Feature-Based Models

🗯 Chrome File Edit	📽 🕏 🖶 🔽 📓 🍪 00:00:00 🖷 🤥 🎓 🥅 🕴 🗘 🖌 💭 💷 Tue 2:15 PM 🔍											
🔿 🔿 🔿 📾 Chost Wars Free Mu X 😁 Remember The Milk – X 🥩 A Tour of Scala: Patter X 🥩 Scala Standard Libran X 🗣 Classes and Objects 🛛 🔞 W3C Link Checker: ht X 🕥 JHU MT course 🛛 X 💩 Coogle Translate 🛛 X 💿												
$\epsilon \rightarrow C \parallel \odot$ translate.google.com												
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now												
+You Search Images Videos	Maps News Sho	pping Gmail	More -									Adam Lopez - 🛱
Google												
Translate	From: Detect languag	• •	To: English 👻	Translate								
English Spanish French	Detect language Afrikaans Albanian Arabic Armenian Azerbaijani Basque	Chinese Croatian Czech Danish Dutch English Estonian Filipino	Georgian German Greek Gujarati Haitian Creole Hebrew Hindi Hungarian	Italian Japanese Kannada Korean Latin Latvian Lithuanian Macedonian	Persian Polish Portuguese Romanian Russlan Serbian Slovak Slovenian	Tamil Telugu Thai Turkish Ukrainian Urdu Vietnamese Welsh	Spanish Arabic					
Type text or a website address	Belarusian Bengali Bulgarian Catalan	Finnish French Galician	Icelandic Indonesian Irish	Malay Maltese Norwegian	Spanish Swahili Swedish	Yiddish	k the words above to view	alternate translations. Dismiss				
			Google Transla	te for Business	8: Translator	Foolkit Websi	e Translator Global Mar	ket Finder				
Turn off instant translation								About Google Translate	Mobile	Privacy	Help	Send feedback

• 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.

