EM with Features

Nov. 19, 2013

Word Alignment

das Haus	ein Buch	das Buch
the house	a book	the book

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences

$$\mathbf{e} = \langle e_1, e_2, \ldots, e_m \rangle$$

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences

$\mathbf{e} = \langle e_1, e_2, \ldots, e_m \rangle$ $\mathbf{f} = \langle f_1, f_2, \ldots, f_n \rangle$

Lexical Translation

- Goal: a model $p(\mathbf{e} \mid \mathbf{f}, m)$
- where $\mathbf{e}$ and $\mathbf{f}$ are complete English and Foreign sentences
- Lexical translation makes the following ***assumptions***:
  - Each word in e_i in $\mathbf{e}$ is generated from exactly one word in $\mathbf{f}$
  - Thus, we have an *alignment* a_i that indicates which word e_i "came from", specifically it came from f_{a_i}.
  - Given the alignments $\mathbf{a}$, translation decisions are conditionally independent of each other and depend *only* on the aligned source word f_{a_i}.

IBM Model 1

- Simplest possible lexical translation model
- Additional assumptions
  - The m alignment decisions are independent
  - The alignment distribution for each a_i is uniform over all source words and NULL

$$
\begin{aligned}
&\text{for each } i \in [1, 2, \ldots, m] \\
&\qquad a_i \sim \text{Uniform}(0, 1, 2, \ldots, n) \\
&\qquad e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})
\end{aligned}
$$

IBM Model 1

$$\boxed{\text{for each } i \in [1, 2, \ldots, m]}$$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m}$$

IBM Model 1

$$\text{for each } i \in [1, 2, \ldots, m]$$

$$\boxed{a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)}$$

$$e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n}$$

IBM Model 1

$$\text{for each } i \in [1, 2, \ldots, m]$$
$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$
$$\boxed{e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})}$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n}\, p(e_i \mid f_{a_i})$$

IBM Model 1

$$\text{for each } i \in [1, 2, \ldots, m]$$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$\boxed{e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})}$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n} \, p(e_i \mid f_{a_i})$$

IBM Model 1

$$\text{for each } i \in [1, 2, \ldots, m]$$

$$a_i \sim \text{Uniform}(0, 1, 2, \ldots, n)$$

$$\boxed{e_i \sim \text{Categorical}(\boldsymbol{\theta}_{f_{a_i}})}$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} \frac{1}{1+n} \, p(e_i \mid f_{a_i})$$

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e}, \mathbf{a} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i, a_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

$$p(a, b, c, d) = p(a)p(b)p(c)p(d)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

Recall our independence assumption: all alignment decisions are independent of each other, and given alignments all translation decisions are independent of each other, so **all translation decisions are independent of each other**.

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(\mathbf{e} \mid \mathbf{f}, m) = \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m)$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$\begin{aligned} p(\mathbf{e} \mid \mathbf{f}, m) &= \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m) \\ &= \prod_{i=1}^{m} \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i}) \end{aligned}$$

Marginal probability

$$p(e_i, a_i \mid \mathbf{f}, m) = \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$p(e_i \mid \mathbf{f}, m) = \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i})$$

$$\begin{aligned} p(\mathbf{e} \mid \mathbf{f}, m) &= \prod_{i=1}^{m} p(e_i \mid \mathbf{f}, m) \\ &= \prod_{i=1}^{m} \sum_{a_i=0}^{n} \frac{1}{1+n} p(e_i \mid f_{a_i}) \\ &= \frac{1}{(1+n)^m} \prod_{i=1}^{m} \sum_{a_i=0}^{n} p(e_i \mid f_{a_i}) \end{aligned}$$

Example

0	1	2	3	4
NULL	das	Haus	ist	klein

___	___	___	___
1	2	3	4

Start with a foreign sentence and a target length.

