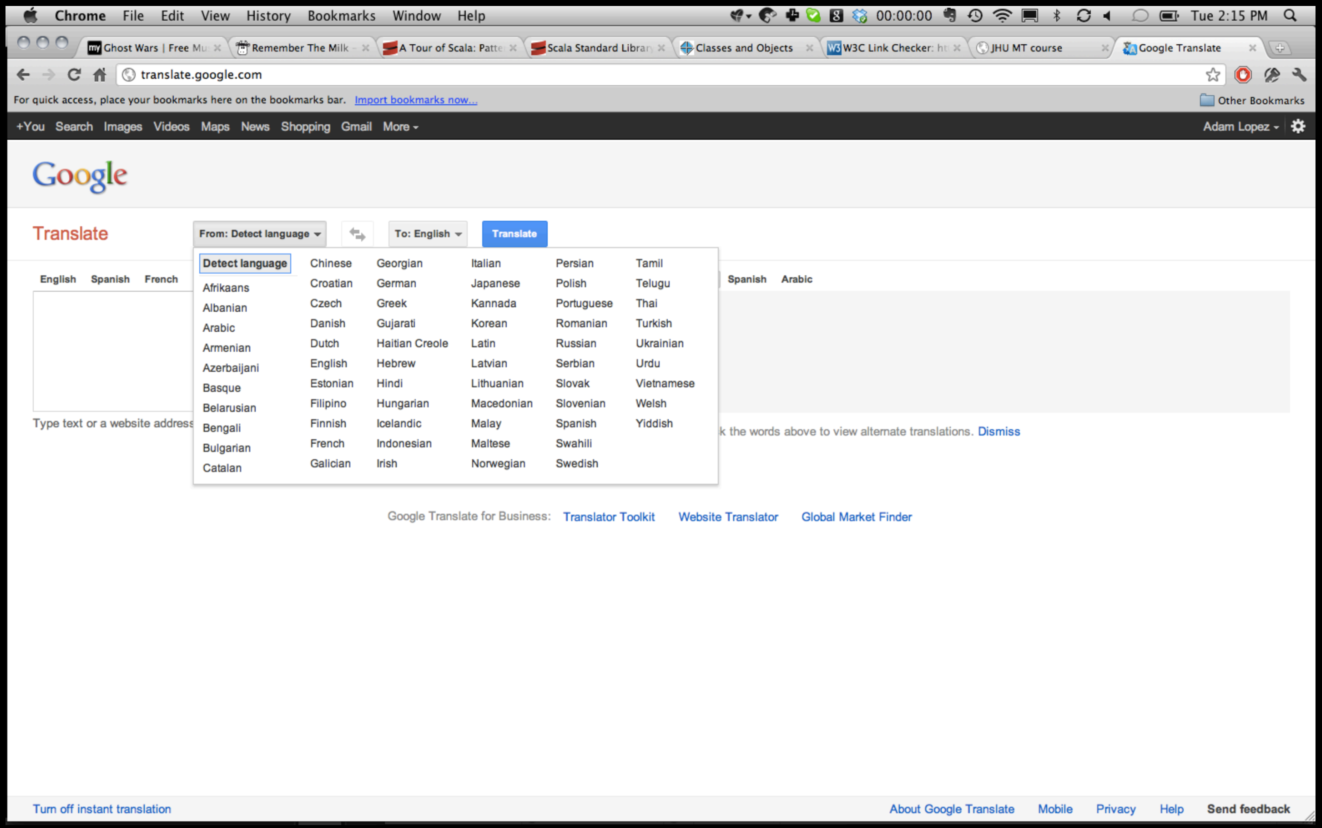
Feature-Based Models



- 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

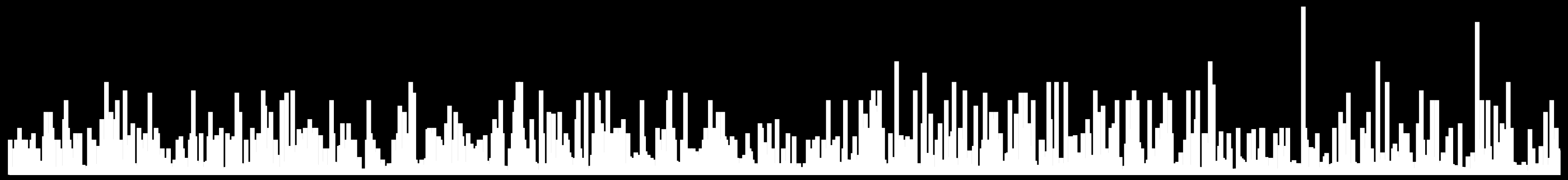
$$p(English|Chinese) \sim$$

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

language model

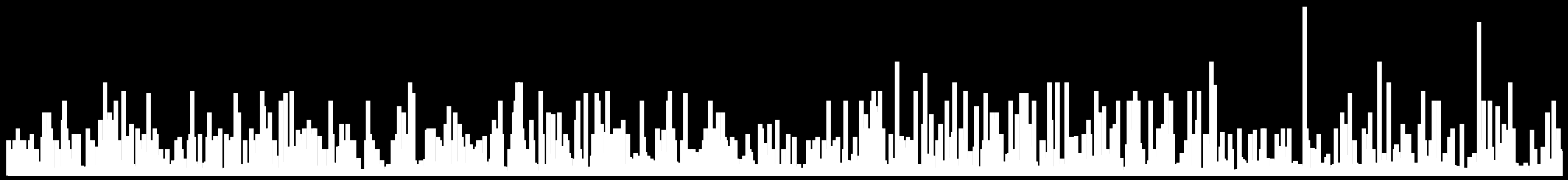
translation model

$p(Chinese|English)$

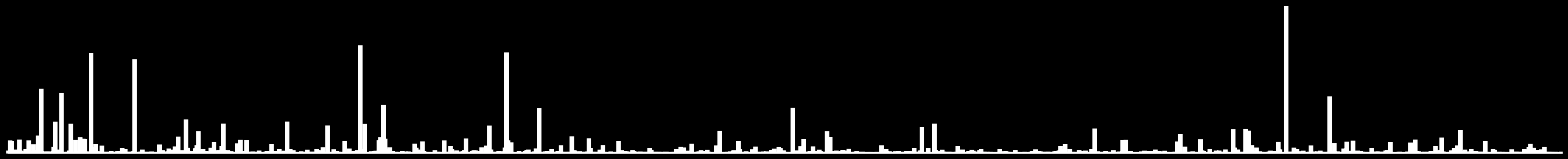


$English$

$p(Chinese|English)$

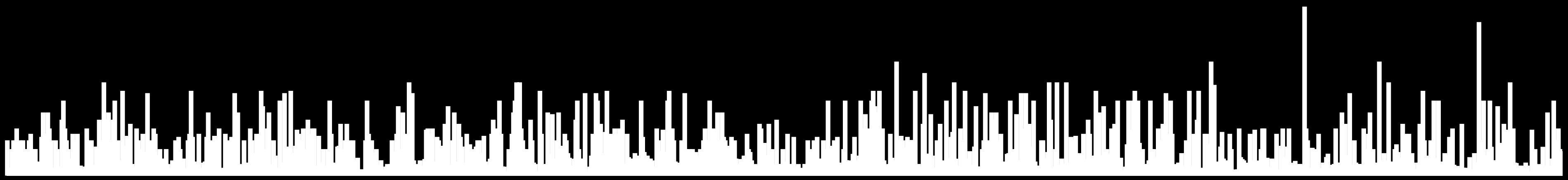


$\times\ p(English)$

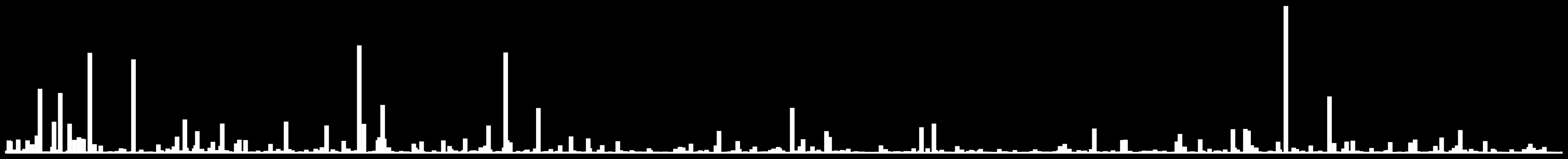


$English$

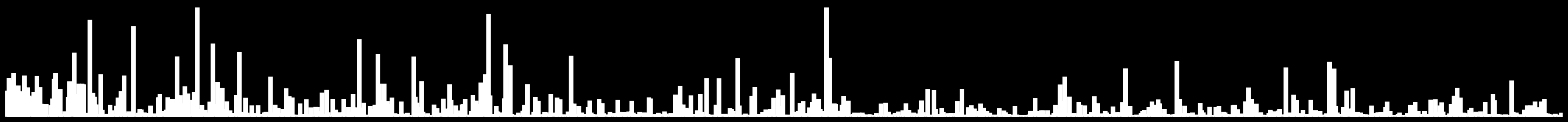
$p(Chinese|English)$



$\times\ p(English)$

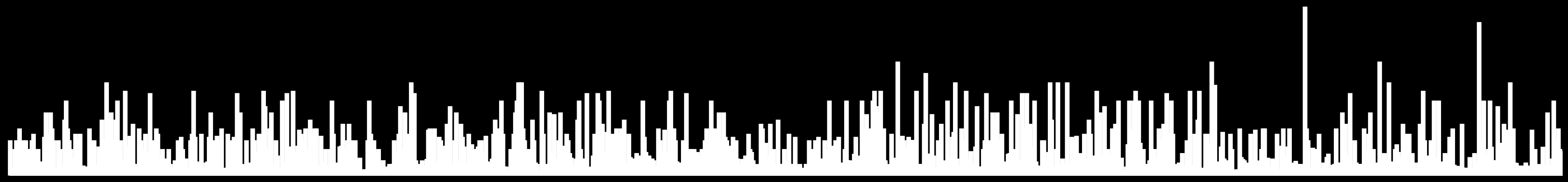


