

Alignment by Agreement

Percy Liang, Ben Taskar, Dan Klein



UC Berkeley
Computer Science Division

Unsupervised word alignment

Goal: learn to map sentence pairs to alignments

the railroad term is “ demand loading ”

le terme ferroviaire est “ chargement sur demande ”

Unsupervised word alignment

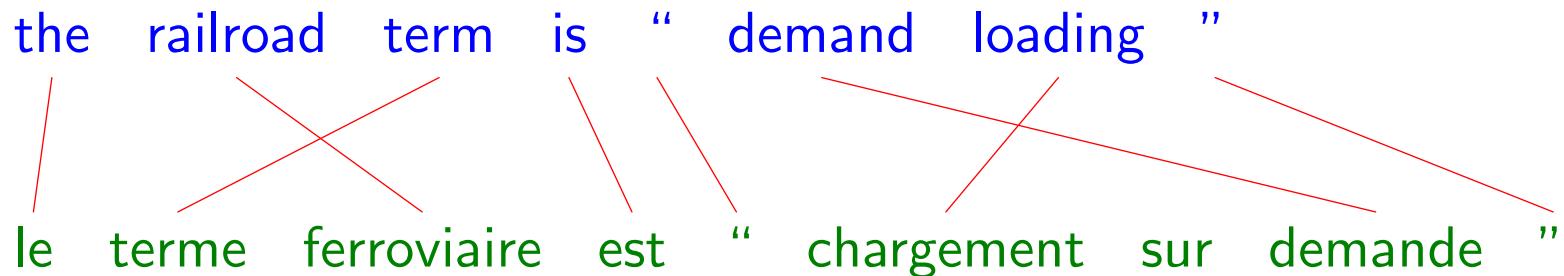
Goal: learn to map sentence pairs to alignments

the railroad term is “ demand loading ”
le terme ferroviaire est “ chargement sur demande ”

The diagram illustrates the process of learning word alignments between two sentences. It shows two rows of words: an English sentence at the top and a French sentence at the bottom. Red lines connect corresponding words between the two languages. Some connections are correct (e.g., 'the' to 'le', 'term' to 'terme'), while others are incorrect or crossed out (e.g., 'railroad' to 'ferroviaire', 'loading' to 'sur', and multiple connections from the final quote mark). This visual representation helps demonstrate the challenge of learning alignments without supervision.

Unsupervised word alignment

Goal: learn to map sentence pairs to alignments



Approach:

jointly train two models to encourage *agreement*

HMM model [Ney, Vogel '96]

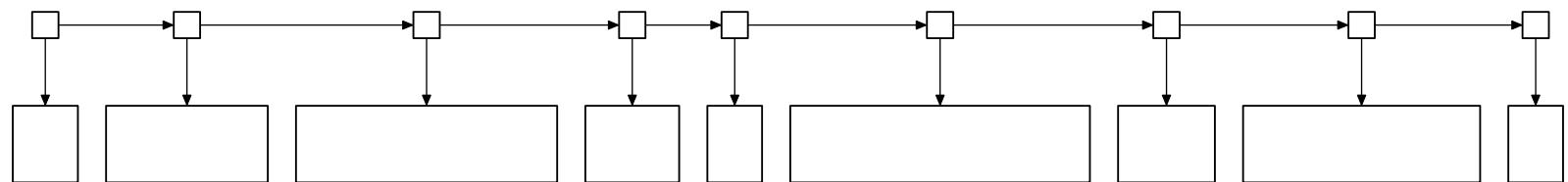
Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

$p(\mathbf{e})$

the railroad term is “ demand loading ”

