</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Unsupervised Word Alignment

- **Input**: A large *bitext* of sentences and their translations

- **Learn**: Alignments and a probabilistic lexicon

<table>
<thead>
<tr>
<th>French (F)</th>
<th>$P(F \mid your)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>votre</td>
<td>0.9</td>
</tr>
<tr>
<td>vous</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Thank you for your trouble
*Merci de vous être dérangé*
Unsupervised Word Alignment

• **Input:** A large *bitext* of sentences and their translations

• **Learn:** Alignments and a probabilistic lexicon

<table>
<thead>
<tr>
<th>French (F)</th>
<th>$P(F \mid \text{your})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>votre</td>
<td>0.9</td>
</tr>
<tr>
<td>vous</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Thank you for **your** trouble

*Merci de **vous** être dérangé*

• **Exciting fact:** Unsupervised methods perform well enough that very few systems use hand-aligned data sets at all
Bilingual Dictionaries are Ambiguous

Example from Douglas Hofstadter
Bilingual Dictionaries are Ambiguous

条件, 抽屉, 速度, "bikini top", "top of your lungs", "top dog", "top brass", "top of the line", "big top", "over the top", "pop top", "top off", "off the top of my head", "take it from the top", "I’m on top of it", ...

Example from Douglas Hofstadter
I declare resumed the session of the European parliament.

Declaro reanudado el periodo de sesiones del parlamento europeo.

adjourned on Friday 17 December 1999, ...

interrumpido el Viernes 17 de Diciembre pasado, ...
I declare resumed the session of the European parliament

Declaro reanudado el periodo de sesiones del parlamento europeo

adjourned on Friday 17 December 1999, ...

interrumpido el Viernes 17 de Diciembre pasado, ...
I declare resumed the session of the European Parliament.

Declaro reanudado el periodo de sesiones del parlamento europeo.

adjourned on Friday 17 December 1999, ...

interrumpido el Viernes 17 de Diciembre pasado, ...
I declare resumed the session of the European parliament

Declaro reanudado el periodo de sesiones del parlamento europeo

adjourned on Friday 17 December 1999, ...

interrumpido el Viernes 17 de Diciembre pasado, ...
I declare resumed the session of the European Parliament

Declaro reanudado el periodo de sesiones del Parlamento Europeo

adjourned on Friday 17 December 1999, ...

interrumpido el Viernes 17 de Diciembre pasado, ...
Properties of Word Alignments

I declare resumed the session of the European Parliament

Declaro reanudado el periodo de sesiones del parlamento europeo

adjourned on Friday 17 December 1999, ...

interrumpido el Viernes 17 de Diciembre pasado, ...

- Often one-to-one or many-to-one (usually over contiguous phrases)
- Occasionally many-to-many, driven by non-literal translations
Idea: Use Word Co-occurrence

- Two words that co-occur regularly are translations
  \[ c(e, f) \quad \text{The number of times } e \text{ and } f \text{ appear together} \]

- Normalize by the word frequencies
  \[ c(f) \quad \text{Count of word } f \quad c(e) \quad \text{Count of word } f \]
  \[ \frac{2 \cdot c(e, f)}{c(e) + c(f)} \quad \text{Dice coefficient} \]

- Enforcing competition across words (e.g., finding a one-to-one or many-to-one mapping) is a good idea
## Probabilistic Modeling 101

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Process or Model</th>
</tr>
</thead>
</table>

<p>| Objective Function |  |</p>
<table>
<thead>
<tr>
<th>Parameters</th>
<th>Process or Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>What we learn</em></td>
<td></td>
</tr>
</tbody>
</table>

**Objective Function**
Probabilistic Modeling 101

Parameters

What we learn

| Foreign (F) | $P(F | \text{your })$ |
|------------|-----------------------|
| nuestro    | ???                   |
| nuestra    | ???                   |
| el         | ???                   |
| queso      | ???                   |

... 

