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 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.

However, the sky remained clear under the strong north wind.
## 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 Inference</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

## Lesson:
- Trade exactness for expressivity

<table>
<thead>
<tr>
<th></th>
<th>Lexical Translation</th>
<th>Local ordering dependency</th>
<th>Fertility</th>
<th>Convex</th>
<th>Tractable Exact Inference</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>
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.

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

Although north wind howls, but sky still very clear.

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

although 北风 呼啸 但是 天空 依然 十分 清澈
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 strong north wind, the sky remained clear under.

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

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.

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.

distortion = 6

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

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

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

Although north wind howls, but sky still very clear.

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

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

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

$p(\text{虽然北风呼啸，但天空依然十分清澈。}) +$

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

$p(\text{虽然北风呼啸，但天空依然十分清澈。}) +$

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

$p(\quad ) +$

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

$p(\quad ) +$

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

$p(\quad ) +$

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

\[
p( \text{虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。} ) + \text{However , the sky remained clear under the strong north wind .} \]

\[
p( \text{虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。} ) + \text{However , the sky remained clear under the strong north wind .} \]

\[
p( \text{虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。} ) + \text{However , the sky remained clear under the strong north wind .} \]

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.

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

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

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

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

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

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

marginal probability of alignments containing link
EM for Model 1

marginal probability of alignments containing link

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

marginal probability of alignments containing link

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

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

marginal probability of all alignments
EM for Model 1

marginal probability of alignments containing link

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

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

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

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

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

\[
p(a_i = j \mid F, E) = \frac{p(a_{i,i'} = \langle j, j' \rangle, F \mid E)}{p(F, E)}
\]
EM for Phrase-Based

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

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

\( p(F, E) \) sums over all possible phrase alignments
Model parameters: $p(E\text{ phrase} \mid F\text{ phrase})$

All we need to do is compute expectations:

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

$p(F,E)$ 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.

Although north wind howls, but sky still very clear.

EM for Phrase-Based

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

However, the sky remained clear under the strong north wind.
EM for Phrase-Based

Although north wind howls, but sky still very clear.

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

Can we compute this quantity?

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

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?
<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>

Phrase Extraction

I open the box

Watashi wa hako wo akemasu
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>I</td>
<td>open</td>
<td>the</td>
<td>box</td>
<td></td>
</tr>
</tbody>
</table>

akemasu / open
I open the box

watashi wa hako wo akemasu
<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>

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>wa</td>
<td>hako</td>
<td>wo</td>
<td>akemasu</td>
</tr>
</tbody>
</table>

I open the box

watashi wa / I
I open the box

Watashi wa hako wo / box

Hako wo / box

Areamasu

Hako wo / 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
I open the box

watashi
wa
hako
wo
akemasu

hako wo / open the box
I open the box

hako wo akemasu / open the box
Phrasal Translation Estimation
Phrasal Translation Estimation

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

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

• Option #2 (Non-global estimation)
  • View phrase pairs as observed, irrespective of context or overlap.
北风呼啸。
北 风 呼啸。

segmentations
substitutions
permutations
segmentations $O(2^n)$ substitutions permutations
北风呼啸。

segmentations $O(2^n)$
substitutions $O(5^n)$
permutations
北风呼啸。

segmentations $O(2^n)$
substitutions $O(5^n)$
permutations $O(n!)$
Key Idea
Key Idea
Key Idea
Key Idea
Key Idea

Dynamic Programming
虽然北风呼啸，但天空依然十分清澈。
虽然北风呼啸，但天空依然十分清澈。
Although crystal clear

然北风呼啸，但天空依然十分清澈。

However

然北风呼啸，但天空依然十分清澈。
Although crystal clear, the north wind was howling, yet the sky remained intensely clear.
wind screamed

虽然北风呼啸，但天空依然十分清澈。

wind shrieked

虽然北风呼啸，但天空依然十分清澈。

north wind

虽然北风呼啸，但天空依然十分清澈。
wind screamed

風呼嘯，但天空依然十分清澈。

wind shrieked

風呼嘯，但天空依然十分清澈。
	north wind

風呼嘯，但天空依然十分清澈。
shrieked,

the sky

, yet
the sky shrieked, yet

although the wind howled, but the sky remained very clear.
虽然北风呼啸，但天空依然十分清澈。
虽然北风呼啸，但天空依然十分清澈。
still quite blue.
clear.

虽然北风呼啸，但天空依然十分清澈。
still quite blue.
clear.

虽然北风呼啸，但天空依然十分清澈。
Although the northern wind shrieked across the sky, but was still very clear.

虽然北风呼啸，但天空依然十分清澈。
Wait a second.
Wait a second.

$O(5n^22^n)$ is still far too much work.
Wait a second.

$O(5n^2 2^n)$ is still far too much work.

Can we do better?
Wait a second.

$O(5n^22^n)$ is still far too much work.

Can we do better?

NO! Knight (1999) shows that this is NP-Complete, by reduction to Hamiltonian Circuit.
Approximation: Pruning
Approximation: Pruning

Idea: prune states by accumulated path length
Approximation: Pruning
Approximation: Pruning

reality: longer paths have lower probability!
Approximation: Pruning
Approximation: Pruning

Solution: Group states by number of covered words.
Approximation: Pruning

Solution: Group states by number of covered words.
Approximation: Pruning

Solution: Group states by number of covered words.
Approximation: Pruning

“Stack” decoding: a linear-time approximation
the sky

虽然北风呼啸，但天空依然十分清澈。
number of vertices: $O(2^n)$

the sky

虽然北风呼啸，但天空依然十分清澈。
number of vertices: \( O(2^n) \)

deleting

the sky

虽然北风呼啸，但天空依然十分清澈。

\[
d = 4
\]

window
number of vertices: \( O(2^n) \)

the sky

虽然北风呼啸，但天空依然十分清澈。

outside window

to left: covered

d = 4

window

to right: uncovered
number of vertices: $O(n2^d)$

the sky

虽然北风呼啸，但天空依然十分清澈。

outside window

to left: covered

d = 4

window

to right: uncovered
Some (not all) key ingredients in Google Translate:
Some (not all) key ingredients in Google Translate:

- Phrase-based translation models
Some (not all) key ingredients in Google Translate:

- Phrase-based translation models
- ... Learned heuristically from word alignments
Some (not all) key ingredients in Google Translate:

- Phrase-based translation models
- ... Learned heuristically from word alignments
- ... Coupled with a huge language model
Some (not all) key ingredients in Google Translate:

- Phrase-based translation models
- ... Learned heuristically from word alignments
- ... Coupled with a huge language model
- ... And very tight pruning heuristics