Conditional Random Fields

Jacob Andreas / MIT 6.864 / Spring 2020

Review: Hidden Markov Models

Part-of-speech tagging

Noun	**Verb**	**Noun**	**Noun**	**Num**	**Noun**
Fed	*raises*	*interest*	*rates*	*0.5*	*percent*

"The Fed has caused interest rates to get .5% bigger"

[example from Greg Durrett]

Part-of-speech tagging

Noun	**Noun**	**Verb**	**Noun**	**Num**	**Noun**
Fed	*raises*	*interest*	*rates*	*0.5*	*percent*

"Rates are interested (but only 0.5%) in Fed raises" (???)

[example from Greg Durrett]

Part-of-speech tagging

Noun	**Noun**	**Verb**	**Noun**	**Num**	**Noun**
Fed	*raises*	*interest*	*rates*	*0.5*	*percent*

We can’t just guess labels in isolation—need to model sentence context!

[example from Greg Durrett]

Named entity recognition

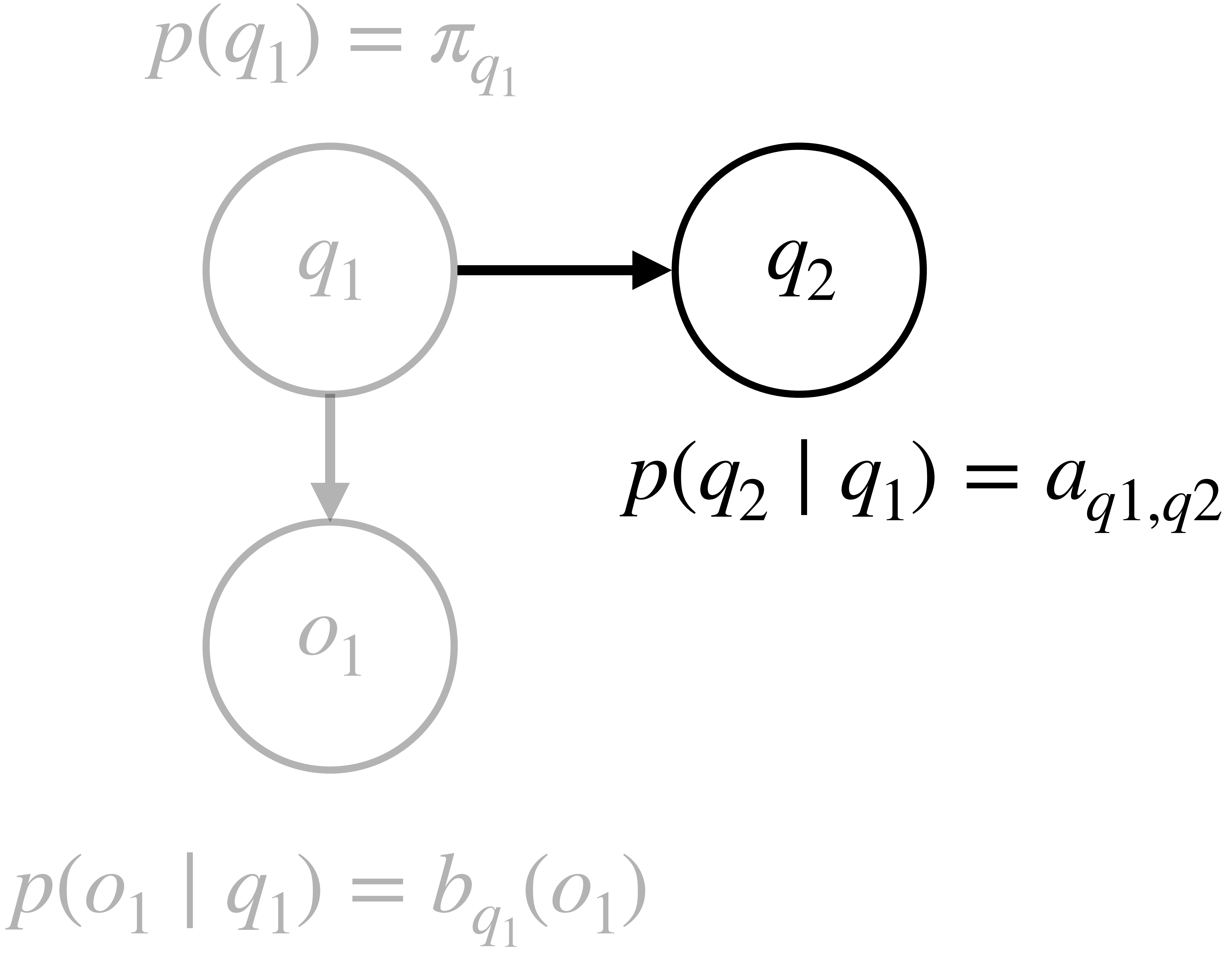
∅	Wake	∅	∅	Action	Arg1	∅	∅	Arg2
hey	*Alexa*	*turn*	*the*	*lights*	*on*	*in*	*the*	*kitchen*