Objective Function
Probabilistic Modeling 101

Parameters

What we learn

<table>
<thead>
<tr>
<th>Foreign (F)</th>
<th>$P(F \mid \text{cheese})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign (F)</td>
<td>$P(F \mid \text{the})$</td>
</tr>
<tr>
<td>Foreign (F)</td>
<td>$P(F \mid \text{your})$</td>
</tr>
<tr>
<td>nuestro</td>
<td>??</td>
</tr>
<tr>
<td>nuestra</td>
<td>??</td>
</tr>
<tr>
<td>el</td>
<td>??</td>
</tr>
<tr>
<td>queso</td>
<td>??</td>
</tr>
</tbody>
</table>

...
### Probabilistic Modeling 101

#### Parameters

**What we learn**

<table>
<thead>
<tr>
<th>Foreign (F)</th>
<th>( P( F \mid \text{cheese} ) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign (F)</td>
<td>( P( F \mid \text{the} ) )</td>
</tr>
<tr>
<td>Foreign (F)</td>
<td>( P( F \mid \text{your} ) )</td>
</tr>
<tr>
<td><em>nuestro</em></td>
<td>???</td>
</tr>
<tr>
<td><em>nuestra</em></td>
<td>???</td>
</tr>
<tr>
<td><em>el</em></td>
<td>???</td>
</tr>
<tr>
<td><em>queso</em></td>
<td>???</td>
</tr>
</tbody>
</table>

...  

#### Process or Model

**How parameters relate to data**

#### Objective Function
### Parameters

**What we learn**

<table>
<thead>
<tr>
<th>Foreign (F)</th>
<th>$P( F \mid \text{cheese} )$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign (F)</td>
<td>$P( F \mid \text{the} )$</td>
</tr>
<tr>
<td>Foreign (F)</td>
<td>$P( F \mid \text{your} )$</td>
</tr>
<tr>
<td>nuestro</td>
<td>???</td>
</tr>
<tr>
<td>nuestra</td>
<td>???</td>
</tr>
<tr>
<td>el</td>
<td>???</td>
</tr>
<tr>
<td>queso</td>
<td>???</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Process or Model

**How parameters relate to data**

I declare resumed the session

### Objective Function
Probabilistic Modeling 101

Parameters

What we learn

<table>
<thead>
<tr>
<th>Foreign (F)</th>
<th>( P(\ F \mid \text{cheese}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign (F)</td>
<td>( P(\ F \mid \text{the}) )</td>
</tr>
<tr>
<td>Foreign (F)</td>
<td>( P(\ F \mid \text{your}) )</td>
</tr>
<tr>
<td>nuestro</td>
<td>???</td>
</tr>
<tr>
<td>nuestra</td>
<td>???</td>
</tr>
<tr>
<td>el</td>
<td>???</td>
</tr>
<tr>
<td>queso</td>
<td>???</td>
</tr>
<tr>
<td>...</td>
<td>???</td>
</tr>
</tbody>
</table>

Process or Model

How parameters relate to data

I declare resumed the session

\[\text{Declaro}\]

Objective Function
Probabilistic Modeling 101

Parameters

What we learn

Process or Model

How parameters relate to data

| Foreign (F) | P( F | cheese ) |
|-------------|----------------|
| Foreign (F) | P( F | the )      |
| Foreign (F) | P( F | your )     |
| nuestro     | ???           |
| nuestra     | ???           |
| el           | ???           |
| queso        | ???           |

...
### Parameters

**What we learn**

<table>
<thead>
<tr>
<th>Foreign (F)</th>
<th>( P( F \mid cheese ) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign (F)</td>
<td>( P( F \mid the ) )</td>
</tr>
<tr>
<td>Foreign (F)</td>
<td>( P( F \mid your ) )</td>
</tr>
<tr>
<td>nuestro</td>
<td>???</td>
</tr>
<tr>
<td>nuestra</td>
<td>???</td>
</tr>
<tr>
<td>el</td>
<td>???</td>
</tr>
<tr>
<td>queso</td>
<td>???</td>
</tr>
</tbody>
</table>

...
Probabilistic Modeling 101

Parameters

What we learn

| Foreign (F) | P(F | cheese) |
|------------|------------|
| Foreign (F) | P(F | the) |
| Foreign (F) | P(F | your) |
| nuestro     | ???       |
| nuestra     | ???       |
| el          | ???       |
| queso       | ???       |
| ...         |           |

