Latent Variable Models and Word Alignment

CS159 Spring 2015

(Based on slides from Adam Lopez, University of Edinburgh, mt-class.org

Probability and Language

Goal

- Write down a ***model*** over sentence pairs.
- ***Learn*** an instance of the model from data.
- Use it to ***predict*** translations of new sentences.

Why probability?

- Formalizes...
  - the concept of ***models***
  - the concept of ***data***
  - the concept of ***learning***
  - the concept of ***inference*** (prediction)
- Derive logical conclusions in the face of ambiguity.

Goal

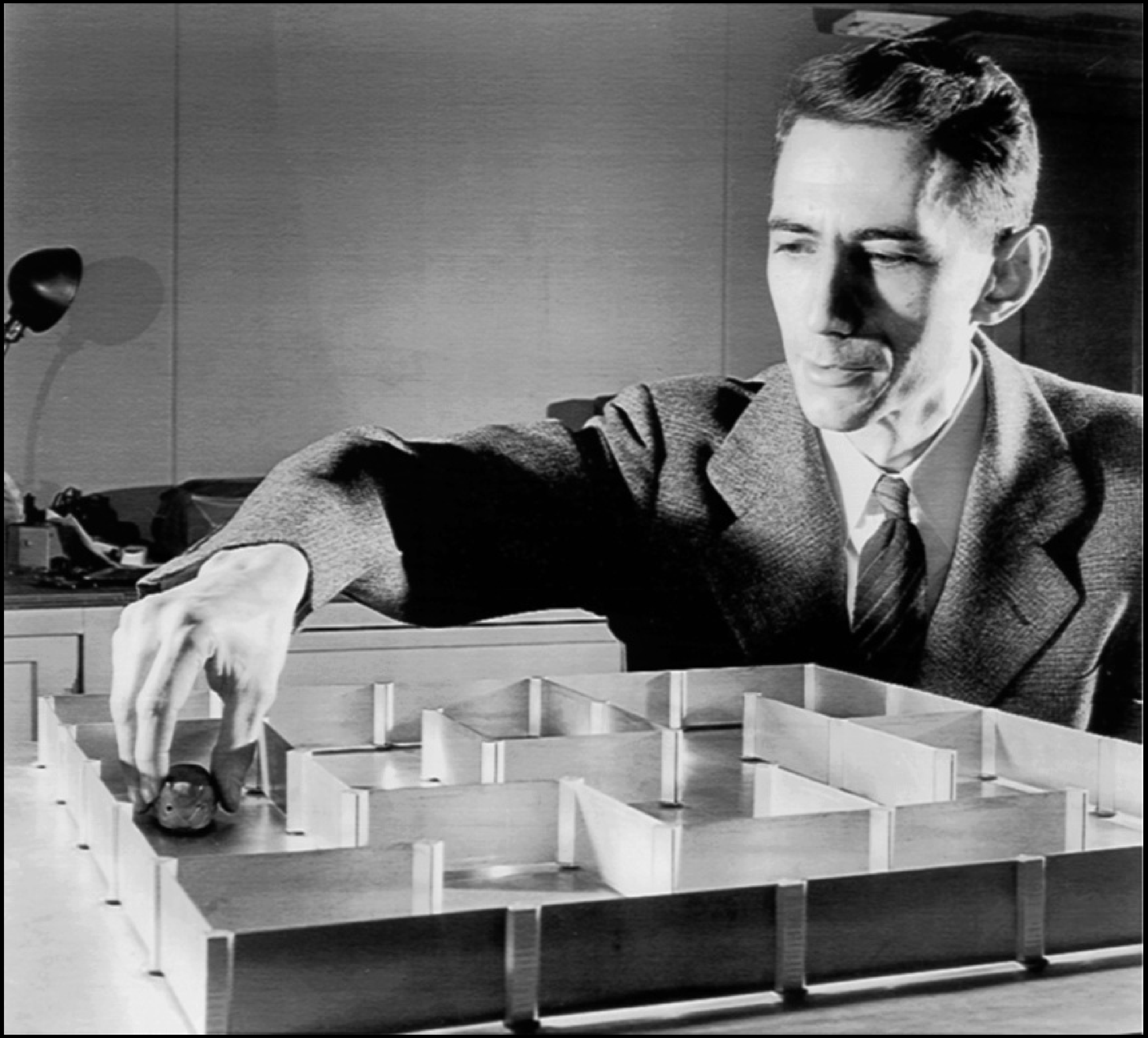
- Write down a ***model*** over sentence pairs.
- ***Learn*** an instance of the model from data.
- Use it to ***predict*** translations of new sentences.

