Phrase-Based Translation
Machine Translation

\[ p(English|Chinese) \sim \]

\[ p(English) \times p(Chinese|English) \]

language model

translation model
Machine Translation

\[ p(\text{English}|\text{Chinese}) \sim p(\text{English}) \times p(\text{Chinese}|\text{English}) \]

- language model
- translation model
The IBM Models
The IBM Models

- Fertility probabilities.
The IBM Models

- Fertility probabilities.
- Word translation probabilities.
The IBM Models

- Fertility probabilities.
- Word translation probabilities.
- Distortion probabilities.
The IBM Models

• Fertility probabilities.
• Word translation probabilities.
• Distortion probabilities.
• Some problems:
  • Weak reordering model -- output is not fluent.
  • Many decisions -- many things can go wrong.
The IBM Models

- Fertility probabilities.
- Word translation probabilities.
- Distortion probabilities.

Some problems:
- Weak reordering model -- output is not fluent.
- Many decisions -- many things can go wrong.
The IBM Models

- Fertility probabilities.
- Word translation probabilities.
- Distortion probabilities.

Some problems:
- Weak reordering model -- output is not fluent.
- Many decisions -- many things can go wrong.
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
IBM Model 4

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.
IBM Model 4

Although north wind howls, but sky still very clear.

$p_f(1 | \text{虽然})$
Although north wind howls, but sky still very clear.
Although north wind howls, but sky still very clear.
Although north wind howls, **but** sky still **very** clear.

虽然北风呼啸，**但**天空**仍然**十分**清澈。
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
IBM Model 4

Although north wind howls, but sky still very clear.

although north wind howls, but sky still very clear.
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.
Although north wind howls, but sky still very clear.

However

IBM Model 4
IBM Model 4

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

$pt(\text{However}|\text{虽然})$
Although north wind howls, but sky still very clear.

However, north wind strong, the sky remained clear. Under the
Although north wind howls, but sky still very clear.

However north wind strong, the sky remained clear. under the
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

However north wind strong, the sky remained clear. under the
IBM Model 4

Although north wind howls, but sky still very clear.

Although north wind strong, the sky remained clear. under the

\[ p_d(0|\text{However}) \]
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

However, north wind strong, the sky remained clear. Under the
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

However, north wind strong, the sky remained clear. Under the

$pd(8|\text{north})$
Although north wind howls, but sky still very clear.

IBM Model 4

However north wind strong, the sky remained clear. under the
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

$p(English, alignment|Chinese) = \prod_{pf} \prod_{pt} \prod_{pd}$
However, the sky remained clear under the strong north wind.

\[ p(\text{English, alignment}|\text{Chinese}) = \Pi_{p_f} \Pi_{p_t} \Pi_{p_d} \]
IBM Model 4

虽然北风吹啸，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.

\[ p(English|Chinese) = \sum_{alignments} \prod_{pf} \prod_{pt} \prod_{pd} \]
p(θ|X)

Objective function

New guess

Current guess

Lower bound

Linear approx.

(from Minka '98)
... and, likelihood is *convex* for IBM Model 1:

But not IBM Models 3-5!
## Tradeoffs: Modeling v. Learning

<table>
<thead>
<tr>
<th></th>
<th>Lexical Translation</th>
<th>Local ordering dependency</th>
<th>Fertility</th>
<th>Convex</th>
<th>Tractable Exact Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IBM Model 1</strong></td>
<td>✔️</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>HMM</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>IBM Model 4</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>
# Tradeoffs: Modeling v. Learning

<table>
<thead>
<tr>
<th></th>
<th>Lexical Translation</th>
<th>Local ordering dependency</th>
<th>Fertility</th>
<th>Convex</th>
<th>Tractable Exact Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Model 1</td>
<td>✔</td>
<td>✘</td>
<td>✘</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>HMM</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
<td>✘</td>
<td>✔</td>
</tr>
<tr>
<td>IBM Model 4</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
<td>✘</td>
</tr>
</tbody>
</table>

**Lesson:** Trade exactness for expressivity
Although north wind howls, but sky still very clear.

However, north wind strong, the sky remained clear under the strong north wind.