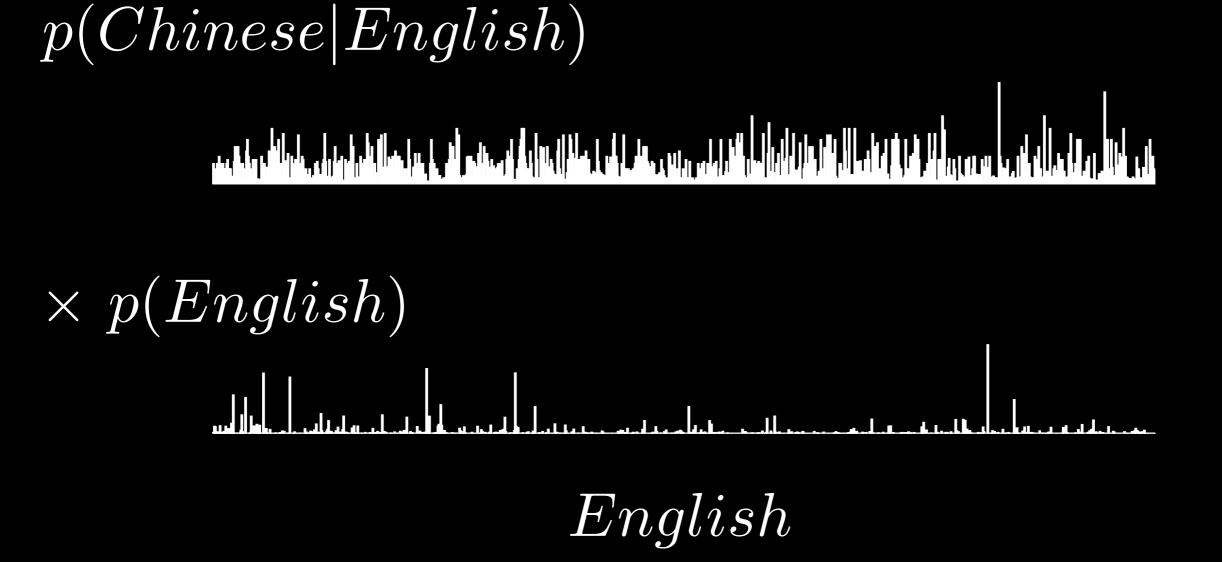
Bayes' Rule

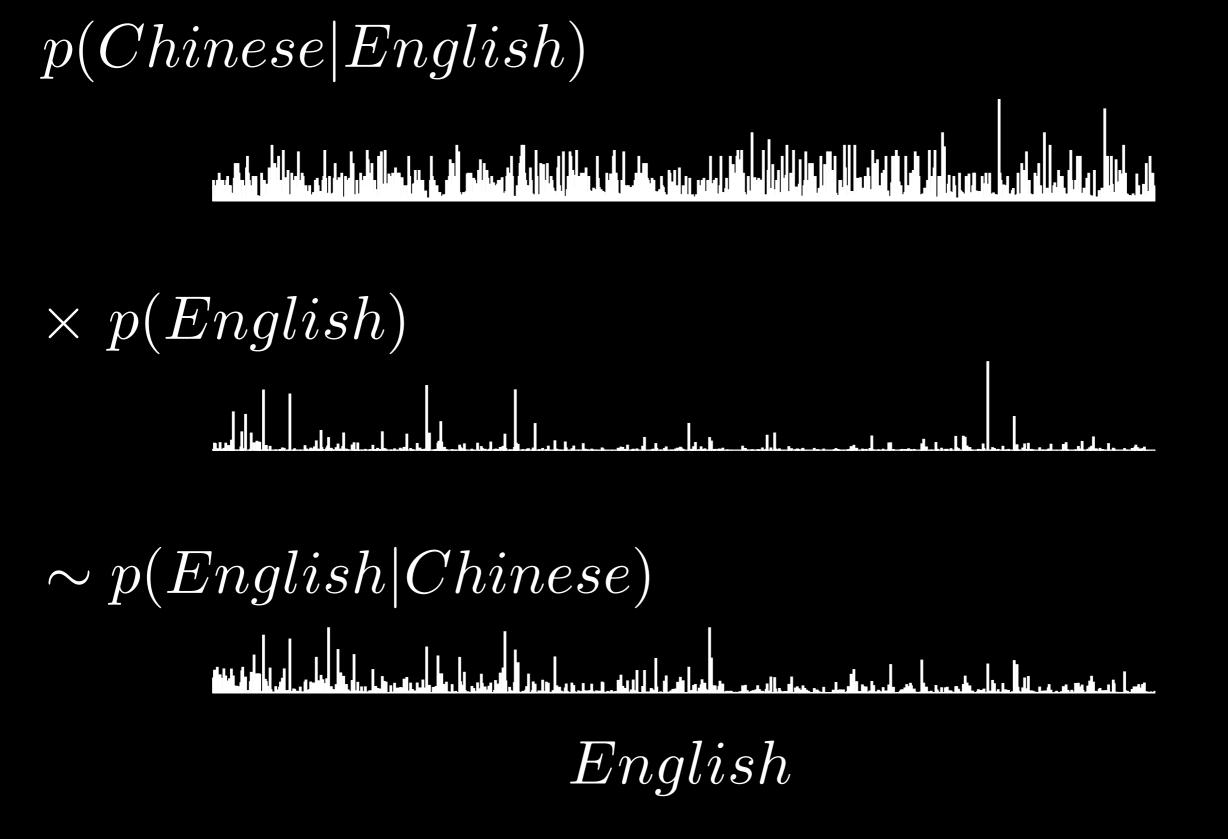
 $p(English|Chinese) \sim$