Example

0	1	2	3	4
NULL	das	Haus	ist	klein

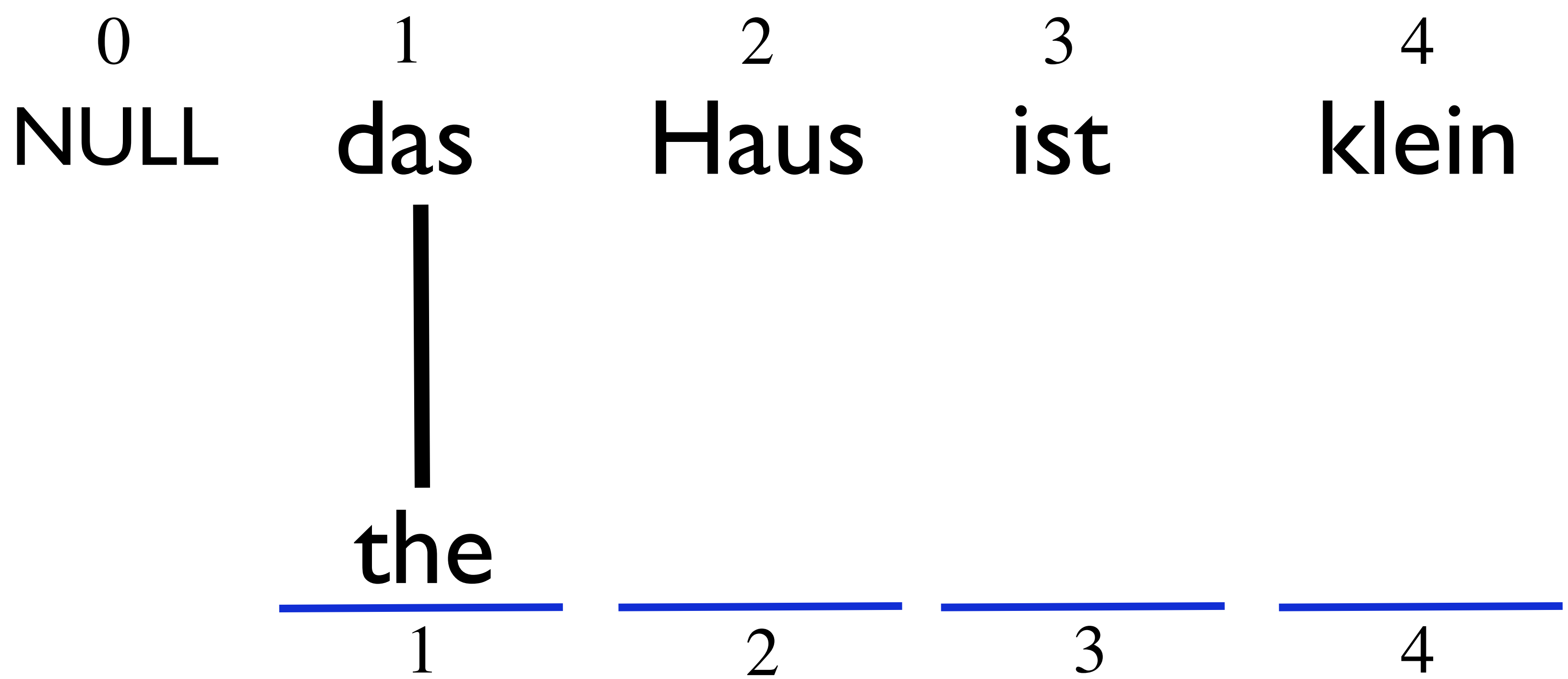
____ ____ ____ ____
1 2 3 4

Example

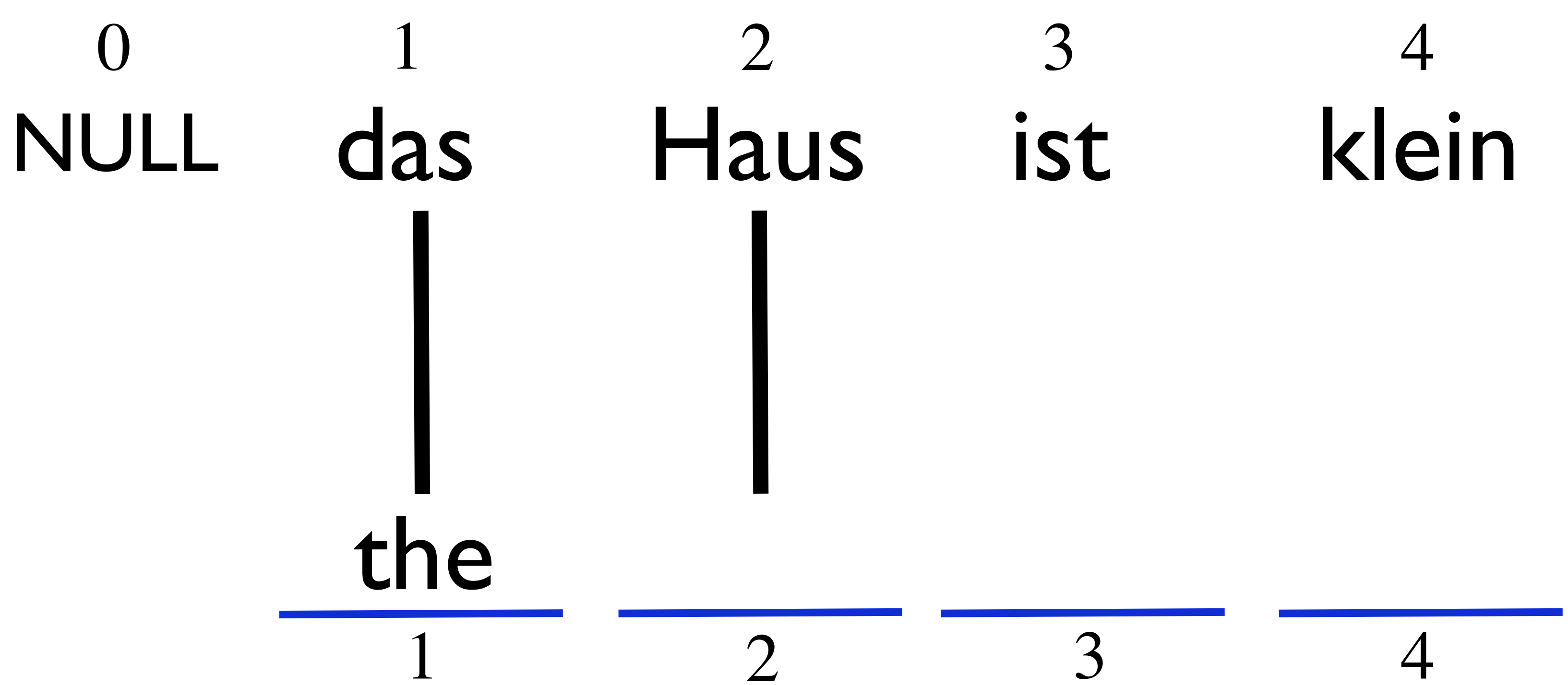
0	1	2	3	4
NULL	das	Haus	ist	klein

1 2 3 4

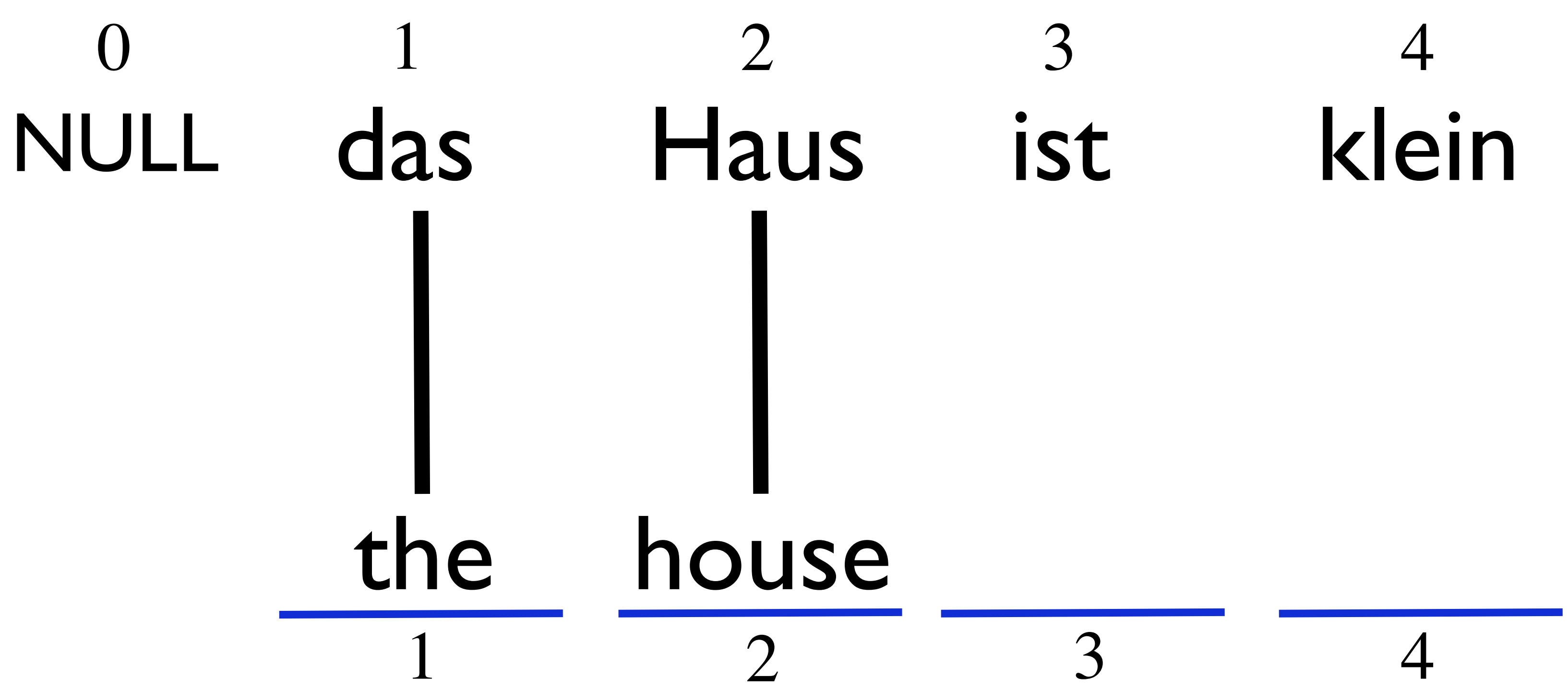
Example



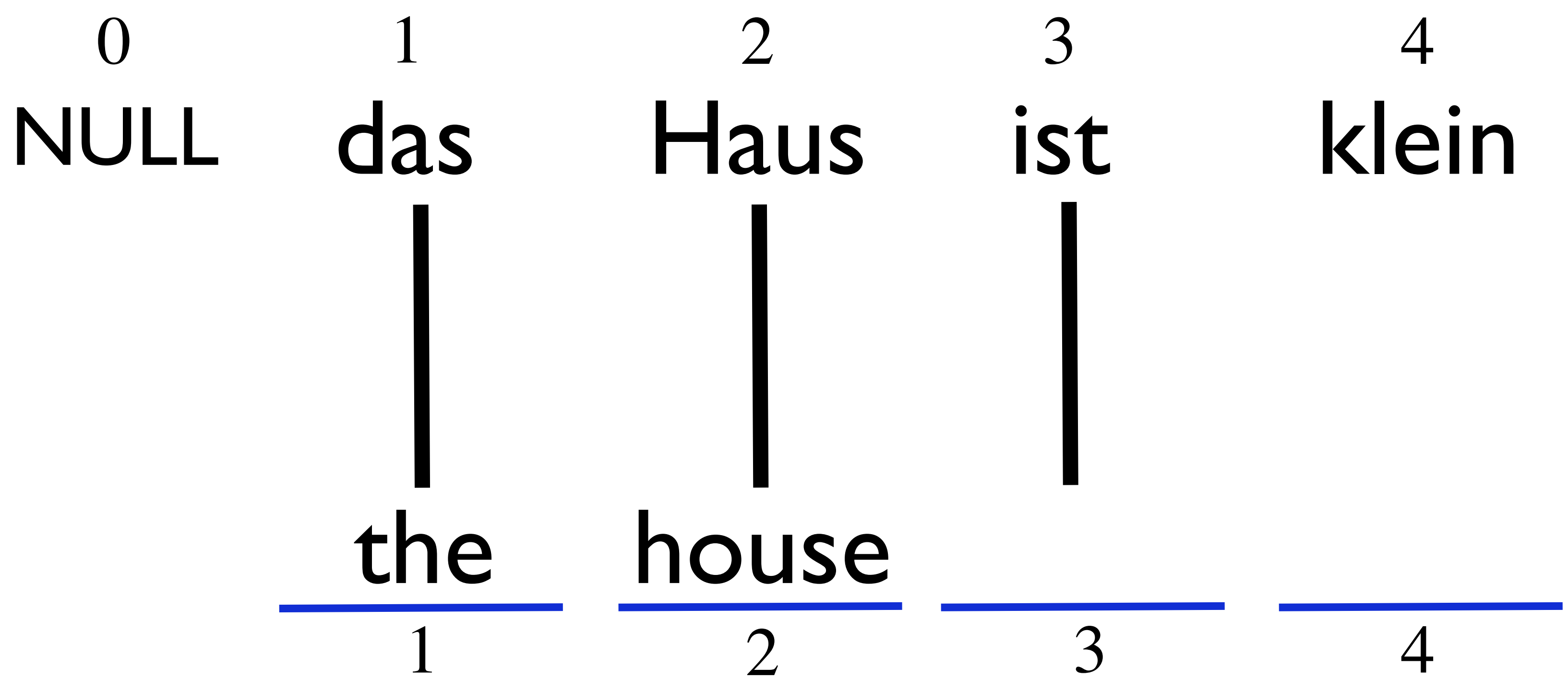