$\sim p(English|Chinese)$

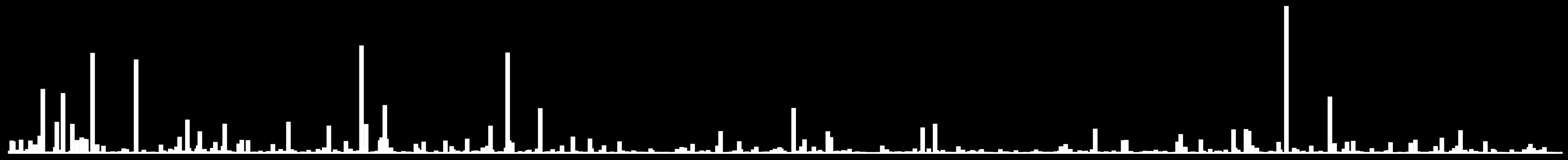


$English$

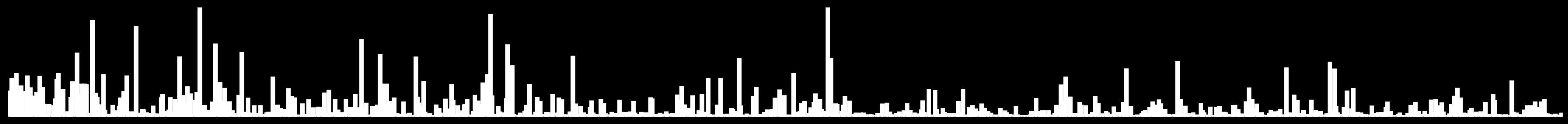
$$p(Chinese|English)^1$$



$$\times\ p(English)^1$$

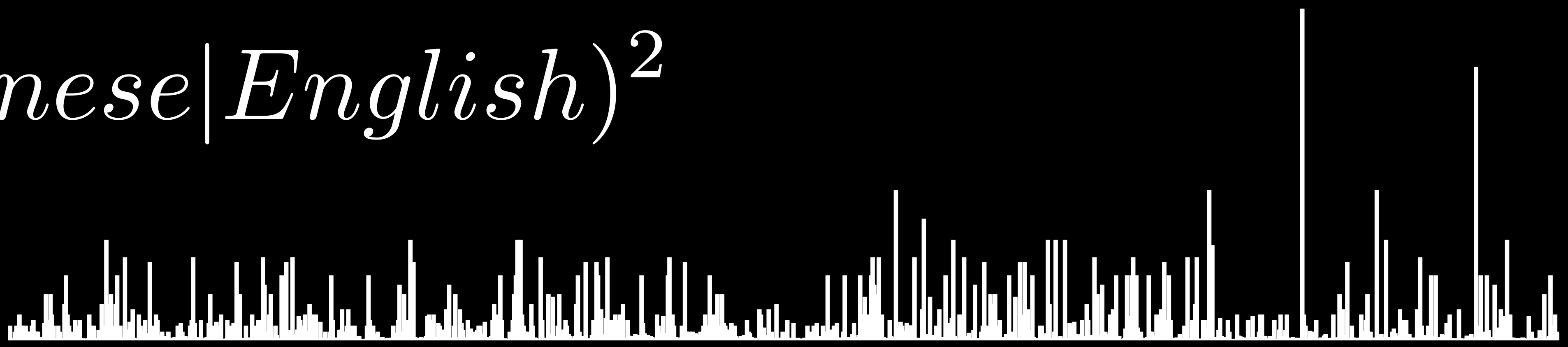


$$\sim p(English|Chinese)$$

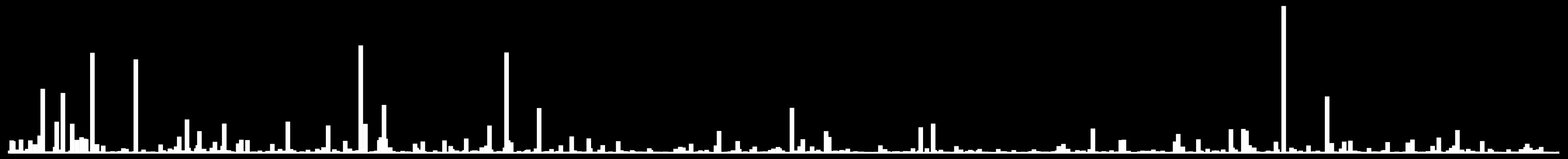


$$English$$

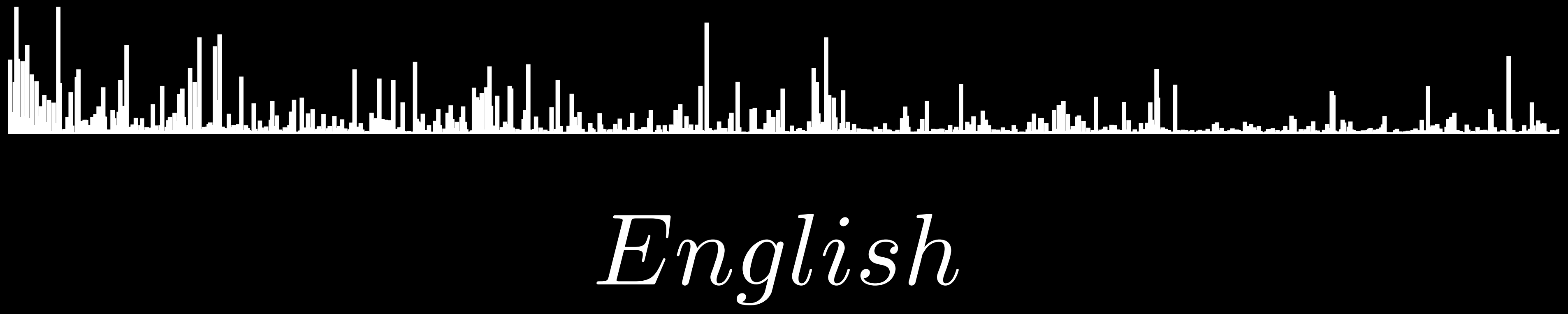
$$p(Chinese|English)^2$$



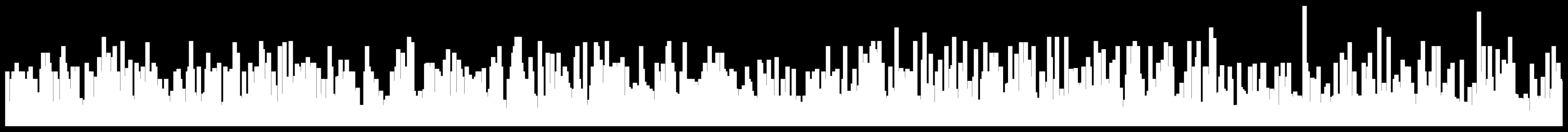
$$\times\ p(English)^1$$



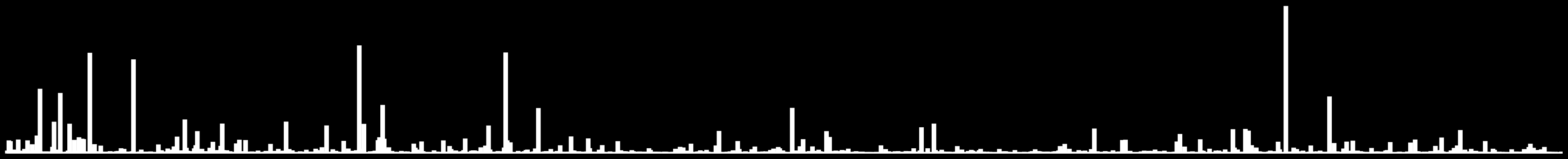
$$\sim p(English|Chinese)$$



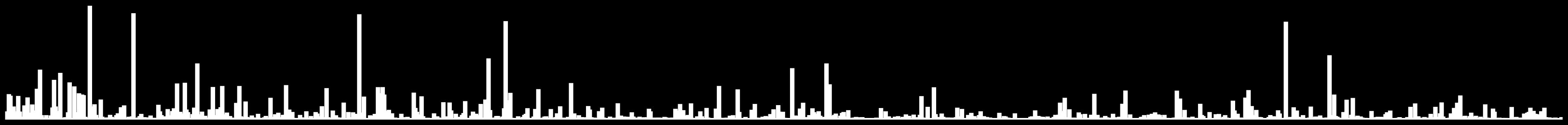
$$p(Chinese|English)^{1/2}$$



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

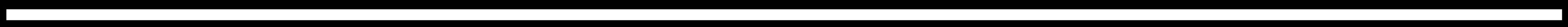