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."

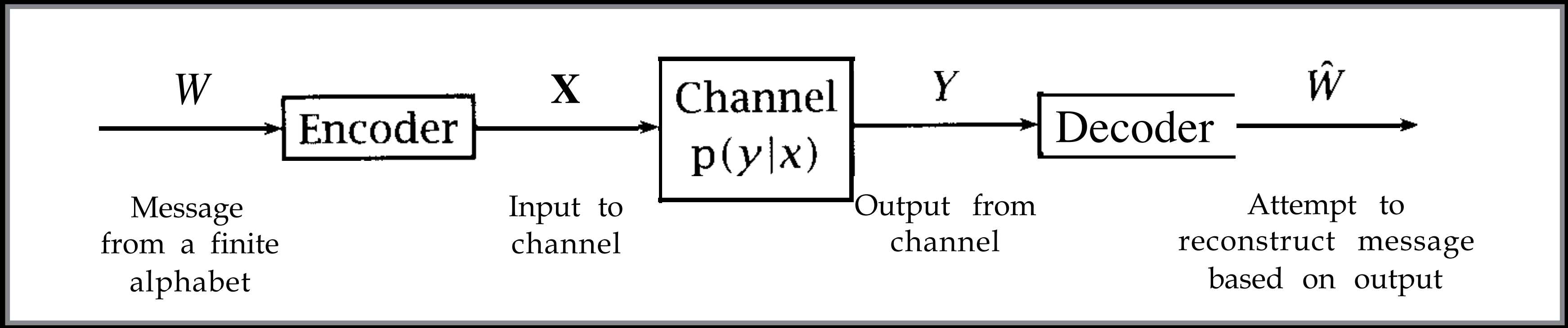
Warren Weaver (1949)

THE MATHEMATICAL THEORY OF COMMUNICATION

by Claude E. Shannon and Warren Weaver



Claude Shannon



“Intended” Language

Observed Language

Bayes' Rule

$$p(English|Chinese) = \frac{\overbrace{p(English)}^{\text{prior}} \times \overbrace{p(Chinese|English)}^{\text{likelihood}}}{\underbrace{p(Chinese)}_{\text{evidence}}}$$

Noisy Channel

$$\underbrace{p(English}_{\boxed{\text{Intended}}}|\underbrace{Chinese}_{\boxed{\text{Observed}}}) =$$

$$\frac{p(English) \times p(Chinese|English)}{p(Chinese)}$$

signal model

channel model

normalization (ensures we're working with valid probabilities).

Machine Translation

$$p(English|Chinese) =$$

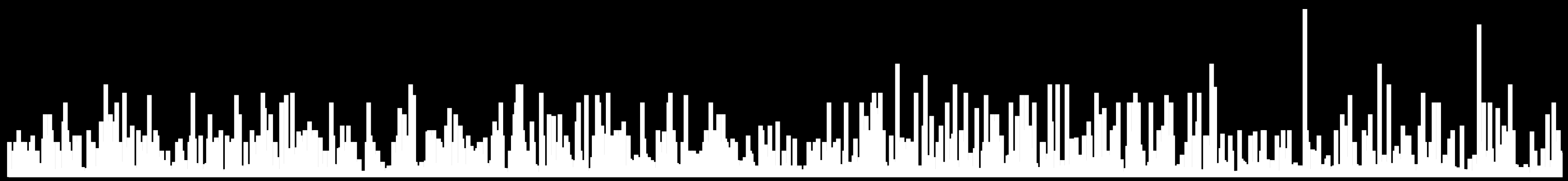
$$\frac{p(English) \times p(Chinese|English)}{p(Chinese)}$$