Example



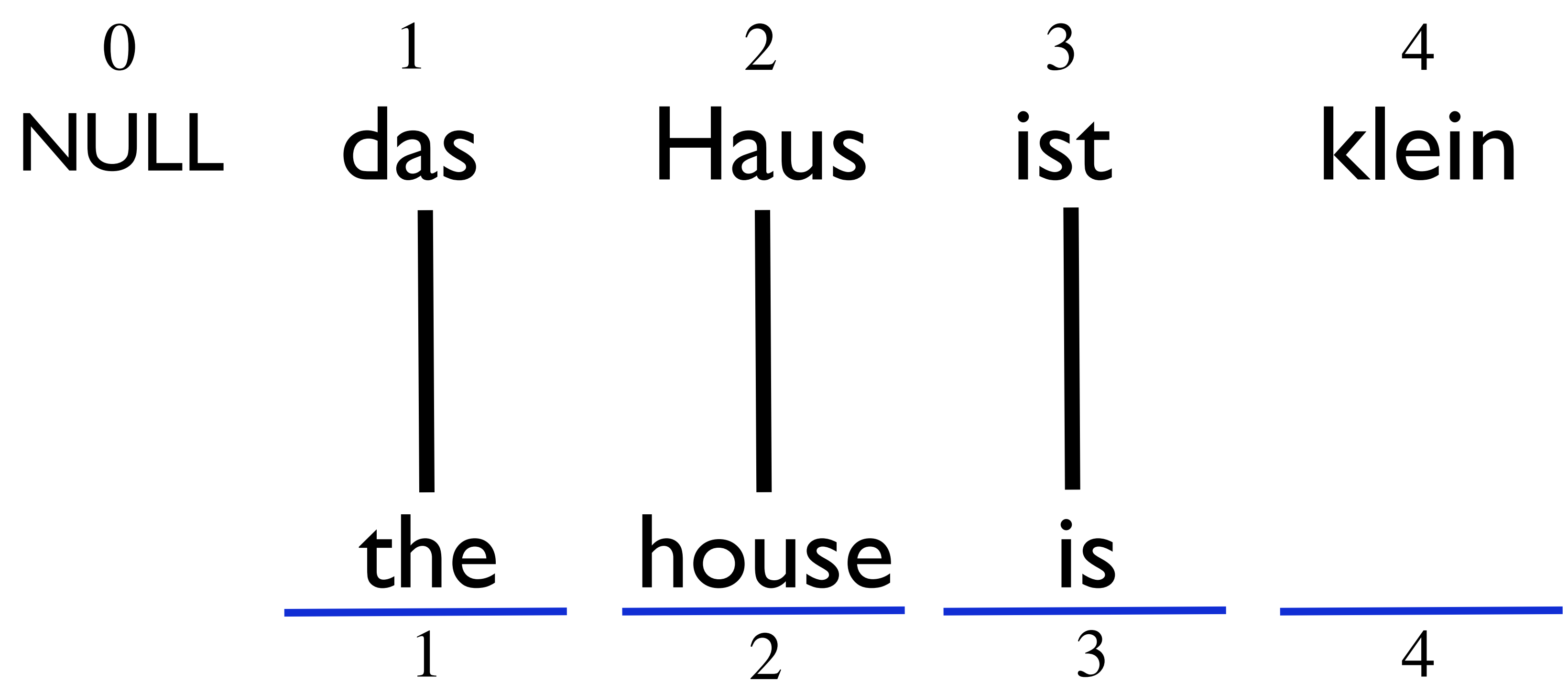
Example



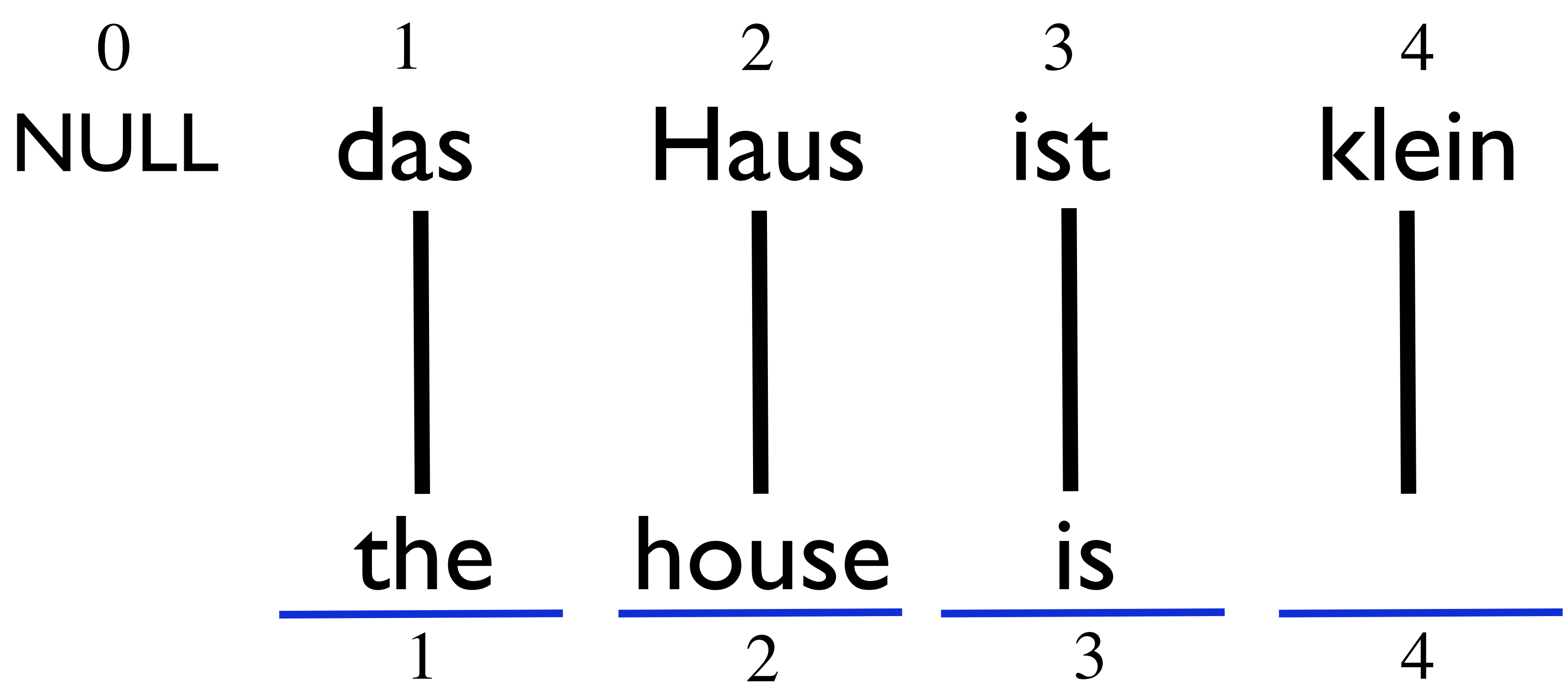
Example



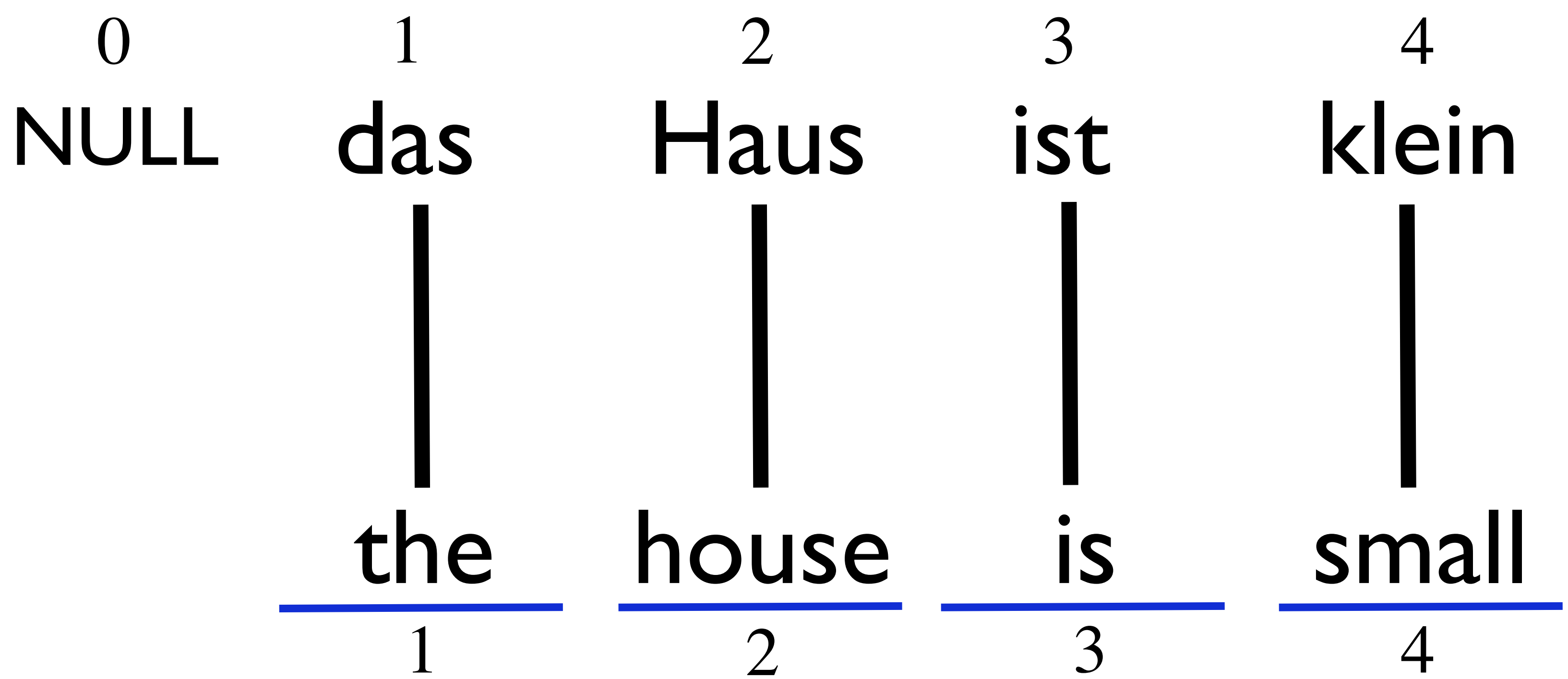
Example



Example



Example



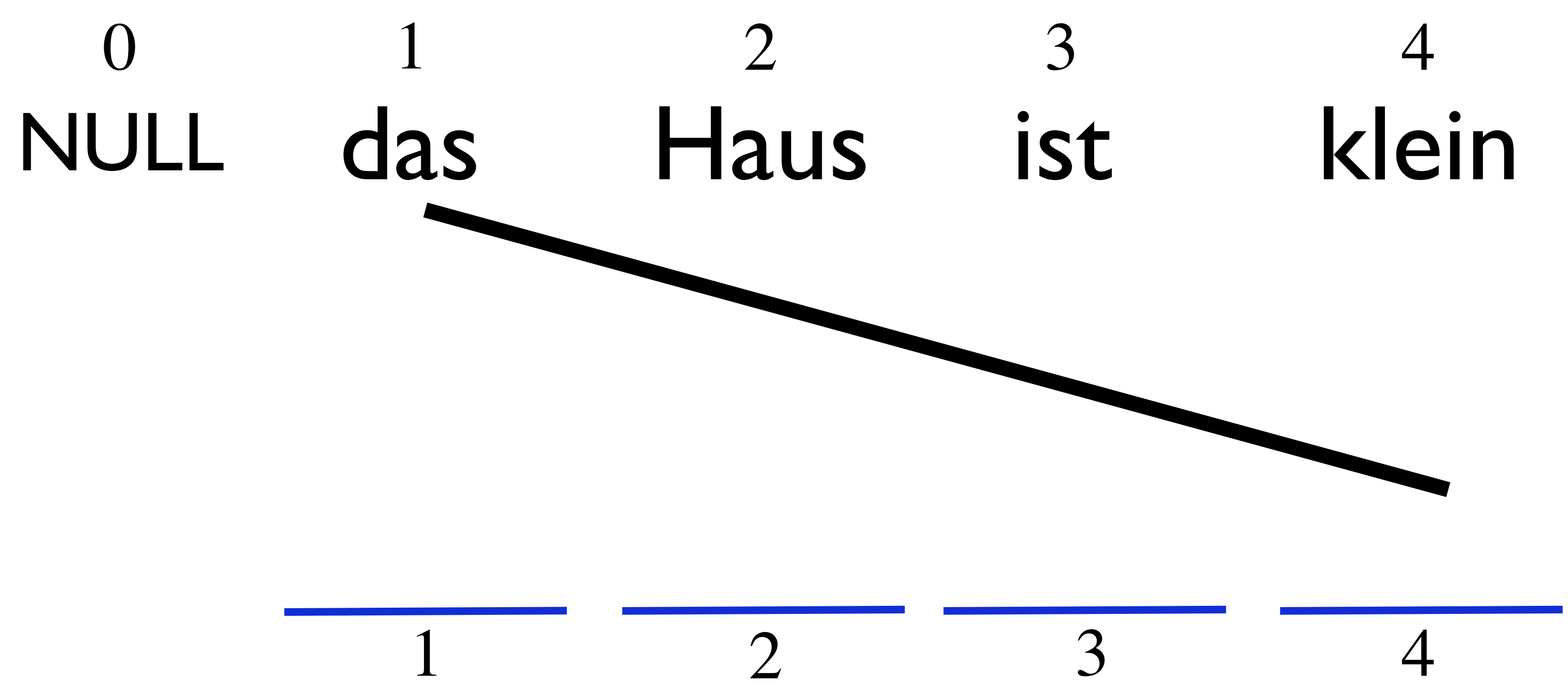
Example

0	1	2	3	4
NULL	das	Haus	ist	klein

____ ____ ____ ____

1	2	3	4

