IBM Model 1
and
Machine Translation
Recap
Expectation Maximization (EM)

0. Assume *some* value for your parameters

Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

   \[ p(z_i) \quad \rightarrow \quad \text{count}(z_i, w_i) \]

2. M-step: maximize log-likelihood, assuming these uncertain counts

   \[ p^{(t)}(z) \quad \leftrightarrow \quad \text{estimated counts} \quad \rightarrow \quad p^{(t+1)}(z) \]
Imagine three coins

Flip 1st coin (penny)

If heads: flip 2nd coin (dollar coin)

If tails: flip 3rd coin (dime)

unobserved: vowel or consonant? part of speech?

observed: a, b, e, etc.
We run the code, vs.
The run failed
Imagine three coins

Flip 1\textsuperscript{st} coin (penny)

\[ p(\text{heads}) = \lambda \quad p(\text{tails}) = 1 - \lambda \]

If heads: flip 2\textsuperscript{nd} coin (dollar coin)

\[ p(\text{heads}) = \gamma \quad p(\text{tails}) = 1 - \gamma \]

If tails: flip 3\textsuperscript{rd} coin (dime)

\[ p(\text{heads}) = \psi \quad p(\text{tails}) = 1 - \psi \]
Machine Translation

https://upload.wikimedia.org/wikipedia/commons/c/ca/Rosetta_Stone_BW.jpeg
Historical Context: World War II

From the National Archives (United Kingdom), via Wikimedia Commons, https://commons.wikimedia.org/wiki/File%3AColossus.jpg
By Antoine Taveneaux (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons https://commons.wikimedia.org/wiki/File%3ATuring-statue-Bletchley_14.jpg
Warren Weaver’s Note

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, 1947)

Noisy Channel Model

written in (clean) English

observed Russian (noisy) text

Decoded translation/decode model

Reranked (clean) language model

English
Noisy Channel Model

\[ p(X \mid Y) \propto p(Y \mid X) \ast p(X) \]
Noisy Channel Model

written in (clean) English

observed Russian (noisy) text

Decode

translation/decode model

Rerank

(clean) language model

\[ p(X | Y) \propto p(Y | X) \times p(X) \]
Translation

Translate French (observed) into English:

Le chat est sur la chaise.

The cat is on the chair.
Translation

Translate French (observed) into English:

Le chat est sur la chaise.

The cat is on the chair.

$p(English|French) \propto p(French|English) \ast p(English)$
Translation

Translate French (observed) into English:

Le chat est sur la chaise.

The cat is on the chair.

\[ p(\text{English}|\text{French}) \propto p(\text{French}|\text{English}) \times p(\text{English}) \]
Alignment

Le chat est sur la chaise.
The cat is on the chair.

$p(English|French) \propto p(French|English) \ast p(English)$
Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world,

Whereas disregard and contempt for human rights have resulted in barbarous acts which have outraged the conscience of mankind, and the advent of a world in which human beings shall enjoy freedom of speech and belief and freedom from fear and want has been proclaimed as the highest aspiration of the common people,

Whereas it is essential, if man is not to be compelled to have recourse, as a last resort, to rebellion against tyranny and oppression, that human rights should be protected by the rule of law,

Whereas it is essential to promote the development of friendly relations between nations,
Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world.

Whereas disregard and contempt for human rights have resulted in barbarous acts which have outraged the conscience of mankind, and the advent of a world in which human beings shall enjoy freedom of speech and belief and freedom from fear and want has been proclaimed as the highest aspiration of the common people,

Whereas it is essential, if man is not to be compelled to have recourse, as a last resort, to rebellion against tyranny and oppression, that human rights should be protected by the rule of law,

Whereas it is essential to promote the development of friendly relations between nations,

...
Alignments

• If we had word-aligned text, we could easily estimate $P(f|e)$.
  – But we don’t usually have word alignments, and they are expensive to produce by hand…
• If we had $P(f|e)$ we could produce alignments automatically.
IBM Model 1 (1993)

- Lexical Translation Model
- Word Alignment Model
- The simplest of the original IBM models
- For all IBM models, see the original paper (Brown et al, 1993):
Simplified IBM 1

• We’ll work through an example with a simplified version of IBM Model 1
• Figures and examples are drawn from *A Statistical MT Tutorial Workbook*, Section 27, (Knight, 1999)
• **Simplifying assumption:** each source word must translate to exactly one target word and vice versa
IBM Model 1 (1993)

- $f$: vector of French words
- $e$: vector of English words
- $a$: vector of alignment indices

Le chat est sur la chaise verte

The cat is on the green chair
IBM Model 1 (1993)

- $f$: vector of French words
- $e$: vector of English words
- $a$: vector of alignment indices
- $t(f_j|e_{a_j})$: translation probability of the word $f_j$ given the word $e_i$