language model

translation model

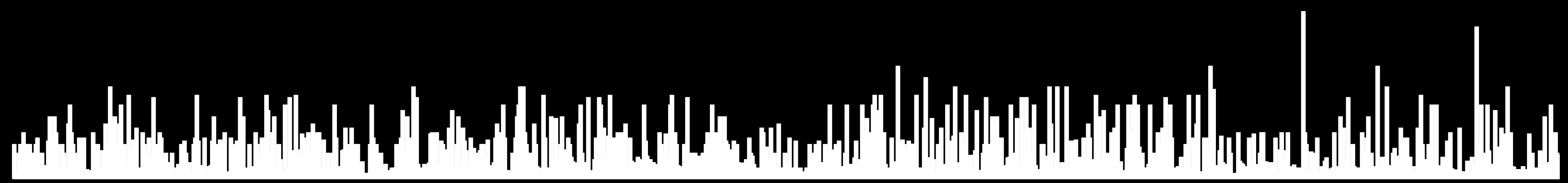
normalization (ensures we're working with valid probabilities).

$$p(Chinese|English)$$

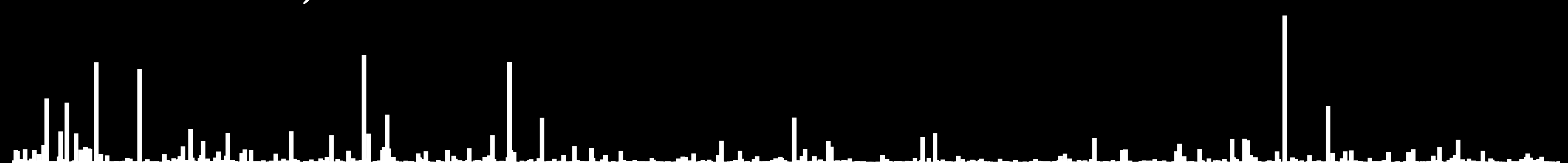


$English$

$p(Chinese|English)$

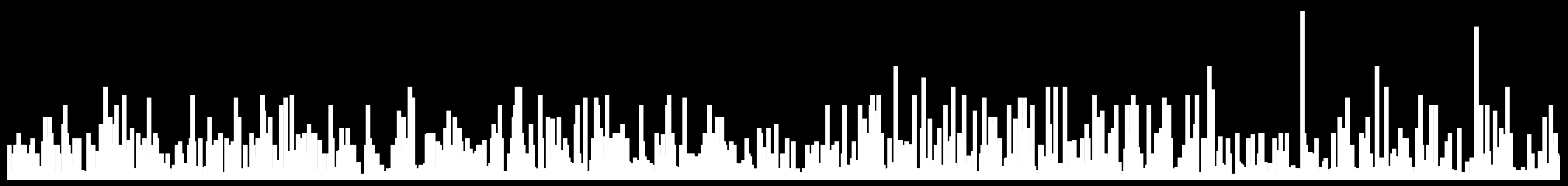


$\times\ p(English)$

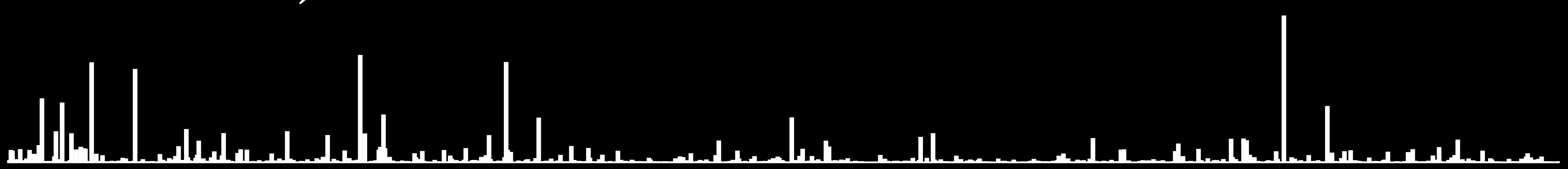


$English$

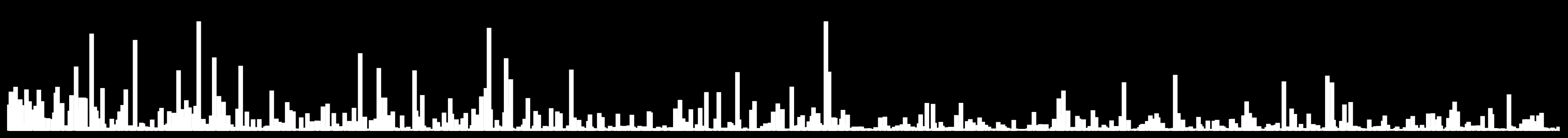
$p(Chinese|English)$



$\times\ p(English)$



$\sim p(English|Chinese)$



$English$

Machine Translation

$$p(English|Chinese) =$$

$$\frac{\underbrace{p(English)}_{\text{language model}} \times \underbrace{p(Chinese|English)}_{\text{translation model}}}{\underbrace{p(Chinese)}_{\text{evidence}}}$$

Machine Translation

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

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

Questions our model must answer:

What is the probability of an English sentence?

What is the probability of a Chinese sentence, given a particular English sentence?