Example

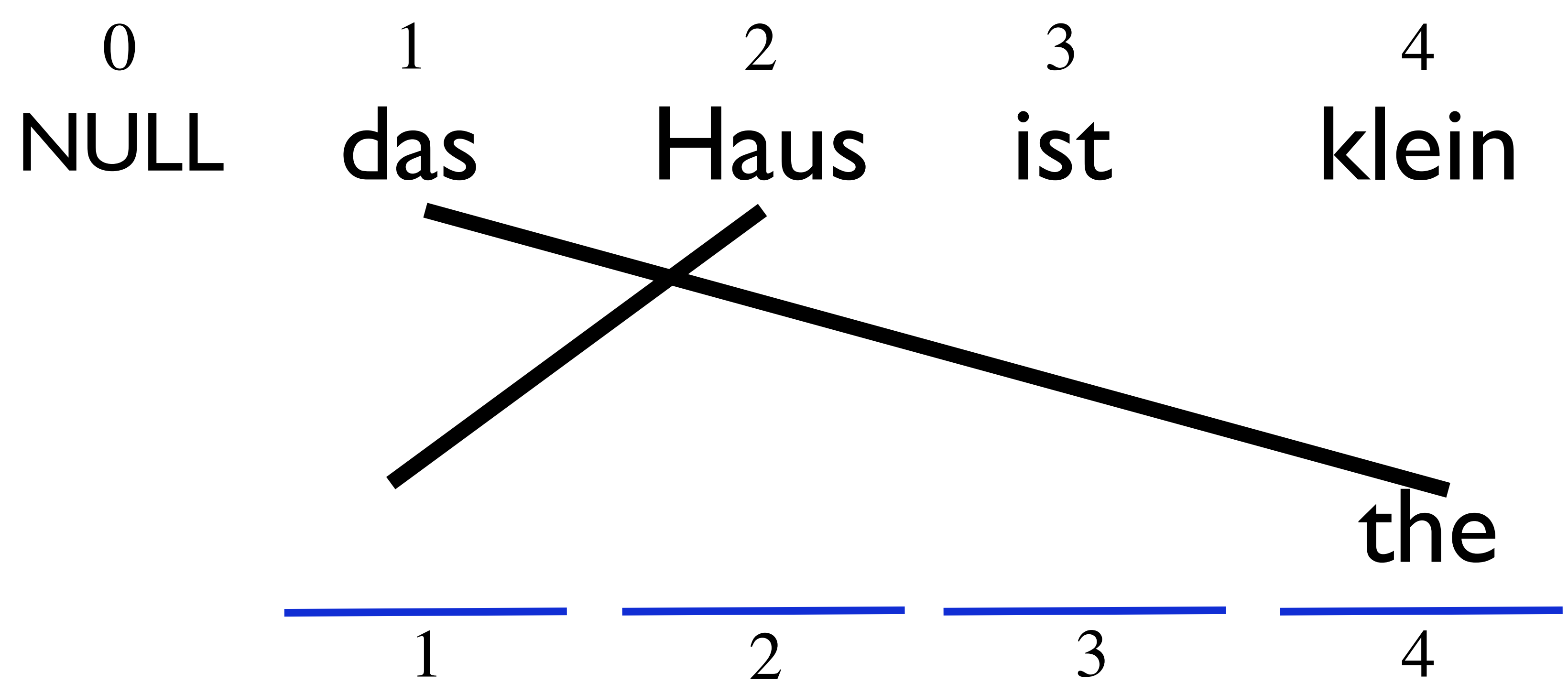


Example

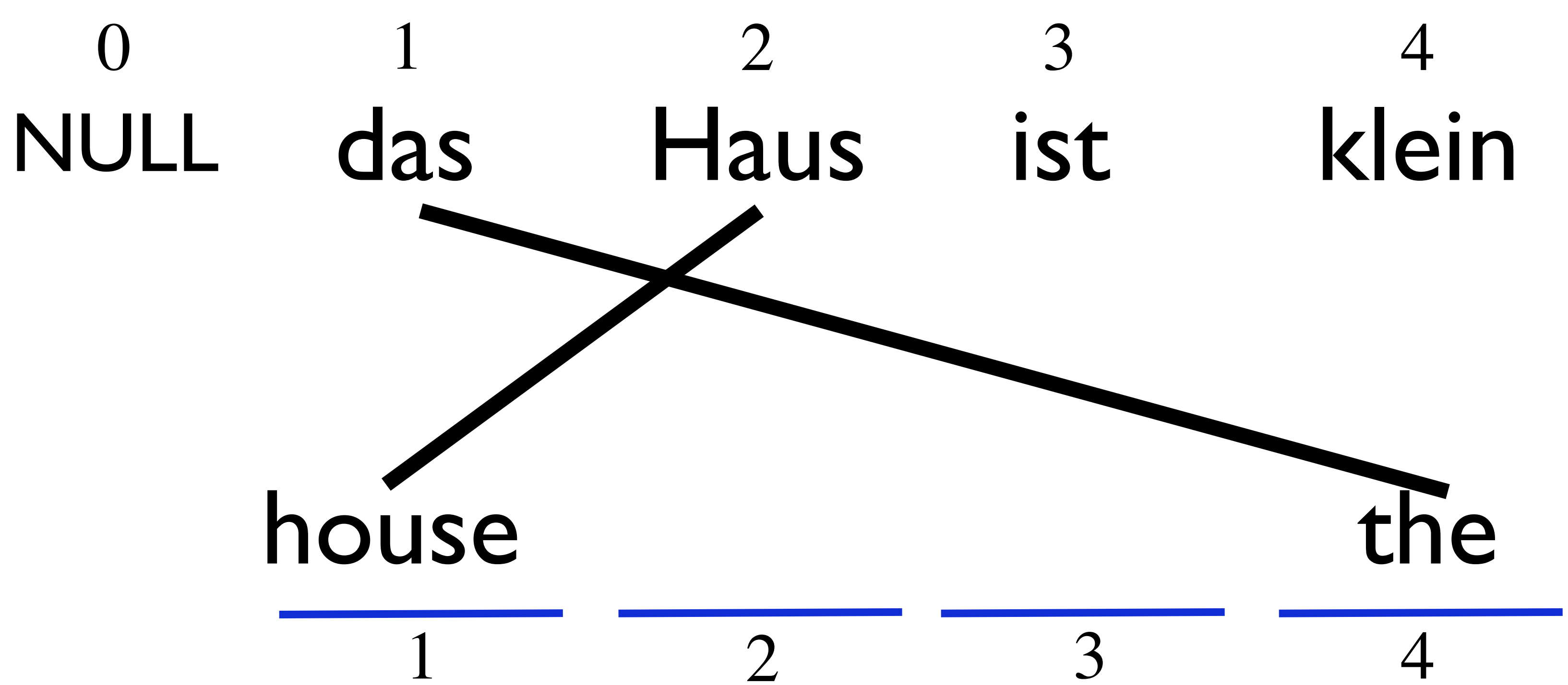
0	1	2	3	4
NULL	das	Haus	ist	klein

			the
___	___	___	___
1	2	3	4

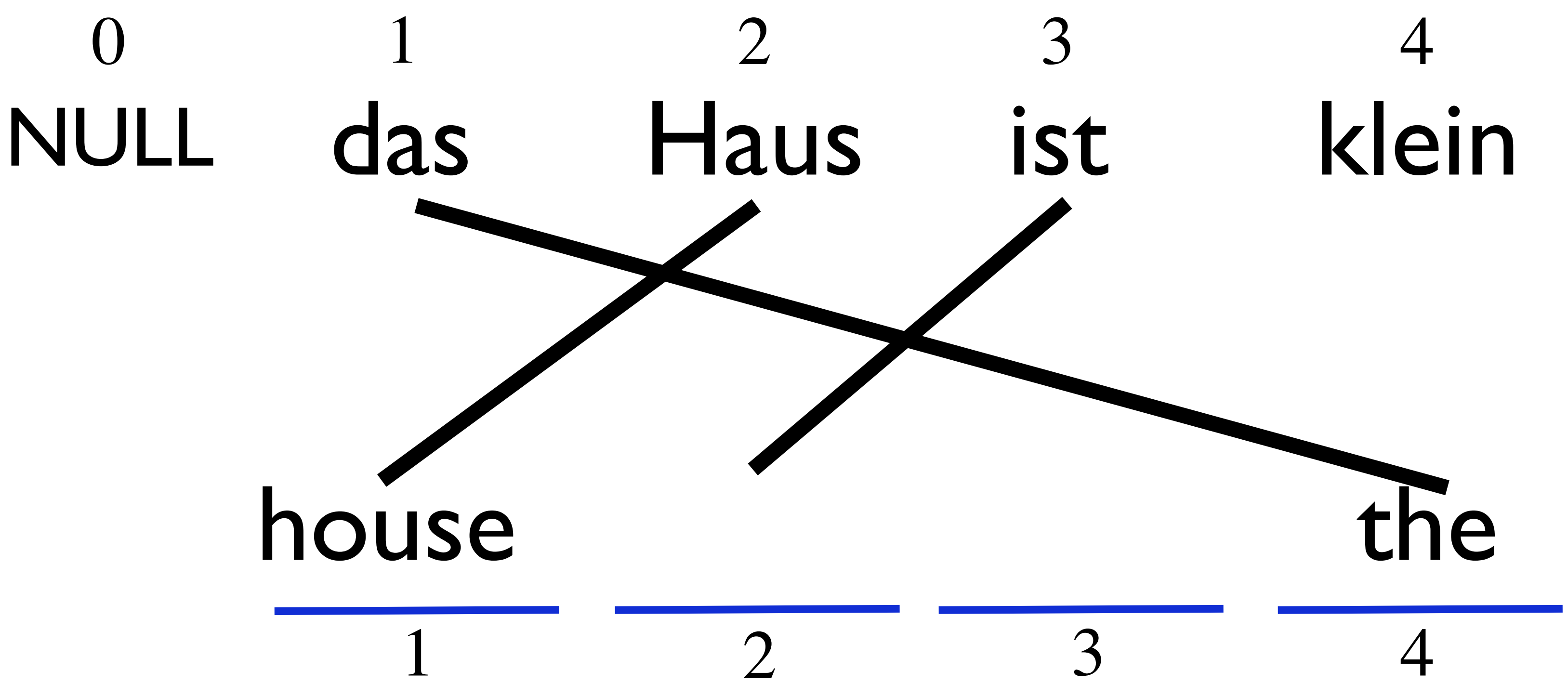
Example



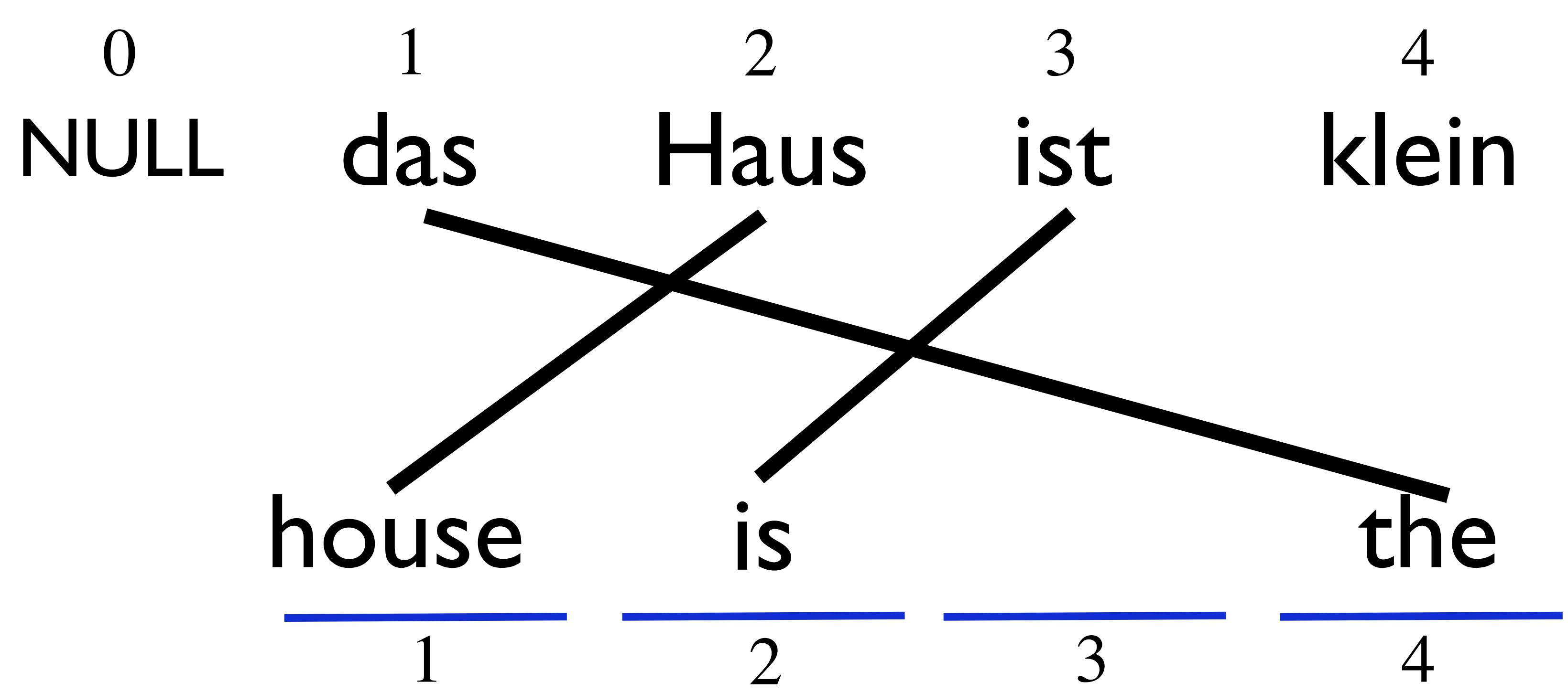
Example