 $p(English) \times p(Chinese|English)$ $\int \int delta delta$

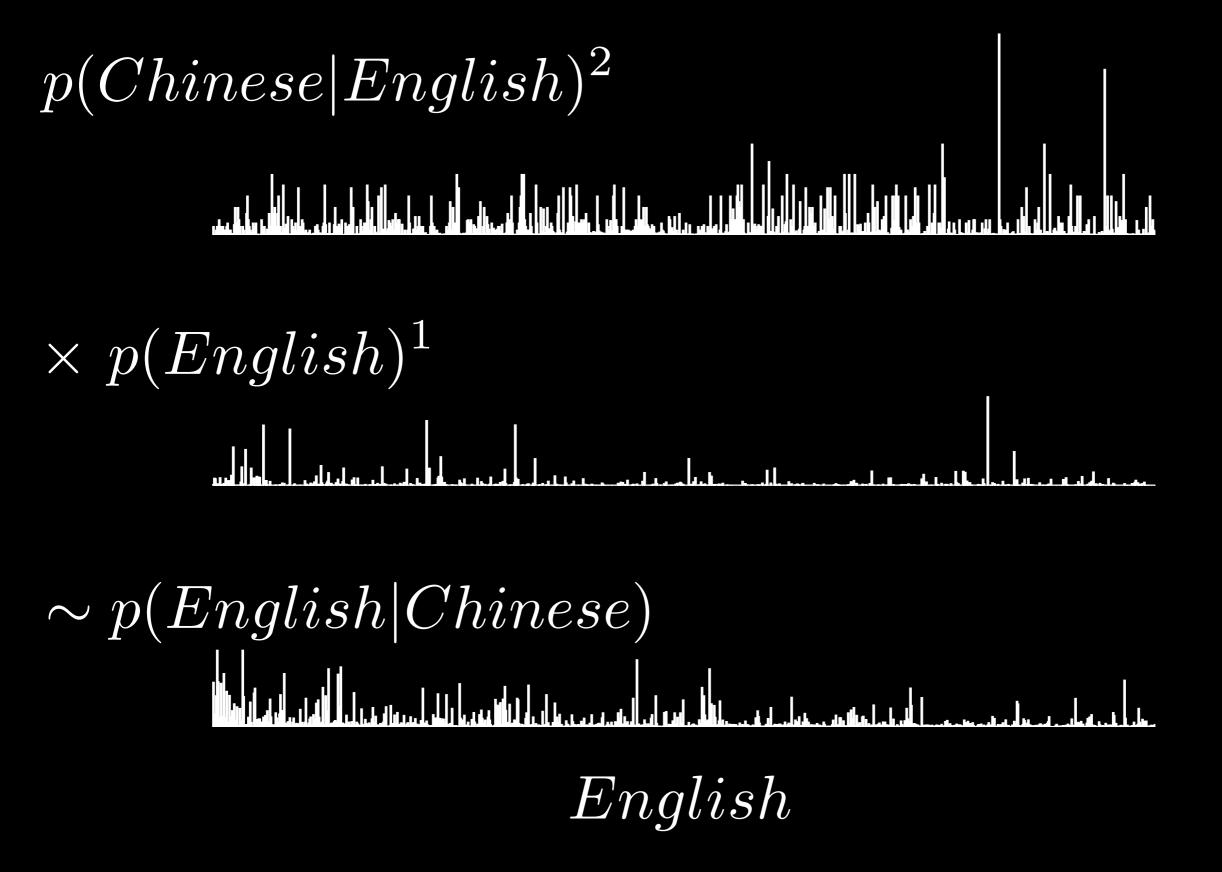
p(Chinese|English)