Lexical Translation

- How to translate a word → look up in dictionary

  Haus — house, building, home, household, shell.

- Multiple translations
  - some more frequent than others
  - for instance: house, and building most common
  - special cases: Haus of a snail is its shell

- Note: In all lectures, we translate from a foreign language into English

Collect Statistics

Look at a parallel corpus (German text along with English translation)

Translation of *Haus*	**Count**
house	8,000
building	1,600
home	200
household	150
shell	50

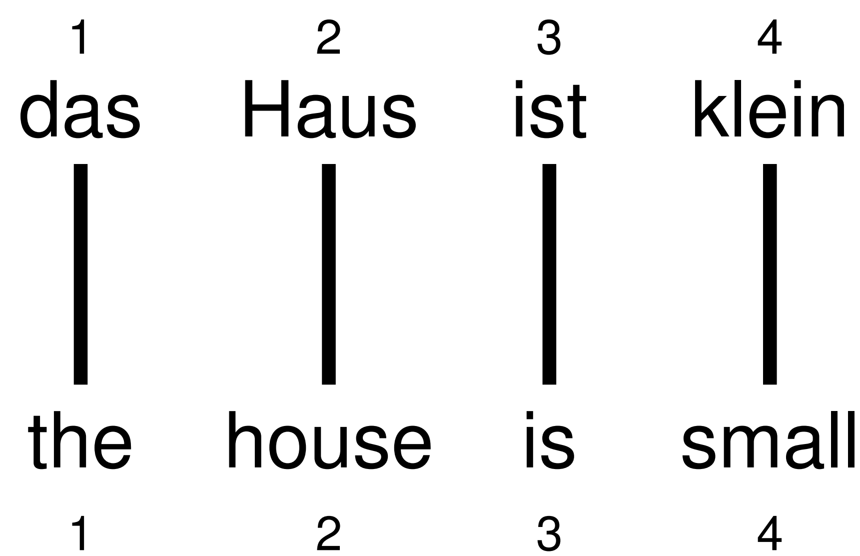
Estimate Translation Probabilities

Maximum likelihood estimation

$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \text{house}, \\ 0.16 & \text{if } e = \text{building}, \\ 0.02 & \text{if } e = \text{home}, \\ 0.015 & \text{if } e = \text{household}, \\ 0.005 & \text{if } e = \text{shell}. \end{cases}$$

Alignment

- In a parallel text (or when we translate), we align words in one language with the words in the other