Grammar Induction

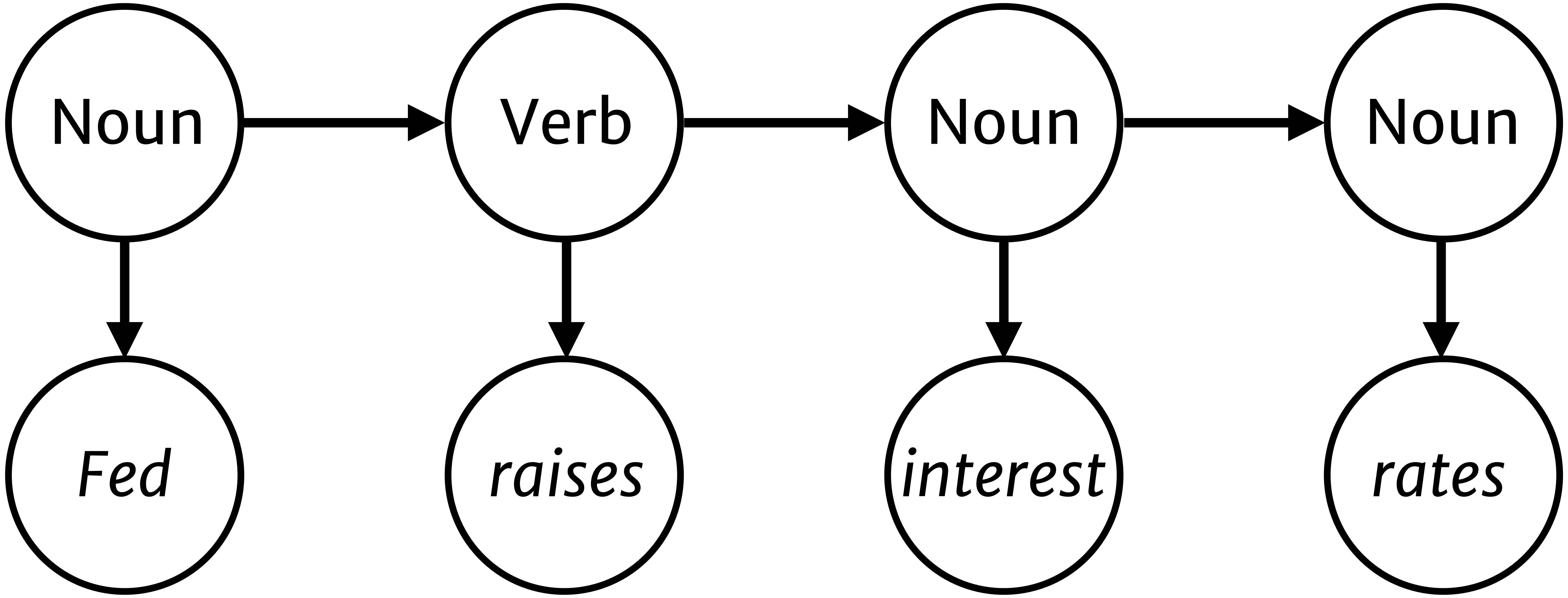
1 2 3 2 3 1 1 3 4 2 5

f84hh4-**l8da4d**-wr-**o40hi**-**eb3**-m8bb-9e8d-**j74**-1e0h3-0i-**0**

HMMs as generative models



HMMs as generative models



HMMs define a joint distribution p(O, Q) over hidden states and observations.