English

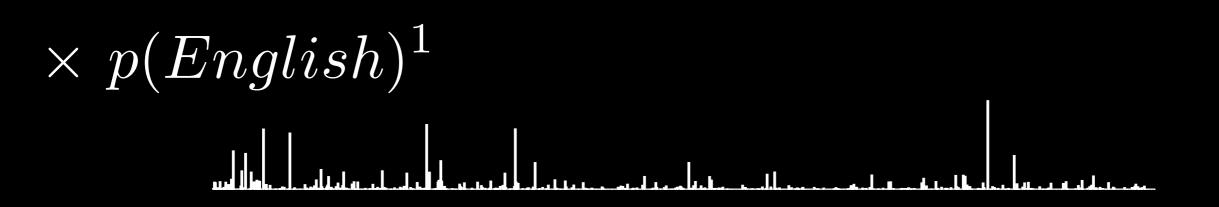




 $p(Chinese|English)^1$ $\times p(English)^1$ $\sim p(English|Chinese)$ English

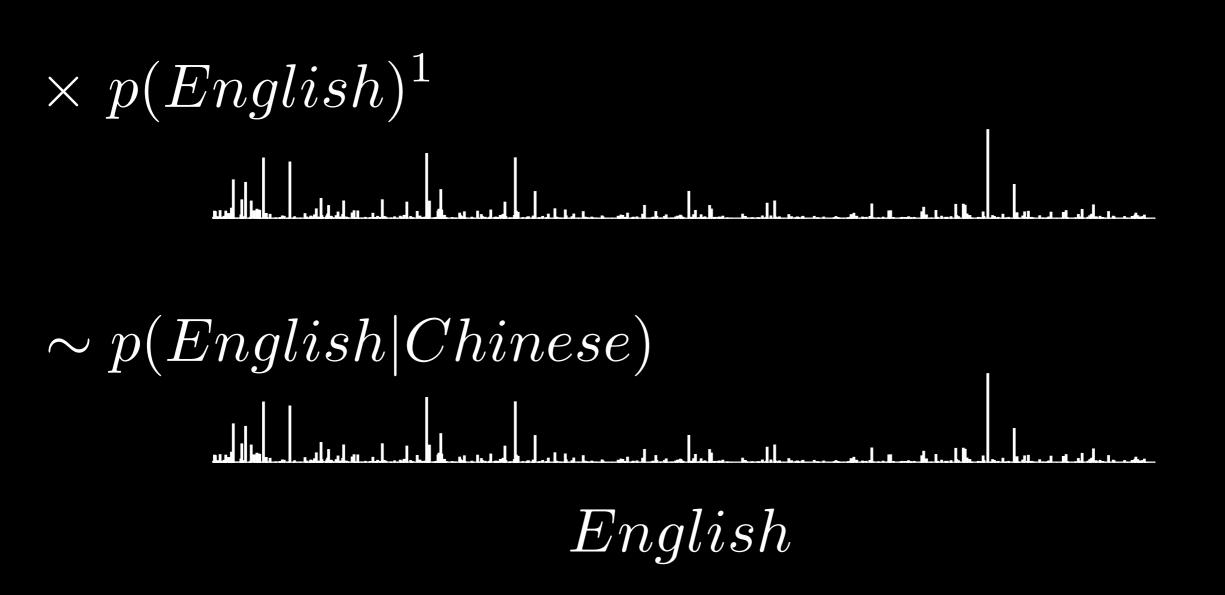


$p(Chinese|English)^{1/2}$

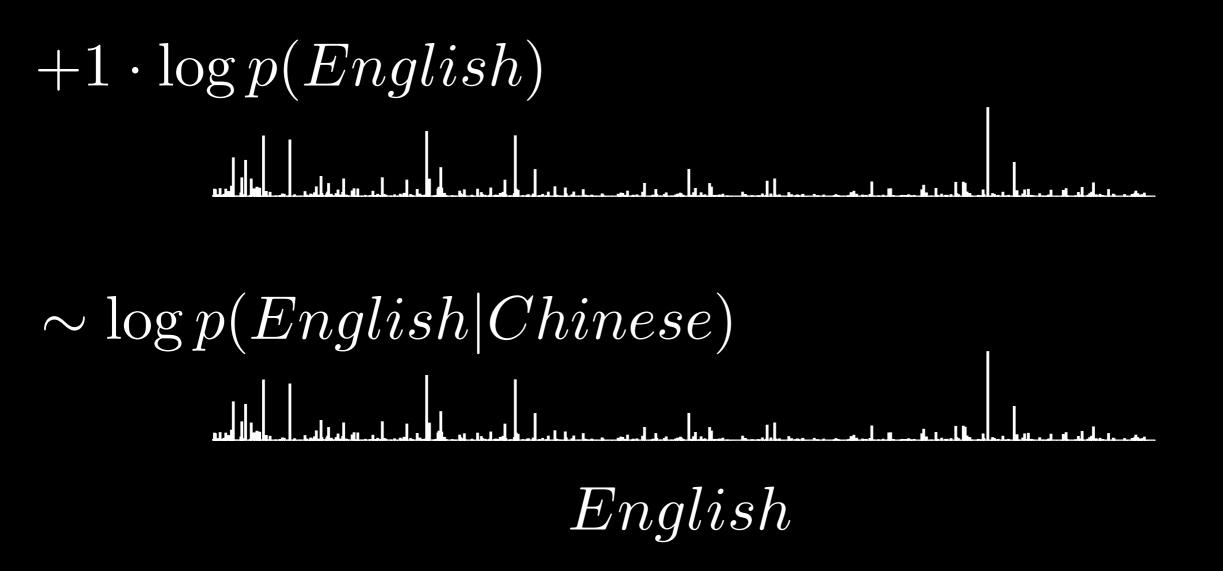


 $\sim p(English|Chinese)$

$p(Chinese|English)^0$

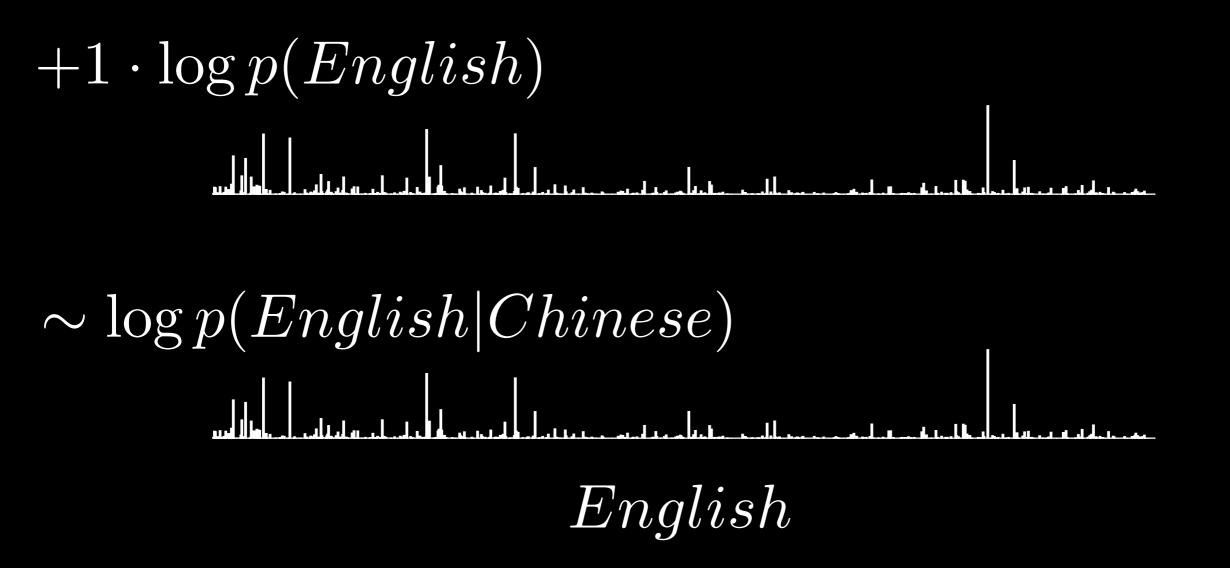


$0 \cdot \log p(Chinese|English)$

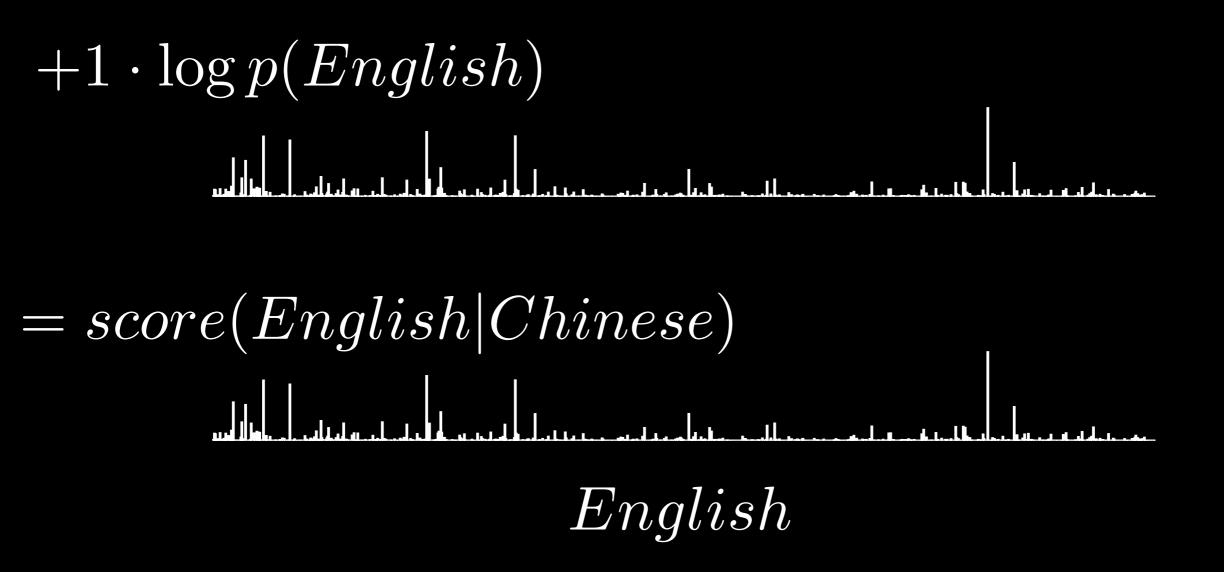


log(x) is monotonic for positive x
(i.e. log(x) > log(y) iff x>y)

 $0 \cdot \log p(Chinese|English)$



$0 \cdot \log p(Chinese|English)$



score(English|Chinese) =

 $\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English)$

score(English|Chinese) =

 $\exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))$

 $\frac{\exp(\lambda_1 \log p(Chinese | English) + \lambda_2 \log p(English))}{\sum_{nglish} \exp(\lambda_1 \log p(Chinese | English) + \lambda_2 \log p(English))}$

 $\frac{\exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))}{\sum_{nglish} \exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))}$

$p(English) \times p(Chinese|English)$

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

 $\frac{\exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))}{\sum_{nglish} \exp(\lambda_1 \log p(Chinese|English) + \lambda_2 \log p(English))}$

$$p(English|Chinese) = \\ \exp\left\{\sum_{k} \lambda_k h_k(English, Chinese)\right\} \\ \sum_{English'} \exp\left\{\sum_{k} \lambda_k h_k(English', Chinese)\right\}$$