- Word positions are numbered 1–4

Alignment Function

- Formalizing alignment with an alignment function

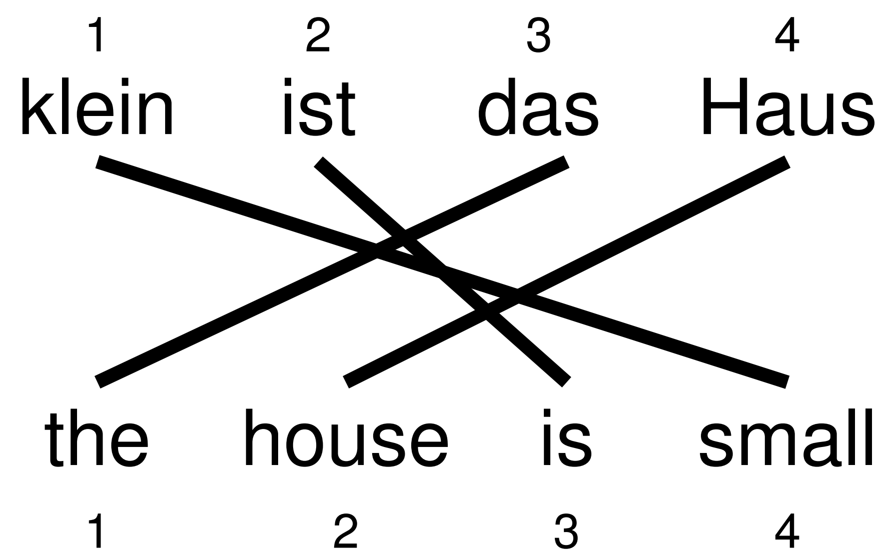
- Mapping an English target word at position i to a German source word at position j with a function $a : i \to j$

- Example

$$a : \{1 \to 1, 2 \to 2, 3 \to 3, 4 \to 4\}$$

Reordering

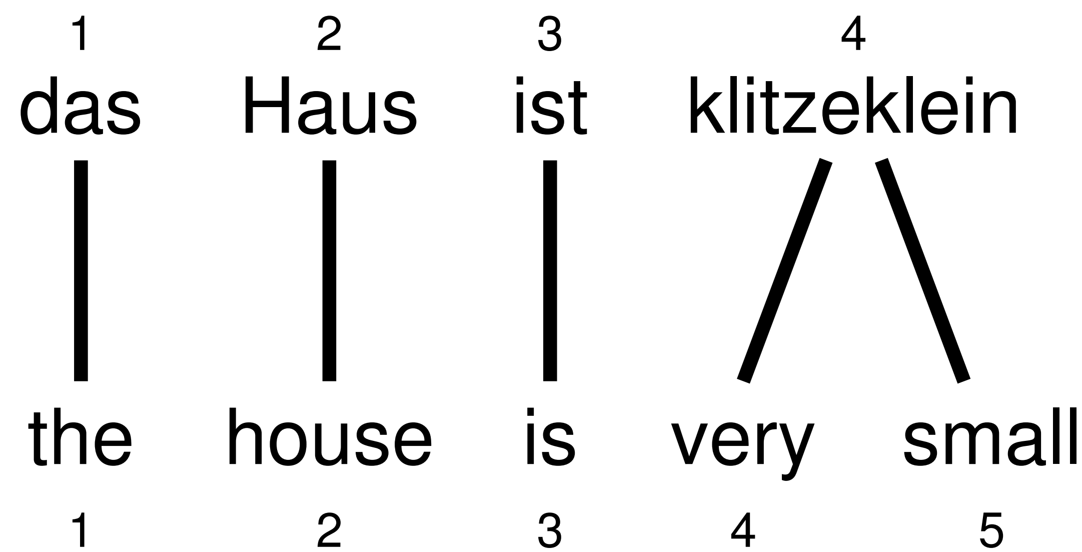
Words may be reordered during translation