Process or Model

How parameters relate to data

I declare resumed the session

Declaro reanudado el periodo

Objective Function
### Parameters

**What we learn**

| Foreign (F) | $P(F | \text{cheese})$ |
|-------------|-----------------------|
| Foreign (F) | $P(F | \text{the})$   |
| Foreign (F) | $P(F | \text{your})$   |
| nuestro     | ???                   |
| nuestra     | ???                   |
| el           | ???                   |
| queso        | ???                   |
| ...          |                       |

### Process or Model

**How parameters relate to data**

I declare resumed the session

Declaro reanudado el periodo de

### Objective Function
Probabilistic Modeling 101

Parameters

What we learn

<table>
<thead>
<tr>
<th>Foreign (F)</th>
<th>$P( F \mid \text{cheese} )$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign (F)</td>
<td>$P( F \mid \text{the} )$</td>
</tr>
<tr>
<td>Foreign (F)</td>
<td>$P( F \mid \text{your} )$</td>
</tr>
<tr>
<td>nuestro</td>
<td>???</td>
</tr>
<tr>
<td>nuestra</td>
<td>???</td>
</tr>
<tr>
<td>el</td>
<td>???</td>
</tr>
<tr>
<td>queso</td>
<td>???</td>
</tr>
</tbody>
</table>

Process or Model

How parameters relate to data

I declare resumed the session

Declaro reanudado el periodo de sesiones

Objective Function
Probabilistic Modeling 101

Parameters

<table>
<thead>
<tr>
<th>Foreign (F)</th>
<th>( P( F \mid \text{cheese}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign (F)</td>
<td>( P( F \mid \text{the}) )</td>
</tr>
<tr>
<td>Foreign (F)</td>
<td>( P( F \mid \text{your}) )</td>
</tr>
<tr>
<td>nuestro</td>
<td>???</td>
</tr>
<tr>
<td>nuestra</td>
<td>???</td>
</tr>
<tr>
<td>el</td>
<td>???</td>
</tr>
<tr>
<td>queso</td>
<td>???</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Process or Model

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I declare resumed the session</td>
</tr>
<tr>
<td>Declaro reanudado el periodo de sesiones</td>
</tr>
</tbody>
</table>

Objective Function

Guides learning
Probabilistic Modeling 101

Parameters

What we learn

| Foreign (F) | \( P( F | \text{cheese} ) \) |
|-------------|-------------------------------|
| Foreign (F) | \( P( F | \text{the} ) \) |
| Foreign (F) | \( P( F | \text{your} ) \) |
| \text{nuestro} | ??? |
| \text{nuestra} | ??? |
| \text{el} | ??? |
| \text{queso} | ??? |

Process or Model

How parameters relate to data

I declare resumed the session

\[ \text{Declaro reanudado el periodo de sesiones} \]

Objective Function

Guides learning

Maximize the probability of what you observe
Probabilistic Modeling 101

Parameters

What we learn

<table>
<thead>
<tr>
<th>Foreign (F)</th>
<th>( P( F \mid \text{cheese}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign (F)</td>
<td>( P( F \mid \text{the}) )</td>
</tr>
<tr>
<td>Foreign (F)</td>
<td>( P( F \mid \text{your}) )</td>
</tr>
</tbody>
</table>
| 0.5
| 0.5
| 0.0
| 0.0

...
• Assume that foreign words are each generated independently
• Assume a hidden alignment vector $a$ encoding which English word generates each foreign word

I declare resumed the session

Declaro reanudado el periodo de sesiones
• Assume that foreign words are each generated independently

• Assume a hidden alignment vector $a$ encoding which English word generates each foreign word

I declare resumed the session

Declaro reanudado el periodo de sesiones

$a_6 = 5$
IBM Model 1

- Assume that foreign words are each generated independently
- Assume a hidden alignment vector $a$ encoding which English word generates each foreign word

$$P(f, a | e) = \prod_{j=1}^{J} P(a_j = i | I, J) P(f_j | e_i)$$

I declare resumed the session

Declaro reanudado el periodo de sesiones

$a_6 = 5$
IBM Model 1

- Assume that foreign words are each generated independently
- Assume a hidden alignment vector $a$ encoding which English word generates each foreign word

\[
P(f, a | e) = \prod_{j=1}^{J} P(a_j = i | I, J) P(f_j | e_i) = \frac{1}{I + 1} P(f_j | e_i)
\]
Estimating Model 1 Parameters

\[ P(f|e) \]
Estimating Model 1 Parameters

- Free parameters in the model: $P(f|e)$
- E-step computes expected alignments (posteriors)
Estimating Model 1 Parameters

