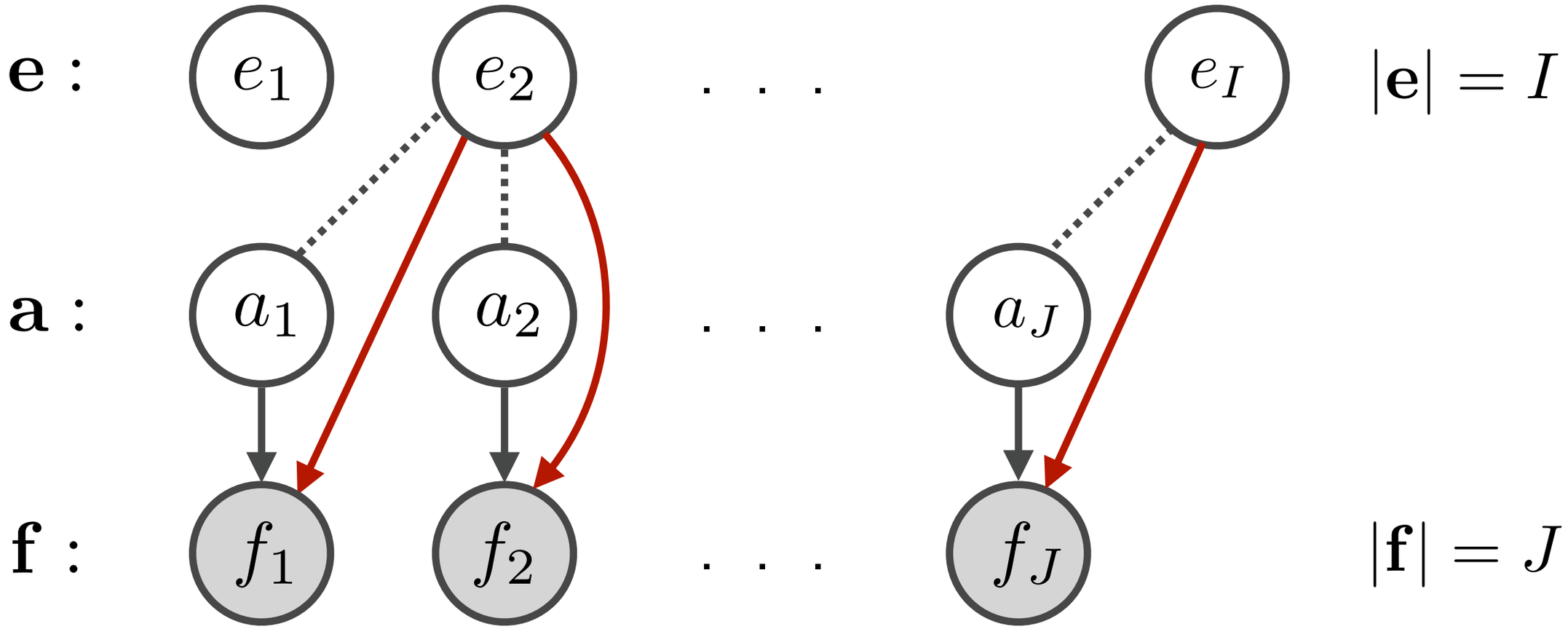
EM for Word Alignment

11-711 recitation

Maria Ryskina

IBM Model 1



$a_1 = 2 \quad a_2 = 2 \qquad\qquad a_J = I$

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{j=1}^{J} p(f_j|e_{a_j})p(a_j)$$

Model 1: parameters

Emission (translation) probabilities: $\theta_{f,e}$

word types: $p(\text{chat}|\text{cat})$

Training objective: $\max_\theta p(\mathbf{f}|\mathbf{e},\theta) = \max_\theta \sum_{\mathbf{a}} p(\mathbf{f},\mathbf{a}|\mathbf{e},\theta)$

$$p(\mathbf{f},\mathbf{a}|\mathbf{e},\theta) = \prod_{j=1}^{J} p(a_j = i)p(f_j|e_i,\theta)$$

$$p(a_j = i) = \frac{1}{I+1}$$

uniform prior

$$p(f_j = f|e_i = e) = \theta_{f,e}$$

translation probability

Model 1: EM algorithm

- E-step: impute missing data $\leftarrow \mathbf{a}$
- M-step: estimate parameters based on imputed complete data $\leftarrow \theta$

E-step: computing expected counts:

$$d_{f,e}(\theta) = \mathbb{E}_{p(\mathbf{a}|\mathbf{f},\mathbf{e},\theta)}[c_{f,e}]$$