$$a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$$

One-to-Many Translation

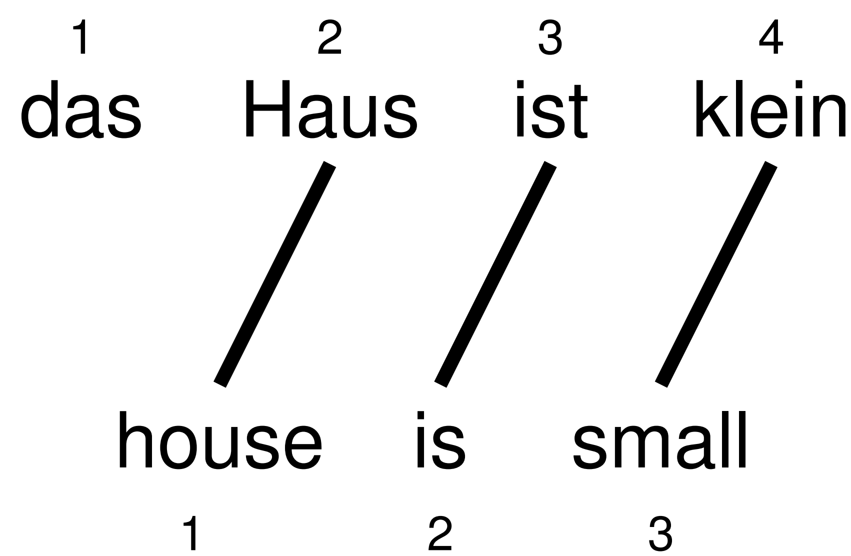
A source word may translate into multiple target words



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

Dropping Words

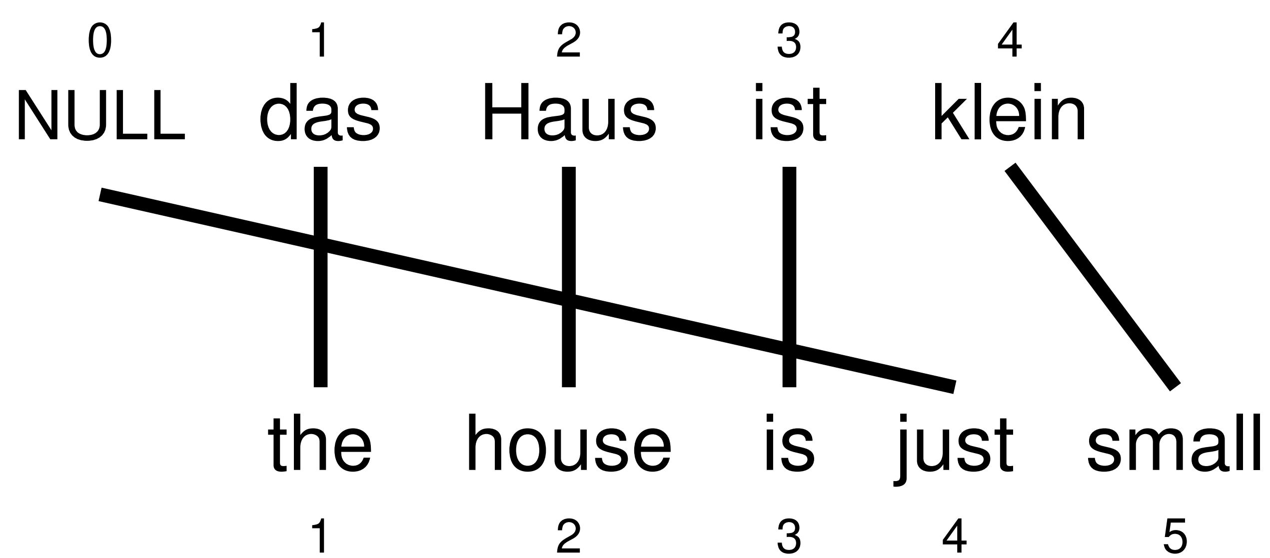
Words may be dropped when translated
(German article das is dropped)



$$a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$$

Inserting Words

- Words may be added during translation
  - The English just does not have an equivalent in German
  - We still need to map it to something: special NULL token



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}$$

IBM Model 1

- Generative model: break up translation process into smaller steps
  - IBM Model 1 only uses lexical translation

- Translation probability
  - for a foreign sentence $\mathbf{f} = (f_1, ..., f_{l_f})$ of length l_f
  - to an English sentence $\mathbf{e} = (e_1, ..., e_{l_e})$ of length l_e
  - with an alignment of each English word e_j to a foreign word f_i according to the alignment function $a : j \to i$