- Free parameters in the model: \( P(f|e) \)
- E-step computes expected alignments (posteriors)

\[
P(a_j = i|e, f) = \frac{\frac{1}{I+1} P(f_j | e_i)}{\sum_{i'} \frac{1}{I+1} P(f_j | e_{i'})}
\]
Estimating Model 1 Parameters

- Free parameters in the model: \( P(f|e) \)

- E-step computes expected alignments (posteriors)

\[
P(a_j = i | e, f) = \frac{1}{I+1} P(f_j | e_i) \sum_{i'} \frac{1}{I+1} P(f_j | e_{i'})
\]

- M-step computes ratios of expected counts
Estimating Model 1 Parameters

- Free parameters in the model: $P(f|e)$
- E-step computes expected alignments (posteriors)

$$P(a_j = i|e, f) = \frac{\frac{1}{I+1} P(f_j|e_i)}{\sum_{i'} \frac{1}{I+1} P(f_j|e_{i'})}$$

- M-step computes ratios of expected counts

$$P(f|e) = \frac{\text{sum of posteriors for } f \text{ aligned to } e}{\text{sum of posteriors of any } f' \text{ aligned to } e}$$
Estimating Model I Parameters

- Free parameters in the model: \( P(f|e) \)

- E-step computes expected alignments (posteriors)

\[
P(a_{ij} = i|e, f) = \frac{\frac{1}{I+1} P(f_j|e_i)}{\sum_{i'} \frac{1}{I+1} P(f_j|e_{i'})}
\]

- M-step computes ratios of expected counts

\[
P(f|e) = \frac{\text{sum of posteriors for } f \text{ aligned to } e}{\text{sum of posteriors of any } f' \text{ aligned to } e}
\]

- Iterate e- and m-step many times (5 or 10)
Aligning Words Under the Model

- **Viterbi**: For every \( j \), select \( i \) that maximizes

\[
P(a_j = i | e, f)
\]

*Gives competition among explanations*

- **Posterior**: Align every \((i,j)\) that has

\[
P(a_j = i | e, f) > \tau
\]

*Gives control over how many alignment links to posit*
Evaluation: Alignment Error Rate

☐ = Sure
☐ = Possible
■ = Predicted

en
1978
,
on
a
enregistré
1,122,000
divorces
sur
le
continent
Evaluation: Alignment Error Rate

☐ = Sure
☐ = Possible
■ = Predicted

**Precision:** fraction of predicted that are sure or possible

en 1978
, on a
enregistré 1,122,000 divorces sur
le continent .
Evaluation: Alignment Error Rate

□ = Sure
☐ = Possible
■ = Predicted

Precision: fraction of predicted that are sure or possible

Recall: fraction of sure that are predicted
Evaluation: Alignment Error Rate

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

**Precision:** fraction of predicted that are sure or possible

**Recall:** fraction of sure that are predicted
Evaluation: Alignment Error Rate

\[ \square = \text{Sure} \]

\[ \bigcirc = \text{Possible} \]

\[ \blacksquare = \text{Predicted} \]

**Precision:** fraction of predicted that are sure or possible

**Recall:** fraction of sure that are 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 IBM Model 1

- Too many alignments to rare words (garbage collection)
- Alignments jump around all over the sentence
Problems with IBM Model 1

- Too many alignments to rare words (garbage collection)

- Alignments jump around all over the sentence
Intersected IBM Model I

le terme ferroviaire est << chargement sur demande >>

the railroad term is << demand loading >>
Intersected IBM Model 1

- Train Model 1 in both directions, align with each, then intersect the output.
- Result is one-to-one with Viterbi alignments.
- Second model filters the first, eliminating mistakes.
Train Model 1 in both directions, align with each, then intersect the output