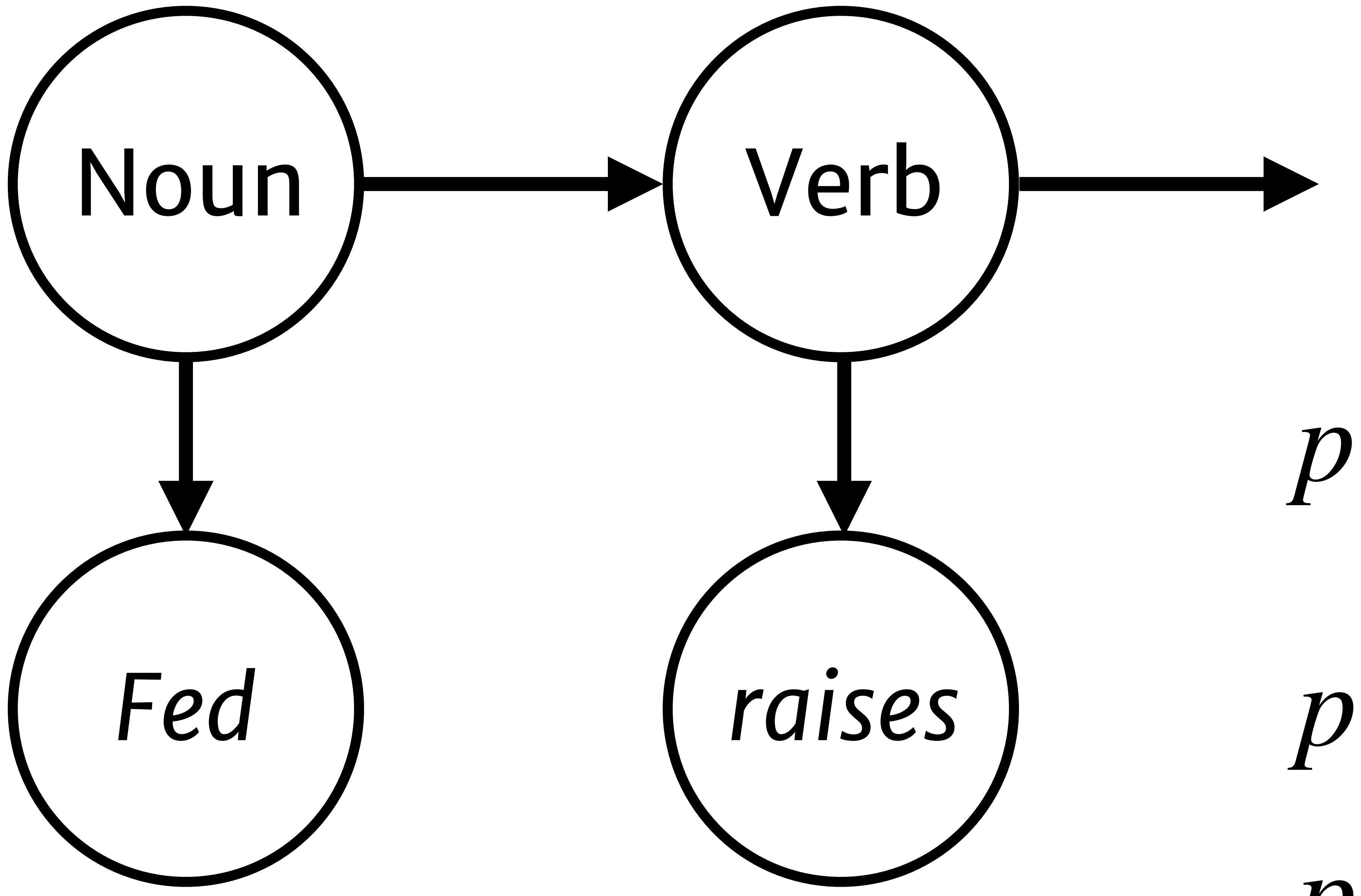
Queries

If we're given the parameters A, B and π, what questions can we answer?

Queries: joint probability

Q1: what is the joint probability of a pair of (observation, tag) sequences?

$$p(O, Q)$$



$$p((\text{Fed, raises}, \ldots), (\text{Noun, Verb}, \ldots)) =$$

$$p(\text{Noun})\ p(\text{Fed} \mid \text{Noun})\ p(\text{Verb} \mid \text{Noun})$$
$$p(\text{raises} \mid \text{Verb})\ \cdots$$

Queries: marginal probability

Q2: what is the **marginal** probability of an observation?

$$p(O)$$

$$p(O) = \sum_{Q} p(O, Q)$$

$\uparrow$

(num tags)$^{\text{(sequence length)}}$
of these!

Queries: marginal probability

Q2: what is the **marginal** probability of an observation?

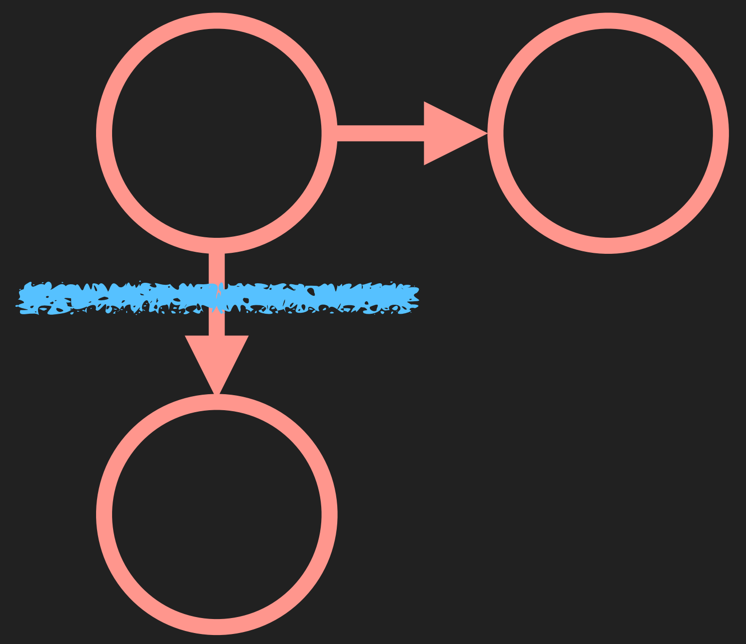
$$p(O)$$

Forward algorithm:
notice that

$$p(O_{:t}, q_t = j) = p(o_t \mid q_t = j) \sum_i p(O_{:t-1}, q_{t-1} = i) p(q_t = j \mid q_{t-1} = i)$$

