Introduction to Machine Translation (Part 1)

Adam Lopez
University of Edinburgh

With content and dumb jokes borrowed from / inspired by: Chris Callison-Burch, Philipp Koehn, Kevin Knight, and Martin Kay
Introduction to Machine Translation (Part 1)

Adam Lopez
University of Edinburgh

With content and dumb jokes borrowed from / inspired by: Chris Callison-Burch, Philipp Koehn, Kevin Knight, and Martin Kay

Errors and bad timing are all mine, though.
虽然北风呼啸，但天空依然十分清澈。
雖然北風呼嘯，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.
The Tower of Babel

Pieter Brueghel the Elder (1563)
Georges Artsrouni’s “mechanical brain”, patented 1933 (France)
ENIAC (1946)
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 (1949)
Because we want to provide everyone with access to all the world's information, including information written in every language, one of the exciting projects at Google Research is machine translation...

Now you can see the results for yourself. We recently launched an online version of our system for Arabic-English and English-Arabic. Try it out!
Although the wind was strong, the sky remained clear.

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

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.

虽然北风呼啸，但天空依然十分清澈。
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.
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.
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.
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.
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.

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

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

Despite the strong northerly winds, the sky remains very clear.
Although north wind howls, but sky still very clear.

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

Despite the strong northerly winds, the sky remains very clear.

The sky was still crystal clear, though the north wind was howling.

Although a north wind was howling, the sky remained clear and blue.
Although north wind howls, but sky still very clear.

Although a north wind was howling, the sky remained clear and blue.
Statistical Machine Translation
Statistical Machine Translation

Develop a *model* of translation that can be *learned* from *data* and used to *predict* the correct English translation of new Chinese sentences.
Statistical Machine Translation

Develop a \textit{model} of translation that can be \textit{learned} from \textit{data} and used to \textit{predict} the correct English translation of new Chinese sentences.

this talk
Statistical Machine Translation

Develop a model of translation that can be learned from data and used to predict the correct English translation of new Chinese sentences.

John’s talk
Statistical Machine Translation

Develop a *model* of translation that can be *learned* from *data* and used to *predict* the correct English translation of new Chinese sentences.
中文简体
Modeling
虽然北风呼啸，但天空依然十分清澈。

\[ p(English | Chinese) \]
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

p(English|Chinese)
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.
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.
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.

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.
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.

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.
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.

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

$\mathcal{p}_t(However \mid 虽然)$
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.

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

However north wind strong, the sky remained clear.

$pd(0|\text{However})$
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.

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.

\[ p(\text{English, alignment}|\text{Chinese}) = \prod_{p_f} \prod_{p_t} \prod_{p_d} \]
However, the sky remained clear under the strong north wind.

\[ p(\text{English, alignment}|\text{Chinese}) = \prod_{p_f} \prod_{p_t} \prod_{p_d} \]
However, the sky remained clear under the strong north wind.

\[ p(\text{English}|\text{Chinese}) = \sum_{\text{alignments}} \Pi_{pf} \Pi_{pt} \Pi_{pd} \]
Digression: Learning
Digression: Learning

\[ pd(0|\text{However}) \]

\[ pt(\text{However}|\text{虽然}) \]

\[ pf(1|\text{虽然}) \]
Digression: Learning

\[ p_d(0|\text{However}) = 0.67 \]

\[ p_t(\text{However}|\text{虽然}) = 0.85 \]

\[ p_f(1|\text{虽然}) = 0.65 \]
<table>
<thead>
<tr>
<th>Induction</th>
<th>Parameter Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_d(0</td>
<td>\text{However}) )</td>
</tr>
<tr>
<td>( p_t(\text{However}</td>
<td>\text{虽然}) )</td>
</tr>
<tr>
<td>( p_f(1</td>
<td>\text{虽然}) )</td>
</tr>
</tbody>
</table>

...
The IBM Models
The IBM Models

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

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

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

Some problems:

- Complexity -- exponential in sentence length.
- 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.
Although *north wind howls*, *but sky still very clear*.  