Result is one-to-one with Viterbi alignments

Second model filters the first, eliminating mistakes

<table>
<thead>
<tr>
<th>Model</th>
<th>P/R</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 $E \rightarrow F$</td>
<td>82/58</td>
<td>30.6</td>
</tr>
<tr>
<td>Model 1 $F \rightarrow E$</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 for IBM Model 1

- We can intersect model predictions during training as well
- Modified alignment posterior: $P_{e\rightarrow f}(a_j = i \mid e, f) \cdot P_{f\rightarrow e}(a_i = j \mid e, f)$
- Models are forced to agree as they select parameters
- Same precision benefits, but higher recall from more agreement

<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 JOINT</td>
<td>93/69</td>
<td>19.5</td>
</tr>
</tbody>
</table>
IBM Model 2

- Words at the beginning of sentences should align
- Words at the end of sentences should align
- Model 2: Alignment probability depends on position
On Tuesday Nov. 4, earthquakes rocked Japan once again.

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

Absolute position isn’t the right quantity
E: Thank you, I shall do so gladly.

F: Gracias, lo haré de muy buen grado.
E: Thank you, I shall do so gladly.

F: Gracias, lo haré de muy buen grado.
E: Thank you, I shall do so gladly.

A: 

F: Gracias, lo haré de muy buen grado.
E: Thank you, I shall do so gladly.

Model Parameters

*Emissions:* $P(F_1 = \text{Gracias} | E_{A1} = \text{Thank})$

*Transitions:* $P(A_2 = 3 | I, J)$
IBM Models 1/2

Thank you, I shall do so gladly.

Emissions: $P(F_1 = \text{Gracias} \mid E_{A_1} = \text{Thank})$

Transitions: $P(A_2 = 3 \mid I, J)$
IBM Models 1/2

E: Thank you, I shall do so gladly.

A: 1 3 7 6 8 8 8 8 9

F: Gracias, lo haré de muy buen grado.

Model Parameters

Emissions: P(F₁ = Gracias | Eₐ₁ = Thank)  Transitions: P(A₂ = 3 | I, J)
The HMM Model

E: Thank you, I shall do so gladly.

F: Gracias, lo haré de muy buen grado.
The HMM Model

E: Thank you, I shall do so gladly.

A: 1 2 3 4 5 6 7 8 9

F: Gracias, lo haré de muy buen grado.
The HMM Model

**A:**

1 → 3 → 7 → 6 → 8 → 8 → 8 → 8 → 9 → 9

**E:** Thank you, I shall do so gladly.

**F:** Gracias, lo haré de muy buen grado.
The HMM Model

E:  Thank you, I shall do so gladly.

A:  1 → 3 → 7 → 6 → 8 → 8 → 8 → 8 → 9

F:  Gracias, lo haré de muy buen grado.

Model Parameters

Emissions: P( F_1 = Gracias | E_{A_1} = Thank )

Transitions: P( A_2 = 3 | A_1 = 1)
HMM Examples

we deemed it advisable to attend the meeting and so informed cojo.

we ne avons pas cru bon de assister

la reunion et en avons inform le cojo en consequence.

nous ne avons pas cru bon de assister

la reunion et en avons inform le cojo en consequence.

Wednesday, June 16, 2010
### Alignment Error Rates for HMMs

**French-English**

<table>
<thead>
<tr>
<th>Model</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 JOINT</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 JOINT</td>
<td>4.7</td>
</tr>
</tbody>
</table>
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Idea: *Don’t assume words are independent*

\[
P(\text{gracias} | \text{you})
\]

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Model-Based Phrase Alignment

Idea: Don’t assume words are independent

\[ P(\text{gracias} | \text{you}) \quad P(\text{gracias, Thank you}) \]

Thank you, I will do it gladly.

Gracias,
lo haré de muy buen grado.
Idea: *Don’t assume words are independent*

\[ P(\text{gracias} | \text{you}) \quad P(\text{gracias, Thank you}) \]

Thank you, I will do it gladly.