$$p(O_{:t}, q_t = j) = p(O_{:t-1}, q_t = j)\ p(o_t \mid q_t = j)$$

HMM definition

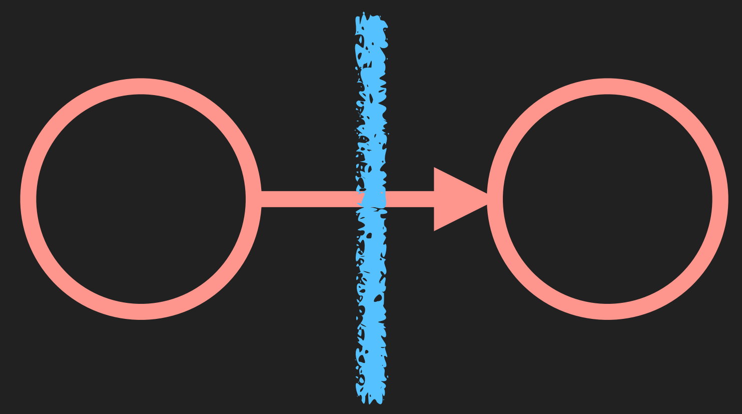


$$= \left(\sum_i p(O_{:t-1}, q_{t-1} = i, q_t = j) \right) p(o_t \mid q_t = j)$$

marginalizing over q_{t-1}

$$= \left(\sum_i p(O_{:t-1}, q_{t-1} = i) p(q_t = j \mid q_{t-1} = i) \right) p(o_t \mid q_t = j)$$

HMM definition



The forward algorithm

Q2: what is the **marginal** probability of an observation?

$$p(O)$$

$$p(O_{:t},\ q_t = j) = p(o_t \mid q_t = j) \sum_i p(O_{:t-1},\ q_{t-1} = i) p(q_t = j \mid q_{t-1} = i)$$

$$\boxed{\begin{aligned} \alpha(t,j) &= b_j(o_t) \sum_i \alpha(t-1,i)\ a_{ij} \\ \alpha(1,j) &= \pi_j\ b_j(o_1) \end{aligned}}$$

← dynamic program!

The forward algorithm

Q2: what is the **marginal** probability of an observation?

$$p(O)$$

Forward algorithm: $\alpha(t,j) = b_j(o_t) \sum_i \alpha(t-1,i)\ a_{ij}$

	1	2	3
Noun	$\pi_{\text{Noun}} b_{\text{Noun}}(\text{Fed})$	$\alpha(1,\text{Noun})$	$\alpha(2,\text{Noun})$
Verb	$\pi_{\text{Verb}} b_{\text{Verb}}(\text{Fed})$	$\alpha(1,\text{Verb})$	
	Fed	*raises*	*interest*

The forward algorithm

Q2: what is the **marginal** probability of an observation?

$$p(O)$$

$$p(O) = \sum_i p(O_{:T}, q_T = i) = \sum_i \alpha(T, i)$$

T := sequence length

The backward algorithm

Q2: what is the **marginal** probability of an observation?

$$p(O)$$

$$p(O_{t+1:} \mid q_t = i) = \sum_j p(q_{t+1} = j \mid q_t = i)\ p(o_{t+1} \mid q_{t+1} = j)\ p(O_{t+2:} \mid q_{t+1} = j)$$

$$\beta(t, i) = \sum_j a_{ij}\ b_j(o_{t+1})\ \beta(t + 1, j)$$

$$\beta(T, i) = 1$$

Same trick!

The forward–backward algorithm

Now we know how to compute:

$$\alpha(t, i) = p(O_{:t},\ q_t = i)$$

$$\beta(t, i) = p(O_{t+1:} \mid\ q_t = i)$$

$$\boxed{\begin{aligned} &\alpha(t, i)\ \beta(t, i) = p(O, q_t = i) \\ &\alpha(t, i)\ a_{i,j}\ b_j(o_{t+1})\ \beta(t+1, j) = p(O, q_t = i, q_{t+1} = j) \end{aligned}}$$

The Viterbi algorithm

Q3: what is the **most probable** assignment of tags to observations?

$$\operatorname{argmax}_Q \; p(O, Q)$$

$$\max_{Q_{t-1:}} \; p(O_{:t},\, Q_{:t-1},\, q_t = j) = \max_i \left(\max_{Q_{t-2:}} \; p(O_{:t-1},\, Q_{t-2:},\, q_{t-1} = i) \right)$$

$$\cdot\, p(q_t = j \mid q_{t-1} = i) \cdot p(o_t \mid q_t = j)$$

$$\max_{Q_{t-1:}} p(O_{:t}, Q_{:t-1}, q_t = j) = \max_{Q_{t-1:}} p(O_{:t-1}, Q_{:t-1}, q_t = j)\, p(o_t \mid q_t = j)$$