However, the *north wind strong*, 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.
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.

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

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

雖然北方呼嘯，但天空依然十分清澈。

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

However, 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.

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

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

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

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

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

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

Some problems:

- Complexity -- exponential in sentence length.
- Weak reordering model -- output is not fluent.
- Many decisions -- many things can go wrong.
\[ p(\text{English, alignment}|\text{Chinese}) = \prod_{\text{decision}} p(\text{decision}) \]
\[ p(English|Chinese) = \sum \prod_{\text{alignments decision}} p(\text{decision}) \]
\hat{p}(English|Chinese) = \prod_{\text{decision}} p(\text{decision})
argmax \hat{p}(English|Chinese) = \prod_{\text{decision}} p(\text{decision})
\[ p(English|Chinese) \]
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 (1949)
...But the probability that an event has happened is the same as the probability I have to guess right if I guess it has happened. Wherefore the following proposition is evident: If there be two subsequent events, the probability of the 2d \( \frac{b}{N} \) and the probability both together \( \frac{P}{N} \), and it being 1st discovered that the 2d event has also happened, the probability I am right is \( \frac{P}{b} \).

Thomas Bayes
Bayes’ Theorem

\[ p(English|Chinese) = \frac{p(English) \times p(Chinese|English)}{p(Chinese)} \]
Bayes' Theorem

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

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

translation model (likelihood)
Bayes’ Theorem

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

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

- **language model (prior)**
- **translation model (likelihood)**
Bayes’ Theorem

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

- **Language model (prior)**
- **Translation model (likelihood)**
- **Normalization term (ensures we’re working with valid probabilities).**
Bayes’ Theorem

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

\[ p(English) \times p(Chinese|English) \]
Bayes’ Theorem

\[ p(English) \]
Bayes’ Theorem

\[ p(\text{English}) \]

However, the sky remained clear under the strong north wind.
Bayes’ Theorem

\[ p(\text{English}) \]

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

\[ p(\text{Chinese}|\text{English}) \]
However, the sky remained clear under the strong north wind.

\[ p(\text{Chinese}|\text{English}) \]
Bayes’ Theorem

\[ p(English) \]
Bayes’ Theorem

\[ p(\text{English}) \]

configuration
Bayes’ Theorem

\[ p(\text{English}) \]

configuration

\[ p(\text{image}|\text{English}) \]
Bayes’ Theorem

\[ p(\text{English}) \]

configuration

\[ p(\text{image}|\text{English}) \]

configuration
Bayes’ Theorem

\[ p(DNA) \]
Bayes’ Theorem

\[ p(DNA) \]
Bayes’ Theorem

\[ p(DNA) \rightarrow p(mutation|DNA) \]
Bayes’ Theorem

\[ p(DNA) \quad \rightarrow \quad p(mutation \mid DNA) \]
Bayes’ Theorem

\[ p(English|Chinese) = \frac{p(English) \times p(Chinese|English)}{p(Chinese)} \]

- language model (prior)
- translation model (likelihood)
- normalization term (ensures we’re working with valid probabilities).
Language Models

However

\[ p(\text{However}|\text{START}) \]
Language Models

$p(, | However)$
Language Models

However, the

\[ p(\text{the}|,) \]
Language Models

However, the sky

\[ p(\text{sky}|\text{the}) \]
Language Models

\[ p(English) = \prod_{i=1}^{\text{length}(English)} p(\text{word}_i | \text{word}_{i-1}) \]
\( p(\text{Chinese}|\text{English}) \)
\[ p(\text{Chinese} | \text{English}) \times p(\text{English}) \]

\[ \text{English} \]
\[ p(\text{Chinese} | \text{English}) \times p(\text{English}) \sim p(\text{English} | \text{Chinese}) \]
\( p(Chinese|English) \)^1
\[ \times \ p(English) \]^1
\[ \sim p(English|Chinese) \]
\[ p(\text{Chinese}|\text{English})^2 \times p(\text{English})^1 \sim p(\text{English}|\text{Chinese}) \]
$p(\text{Chinese} | \text{English})^{1/2}$