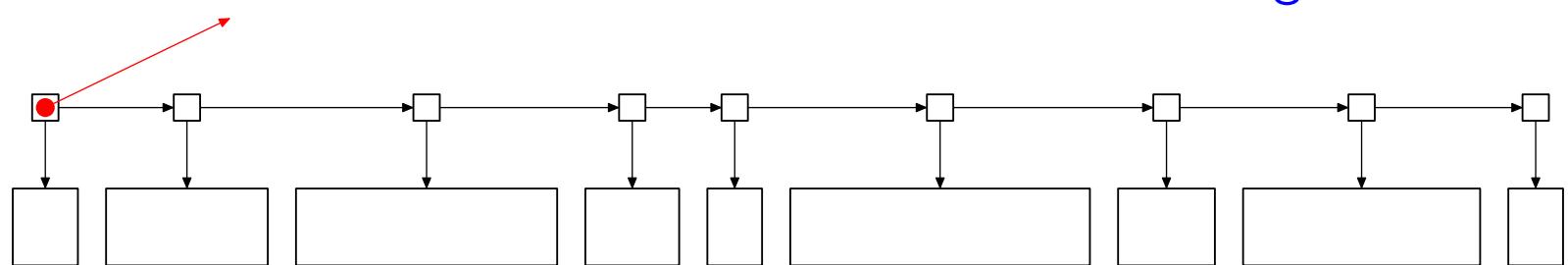


HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

$p(\mathbf{e})$

the railroad term is “ demand loading ”

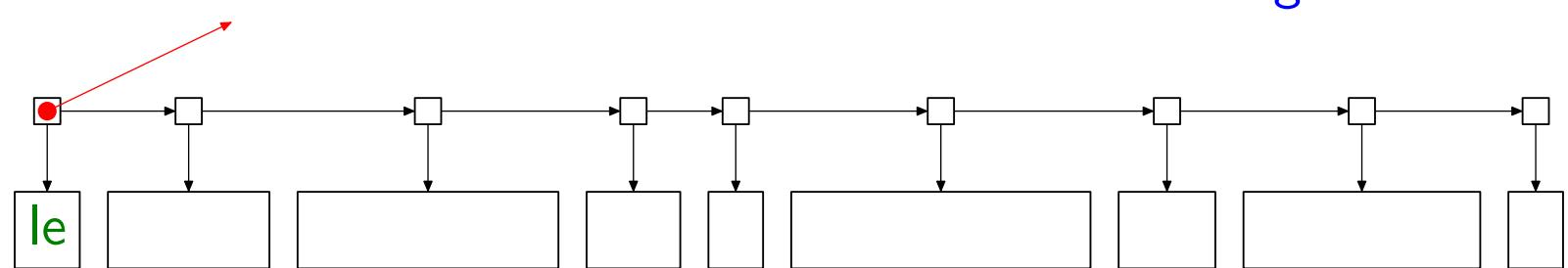


HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

$p(\mathbf{e})$

the railroad term is “ demand loading ”