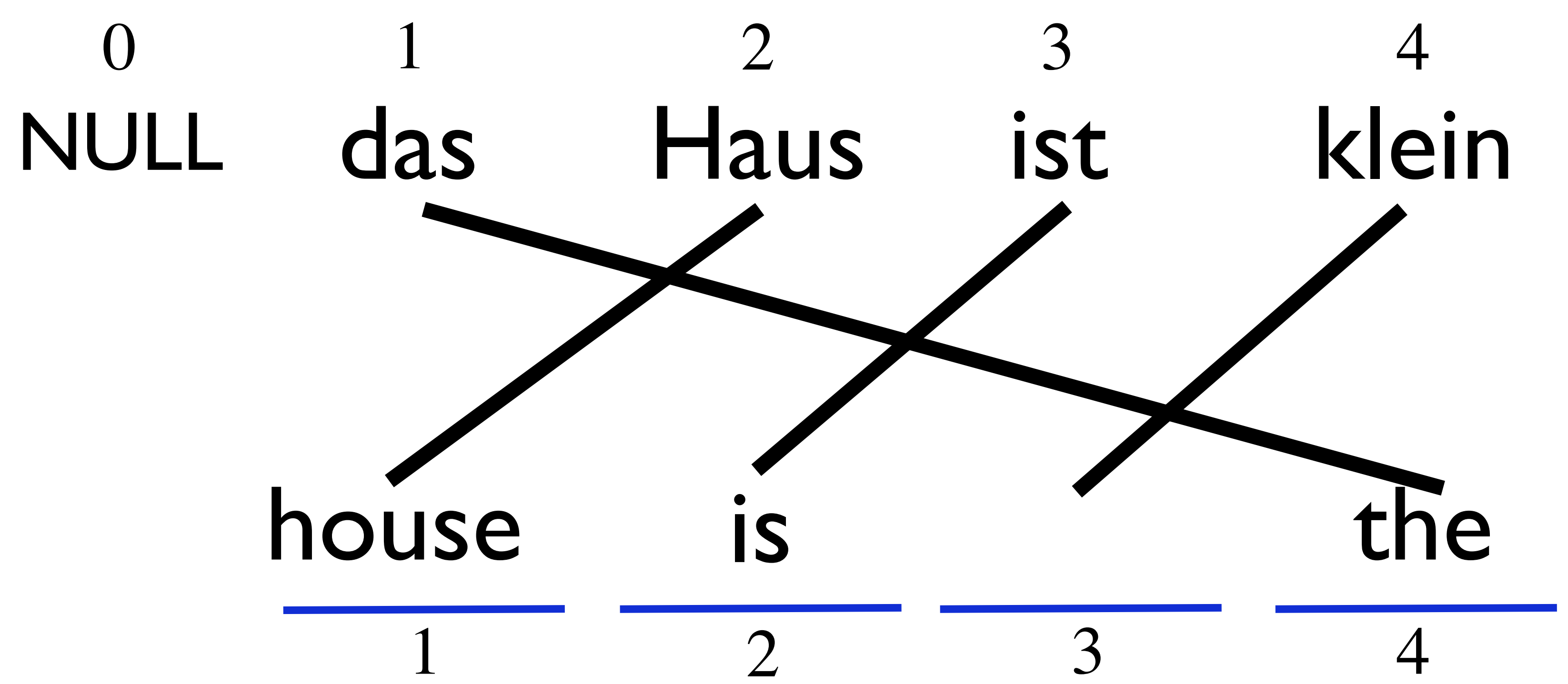
Example



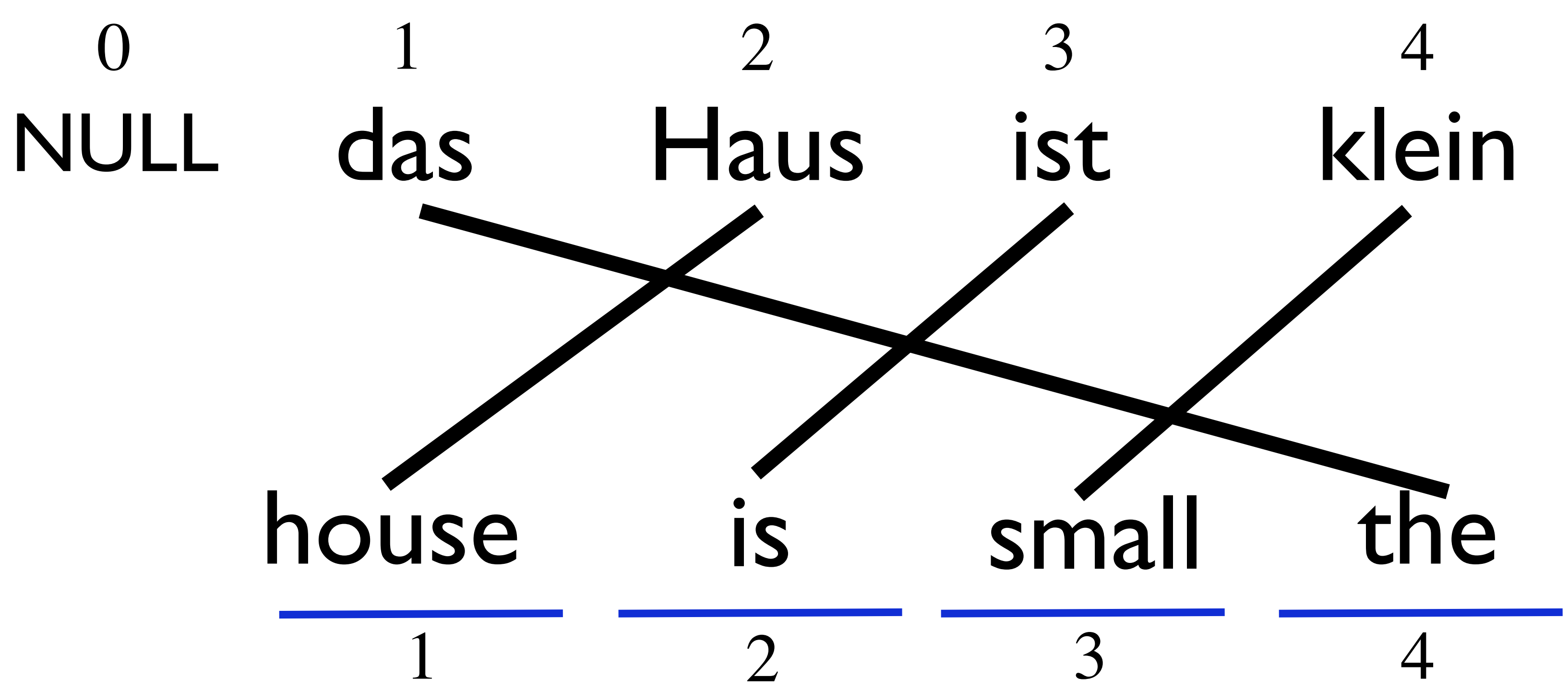
Example

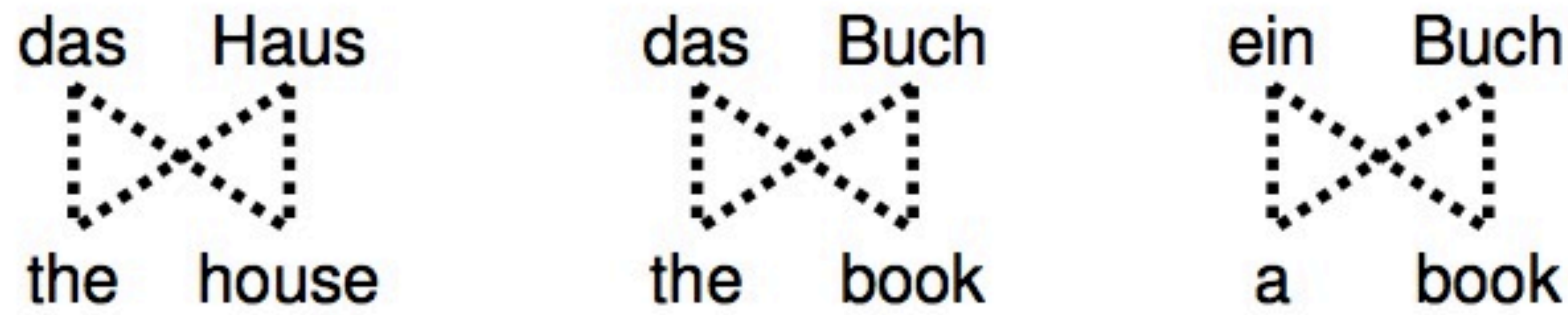


Example



Example





e	f	initial
the	das	0.25
book	das	0.25
house	das	0.25
the	buch	0.25
book	buch	0.25
a	buch	0.25
book	ein	0.25
a	ein	0.25
the	haus	0.25
house	haus	0.25