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j | f_{a(j)})$$

  - parameter ϵ is a normalization constant

Example

das

e	$t(e\|f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus

e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist

e	$t(e\|f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein

e	$t(e\|f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

$$\begin{aligned} p(e,a|f) &= \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein}) \\ &= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4 \\ &= 0.0028\epsilon \end{aligned}$$

IBM Model 1

alignment of French
word at position i

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = p(I|J) \prod_{i=1}^{I} p(a_i|J) \cdot p(f_i|e_{a_i})$$

French, English
sentence lengths

French, English
word pair

IBM Model 1

$$p(despite|\text{虽然})$$

$$p(however|\text{虽然})$$

$$p(although|\text{虽然})$$

$$p(northern|\text{北})$$

$$p(north|\text{北})$$

IBM Model 1

$p(despite|\text{虽然})$???

$p(however|\text{虽然})$???

$p(although|\text{虽然})$???

$p(northern|\text{北})$???

$p(north|\text{北})$???

IBM Model 1

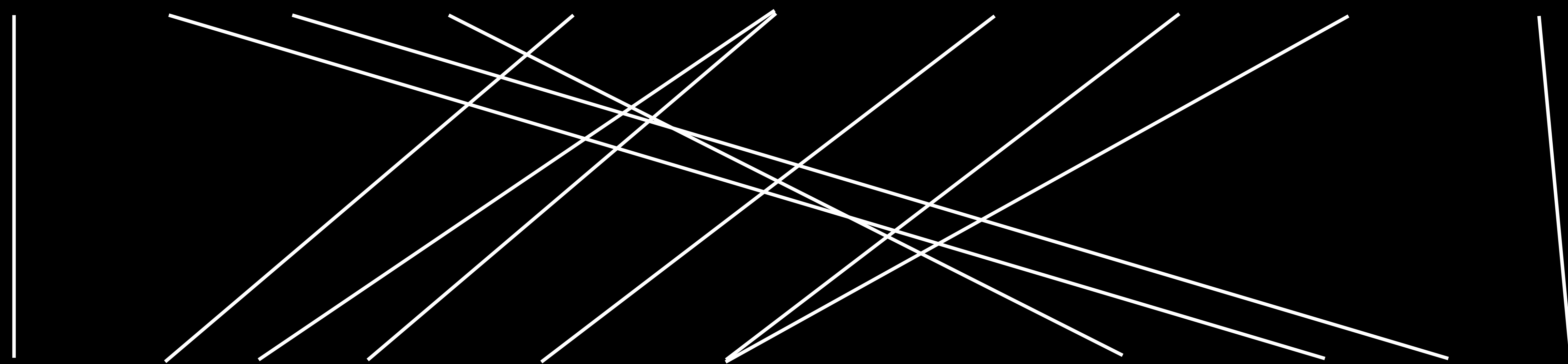
$$\theta \left\{ \begin{array}{ll} p(despite|\text{虽然}) & ??? \\ p(however|\text{虽然}) & ??? \\ p(although|\text{虽然}) & ??? \\ \\ p(northern|\text{北}) & ??? \\ p(north|\text{北}) & ??? \end{array} \right.$$

- Optimization

MLE for IBM Model 1 (observed)

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 .

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{f}, \mathbf{a} | \mathbf{e})$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(p(I^{(n)}|J^{(n)}) \prod_{i=1}^{I^{(n)}} p(a_i^{(n)}|J^{(n)}) \cdot p(f_i^{(n)}|e_{a_i}^{(n)}) \right)$$

MLE for IBM Model 1 (observed)

number of sentences

alignment of French word at position i

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(p(I^{(n)} | J^{(n)}) \prod_{i=1}^{I^{(n)}} p(a_i^{(n)} | J^{(n)}) \cdot p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