$$\sim p(English|Chinese)$$

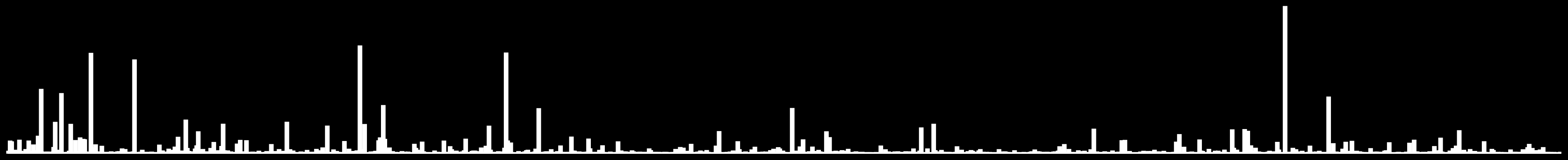


$$English$$

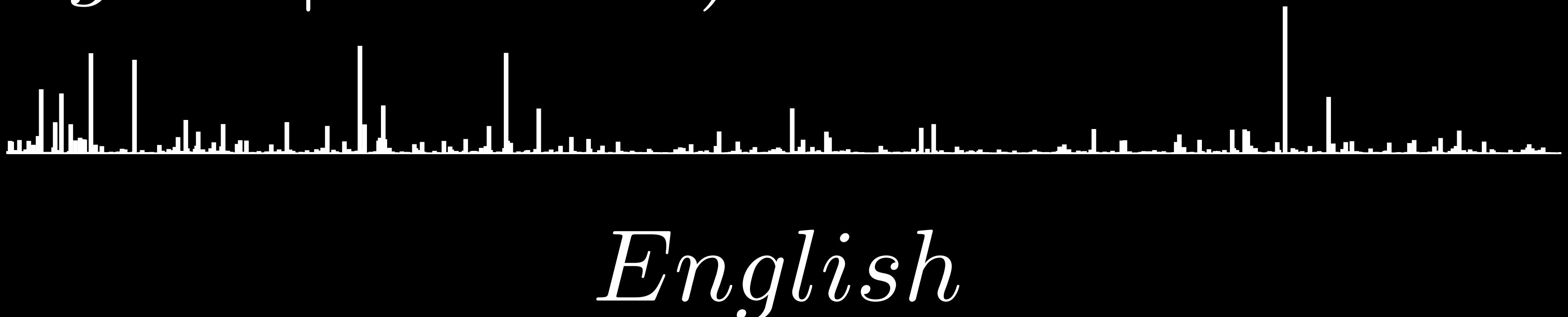
$$p(Chinese|English)^0$$



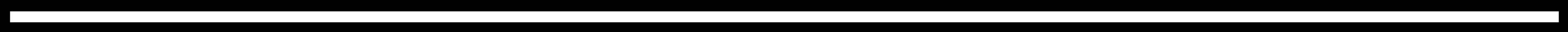
$$\times\ p(English)^1$$



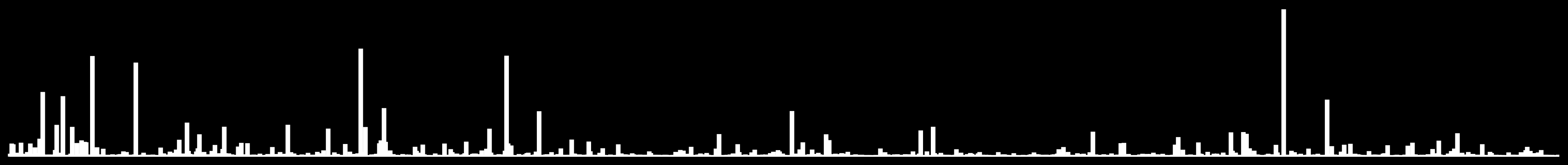
$$\sim p(English|Chinese)$$



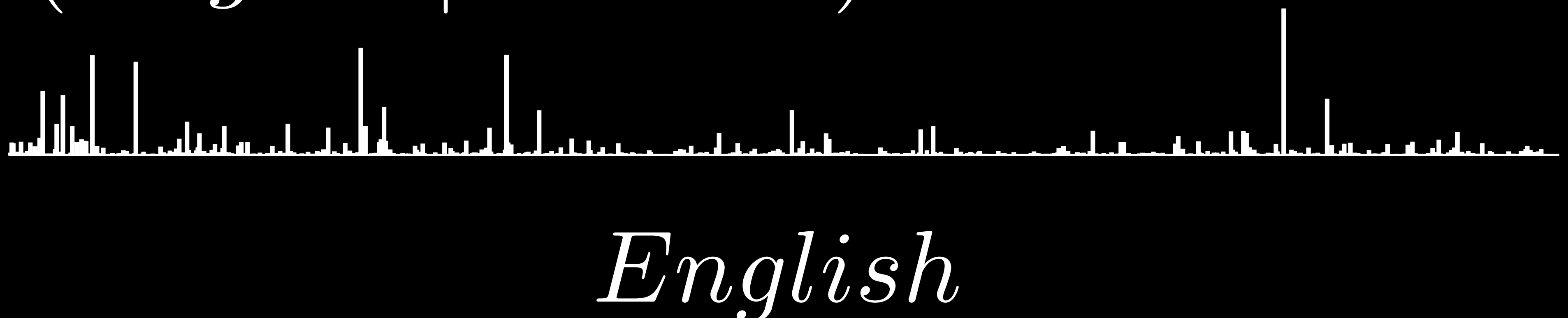
$$0 \cdot \log p(Chinese|English)$$



$$+1 \cdot \log p(English)$$



$$\sim \log p(English|Chinese)$$

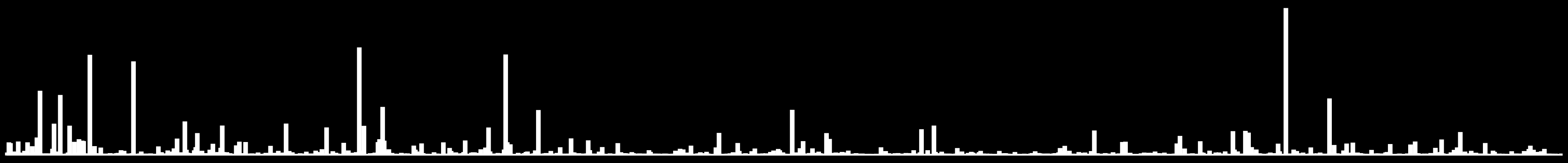


log(x) is monotonic for positive *x*

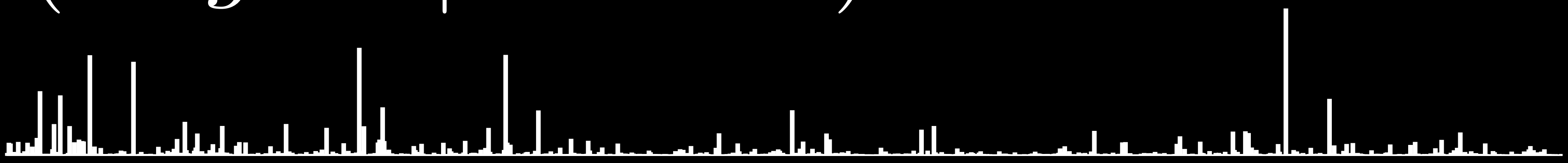
(i.e. *log(x)* > *log(y)* iff *x*>*y*)

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