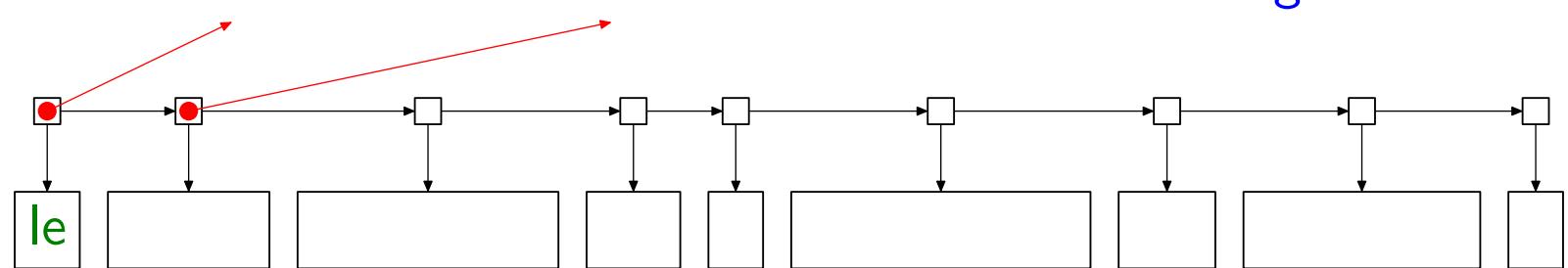


HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

$p(\mathbf{e})$

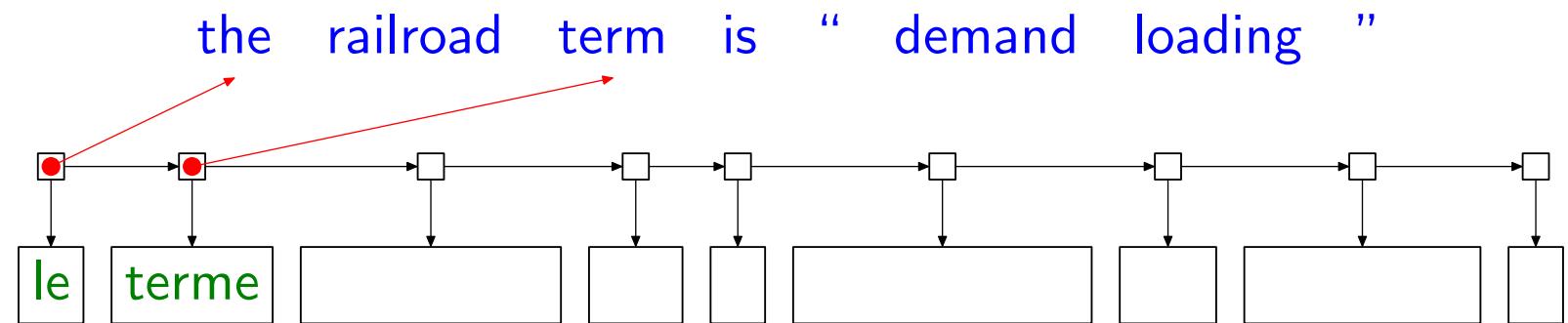
the railroad term is “ demand loading ”



HMM model [Ney, Vogel '96]

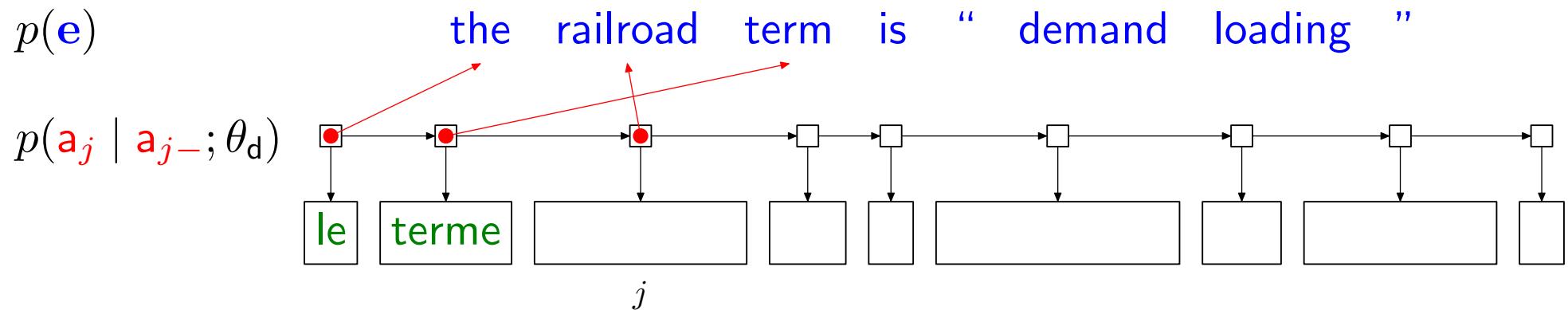
Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$

$p(\mathbf{e})$



HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$



Distortion θ_d

$$p(\begin{array}{c} \uparrow \\ \bullet \\ \downarrow \end{array}, \begin{array}{c} \uparrow \\ \bullet \\ \downarrow \end{array}) = 0.6$$