What are some things this model doesn’t account for?
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

What are some things this model doesn't account for?
Phrase-based Models

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Phrase-based Models

Although north wind howls, but sky still very clear.
Phrase-based Models

Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

However
Although north wind howls, but sky still very clear.

Phrase-based Models

虽然 北风 呼啸，但 天空 依然 十分 清澈。

However the strong north wind, the sky remained clear under.
Although north wind howls, but sky still very clear.

However the strong north wind, the sky remained clear under.

Phrase-based Models
Although north wind howls, but sky still very clear.

However, the strong north wind, the sky remained clear under.

Phrase-based Models
Although north wind howls, but sky still very clear.

However, the strong north wind, the sky remained clear under.

Phrase-based Models
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[ p(\text{English, alignment} | \text{Chinese}) = p(\text{segmentation}) \cdot p(\text{translations}) \cdot p(\text{reorderings}) \]
Although north wind howls, but sky still very clear.

$p(\text{English, alignment}|\text{Chinese}) = p(\text{segmentation}) \cdot p(\text{translations}) \cdot p(\text{reorderings})$
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[ p(\text{English, alignment}|\text{Chinese}) = p(\text{segmentation}) \cdot p(\text{translations}) \cdot p(\text{reorderings}) \]

\text{distortion} = 6
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Phrase-based Models

Although north wind howls, but sky still very clear.

Despite the strong north wind, the sky remained clear under.

$p(\text{English, alignment}|\text{Chinese}) = p(\text{segmentation}) \cdot p(\text{translations}) \cdot p(\text{reorderings})$

distortion = 6
Although north wind howls, but sky still very clear.

\[
p(\text{English, alignment}|\text{Chinese}) = p(\text{segmentation}) \cdot p(\text{translations}) \cdot p(\text{reorderings})
\]

distortion = 6
Phrase-based Models
Phrase-based Models

- Segmentation probabilities.
Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.
Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.

Some problems:
- Weak reordering model -- output is not fluent.
- Many decisions -- many things can go wrong.
Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.

Some problems:
- Weak reordering model -- output not fluent.
- Many decisions -- many things can go wrong.
Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.

Some problems:

- Weak reordering model -- output is not fluent.
- Many decisions -- many things can go wrong.
Phrase-based Models

- Segmentation probabilities: fixed (uniform)
- **Phrase translation probabilities.**
- Distortion probabilities: fixed (decaying)
Learning $p(\text{Chinese} \mid \text{English})$

- Reminder: (nearly) every problem comes down to computing either:
  - Sums: MLE or EM (learning)
  - Maximum: most probable (decoding)
Recap: Expectation Maximization

- Arbitrarily select a set of parameters (say, uniform).
- Calculate *expected counts* of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate.
- Guaranteed that likelihood is monotonically nondecreasing.
Marginalize: sum all alignments containing the link

\[ p( \text{虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。} ) + \]

However, the sky remained clear under the strong north wind.
Divide by sum of all possible alignments

$p(\text{虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。}) +

$p(\text{然而, 晴空万里, 北风呼啸。}) +

$p(\text{虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。})$

However, the sky remained clear under the strong north wind.

However, the sky remained clear under the strong north wind.
Divide by sum of all possible alignments

$p( ) +$

However, the sky remained clear under the strong north wind.

$p( ) +$

However, the sky remained clear under the strong north wind.

$p( )$

However, the sky remained clear under the strong north wind.

We have to sum over exponentially many alignments!
EM for Model 1

probability of an alignment.

\[ p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j)p(f_i|e_j) \]
EM for Model 1

probability of an alignment.

\[ p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j)p(f_i|e_j) \]

observed uniform
EM for Model 1

probability of an alignment.

\[ p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j) p(f_i|e_j) \]

factors across words.