Le chat est sur la chaise verte

The cat is on the green chair

$P(a, f|e) = \prod_{j=1}^{m} t(f_j|e_{a_j}) = t(f_1|e_{a_1}) \cdots t(f_m|e_{a_m})$

$P(f|e) = \sum_{a} P(a, f|e)$
Model and Parameters

**Want:** \( P(f|e) \)

But don’t know how to train this directly…

**Solution:** Use \( P(a, f|e) \), where \( a \) is an alignment

Remember \( P(f|e) = \sum_a P(a, f|e) \)
Model and Parameters: Intuition

Translation prob.: \( t(f_j \mid e_i) \)

Example:
\[ t(\text{chaise} \mid \text{chair}) > t(\text{chaise} \mid \text{the}) \]

Interpretation:
How probable is it that we see \( f_j \) given \( e_i \)
Model and Parameters: Intuition

Alignment/translation prob.: \( P(a, f | e) \)

Example (visual representation of \( a \)):

\[
P( \times | \text{“the cat”} ) < P( \| | \text{“the cat”} )
\]

Interpretation:
How probable are the alignment \( a \) and the translation \( f \) (given \( e \))
Model and Parameters: Intuition

Alignment prob.: $P(a | e, f)$

Example:

$P(\times | “le chat”, “the cat”) < P(\ | \ | “le chat”, “the cat”)$

Interpretation:

How probable is alignment $a$ (given $e$ and $f$)
Model and Parameters

How to compute:

\[
P(a, f|e) = \prod_{j=1}^{m} t(f_j|e_{a_j}) = t(f_1|e_{a_1}) \cdots t(f_m|e_{a_m})
\]

\[
P(f|e) = \sum_{a} P(a, f|e)
\]

\[
P(a|e, f) = \frac{P(a, f|e)}{\sum_{a'} P(a', f|e)}
\]
Parameters

In the coin example, we had 3 parameters from which we could compute all others:

\[ p(\text{heads}) = \lambda \]

\[ p(\text{heads}) = \gamma \]

\[ p(\text{heads}) = \psi \]
Parameters

For IBM model 1, we can compute all parameters given translation parameters:

$$t(f_j|e_i)$$

How many of these are there?
Parameters

For IBM model 1, we can compute all parameters given translation parameters:

\[ t(f_j | e_i) \]

How many of these are there?

\[ |French \ vocabulary| \times |English \ vocabulary| \]
## Data

Two sentence pairs:

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>b c</td>
<td>x y</td>
</tr>
<tr>
<td>b</td>
<td>y</td>
</tr>
</tbody>
</table>
All Possible Alignments

(French: x, y)  

(English: b, c)  

Remember:
- simplifying assumption that each word must be aligned exactly once
Expectation Maximization (EM)

0. Assume \textit{some} value for $t(f_j|e_i)$ and compute other parameter values

\textbf{Two step, iterative algorithm}

1. \textbf{E-step}: count alignments and translations under uncertainty, assuming these parameters

\[ t(f_j|e_i) \quad P(a,f|e) \]

2. \textbf{M-step}: maximize log-likelihood (update parameters), using uncertain counts
EM Step 0: Initialize

Set parameter values uniformly.
All translations have an equal chance of happening.

\[
\begin{align*}
  t(x|b) &= 1/2 \\
  t(y|b) &= 1/2 \\
  t(x|c) &= 1/2 \\
  t(y|c) &= 1/2
\end{align*}
\]
E-step: Compute $P(a,f|e)$

For all alignments, compute $P(a,f|e)$

Remember:

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

$$t(x|b) = 1/2$$
$$t(y|b) = 1/2$$
$$t(x|c) = 1/2$$
$$t(y|c) = 1/2$$

$$P(\begin{array}{c} x \\ y \end{array} | b c) = ?$$

$$P(\begin{array}{c} x \\ y \end{array} \times | b c) = ?$$

$$P(\begin{array}{c} y \\ b \end{array}) = \frac{1}{2}$$
E-step: Compute $P(a, f | e)$

For all alignments, compute $P(a, f | e)$

Remember:

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

- $t(x|b) = 1/2$
- $t(y|b) = 1/2$
- $t(x|c) = 1/2$
- $t(y|c) = 1/2$

$$P(\quad | b c) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(\quad | b c) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(\quad | b ) = \frac{1}{2}$$
E-step: Compute $P(a \mid e, f)$

$$P(a \mid e, f) = \frac{P(a, f \mid e)}{\sum_{a'} P(a', f \mid e)}$$

- $P(\_\_\_ \mid b, c) = \frac{1}{4}$
- $P(\_\_\_ \mid b, c) = \frac{1}{4}$
- $P(\_\_\_ \mid b, c) = \frac{1}{4}$
- $P(\_\_\_ \mid b ) = \frac{1}{2}$