$\times p(\text{English})^1$

$\sim p(\text{English} | \text{Chinese})$

$\text{English}$
\[ p(\text{Chinese}|\text{English})^0 \]

\[ \times p(\text{English})^1 \]

\[ \sim p(\text{English}|\text{Chinese}) \]
\[ 0 \cdot \log p(Chinese|English) \]

\[ +1 \cdot \log p(English) \]

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

\[ \text{English} \]
\[
p(English|Chinese) = \\
\frac{1}{Z} \left[ \lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English) \right]
\]
\[ p(English|Chinese) = \frac{1}{\mathcal{Z}} \left[ \sum_i \lambda_i h_i(Chinese, English) \right] \]
\[
\arg\max_{English} \ p(English|Chinese) = \frac{1}{Z} \left[ \sum_i \lambda_i h_i(Chinese, English) \right]
\]
• Optimization
Prediction
\[ p(\text{English}|\text{Chinese}) = \]
\[ \frac{1}{Z} \left[ \frac{1}{Z} \left[ \lambda_1 \log \frac{1}{Z} \left[ \sum_i p(\text{Chinese}|\text{English}, \text{English}) \right] \right] \right] \]
\[
\text{argmax } p(\text{English}|\text{Chinese}) = \\
\text{English } \frac{1}{Z} [\lambda_1 \log \frac{1}{Z} \left[ \sum_i \log p(C_{\text{Chinese}}|\text{English}, \lambda \log p(\text{English})) \right]]
\]
• Optimization
北风呼啸。
北风呼啸。

segmentations
substitutions
permutations
北风呼啸。

segmentations $O(2^n)$
substitutions
permutations
segmentations $O(2^n)$
substitutions $O(5^n)$
permutations
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

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

虽然北风呼啸，但天空依然十分清澈。
wind screamed

雖然北風呼嘯，但天空依然十分清澈。

wind shrieked

雖然北風呼嘯，但天空依然十分清澈。

north wind

雖然北風呼嘯，但天空依然十分清澈。
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。
shrieked, yet the sky

Although the wind screamed, yet the sky remained very clear.

虽然北风呼啸，但天空依然十分清澈。
shrieked,

the sky

, yet
sky,

虽然北风呼啸，但天空依然十分清澈。
sky, 

虽然北风呼啸，但天空依然十分清澈。
still quite blue.

clear.

雖然北風呼嘯，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。
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: $O(2^n)$

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

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

outside window to left: covered

outside window to right: uncovered

d = 4
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.

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.

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.

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

$X_1 X_2$
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 strong north wind, the sky remained clear under.

X1 X2

X2 X1
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.

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.

然而，强劲的北风下，天空依然清澈。
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.

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

\[ X \rightarrow X_1X_2/X_1X_2 \]
\[ X \rightarrow X_1X_2/X_2X_1 \]
\[ X \rightarrow \text{Chinese phrase}/\text{English phrase} \]
Although north wind howls, but sky still very clear.

However, the strong north wind, the sky remained clear under.
je pense que, indépendamment de notre parti, nous trouvons cela unacceptable, que, I believe, we all find unacceptable, regardless of political party.
je pense que, indépendamment de notre parti, nous trouvons cela unacceptable.
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.

虽然北风呼啸，但天空依然十分清澈。

the strong north wind

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

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

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

$X \rightarrow \text{虽然} \ X_1, \text{但} \ X_2/H\text{owever} \ , \ X_2X_1$
Synchronous CFG
Synchronous CFG

- Phrase translation probabilities.
- Hierarchical rewriting probabilities.
- Some solutions?
  - Complexity -- polynomial in sentence length.
  - Can it help with fluency?
These people include astronauts coming from France.
Evaluation
More has been written about machine translation evaluation than about machine translation itself.