$$+1 \cdot \log p(English)$$

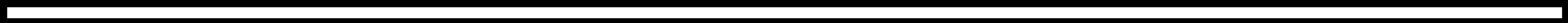


$$\sim \log p(English|Chinese)$$

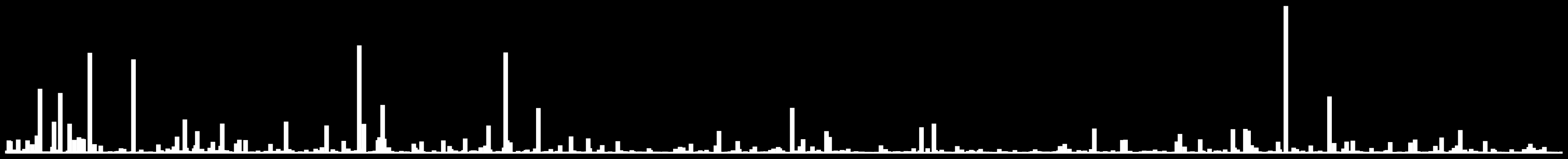


$English$

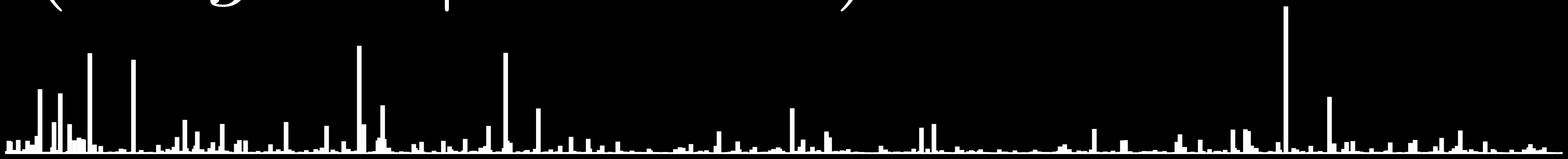
$$0 \cdot \log p(Chinese|English)$$



$$+1 \cdot \log p(English)$$



$$= score(English|Chinese)$$



$English$

$$score(English|Chinese) =$$

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

$$score(English|Chinese) =$$

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

$$p(English|Chinese) =$$

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

$$p(English|Chinese) =$$

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

log-linear model
maximum entropy model
conditional model
undirected model

$$p(English|Chinese) =$$

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

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

log-linear model
maximum entropy model
conditional model
undirected model

$$p(English|Chinese) =$$

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

log-linear model

maximum entropy model

conditional model

undirected model

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

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

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

Z is the normalization term or *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 λ_k terms.

What’s a Feature?

What’s a Feature?

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

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

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:

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

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

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

What's a Feature?

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

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

- 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.

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 | 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.

What’s a Feature?

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

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

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(Chinese | 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(Chinese | 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(Chinese | 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:

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

- 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.

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 北 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.

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?