$P(\_\_\_ \mid b, c, x, y) = \frac{\frac{1}{4}}{2/4} = \frac{1}{2}$

$P(\_\_\_ \mid b, c, x, y) = \frac{\frac{1}{4}}{2/4} = \frac{1}{2}$

$P(\_\_\_ \mid b, c, x, y) = \frac{\frac{1}{4}}{2/4} = \frac{1}{2}$

$P(\_\_\_ \mid b ) = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$
Collect Counts: Example

\[ P( | \parallel b c, x y) = \frac{1}{2} \]

Count instances where \( b \) and \( y \) are aligned:

\[ ct(y|b) = \frac{1}{2} + 1 \]

\[ P( \times | b c, x y) = \frac{1}{2} \]

\[ P( | | b, y ) = 1 \]
Collect Counts

\[
P(|| b, c, x, y) = \frac{1}{2}
\]

\[
P(\times | b, c, x, y) = \frac{1}{2}
\]

\[
P( | b, y ) = 1
\]

\[
tc(x|b) = 1/2
\]

\[
tc(y|b) = 1/2 + 1 = 3/2
\]

\[
tc(x|c) = 1/2
\]

\[
tc(y|c) = 1/2
\]
M-step: Normalize

\[
t_c(x|b) = \frac{1}{2}
\]

\[
t_c(y|b) = \frac{1}{2} + 1 = \frac{3}{2}
\]

\[
t_c(x|c) = \frac{1}{2}
\]

\[
t_c(y|c) = \frac{1}{2}
\]

\[
t(x|b) = \frac{\frac{1}{2}}{\frac{4}{2}} = \frac{1}{4}
\]

\[
t(y|b) = \frac{\frac{3}{2}}{\frac{4}{2}} = \frac{3}{4}
\]

\[
t(x|c) = \frac{\frac{1}{2}}{1} = \frac{1}{2}
\]

\[
t(y|c) = \frac{\frac{1}{2}}{1} = \frac{1}{2}
\]
E-step (again!): $P(a, f | e)$

Compute $P(a, f | e)$ using new parameters:

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

- $t(x|b) = \frac{1/2}{4/2} = 1/4$
- $t(y|b) = \frac{3/2}{4/2} = 3/4$
- $t(x|c) = \frac{1/2}{1} = 1/2$
- $t(y|c) = \frac{1/2}{1} = 1/2$

$$P(\begin{array}{c} x \ \ y \end{array} | b, c) = \frac{1}{4} \times \frac{3}{4} = \frac{3}{8}$$

$$P(\begin{array}{c} x \ \ y \end{array} | b, c) = ?$$

$$P(\begin{array}{c} y \end{array} | b) = 3/4$$
E-step (again!): $P(a,f|e)$

Compute $P(a,f|e)$ using new parameters:

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

$$t(x|b) = \frac{1/2}{4/2} = 1/4$$
$$t(y|b) = \frac{3/2}{4/2} = 3/4$$
$$t(x|c) = \frac{1/2}{1} = 1/2$$
$$t(y|c) = \frac{1/2}{1} = 1/2$$

$$P(\begin{array}{cc} x & y \\ \end{array} | b \ c) = \frac{1}{4} \times \frac{1}{2} = \frac{1}{8}$$

$$P(\begin{array}{cc} x & y \\ \end{array} | b \ c) = \frac{1}{2} \times \frac{3}{4} = 3/8$$

$$P(\begin{array}{c} y \\ \end{array} | b ) = 3/4$$
E-step (again): Compute $P(a | e, f)$

$$P(a | e, f) = \frac{P(a, f | e)}{\sum_{a'} P(a', f | e)}$$

$P(\begin{array}{l}x \\ y \end{array} | b, c) = \frac{1}{8}$

$P(\begin{array}{ll}x & y \\ x & y \end{array} | b, c) = \frac{3}{8}$

$P(\begin{array}{l}x \\ y \end{array} | b, c) = \frac{3}{4}$

$P(\begin{array}{c}y \end{array} | b) = \frac{3}{4}$
Collect Counts (again)

\[ P(\big|\big| b, c, x, y) = \frac{1}{4} \]

\[ P(\times | b, c, x, y) = \frac{3}{4} \]

\[ P(\big| b, y) = 1 \]

\[ tc(x|b) = \frac{1}{4} \]
\[ tc(y|b) = \frac{3}{4} + 1 = \frac{7}{4} \]
\[ tc(x|c) = \frac{3}{4} \]
\[ tc(y|c) = \frac{1}{4} \]
M-step (again): Normalize Counts

