Feature-Based Models
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
- Today: more flexible parameterizations.
Bayes’ Rule

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

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

language model  translation model
$p(\text{Chinese} | \text{English})$

*English*
\[ p(\text{Chinese} \mid \text{English}) \times p(\text{English}) \]
\[ p(\text{Chinese} \mid \text{English}) \]

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

\[ \sim p(\text{English} \mid \text{Chinese}) \]
\[ p(\text{Chinese}|\text{English})^1 \]

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

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

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

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

English
\[ p(\text{Chinese}|\text{English})^{1/2} \times p(\text{English}) \sim p(\text{English}|\text{Chinese}) \]
\[ p(Chinese | English)^0 \]

\[ \times p(English)^1 \]

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

\[ + 1 \cdot \log p(\text{English}) \]

\[ \sim \log p(\text{English} | \text{Chinese}) \]
\[ \log(x) \text{ is monotonic for positive } x \]
\[ \text{(i.e. } \log(x) > \log(y) \text{ iff } x > y\text{)} \]

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

\[ +1 \cdot \log p(\text{English}) \]

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

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

+ 1 \cdot \log p(\text{English})

= \text{score}(\text{English}|\text{Chinese})

\text{English}
score(English|Chinese) =

\[ \lambda_1 \log p(\text{Chinese}|\text{English}) + \lambda_2 \log p(\text{English}) \]
score(\textit{English} | \textit{Chinese}) = \\
\exp(\lambda_1 \log p(\textit{Chinese} | \textit{English}) + \lambda_2 \log p(\textit{English}))
\[ p(\text{English}|\text{Chinese}) = \frac{\exp(\lambda_1 \log p(\text{Chinese}|\text{English}) + \lambda_2 \log p(\text{English}))}{\sum_{\text{English}} \exp(\lambda_1 \log p(\text{Chinese}|\text{English}) + \lambda_2 \log p(\text{English}))} \]
\[ p(English|Chinese) = \frac{\exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))}{\sum_{English} \exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))} \]

log-linear model
maximum entropy model
conditional model
undirected model
\[ p(\text{English}|\text{Chinese}) = \]

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

Note: Original model is a special case of this model!

log-linear model
maximum entropy model
conditional model
undirected model
\[ p(English|Chinese) = \frac{\exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))}{\sum_{English} \exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))} \]
\[ p(English \mid Chinese) = \frac{\exp \left\{ \sum_{k} \lambda_k h_k(English, Chinese) \right\}}{\sum_{English'} \exp \left\{ \sum_{k} \lambda_k h_k(English', Chinese) \right\}} \]

log-linear model
maximum entropy model
conditional model
undirected model
\[ p(English|Chinese) = \frac{1}{Z} \exp \left\{ \sum_k \lambda_k h_k(English, Chinese) \right\} \]

log-linear model
maximum entropy model
conditional model
undirected model
\[ p(\text{English}|\text{Chinese}) = \frac{1}{Z} \exp \left\{ \sum_k \lambda_k h_k(\text{English}, \text{Chinese}) \right\} \]

\( Z \) is the normalization term or \textit{partition function}

log-linear model
maximum entropy model
conditional model
undirected model
\[ p(English|Chinese) = \frac{1}{Z} \exp \left\{ \sum_k \lambda_k h_k(English, Chinese) \right\} \]

\( Z \) is the normalization term or partition function

The functions \( h_k \) are features or feature functions

They are deterministic (fixed) functions of the input/output pair.

The parameters of the model are the \( \lambda_k \) terms.
What’s a Feature?
What’s a Feature?

A feature can be any function in the form:

\[ h_k : \text{English} \times \text{Chinese} \rightarrow \mathbb{R}_+ \]
What’s a Feature?

A feature can be any function in the form:

$$h_k : English \times Chinese \rightarrow \mathbb{R}_+$$

- Language model: $p(English)$
What’s a Feature?

A feature can be any function in the form:

\[ h_k : \text{English} \times \text{Chinese} \rightarrow \mathbb{R}_+ \]

- Language model: \( p(\text{English}) \)
- Translation model: \( p(\text{Chinese} | \text{English}) \)
What’s a Feature?

A feature can be any function in the form:

\[ h_k : \text{English} \times \text{Chinese} \rightarrow \mathbb{R}_+ \]

- Language model: \( p(\text{English}) \)
- Translation model: \( p(\text{Chinese} \mid \text{English}) \)
- Reverse translation model: \( p(\text{English} \mid \text{Chinese}) \)
What's a Feature?

A feature can be any function in the form:

\[ h_k : English \times Chinese \rightarrow \mathbb{R}_+ \]

- Language model: \( p(English) \)
- Translation model: \( p(Chinese \mid English) \)
- Reverse translation model: \( p(English \mid Chinese) \)
- The number of words in the English sentence.
What’s a Feature?