$$\mathit{freq}(\mathsf{Buch}, \mathsf{book}) = ?$$

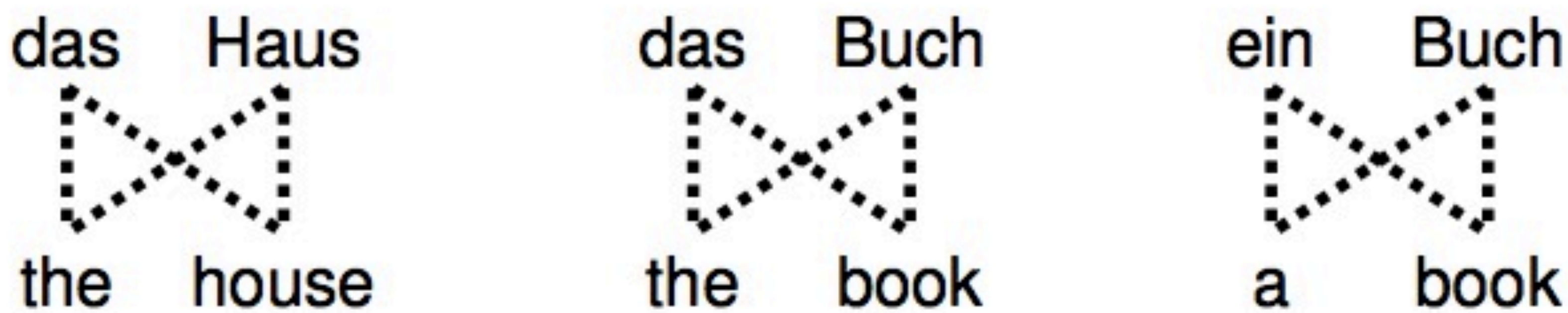
$$\mathit{freq}(\mathsf{das}, \mathsf{book}) = ?$$

$$\mathit{freq}(\mathsf{ein}, \mathsf{book}) = ?$$

$$\mathit{freq}(\mathsf{Buch}, \mathsf{book}) = \sum_i \mathbb{I}(\tilde{e}_i = \mathsf{book}, \tilde{f}_{a_i} = \mathsf{Buch})$$

$$\mathbb{E}_{p_{\mathbf{w}^{(1)}}(\mathbf{a} | \mathbf{f} = \mathsf{das\ Buch}, \mathbf{e} = \mathsf{the\ book})} \sum_i \mathbb{I}[e_i = \mathsf{book}, f_{a_i} = \mathsf{Buch}]$$

Convergence



e	f	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

Evaluation

- Since we have a probabilistic model, we can evaluate **perplexity**.

$$\mathrm{PPL} = 2^{-\frac{1}{\sum_{(\mathbf{e},\mathbf{f})\in\mathcal{D}}|\mathbf{e}|}\log\prod_{(\mathbf{e},\mathbf{f})\in\mathcal{D}} p(\mathbf{e}|\mathbf{f})}$$

	Iter 1	Iter 2	Iter 3	Iter 4	...	Iter ∞
-log likelihood	-	7.66	7.21	6.84	...	-6
perplexity	-	2.42	2.30	2.21	...	2

Hidden Markov Models

- GMMs, the aggregate bigram model, and Model 1 don't have conditional dependencies between random variables
- Let's consider an example of a model where this is not the case

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{y}' \in \mathcal{Y}_{\boldsymbol{x}}} \eta(y_{|\boldsymbol{x}|} \to \text{STOP}) \prod_{i=1}^{|\boldsymbol{x}|} \eta(y_{i-1} \to y_i) \times \gamma(y_i \downarrow x_i)$$

EM for HMMs

- What statistics are sufficient to determine the parameter values?

$$\begin{aligned} \mathit{freq}(q \downarrow x) & \quad \text{How often does } q \text{ emit } x? \\ \mathit{freq}(q \rightarrow r) & \quad \text{How often does } q \text{ transition to } r? \\ \mathit{freq}(q) & \quad \text{How often do we visit } q? \end{aligned}$$

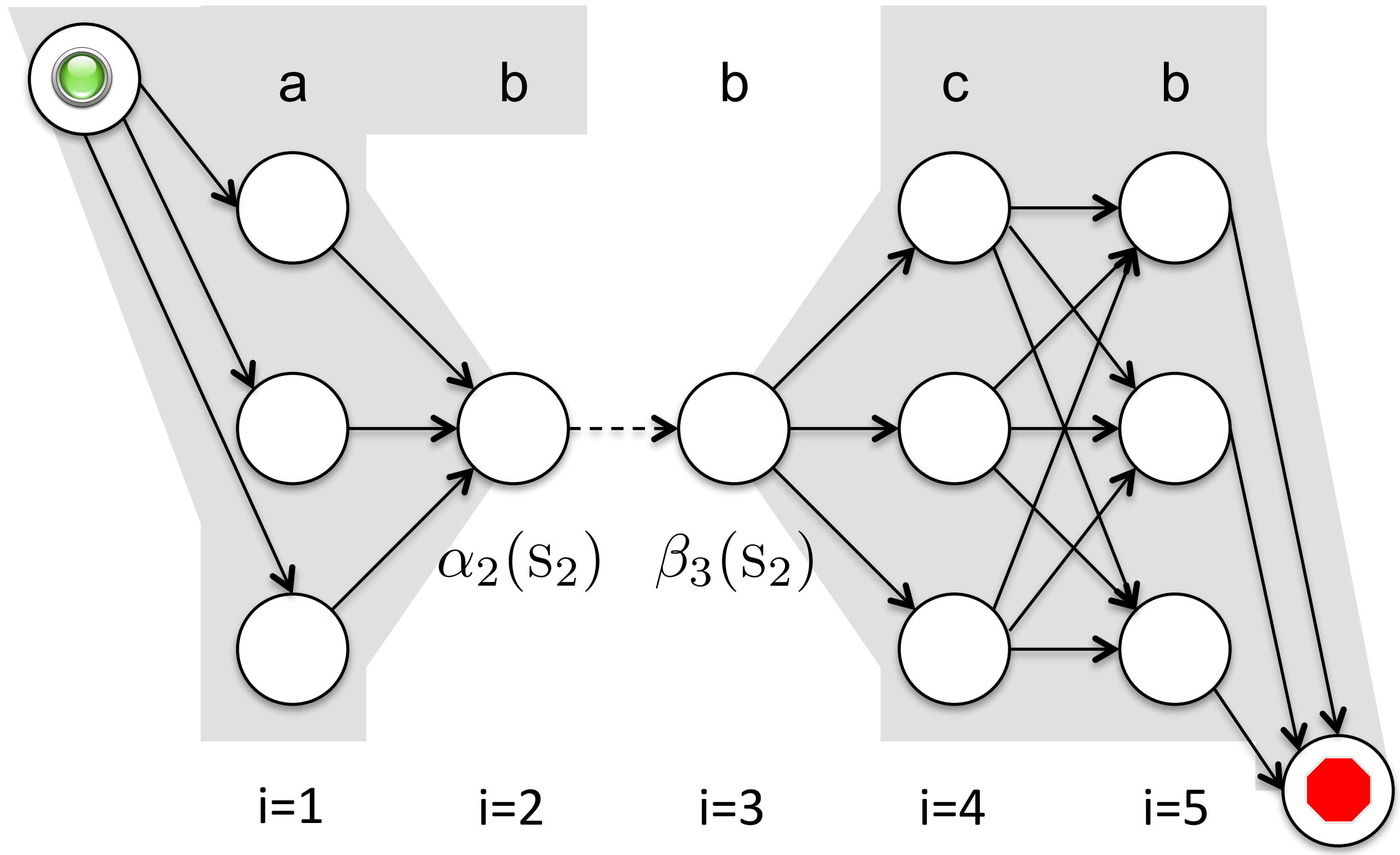
EM for HMMs

- What statistics are sufficient to determine the parameter values?

$$\begin{array}{rl} \mathit{freq}(q \downarrow x) & \text{How often does } q \text{ emit } x? \\ \mathit{freq}(q \to r) & \text{How often does } q \text{ transition to } r? \\ \mathit{freq}(q) & \text{How often do we visit } q? \end{array}$$

And of course...

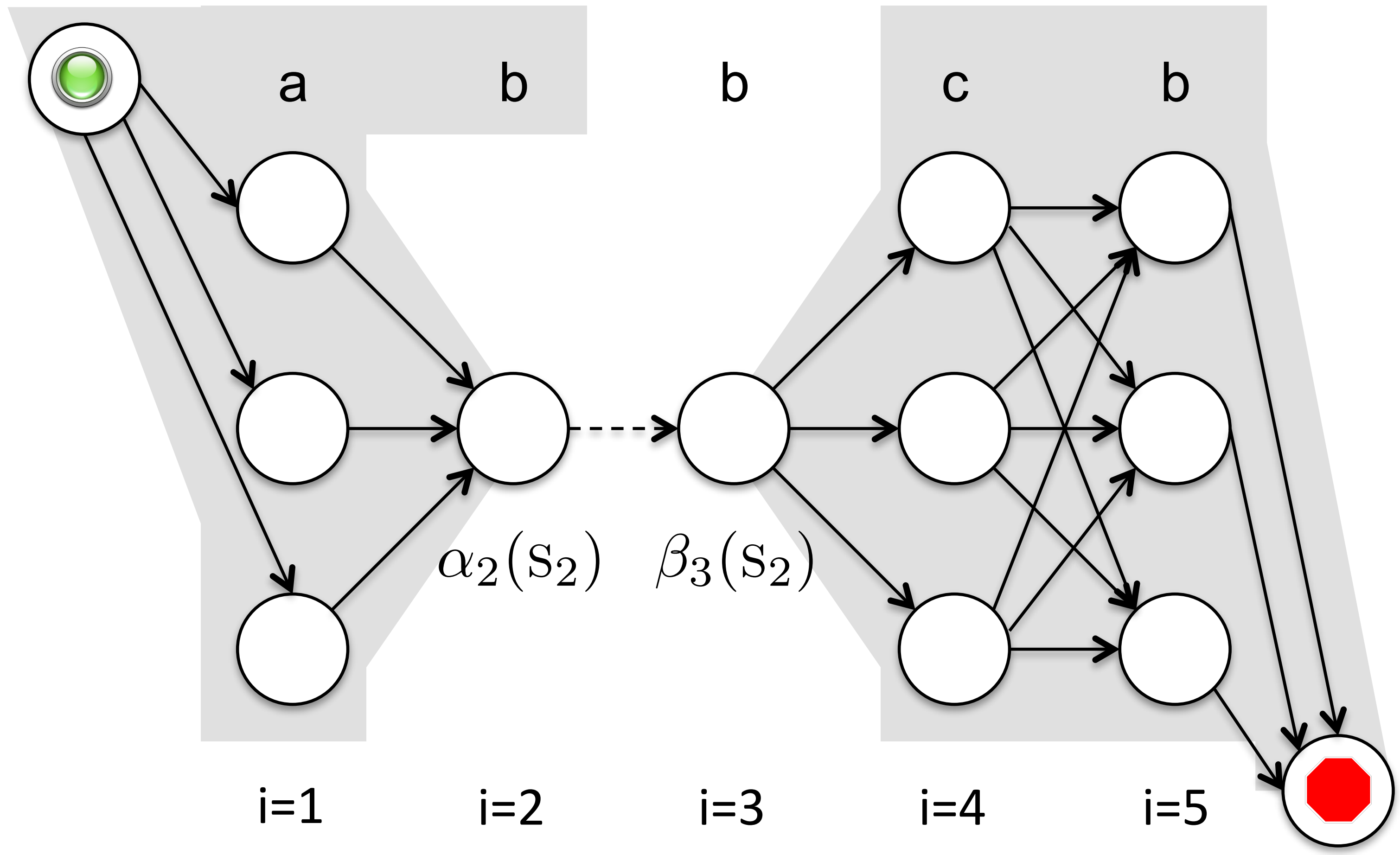
$$\mathit{freq}(q) = \sum_{r \in \mathcal{Q}} \mathit{freq}(q \to r)$$



$$p(y_2 = q, y_3 = r \mid \boldsymbol{x}) \propto p(y_2 = q, y_3 = r, \boldsymbol{x})$$