Collected counts:
\[ tc(x|b) = \frac{1}{4} \]
\[ tc(y|b) = \frac{3}{4} + 1 = \frac{7}{4} \]
\[ tc(x|c) = \frac{3}{4} \]
\[ tc(y|c) = \frac{1}{4} \]

Normalized counts:
\[ t(x|b) = \frac{1}{8} \]
\[ t(y|b) = \frac{7}{8} \]
\[ t(x|c) = \frac{3}{4} \]
\[ t(y|c) = \frac{1}{4} \]
What is happening to $t(f_j | e_i)$?

\[

t(x | b) = \frac{1/2}{4/2} = \frac{1}{4}
\]
\[

t(y | b) = \frac{3/2}{4/2} = \frac{3}{4}
\]
\[

t(x | c) = \frac{1/2}{1} = \frac{1}{2}
\]
\[

t(y | c) = \frac{1/2}{1} = \frac{1}{2}
\]
What does that mean?

Which alignments are more likely to be correct?

\[
\begin{align*}
  t(x|b) &= 1/8 \\
  t(y|b) &= 7/8 \\
  t(x|c) &= 3/4 \\
  t(y|c) &= 1/4
\end{align*}
\]
What does that mean?

Which alignments are more likely to be correct?

\[
\begin{align*}
  t(x|b) &= 1/8 \\
  t(y|b) &= 7/8 \\
  t(x|c) &= 3/4 \\
  t(y|c) &= 1/4
\end{align*}
\]
What would happen to $t(f_j|e_i)$... 

if we repeated these steps many times?
Many Iterations of EM:

\[
\begin{align*}
    t(x|b) &= 0.0001 \\
    t(y|b) &= 0.9999 \\
    t(x|c) &= 0.9999 \\
    t(y|c) &= 0.0001
\end{align*}
\]
Review of IBM Model 1 & EM

• Iteratively learned an alignment/translation model from sentence-aligned text (without “gold standard” alignments)
• Model can now be used for alignment and/or word-level translation
• We explored a simplified version of this; IBM Model 1 allows more types of alignments
Uses for Alignments

• Component of machine translation systems
• Produce a translation lexicon automatically
• Cross-lingual projection/extraction of information
• Supervision for training other models (for example, neural MT systems)
Alignment Examples
(English-> German)
Why is Model 1 insufficient?

• Why won’t this produce great translations?
Why is Model 1 insufficient?

• Why won’t this produce great translations?
  – Indifferent to order (language model may help?)
  – Translates one word at a time
  – Translates each word in isolation
  – ...

## Phrases

**English to Spanish**

<table>
<thead>
<tr>
<th>Maria no</th>
<th>daba una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Spanish to English**

<table>
<thead>
<tr>
<th>Maria no</th>
<th>daba una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Intersection**

<table>
<thead>
<tr>
<th>Maria no</th>
<th>daba una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phrases

Mary
did
not
slap
the
green
witch

Maria no daba una bofetada a la bruja verde
Decoding

What have we done so far?

• We can score alignments.
• We can score translations.

How do we *generate* translations?

• Decoding!
Decoding

Why can’t we just score all possible translations?

What do we do instead?
Decoding

• Many translation options for a word/phrase.
• Decoding is NP-complete (can verify solutions in polynomial time; can’t locate solutions efficiently).
• We use heuristics to limit the search space.
• See: statmt.org/book/slides/06-decoding.pdf

The role of the decoder is to:

• Choose “good” translation options
• Arrange them in a “good” order
How can this go wrong?

• Search doesn’t find the best translation
  – Need to fix the search

• The best translation found is not good
  – Need to fix the model
Decoding Options

In this example from Koehn (2017) slides, there might be >2700 phrase pairs for this (short!) sentence. ([http://mt-class.org/jhu/slides/lecture-decoding.pdf](http://mt-class.org/jhu/slides/lecture-decoding.pdf))
Er hat seit Monaten geplant.
Evaluating Machine Translation

Human evaluations:

• Test set (source, human reference translations, MT output)

• Humans judge the quality of MT output (in one of several possible ways)

Evaluating Machine Translation

Automatic evaluations:
• Test set (source, human reference translations, MT output)
• Aim to mimic (correlate with) human evaluations

Many metrics:
• TER (Translation Error/Edit Rate)
• HTER (Human-Targeted Translation Edit Rate)
• BLEU (Bilingual Evaluation Understudy)
• METEOR (Metric for Evaluation of Translation with Explicit Ordering)
In addition to this, there are more than 18 tailing heaps located right in the city, which has caused serious health impacts:

Zusätzlich zu diesen gibt es mehr als 18

Post-Editing
Questions?