Gracias, lo haré de muy buen grado.
Alignment by Synchronous Grammars

Idea: *Divide and conquer to align sentences*

In the past two years
Idea: *Divide and conquer to align sentences*

In the past two years
Alignment by Synchronous Grammars

Idea: *Divide and conquer to align sentences*

In the past two years

过去 [past]
两 [two]
年 [year]
中 [in]
Alignment by Synchronous Grammars

Idea: *Divide and conquer to align sentences*

In the past two years

过去 [past]
两 [two]
年 [year]
中 [in]
Alignment by Synchronous Grammars

Idea: Divide and conquer to align sentences
Alignment by Synchronous Grammars

Idea: *Divide and conquer to align sentences*

- Language is hierarchically compositional
Alignment by Synchronous Grammars

Idea: *Divide and conquer to align sentences*

- Language is hierarchically compositional
- Different Languages share constituent types
Alignment by Synchronous Grammars

Idea: *Divide and conquer to align sentences*

- Language is hierarchically compositional
- Different Languages share constituent types
- *Alignment* and *syntax* should be modeled jointly
Estimating a Model that Translates

Yo lo haré después  Model of translation  I will do it later
Estimating a Model that Translates

Yo lo haré después

Model of translation

I will do it later

- We have many indicators of translation quality:
  \[ P(f|e), P(e|f), P(e), \ldots \]
Estimating a Model that Translates

- We have many indicators of translation quality: $P(f|e), P(e|f), P(e), ...$
- We call these *features* and we learn their weights

Yo lo haré después  →  I will do it later
Estimating a Model that Translates

• We have many indicators of translation quality: $P(f|e), P(e|f), P(e), ...$

• We call these features and we learn their weights

• Typically called the tuning stage in an MT pipeline
Estimating a Model that Translates

- We have many indicators of translation quality: $P(f|e), P(e|f), P(e), ...$
- We call these *features* and we learn their weights
- Typically called the *tuning* stage in an MT pipeline
- Discriminative training for structured problems:
  - Need an output scoring function
  - Need another optimization procedure
Automatic Translation Evaluation

• Scores how similar an automatically generated hypothesis is to human-generated references

• Dozens of variants — most common is BLEU

Reference: Al - baz declined to make any statement

Hypothesis: Al - baz declined to give any statement
Automatic Translation Evaluation

• Scores how similar an automatically generated hypothesis is to human-generated references

• Dozens of variants — most common is BLEU

Reference: Al - baz declined to make any statement

Hypothesis: Al - baz declined to give any statement

2/5
Automatic Translation Evaluation

- Scores how similar an automatically generated hypothesis is to human-generated references
- Dozens of variants — most common is BLEU

**Reference:** Al-baz declined to make any statement

**Hypothesis:** Al-baz declined to give any statement

2/5

3/6
Automatic Translation Evaluation

- Scores how similar an automatically generated hypothesis is to human-generated references
- Dozens of variants — most common is BLEU

Reference: Al - baz declined to make any statement
Hypothesis: Al - baz declined to give any statement

Comparing references and hypotheses with visual alignment:
- Reference: Al - baz declined to make any statement
- Hypothesis: Al - baz declined to give any statement

Alignment scores:
- 2/5
- 3/6
- 5/7
- 7/8
Automatic Translation Evaluation

- Scores how similar an automatically generated hypothesis is to human-generated references
- Dozens of variants — most common is BLEU

Reference: AI - baz declined to make any statement
Hypothesis: AI - baz declined to give any statement

Systems are trained to optimize this metric

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2/5</td>
<td>3/6</td>
<td>5/7</td>
<td>7/8</td>
<td></td>
</tr>
</tbody>
</table>

Wednesday, June 16, 2010
Minimum Error Rate Training (Och ’02)

Maximize the BLEU score of the highest scoring translation
Minimum Error Rate Training (Och ’02)

Maximize the BLEU score of the highest scoring translation
Max Margin Training (Chiang et al ’08)

Make the margin larger than the loss
Max Margin Training (Chiang et al ’08)

Make the margin larger than the loss
Model Score Predicts Accuracy

- + Samples from output space
- × Samples near maximum
- ◇ Highest scoring translation

Translation Quality (BLEU)

Total model score for 1000 sentences

Wednesday, June 16, 2010
Current Research on MT Estimation

• Add linguistic knowledge to the pipeline (syntax, disambiguation models, etc.)

• Use synchronous grammars and phrase models for alignment (instead of words)

• Condition on more context when translating a word or phrase

• Add lots of features to the log-linear model