French, English sentence lengths

French, English word pair

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(\underbrace{p(I^{(n)}|J^{(n)}) \prod_{i=1}^{I^{(n)}} p(a_i^{(n)}|J^{(n)})}_{\text{constant (w.r.t. words)!}} \cdot p(f_i^{(n)}|e_{a_i}^{(n)}) \right)$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)})$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

$$\log(a) < \log(b) \iff a < b$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{count(\langle f,e \rangle)} \right)$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e)$$

log of product = sum of logs

MLE for IBM Model 1 (observed)

$$\Lambda(\theta, \lambda) = \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e) \underbrace{- \sum_{e} \lambda_e \left(\sum_{f} p(f|e) - 1 \right)}_{\text{Lagrange multiplier expresses normalization constraint}}$$

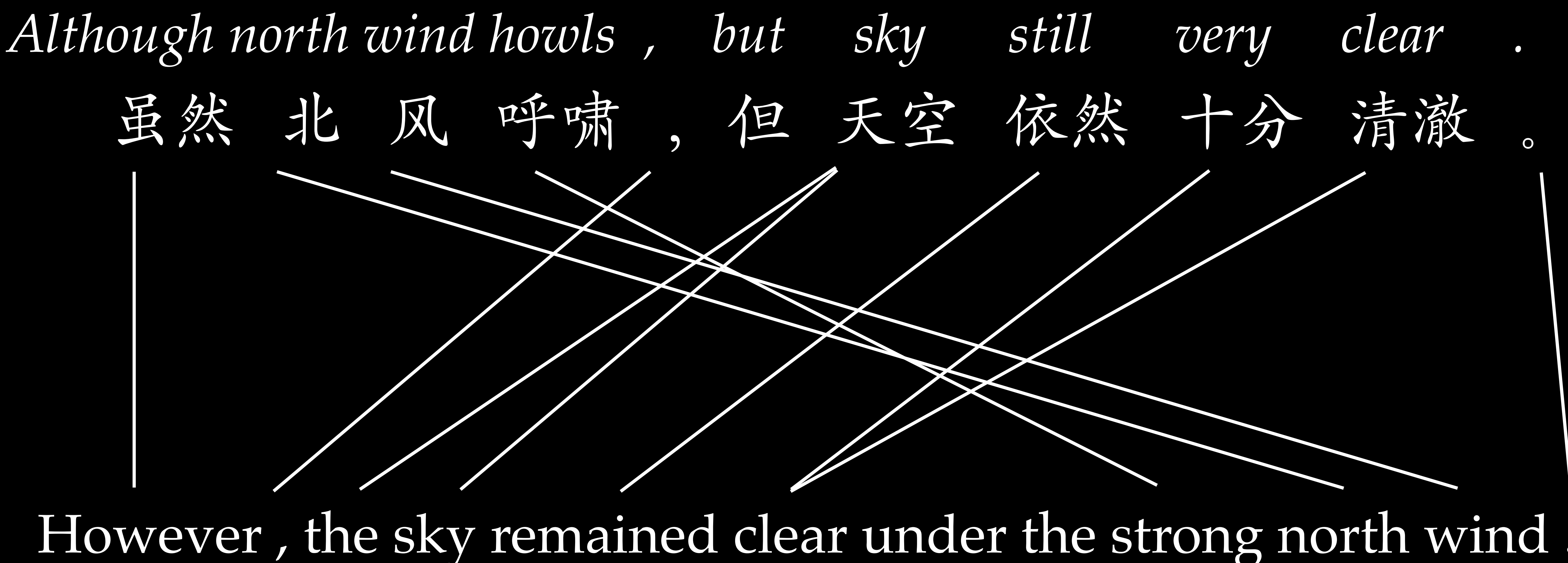
MLE for IBM Model 1 (observed)

$$\Lambda(\theta, \lambda) = \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e)$$

$$- \sum_{e} \lambda_e \left(\sum_{f} p(f|e) - 1 \right)$$

derivative $$\frac{\partial \Lambda(\theta, \lambda)}{\partial p(f|e)} = \frac{count(\langle f, e \rangle)}{p(f|e)} - \lambda_e$$

MLE for IBM Model 1 (observed)



$$p(however|\text{虽然}) = \frac{\#\text{ of times 虽然 aligns to However}}{\#\text{ of times 虽然 aligns to any word}}$$

MLE for IBM Model 1 (unobserved)

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 .

$p(however|$ 虽然 $)$ = ???

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$