HMM definition

$$= \max_{Q_{t-2:},\, i} p(O_{:t-1}, Q_{:t-2}, q_{t-1} = i, q_t = j)\, p(o_t \mid q_t = j)$$

separating $Q_{t-2:}$ and q_{t-1}

$$= \max_{Q_{t-2:},\, i} p(O_{:t-1}, Q_{:t-2}, q_{t-1} = i)\, p(q_t = j \mid q_{t-1} = i)\, p(o_t \mid q_t = j)$$

HMM definition

$$= \max_{i} \left(\max_{Q_{t-2:}} p(O_{:t-1}, Q_{t-2:}, q_{t-1} = i) \right) p(q_t = j \mid q_{t-1} = i)\, p(o_t \mid q_t = j)$$

separating args to max

The Viterbi algorithm

Q3: what is the **most probable** assignment of tags to observations?

$$\operatorname{argmax}_Q \; p(O, Q)$$

$$\max_{Q_{t-1:}} \; p(O_{:t},\, Q_{:t-1},\, q_t = j) = \max_i \left(\max_{Q_{t-2:}} \; p(O_{:t-1},\, Q_{t-2:},\, q_{t-1} = i) \right) \cdot p(q_t = j \mid q_{t-1} = i) \cdot p(o_t \mid q_t = j)$$

The Viterbi algorithm

Q3: what is the **most probable** assignment of tags to observations?

$$\operatorname{argmax}_Q \, p(O, Q)$$

best length-t-1 tag seq. ending in i

$$\boxed{\max_{Q_{t-1:}} p(O_{:t}, Q_{:t-1}, q_t = j)} = \max_i \left(\boxed{\max_{Q_{t-2:}} p(O_{:t-1}, Q_{t-2:}, q_{t-1} = i)} \cdot p(q_t = j \mid q_{t-1} = i) \cdot p(o_t \mid q_t = j) \right)$$

best length-t tag seq. ending in j

The Viterbi algorithm

Q3: what is the **most probable** assignment of tags to observations?

$$\mathrm{argmax}_Q \; p(O, Q)$$

$$\max_{Q_{t-1:}} \; p(O_{:t},\, Q_{:t-1},\, q_t = j) = \max_i \left(\max_{Q_{t-2:}} \; p(O_{:t-1},\, Q_{t-2:},\, q_{t-1} = i) \right) \cdot p(q_t = j \mid q_{t-1} = i) \cdot p(o_t \mid q_t = j)$$

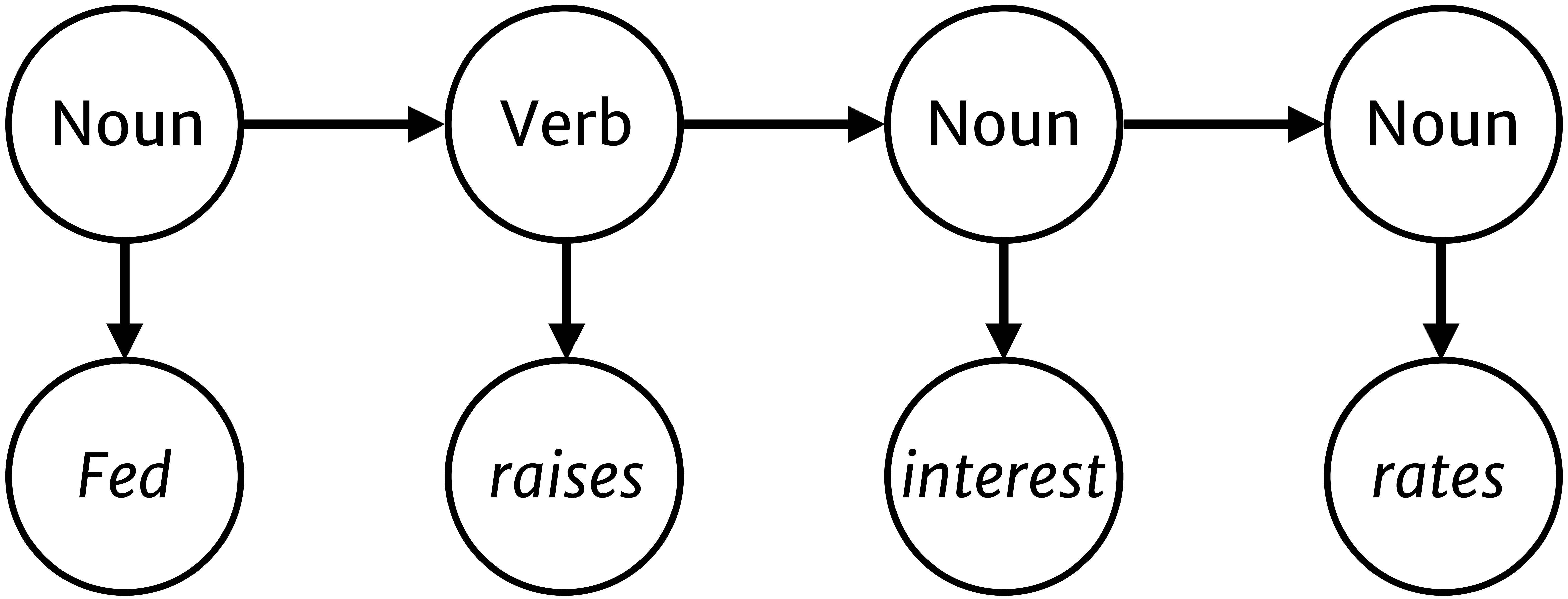
$$\boxed{\delta(t,j) = b_j(o_t) \max_i \delta(t-1,i) \; a_{ij} \qquad \delta(1,j) = \pi(j) \; b_j(o_1)}$$