A feature can be *any* function in the form:

$$h_k : \text{English} \times \text{Chinese} \rightarrow \mathbb{R}_+$$

- Language model: \( p(\text{English}) \)
- Translation model: \( p(\text{Chinese} \mid \text{English}) \)
- Reverse translation model: \( p(\text{English} \mid \text{Chinese}) \)
- The number of words in the English sentence.
- The number of verbs in the English sentence.
What’s a Feature?

A feature can be any function in the form:

\[ h_k : \text{English} \times \text{Chinese} \rightarrow \mathbb{R}_+ \]

- Language model: \( p(\text{English}) \)
- Translation model: \( p(\text{Chinese} \mid \text{English}) \)
- Reverse translation model: \( p(\text{English} \mid \text{Chinese}) \)
- The number of words in the English sentence.
- The number of verbs in the English sentence.
- 1 if the English sentence has a verb, 0 otherwise.
What’s a Feature?

A feature can be any function in the form:

\[ h_k : \text{English} \times \text{Chinese} \rightarrow \mathbb{R}_+ \]
What’s a Feature?

A feature can be *any* function in the form:

\[ h_k : English \times Chinese \rightarrow \mathbb{R}_+ \]

- A word-based translation model: \( p(\text{Chinese} \mid \text{English}) \)
What’s a Feature?

A feature can be any function in the form:

\[ h_k : English \times Chinese \rightarrow \mathbb{R}_+ \]

- A word-based translation model: \( p(\text{Chinese} \mid \text{English}) \)
- Agreement features in the English sentence.
What’s a Feature?

A feature can be any function in the form:

\[ h_k : English \times Chinese \rightarrow \mathbb{R}_+ \]

- A word-based translation model: \( p(\text{Chinese} \mid \text{English}) \)
- Agreement features in the English sentence.
- Features over part-of-speech sequences in the English sentence.
What’s a Feature?

A feature can be any function in the form:

\[ h_k : \text{English} \times \text{Chinese} \rightarrow \mathbb{R}_+ \]

• A word-based translation model: \( p(\text{Chinese} \mid \text{English}) \)

• Agreement features in the English sentence.

• Features over part-of-speech sequences in the English sentence.

• How many times the sentence pair includes the English word *north* and Chinese word 北.
What’s a Feature?

A feature can be any function in the form:

\[ h_k : English \times Chinese \rightarrow \mathbb{R}_+ \]

- A word-based translation model: \( p(Chinese \mid English) \)
- Agreement features in the English sentence.
- Features over part-of-speech sequences in the English sentence.
- How many times the sentence pair includes the English word \textit{north} and Chinese word 北.
- Do words \textit{north} and 北 appear in a dictionary?
Learning

$$\arg\max_{\theta} \frac{1}{Z} \exp \left\{ \sum_{k} \lambda_k h_k(\text{English, Chinese}) \right\}$$

where:

$$\theta = \langle \lambda_1, ..., \lambda_K \rangle$$
Learning

$$\arg \max_{\theta} \frac{1}{Z} \exp \left\{ \sum_{k} \lambda_k h_k(English, Chinese) \right\}$$

where:

$$\theta = \langle \lambda_1, \ldots, \lambda_K \rangle$$

Techniques: SGD, L-BFGS
Learning

$$\arg\max_\theta \frac{1}{Z} \exp \left\{ \sum_k \lambda_k h_k(\text{English, Chinese}) \right\}$$

where:

$$\theta = \langle \lambda_1, \ldots, \lambda_K \rangle$$

Techniques: SGD, L-BFGS

Require computing derivatives (expectations!), iterating.
Problems
Problems

- Inference is high-order polynomial!
Problems

• Inference is high-order polynomial!
• Compute over $n$-best lists of outputs.
Problems

- Inference is high-order polynomial!
- Compute over $n$-best lists of outputs.
- Compute over pruned search graphs.
Problems

• Inference is high-order polynomial!
• Compute over $n$-best lists of outputs.
• Compute over pruned search graphs.
• Reachability: what if data likelihood is zero?
Problems

- Inference is high-order polynomial!
- Compute over $n$-best lists of outputs.
- Compute over pruned search graphs.
- Reachability: what if data likelihood is zero?
- Throw away data.
Problems

• Inference is high-order polynomial!
  • Compute over $n$-best lists of outputs.
  • Compute over pruned search graphs.
• Reachability: what if data likelihood is zero?
  • Throw away data.
• Pretend sentence with highest BLEU score is observed.
Problems
Problems

• Why maximize likelihood if we care about BLEU?