p(English|Chinese) = $\frac{1}{Z} \exp\left\{\sum_{k} \lambda_k h_k(English, Chinese)\right\}$

$$p(English|Chinese) = \frac{1}{Z} \exp\left\{\sum_{k} \lambda_{k} h_{k}(English, Chinese)\right\}$$

Z is the normalization term or *partition function*

$$p(English|Chinese) = \frac{1}{Z} \exp\left\{\sum_{k} \lambda_{k} h_{k}(English, Chinese)\right\}$$

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 λ_k terms.

A feature can be *any* function in the form: $h_k : English \times Chinese \rightarrow \mathbb{R}_+$

• Language model: *p*(*English*)

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* | *English*)

- Language model: *p*(*English*)
- Translation model: *p*(*Chinese* | *English*)
- Reverse translation model: *p*(*English* | *Chinese*)

- Language model: *p*(*English*)
- Translation model: *p*(*Chinese* | *English*)
- Reverse translation model: *p*(*English* | *Chinese*)
- The number of words in the English sentence.

- Language model: *p*(*English*)
- Translation model: *p*(*Chinese* | *English*)
- Reverse translation model: *p*(*English* | *Chinese*)
- The number of words in the English sentence.
- The number of verbs in the English sentence.

- Language model: *p*(*English*)
- Translation model: *p*(*Chinese* | *English*)
- Reverse translation model: *p*(*English* | *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.

A feature can be *any* function in the form: $h_k : English \times Chinese \rightarrow \mathbb{R}_+$

• A word-based translation model: *p*(*Chinese* | *English*)

A feature can be *any* function in the form: $h_k : English \times Chinese \rightarrow \mathbb{R}_+$

A word-based translation model: *p(Chinese | English)*Agreement features in the English sentence.

A feature can be *any* function in the form: $h_k : English \times Chinese \rightarrow \mathbb{R}_+$

• A word-based translation model: *p*(*Chinese* | *English*)

Agreement features in the English sentence.

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

- A word-based translation model: *p*(*Chinese* | *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* | *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.
- Do words *north* and $\exists k$ appear in a dictionary?

Learning

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

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

Techniques: SGD, L-BFGS

Learning

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

where: $\theta = \langle \lambda_1, ..., \lambda_K \rangle$

Techniques: SGD, L-BFGS

Require computing derivatives (expectations!), iterating.

• Inference is high-order polynomial!

- Inference is high-order polynomial!
 - Compute over *n*-best lists of outputs.

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

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

- 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.

- 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.

• Why maximize likelihood if we care about BLEU?