Supervised training

Where do π, A and B come from?

Supervised training

Where do π, A and B come from?



If we have labeled sequences, just count.

Supervised training

Where do π, A and B come from?

$$\pi_i = p(q_1 = i) = \frac{\#(q_1 = i)}{\#\text{sequences}}$$

$$a_{ij} = p(q_t = j \mid q_{t-1} = i) = \frac{\#(q_{t-1} = i,\ q_t = j)}{\#(q_{t-1} = i, q_t = *\ )}$$

$$b_i(w) = p(o_t = w \mid q_t = i) = \frac{\#(q_t = i,\ o_t = w)}{\#(q_t = i)}$$

If we have labeled sequences, just count.

Unsupervised training

$$a_{ij} = p(q_t = j \mid q_{t-1} = i) = \frac{\#(q_{t-1} = i,\ q_t = j)}{\#(q_{t-1} = i, q_t = *\ )}$$

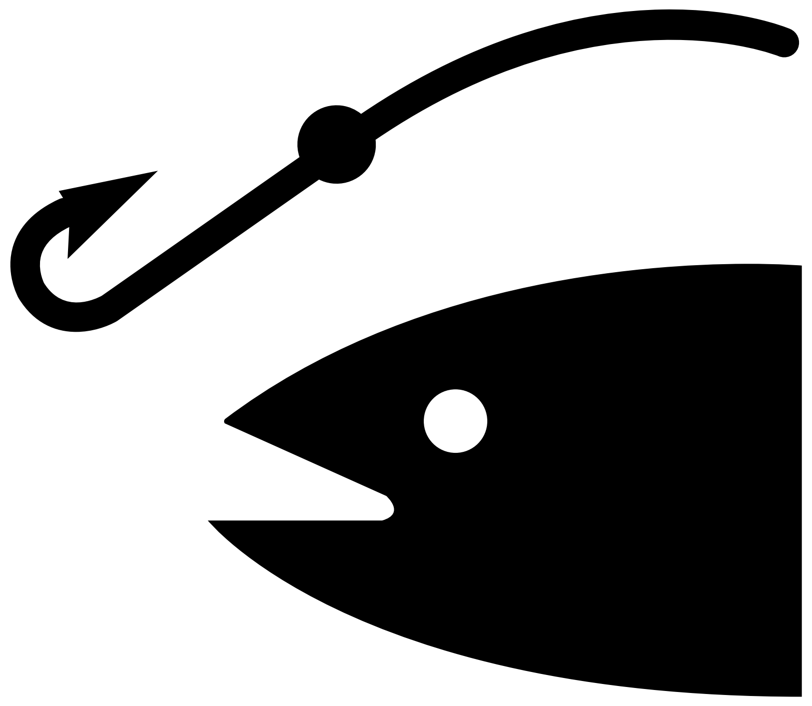
$$a_{ij} = p(q_t = j \mid q_{t-1} = i) = \frac{\sum_O \sum_t p(q_{t-1} = i,\ q_t = j \mid O)}{\sum_O \sum_t p(q_{t-1} = i,\ q_t = * \mid O)}$$

If we don't have labeled sequences,
compute expected labelings under current parameters,
then re-estimate parameters.