M-step: reestimating parameters:

$$\theta_{f,e} \propto d_{f,e}(\theta)$$

Model 1: E-step

Expected counts:

$$d_{f,e}(\theta) = \mathbb{E}_{p(\mathbf{a}|\mathbf{f},\mathbf{e},\theta)}[c_{f,e}]$$

At iteration t :

french token of type f | aligned to | english token of type e

$$d^{(t)}_{f,e}(\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} \mathbb{1}[f_j = f]\mathbb{1}[a_j = i]\mathbb{1}[e_i = e] \cdot p(a_j = i|\mathbf{f},\mathbf{e},\theta^{(t)})$$

Model 1: E-step

Computing posteriors:

$$p(\mathbf{a}|\mathbf{f},\mathbf{e},\theta) = \frac{p(\mathbf{f},\mathbf{a}|\mathbf{e},\theta)}{p(\mathbf{f}|\mathbf{e},\theta)} = \prod_{j=1}^{J} \frac{p(f_j,a_j|\mathbf{e},\theta)}{p(f_j|\mathbf{e},\theta)}$$

$$p(a_j|\mathbf{f},\mathbf{e},\theta) = \frac{p(f_j,a_j|\mathbf{e},\theta)}{p(f_j|\mathbf{e},\theta)}$$

$$p(a_j = i|\mathbf{f},\mathbf{e},\theta^{(t)}) = \frac{\theta^{(t)}_{f_j,e_i} \cdot p(a_j = i)}{\sum_{k=1}^{I} \theta^{(t)}_{f_j,e_k} \cdot p(a_j = k)}$$

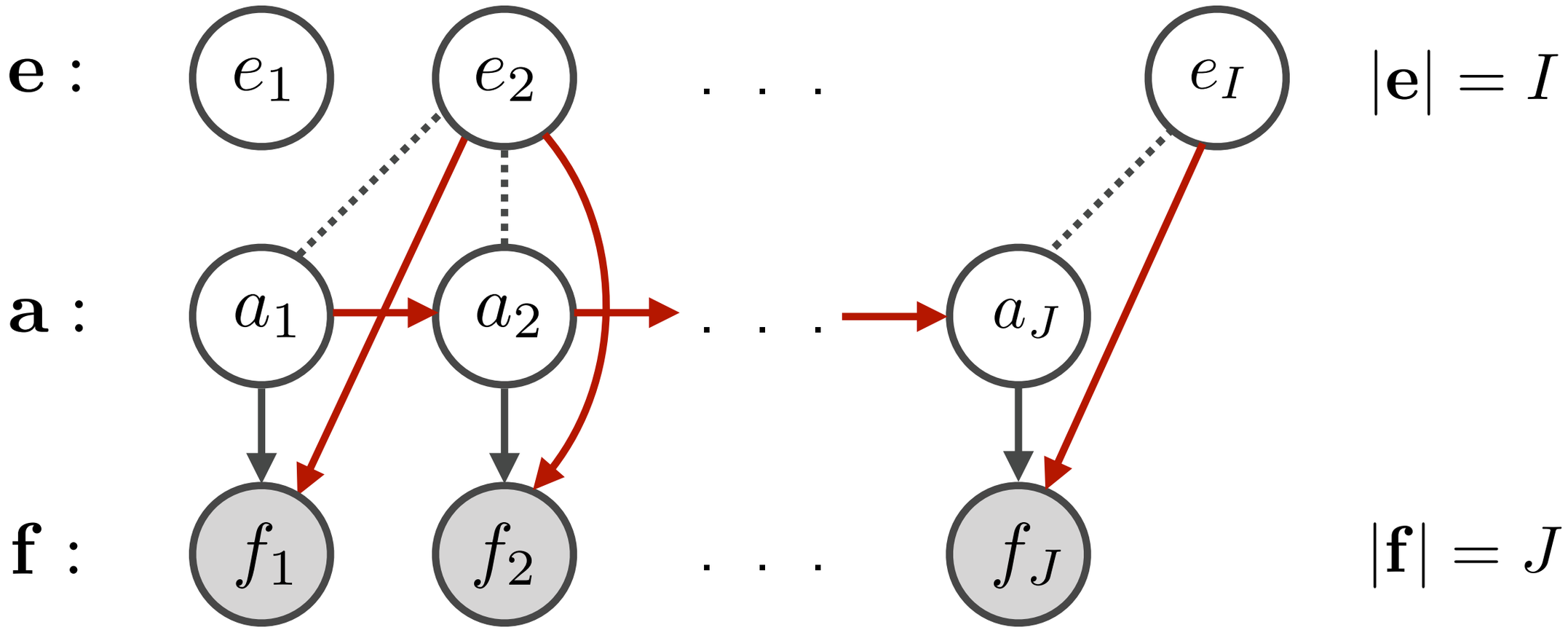
Model 1: M-step

Reestimating parameters:

$$\theta_{f,e}^{(t+1)} \propto d_{f,e}^{(t)}(\theta^{(t)})$$

$$\theta_{f,e}^{(t+1)} = \frac{d_{f,e}^{(t)}(\theta^{(t)})}{\sum_{\tilde{f}} d_{\tilde{f},e}^{(t)}(\theta^{(t)})}$$

HMM Model



$a_1 = 2 \quad a_2 = 2 \qquad\qquad a_J = I$

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{j=1}^{J} p(f_j|e_{a_j}) p(a_j|a_{j-1})$$

HMM: parameters

Emission (translation) probabilities: $\theta_{f,e}$

Transition ("jump") probabilities: ψ_k

$$p(a_j = i | a_{j-1} = i', \psi) = \psi_{|i-i'|}$$

Try other bucketing strategies!

Transition probabilities can be:

- fixed: e.g. $\psi_k \propto \exp(-\lambda(k-1))$
- learned with EM algorithm

HMM: EM algorithm

E-step: computing expected counts:

$$d_{f,e}(\theta) = \mathbb{E}_{p(\mathbf{a}|\mathbf{f},\mathbf{e},\psi,\theta)}[c_{f,e}]$$

$$d_k(\psi) = \mathbb{E}_{p(\mathbf{a}|\mathbf{f},\mathbf{e},\psi,\theta)}[c_{|a_j - a_{j-1}|=k}]$$

M-step: reestimating parameters:

$$\theta_{f,e} \propto d_{f,e}(\theta)$$

$$\psi_k \propto d_k(\psi)$$

HMM: E-step

Expected counts:

$$d_{f,e}^{(t)}(\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} \mathbb{1}[f_j = f]\mathbb{1}[a_j = i]\mathbb{1}[e_i = e]\times$$

$$\times\, p(a_j = i|\mathbf{f}, \mathbf{e}, \theta^{(t)}, \psi^{(t)})$$

$$d_{k}^{(t)}(\psi) = \sum_{i=1}^{I}\sum_{i'=1}^{I}\sum_{j=1}^{J} \mathbb{1}[a_j = i]\mathbb{1}[|i - i'| = k]\mathbb{1}[a_{j-1} = i']\times$$

$$\times\, p(a_j = i, a_{j-1} = i'|\mathbf{f}, \mathbf{e}, \theta^{(t)}, \psi^{(t)})$$

HMM: E-step

Computing posteriors: forward-backward

α_j^i — sum of all paths up to $a_j = i$

β_j^i — sum of all paths starting from $a_j = i$

$$p(a_j = i|\mathbf{f}, \mathbf{e}, \theta^{(t)}, \psi^{(t)}) = \frac{\alpha_j^i \beta_j^i}{Z}$$

$$p(a_j = i, a_{j-1} = i'|\mathbf{f}, \mathbf{e}, \theta^{(t)}, \psi^{(t)}) = \frac{\alpha_{j-1}^{i'} \cdot \beta_j^i \cdot \psi_{|i-i'|}^{(t)} \cdot \theta_{f_j, e_i}}{Z}$$

HMM: M-step

Reestimating parameters:

$$\theta_{f,e}^{(t+1)} = \frac{d_{f,e}^{(t)}(\theta^{(t)})}{\sum_{\tilde{f}} d_{\tilde{f},e}^{(t)}(\theta^{(t)})}$$

$$\psi_k^{(t+1)} = \frac{d_k(\psi^{(t)})}{\sum_l d_l(\psi^{(t)})}$$

Possible solutions for null

- Fixed transition to null:

$$p(a_j = \text{null} | a_{j-1} = i') = \epsilon$$

$$p(a_j = i | a_{j-1} = i') = (1 - \epsilon)\psi_{|i-i'|}$$

- Uniform transition from null:

$$p(a_j = i | a_{j-1} = \text{null}) = \frac{1 - \epsilon}{I}$$

- Smarter: insert a null for every target word (Och & Ney '03)

- Special prior on null:

$$p(a_j = \text{null}) = \epsilon$$

$$p(a_j = i) = \frac{1 - \epsilon}{I}$$

Questions?