observed uniform
EM for Model 1

\[ p(\alpha_i = j | F, E) = \frac{p(\alpha_i = j, F | E)}{p(F, E)} = \]
EM for Model 1

\[ p(a_i = j \mid F, E) = \frac{p(a_i = j, F \mid E)}{p(F, E)} = p(F, E) \]

\[ \sum_{a \in A: \text{北} \leftrightarrow \text{north}} p(\text{north} \mid \text{北}) \cdot p(\text{rest of } a) \]
EM for Model 1

\[ p(a_i = j | F, E) = \frac{p(a_i = j, F | E)}{p(F, E)} = \sum_{a \in A: 北 \leftrightarrow north} p(\text{north} | 北) \cdot p(\text{rest of } a) \]

marginal probability of alignments containing link
EM for Model 1

marginal probability of alignments containing link

\[ p(north \mid 北) \sum_{a \in A: 北 \leftrightarrow north} p(rest \ of \ a) \]
EM for Model 1

marginal probability of alignments containing link

\[
\frac{p(north|\text{北}) \sum_{a \in A: \text{北} \leftrightarrow \text{north}} p(\text{rest of } a)}{\sum_{c \in \text{Chinese words}} p(north|c) \sum_{a \in A: \leftarrow \text{north}} p(\text{rest of } a)}
\]

marginal probability of all alignments
EM for Model 1

marginal probability of alignments containing link

\[ p(\text{north} \mid \text{北}) \sum_{\substack{a \in A: \text{北} \leftrightarrow \text{north}}} p(\text{rest of } a) \]

\[ \sum_{c \in \text{Chinese words}} p(\text{north} \mid c) \sum_{\substack{a \in A: c \leftrightarrow \text{north}}} p(\text{rest of } a) \]

marginal probability of all alignments
EM for Model 1

marginal probability of alignments containing link

\[
p(\text{north} | \text{北}) \sum_{a \in A: \text{北} \leftrightarrow \text{north}} p(\text{rest of } a) \]

\[
\sum_{c \in \text{Chinese words}} p(\text{north} | c) \sum_{a \in A: c \leftrightarrow \text{north}} p(\text{rest of } a) \]

identical!

marginal probability of all alignments
EM for Model 1

\[
p(north | 北) \frac{1}{\sum_{c \in \text{Chinese words}} p(north | c)}
\]
EM for Phrase-Based

- Model parameters: $p(E\ phrase \mid F\ phrase)$
- All we need to do is compute expectations:

$$p(a_{i,i'} = \langle j, j' \rangle \mid F, E) = \frac{p(a_{i,i'} = \langle j, j' \rangle, F \mid E)}{p(F, E)}$$
Model parameters: $p(E\ \text{phrase} \mid F\ \text{phrase})$

All we need to do is compute expectations:

$$p(a_{i,i'} = \langle j, j' \rangle \mid F, E) = \frac{p(a_{i,i'} = \langle j, j' \rangle, F \mid E)}{p(F, E)}$$

$p(F,E)$ sums over all possible phrase alignments
EM for Phrase-Based

- Model parameters: $p(E \text{ phrase} \mid F \text{ phrase})$
- All we need to do is compute expectations:

$$p(a_{i,i'} = \langle j, j' \rangle \mid F, E) = \frac{p(a_{i,i'} = \langle j, j' \rangle, F \mid E)}{p(F, E)}$$

$p(F,E)$ sums over all possible phrase alignments
...which are one-to-one by definition.
EM for Phrase-Based

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[ p(a_{i,i'} = \langle j, j' \rangle | F, E) = \frac{p(a_{i,i'} = \langle j, j' \rangle, F | E)}{p(F, E)} \]
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

\[ p(a_{i,i'} = \langle j, j' \rangle | F, E) = \frac{p(a_{i,i'} = \langle j, j' \rangle, F | E)}{p(F, E)} \]

Can we compute this quantity?
EM for Phrase-Based

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.

虽然北风呼啸，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.

\[
p(a_{i,i'} = \langle j, j' \rangle | F, E) = \frac{p(a_{i,i'} = \langle j, j' \rangle, F|E)}{p(F, E)}
\]

Can we compute this quantity?

How many 1-to-1 alignments are there of the remaining 8 Chinese and 8 English words?
Recap: Expectation Maximization

- Arbitrarily select a set of parameters (say, uniform).
- Calculate *expected counts* of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate.

- Guaranteed that likelihood is monotonically nondecreasing.
Recap: Expectation Maximization

- Arbitrarily select a set of parameters (say, uniform).
- Calculate *expected counts* of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate.
- Guaranteed that likelihood is monotonically nondecreasing.