Conditional Random Fields

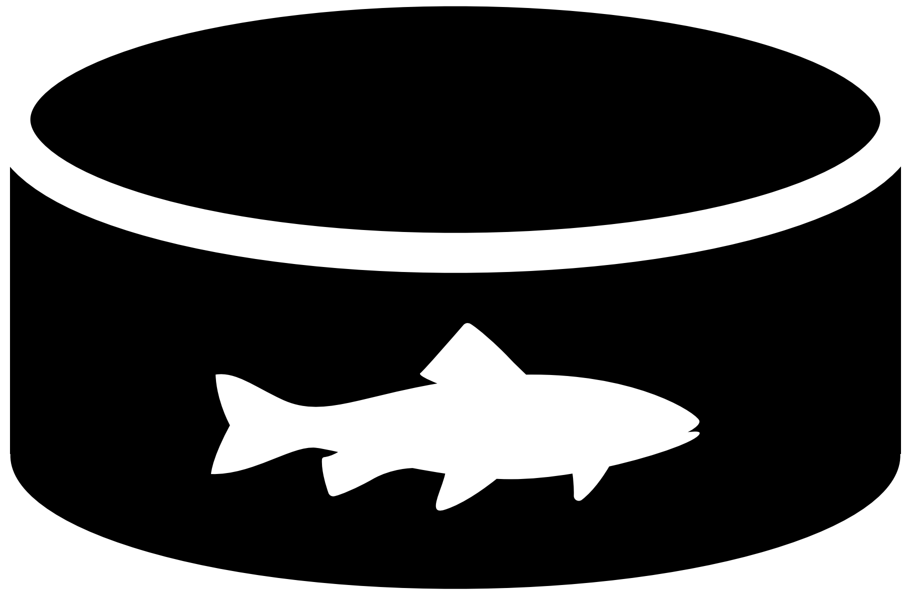
Uncertainty and context

Noun	**Modal**	**Verb**
People	*can*	*fish*



Uncertainty and context

Noun	**Verb**	**Noun**
People	*can*	*fish*



Uncertainty and context

Modal Verb

On my boat, people can fish

Uncertainty and context

Verb Noun

In my factory, people can fish

Uncertainty and context

While aboard my floating tuna

??   ??

cannery, people can fish.

Uncertainty and context

Modal Verb

*On my **boat**, people can fish*

HMMs make it very hard to model
this kind of long-distance dependency.

Tagging as classification?

$\boxed{\textbf{Modal}}$

On my boat, people $\boxed{\textit{can}}$ fish

$$p(\text{Modal} \mid \text{can}, O) \propto \exp\{w_{\text{Modal}}^{\top} f(\text{can}, O)\}$$

Tagging as classification?

Modal

On my boat, people $\boxed{\textit{can}}$ *fish*

$$p(\textsf{Modal} \mid \textsf{can}, O) \propto \exp\{w_{\textsf{Modal}}^{\top} f(\textsf{can}, O)\}$$

Training a discriminative classifier would let us incorporate lots of long-range context features.

Uncertainty and context

Noun 0.5

Verb 0.5

on my floating cannery, people can fish

0.5 Modal

0.5 Verb

but no way to tell that p(Modal, Noun) = 0!

Uncertainty and context

How do we simultaneously support:

structured queries about relationships between tags?

(like an HMM)

rich context features?

(like a discriminative classifier)

Conditional random fields

Define:

$$p(Q \mid O) = \frac{\exp\{\sum_t a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O)\}}{Z(O)}$$

Conditional random fields

Define:

$$p(Q \mid O) = \frac{\exp\{\sum_t a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O)\}}{Z(O)}$$

Looks like a classifier! Scores are log-proportional to a sum of dot products between feature vectors and weights.

Conditional random fields

Define: **Looks like an HMM!** Probability of a sequence factors along (state, state) and (state, obs) pairs.

$$p(Q \mid O) = \frac{\exp\{\sum_t a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O)\}}{Z(O)}$$

(but now we can use the whole context, not just o_t)

Normalizing the model

$$p(Q \mid O) = \frac{\exp\{\sum_t a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O)\}}{Z(O)}$$

What is Z? For this to be a proper distribution, needs to sum to 1 over all Q, *i.e.*:

$$Z(O) = \sum_{Q'} \exp\left\{ \sum_t a^\top \phi_a(q'_{t-1}, q'_t) + b^\top \phi_b(q'_t, O) \right\}$$

"partition function"

Queries

If we're given the parameters A, B and π, what questions can we answer?

Queries: joint probability?

~~Q1: what is the joint probability of a pair of (observation, tag) sequences?~~ $\cancel{p(O,Q)}$

In HMMs, this is easy (but $P(O)$ and $P(Q|O)$ are harder)

In CRFs, there is no generative model of O and **no joint probability!**

Queries: conditional probability

Q2: what is the conditional probability of tags Q given observations O?

$$p(Q \mid O)$$

$$p(Q \mid O) = \frac{\exp\{\sum_t a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O)\}}{Z}$$

Just need to compute Z!

Computing the partition function

$$Z(T,j,O) = \sum_{Q:\ |Q|=T,\ q_T=j} \exp\left\{ \sum_{t=1}^{T} a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O) \right\}$$

length-T sequences that end in i

Claim:

$$Z(T,j,O) = \sum_i Z(T-1,i) \cdot \exp\{a^\top \phi_a(i,j) + b^\top \phi_b(j,O)\}$$

$$Z(T,j,O) = \sum_{Q:\ |Q|=T,\ q_T=j} \exp\left\{ \sum_{t=1}^{T} a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O) \right\}$$

by definition

$$= \sum_{i} \sum_{\substack{Q':\ |Q'|=T-1 \\ q_{T-1}=i,\, q_T=j}} \exp\left\{ \sum_{t=1}^{T} a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O) \right\}$$