Yorick Wilks
Why evaluate?
- Rank systems
- Evaluate incremental changes
- Assess new ideas objectively
US willing to negotiate with North Korea but not to pay more compensation.
US willing to negotiate with North Korea but not to pay more compensation.

The United States is willing to hold talks with North Korea but refused to pay remuneration.
Launch of “Endeavour” delayed by robotic arm problems.
Launch of “Endeavour” delayed by robotic arm problems.

“Progress” postponed because of mechanical hand into the sky.
Chinese people in the traditional Spring Festival is approaching, the CPC Central Committee this afternoon in Zhongnanhai on the 22nd non-Party personages to convene a forum in Spring Festival, invited the central committees of democratic parties, the leadership of the National Federation of Industry and Commerce and personages without party affiliation on behalf of comrades gathered together State yes, talked in length about the friendship, to greet the Chinese New Year. CPC Central Committee General Secretary and State President and Central Military Commission Chairman Hu Jintao on behalf of the CPC Central Committee, the State Council, to the central committees of democratic parties, leaders of the National Federation of Industry and Commerce and personages without party affiliation, to members of the united front, to extend my New Year's blessing.
Although the northern wind shrieked across the sky, it was still very clear.
Although the northern wind shrieked across the sky, it was still very clear.

However, the sky remained clear under the strong north wind.
Although the northern wind shrieked across the sky, it was still very clear.

However, the sky remained clear under the strong north wind.
Although the northern wind shrieked across the sky, it was still very clear.

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

Although a north wind was howling, the sky remained clear and blue.

The sky was still crystal clear, though the north wind was howling.

Despite the strong northerly winds, the sky remains very clear.
Although the northern wind shrieked across the sky, it was still very clear.

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

Although a north wind was howling, the sky remained clear and blue.

The sky was still crystal clear, though the north wind was howling.

Despite the strong northerly winds, the sky remains very clear.
Although the northern wind shrieked across the sky, it was still very clear.

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

Although a north wind was howling, the sky remained clear and blue.

The sky was still crystal clear, though the north wind was howling.

Despite the strong northerly winds, the sky remains very clear.
However, the sky remained clear under the strong north wind.

Although a north wind was howling, the sky remained clear and blue.
The sky was still crystal clear, though the north wind was howling.

Despite the strong northerly winds, the sky remains very clear.
However, the sky remained clear under the strong north wind.

Although a north wind was howling, the sky remained clear and blue. The sky was still crystal clear, though the north wind was howling.

Despite the strong northerly winds, the sky remains very clear.
However, the sky remained clear under the strong north wind.

Although a north wind was howling, the sky remained clear and blue.

The sky was still crystal clear, though the north wind was howling.

Despite the strong northerly winds, the sky remains very clear.
However, the sky remained clear under the strong north wind.

Although a north wind was howling, the sky remained clear and blue.

The sky was still crystal clear, though the north wind was howling.

Despite the strong northerly winds, the sky remains very clear.
BiLingual Evaluation Understudy
BLEU(MT output)
$BLEU(\text{argmax score}(\text{English}|\text{Chinese}))$
\[ \sum_{\text{Chinese} \in \text{Test}} \text{BLEU}(\text{argmax score}(\text{English} | \text{Chinese})) \]
• Optimization
Next Up: Learning

http://homepages.inf.ed.ac.uk/alopez

• Thanks: Necip Fazil Ayan, Phil Blunsom, Ondrej Bojar, Chris Callison-Burch, David Chiang, Trevor Cohn, Bonnie Dorr, Chris Dyer, Martin Kay, Philipp Koehn, Nitin Madnani, Christof Monz, Miles Osborne, Philip Resnik, and Michael Subotin