$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{\sum_{q', r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \to r') \times \eta(r' \downarrow x_3)}$$

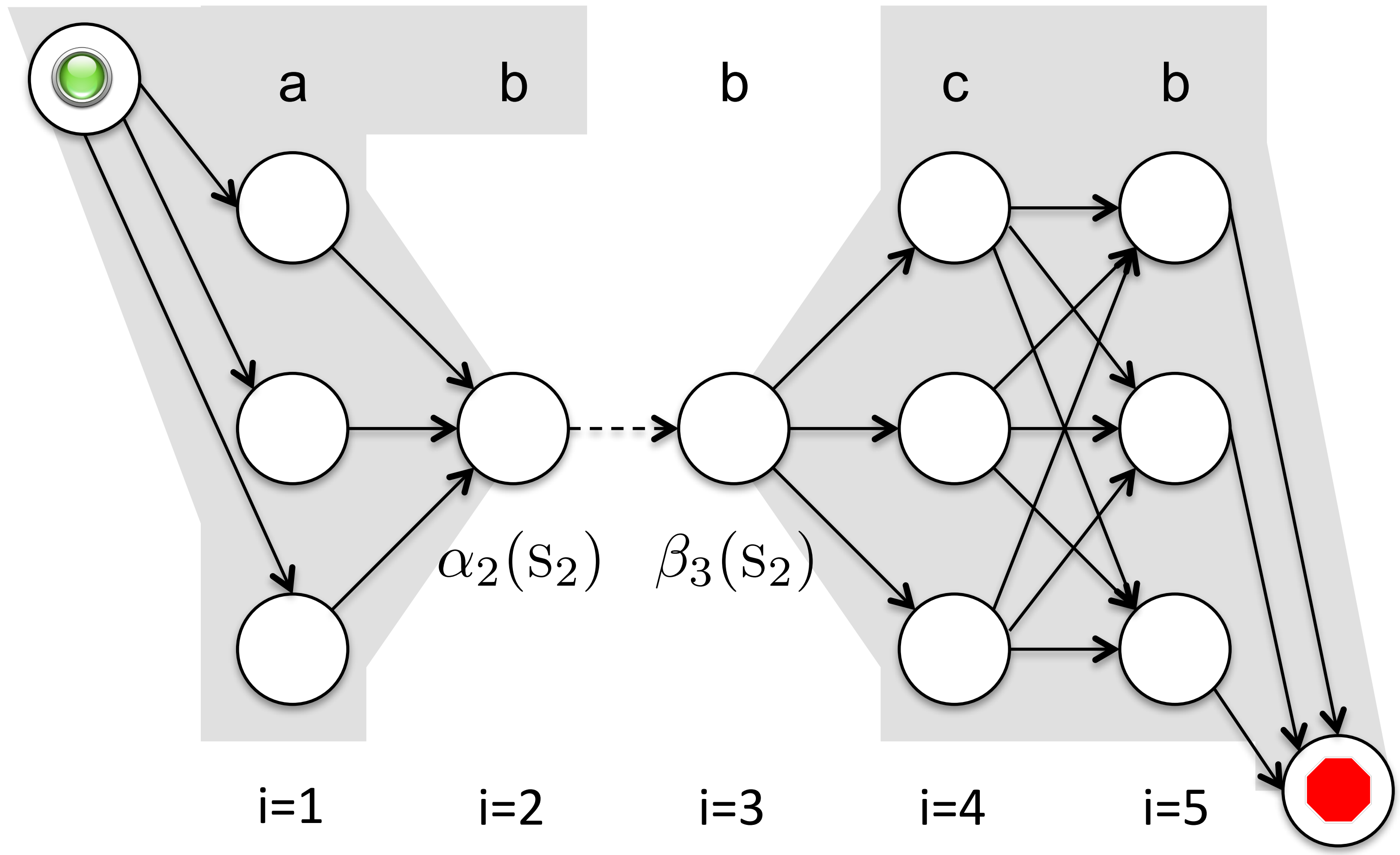
$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\mathrm{STOP})}$$



$$p(y_2 = q, y_3 = r \mid \boldsymbol{x}) \propto p(y_2 = q, y_3 = r, \boldsymbol{x})$$

$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{\sum_{q', r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \to r') \times \eta(r' \downarrow x_3)}$$

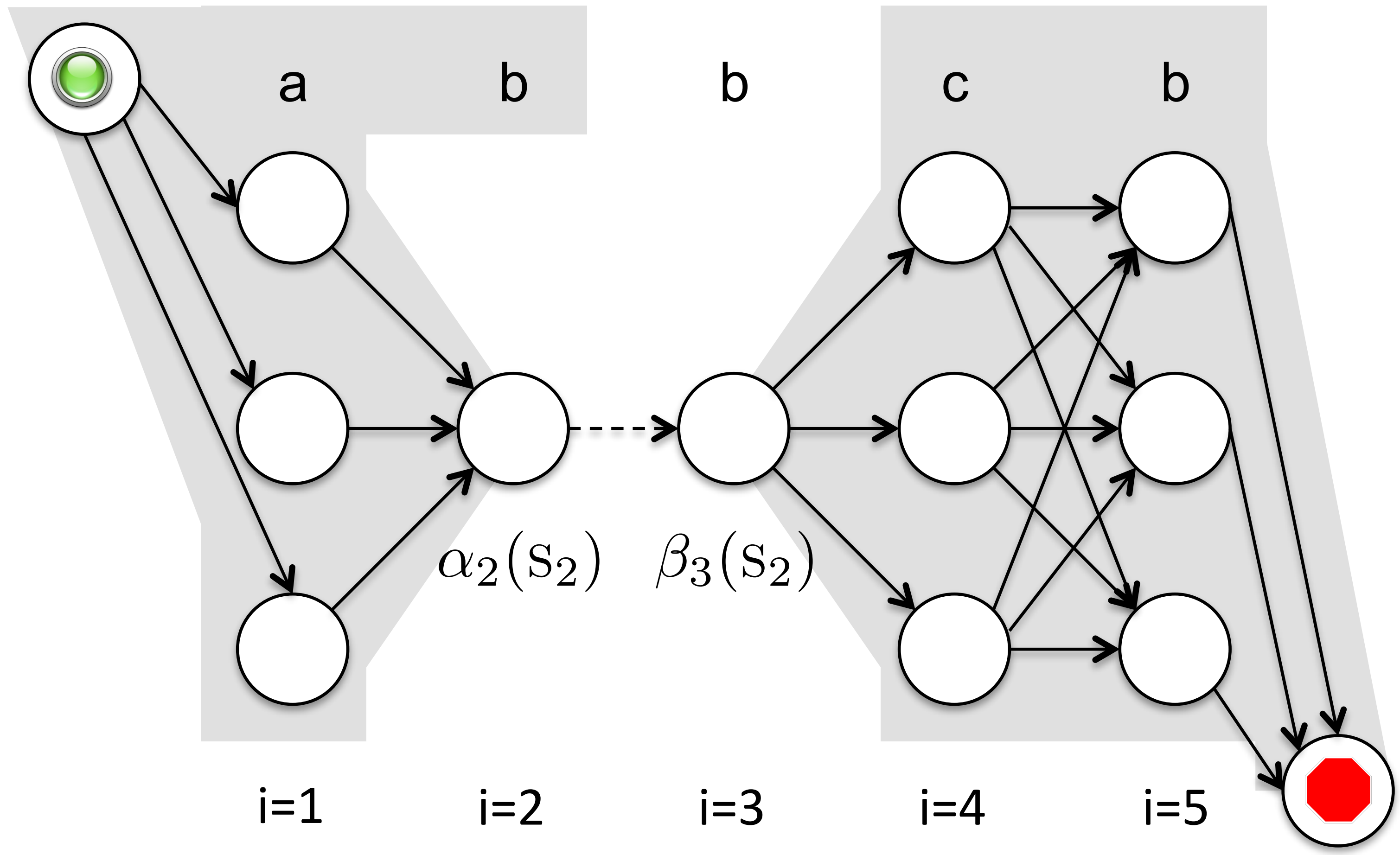
$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\mathrm{STOP})}$$



$$p(y_2 = q, y_3 = r \mid \boldsymbol{x}) \propto p(y_2 = q, y_3 = r, \boldsymbol{x})$$

$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{\sum_{q', r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \to r') \times \eta(r' \downarrow x_3)}$$

$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\mathrm{STOP})}$$



$$p(y_2 = q, y_3 = r \mid \boldsymbol{x}) \propto p(y_2 = q, y_3 = r, \boldsymbol{x})$$

$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{\sum_{q', r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \to r') \times \eta(r' \downarrow x_3)}$$

$$= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\mathrm{STOP})}$$

$$
\begin{aligned}
p(y_2 = q, y_3 = r \mid \boldsymbol{x}) &\propto p(y_2 = q, y_3 = r, \boldsymbol{x}) \\
&= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{\sum_{q', r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \to r') \times \eta(r' \downarrow x_3)} \\
&= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\text{STOP})}
\end{aligned}
$$

The expectation over the full structure is then

$$
\mathbb{E}[freq(q \to r)] = \sum_{i=1}^{|\boldsymbol{x}|} p(y_i = q, y_{i+1} = r \mid \boldsymbol{x})
$$

$$
\begin{aligned}
p(y_2 = q, y_3 = r \mid \boldsymbol{x}) \propto p(y_2 = q, y_3 = r, \boldsymbol{x}) \\
&= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{\sum_{q',r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \to r') \times \eta(r' \downarrow x_3)} \\
&= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\text{STOP})}
\end{aligned}
$$

The expectation over the full structure is then

$$
\mathbb{E}[\mathit{freq}(q \to r)] = \sum_{i=1}^{|\boldsymbol{x}|} p(y_i = q, y_{i+1} = r \mid \boldsymbol{x})
$$