Computing expectations from a phrase-based model, given a sentence pair, is \#P-Complete (by reduction to counting perfect matchings; DeNero & Klein, 2008)
Now What?

- Option #1: approximate expectations
  - Restrict computation to some tractable subset of the alignment space (arbitrarily biased).
  - Markov chain Monte Carlo (very slow).
Now What?

- Change the problem definition

- We already know how to learn word-to-word translation models efficiently.

- Idea: learn word-to-word alignments, extract most probable alignment, then treat it as observed.

- Learn phrase translations consistent with word alignments.

- Decouples alignment from model learning -- is this a good thing?
I open the box

watashi
wa
hako
wo
akemasu
I open the box

<table>
<thead>
<tr>
<th>watashi</th>
<th>open</th>
<th>the</th>
<th>box</th>
</tr>
</thead>
<tbody>
<tr>
<td>wa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hako</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>akemasu</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

akemasu / open
Watashi wa / I

<table>
<thead>
<tr>
<th>watashi</th>
<th>open</th>
<th></th>
<th>the</th>
<th>box</th>
</tr>
</thead>
<tbody>
<tr>
<td>wa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hako</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>akemasu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I open the box
I open the box

<table>
<thead>
<tr>
<th>watashi</th>
<th>wa</th>
<th>hako</th>
<th>wo</th>
<th>akemasu</th>
</tr>
</thead>
<tbody>
<tr>
<td>watashi</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

watashi / I
<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>open</th>
<th>the</th>
<th>box</th>
</tr>
</thead>
<tbody>
<tr>
<td>watashi</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hako</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>akemasu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**I open the box**

**watashi wa / I**
<table>
<thead>
<tr>
<th>watashi</th>
<th>wa</th>
<th>hako</th>
<th>wo</th>
<th>akemasu</th>
<th>I</th>
<th>open</th>
<th>the</th>
<th>box</th>
</tr>
</thead>
<tbody>
<tr>
<td>watashi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>open</td>
<td>the</td>
<td>box</td>
</tr>
<tr>
<td>wa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hako</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>akemasu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

hako wo / box
I open the box

<table>
<thead>
<tr>
<th>watashi</th>
<th>open</th>
<th>the</th>
<th>box</th>
</tr>
</thead>
<tbody>
<tr>
<td>wa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hako</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wo</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>akemasu</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

hako wo / the box
I open the box

すると、私が箱を開けることになります。

ただし、このフレーズは文法上、正しいです。
<table>
<thead>
<tr>
<th>watashi</th>
<th>wa</th>
<th>hako</th>
<th>wo</th>
<th>akemasu</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>open</td>
<td>the</td>
<td>box</td>
<td></td>
</tr>
</tbody>
</table>

hako wo / open the box
### Phrase Extraction

<table>
<thead>
<tr>
<th>watashi</th>
<th>wa</th>
<th>hako</th>
<th>wo</th>
<th>akemasu</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>open</td>
<td>the</td>
<td>box</td>
<td></td>
</tr>
</tbody>
</table>

- **I open the box**
- **hako wo akemasu** / open the box
Phrasal Translation Estimation
Phrasal Translation Estimation

• Approximation #1 (EM over restricted space)

• Align with a word-based model.

• Compute expectations only over alignments consistent with the alignment grid.
Phrasal Translation Estimation

- Approximation #1 (EM over restricted space)
  - Align with a word-based model.
  - Compute expectations only over alignments consistent with the alignment grid.

- Approximation #2 (heuristic estimation)
  - View phrase pairs as observed, irrespective of context or overlap.
  - By far the most common approach.
Phrasal Translation Estimation

- Approximation #1 (EM over restricted space)
  - Align with a word-based model.
  - Compute expectations only over alignments consistent with the alignment grid.

- Approximation #2 (heuristic estimation)
  - View phrase pairs as observed, irrespective of context or overlap.
  - By far the most common approach.
  - Many other possible approximations!
• Some key ingredients in Moses/Google Translate:
• Some key ingredients in Moses/Google Translate:
  • Phrase-based translation models
Some key ingredients in Moses/Google Translate:

- Phrase-based translation models
- ... Learned heuristically from word alignments