rewrite Q as concat. of Q' (ending in i) and $q_T = j$

$$= \sum_{i} \sum_{\substack{Q':\ |Q'|=T-1 \\ q_{T-1}=i,\, q_T=j}} \exp\left\{ a^\top \phi_a(i,j) + b^\top \phi_b(j, O) + \sum_{t=1}^{T-1} a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O) \right\}$$

pull timestep T for inner sum to the front

$$Z(T,j,O) = \sum_i \Bigg[\exp\left\{a^\top \phi_a(i,j) + b^\top \phi_b(j,O)\right\} \times \sum_{\substack{Q':\ |Q'|=T-1 \\ q_{T-1}=i}} \exp \sum_{t=1}^{T-1} \left\{a^\top \phi_a(q_{t-1}, q_t) + b^\top \phi_b(q_t, O)\right\}\Bigg]$$

and then factor it out

$$\sum_i \exp\left\{a^\top \phi_a(i,j) + b^\top \phi_b(j,O)\right\} \cdot Z(T-1,i,O)$$

by definition

The forward recurrence

Same recurrence relation!

$$Z(T,j,O) = \sum_i Z(T-1,i) \cdot \exp\{a^\top \phi_a(i,j) + b^\top \phi_b(j,O)\}$$

$$= \exp\{b^\top \phi_b(j,O)\} \sum_i Z(T-1,i) \cdot \exp\{a^\top \phi_a(i,j)\}$$

$$\alpha(t,j) = b_j(o_t) \sum_i \alpha(t-1,i)\ a_{ij}$$

The forward algorithm (CRF-style)

Q2: what is the partition function for tag sequences of length T and obs. O?

$$Z(O)$$

$$\boxed{\begin{aligned}\alpha(t,j) &= \exp\{b^\top \phi_b(j,O)\} \sum_i \alpha(t-1,i)\ \exp\{a^\top \phi_a(i,j)\} \\ \alpha(1,j) &= \exp\{b^\top \phi_b(j,O)\}\end{aligned}}$$

$$Z(O) = \sum_j Z(T,j,O)$$

The Viterbi Algorithm (CRF-style)

Q2: what is the highest-scoring tag sequence?

$$\max_{Q} p(Q \mid O)$$

$$\delta(t,j) = \exp\{b^\top \phi_b(j,O)\} \max_i \delta(t-1,i) \ \exp\{a^\top \phi_a(i,j)\}$$

$$\delta(1,j) = \exp\{b^\top \phi_b(j,O)\}$$

Supervised training

This looks exactly like text classification.

But, by designing our features carefully, we can do "classification" with an $O(|Q|^T)$-sized output space in $O(|Q|^2T)$ time!

Maximum likelihood estimation: $$\min_{a,b} - \sum_{(Q,O)} \log p(Q \mid O; a, b)$$

SGD: $$a^{(t+1)} = a^{(t)} + \nabla_a \log P(Q \mid O; a, b)$$ (just use autograd!)

Unsupervised training

~~Q1: what is the joint probability of a pair of (observation, tag) sequences?~~ $\cancel{p(O, Q)}$

In CRFs, there is no generative model of O and no joint probability.

Nothing to optimize!

Actually, what is $\nabla_a \log P(Q \mid O; a, b)$?

stuff that's multiplied by a

other stuff

$$\nabla_a \log p(Q \mid O; a, b) = \nabla_a \log \frac{\exp\{a^\top \Phi(Q) + \ldots\}}{\sum_{Q'} \exp\{a^\top \Phi(Q') + \ldots\}}$$

$$\nabla_a \log p(Q \mid O; a, b) = \nabla_a \log \frac{\exp\{a^\top \Phi(Q) + \ldots\}}{\sum_{Q'} \exp\{a^\top \Phi(Q') + \ldots\}}$$

$$= \nabla_a (a^\top \Phi(Q) + \ldots) - \nabla_a \log \sum_{Q'} \exp\{a^\top \Phi(Q') + \ldots\}$$

$$= \Phi(Q) - \frac{\nabla_a \sum_{Q'} \exp\{a^\top \Phi(Q) + \ldots\}}{\sum_{Q'} \exp\{a^\top \Phi(Q') + \ldots\}}$$

$$= \Phi(Q) - \frac{\sum_{Q'} \Phi(Q') \exp\{a^\top \Phi(Q') + \ldots\}}{\sum_{Q'} \exp\{a^\top \Phi(Q') + \ldots\}} \boxed{= \Phi(Q) - \mathbf{E}_{p(Q' \mid O; a, b)} \Phi(Q')}$$

Actually, what is $\nabla_a \log P(Q \mid O; a, b)$?

$$\nabla_a \log p(Q \mid O; a, b) = \nabla_a \log \frac{\exp\{a^\top \Phi(Q) + \ldots\}}{\sum_{Q'} \exp\{a^\top \Phi(Q) + \ldots\}}$$

$$\boxed{= \Phi(Q) - \mathbf{E}_{p(Q' \mid O; a, b)} \Phi(Q')}$$

The gradient of the log-partition function is the expected feature vector under the current predictive distribution (!)

Next class: recurrent neural networks