The expectation over state occupancy is

$$
\mathbb{E}[\mathit{freq}(q)] = \sum_{r \in \mathcal{Q}} \mathbb{E}[\mathit{freq}(q \to r)]
$$

$$
\begin{aligned}
p(y_2 = q, y_3 = r \mid \boldsymbol{x}) &\propto p(y_2 = q, y_3 = r, \boldsymbol{x}) \\
&= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{\sum_{q', r' \in \mathcal{Q}} \alpha_2(q') \times \beta_3(r') \times \eta(q' \to r') \times \eta(r' \downarrow x_3)} \\
&= \frac{\alpha_2(q) \times \beta_3(r) \times \eta(q \to r) \times \eta(r \downarrow x_3)}{p(\boldsymbol{x}) = \alpha_{|\boldsymbol{x}|}(\text{STOP})}
\end{aligned}
$$

The expectation over the full structure is then

$$
\mathbb{E}[freq(q \to r)] = \sum_{i=1}^{|\boldsymbol{x}|} p(y_i = q, y_{i+1} = r \mid \boldsymbol{x})
$$

The expectation over state occupancy is

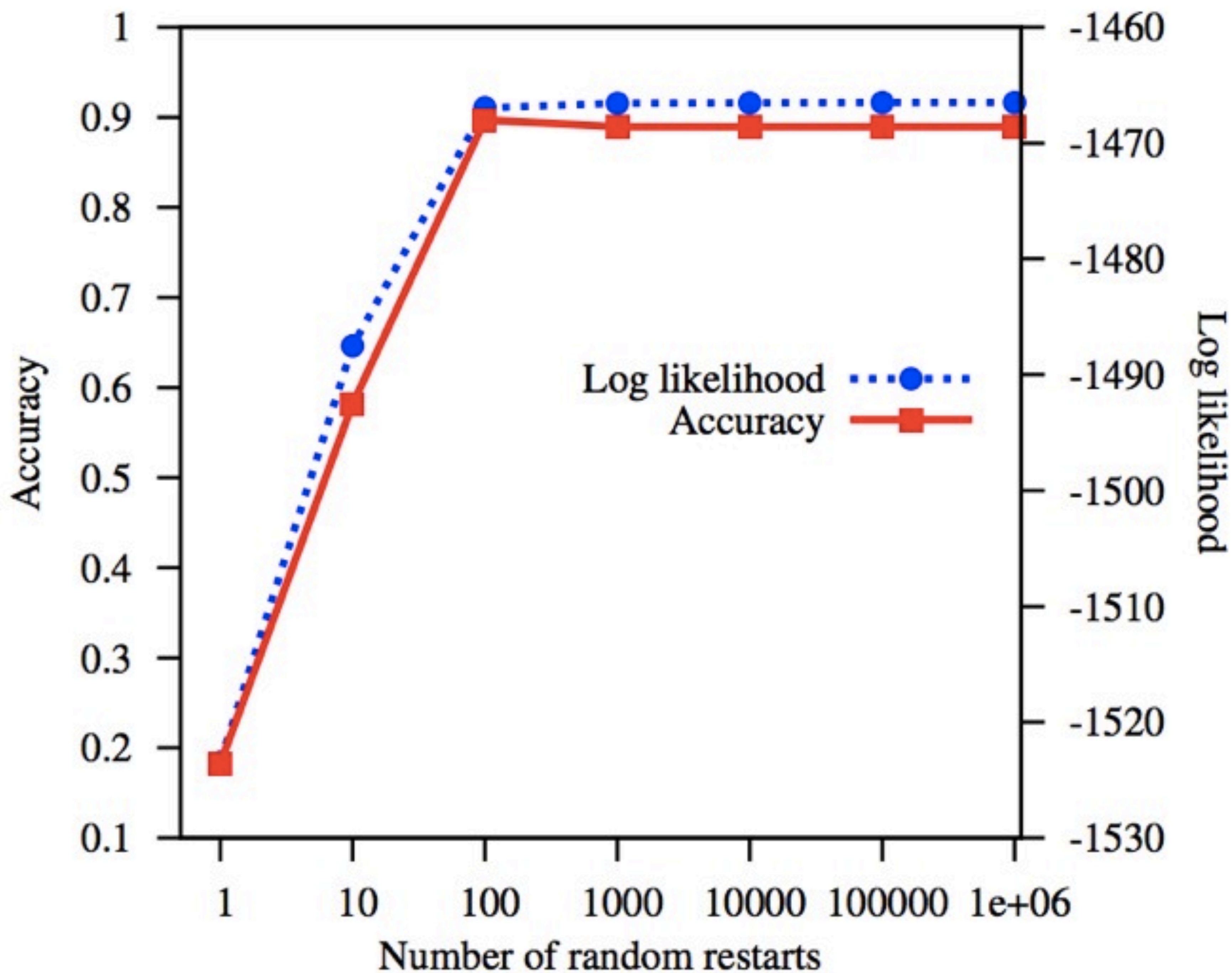
$$
\mathbb{E}[freq(q)] = \sum_{r \in \mathcal{Q}} \mathbb{E}[freq(q \to r)]
$$

What is $\mathbb{E}[freq(q \downarrow x)]$?

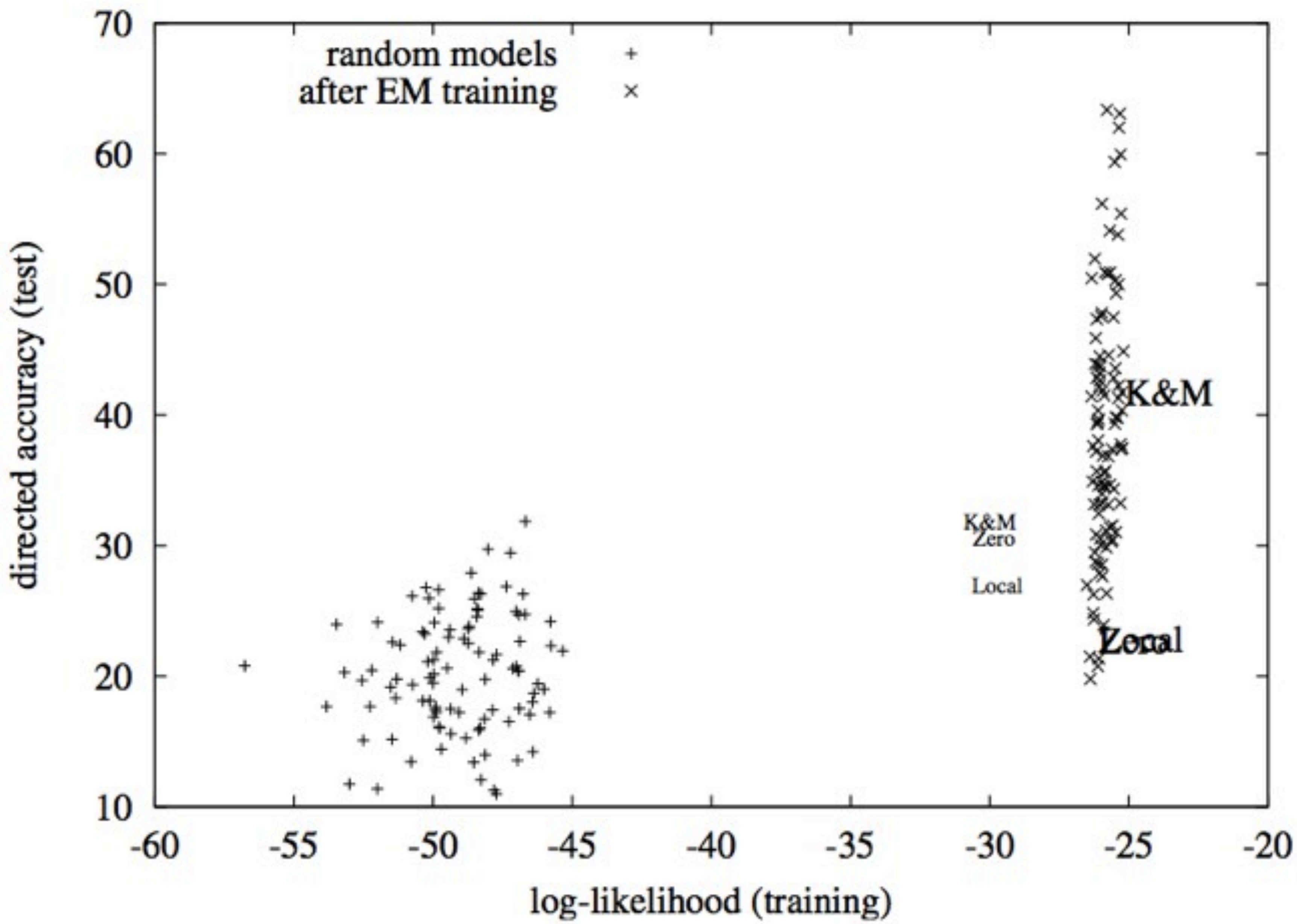
Random Restarts

- Non-convex optimization only finds a local solution
- Several strategies
  - Random restarts
  - Simulated annealing

Decipherment



Grammar Induction



Inductive Bias

- A model can learn nothing without inductive bias ... whence inductive bias?
  - Model structure
  - Priors (next week)
  - Posterior regularization (Google it)
- **Features** provide a very flexible means to bias a model

EM with Features

- Let's replace the multinomials with log linear distributions

$$\eta(q \to r) = \theta_{q,r}$$

$$= \frac{\exp \mathbf{w}^\top \boldsymbol{f}(q,r)}{\sum_{q' \in \mathcal{Q}} \exp \mathbf{w}^\top \boldsymbol{f}(q',r)}$$

How will the likelihood of this model compare to the likelihood of the previous model?

EM with Features

- Let's replace the multinomials with log linear distributions

$$\eta(q \to r) = \theta_{q,r}$$
$$= \frac{\exp \mathbf{w}^\top \boldsymbol{f}(q, r)}{\sum_{q' \in \mathcal{Q}} \exp \mathbf{w}^\top \boldsymbol{f}(q', r)}$$

How will the likelihood of this model compare to the likelihood of the previous model?

Learning Algorithm I

- E step
  - given model parameters, compute posterior distribution over transitions (states, etc)
  - compute $\mathbb{E}_{q(\boldsymbol{y})} \sum_{q,r} \boldsymbol{f}(q, r)$
  - These are your "empirical" expectations

Learning Algorithm I

- M step
  - The gradient of the expected log likelihood of **x**,**y** under q(**y**) is

$$\nabla \mathbb{E}_{q(\boldsymbol{y})} \log p(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_q \sum_{q,r} \boldsymbol{f}(q,r) - \sum_{q,r} \mathbb{E}_q[freq(q)] \mathbb{E}_{p(r|q;\boldsymbol{w})} \boldsymbol{f}(q,r)$$

  - Use LBFGS or gradient descent to solve