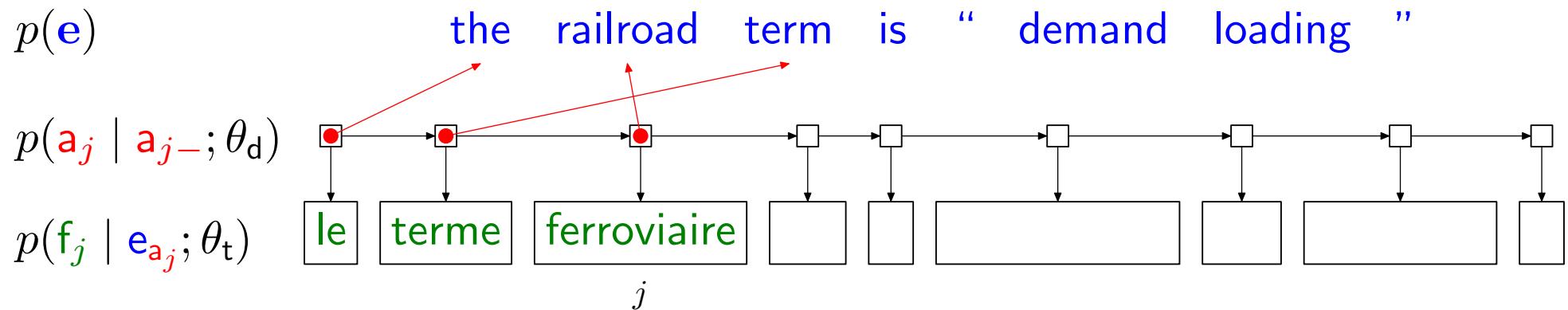
$$p(\begin{array}{c} \uparrow \\ \bullet \\ \downarrow \end{array}, \begin{array}{c} \nearrow \\ \bullet \\ \downarrow \end{array}) = 0.2$$

$$p(\begin{array}{c} \times \\ \bullet \\ \times \end{array}, \begin{array}{c} \times \\ \bullet \\ \times \end{array}) = 0.1$$

...

HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$



Distortion θ_d

$$p(\begin{array}{c} \uparrow \\ \cdot \\ \uparrow \end{array}) = 0.6$$
$$p(\begin{array}{c} \uparrow \\ \cdot \\ \cdot \end{array}) = 0.2$$
$$p(\begin{array}{c} \times \\ \cdot \\ \cdot \end{array}) = 0.1$$

...

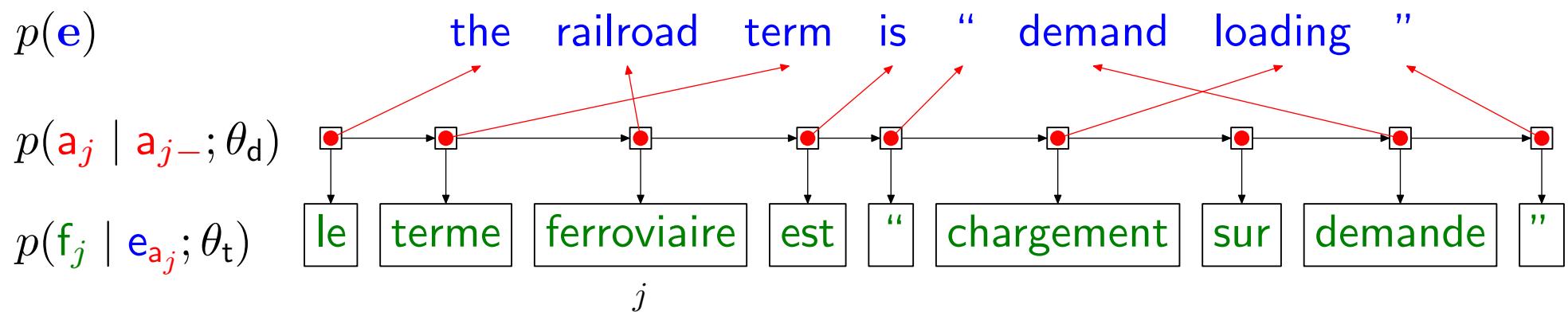
Translation θ_t

$$p(\text{the} \rightarrow \text{le}) = 0.53$$
$$p(\text{the} \rightarrow \text{la}) = 0.24$$
$$p(\text{railroad} \rightarrow \text{ferroviaire}) = 0.19$$
$$p(\text{NULL} \rightarrow \text{le}) = 0.12$$

...

HMM model [Ney, Vogel '96]

Generative model: $p(\mathbf{a}, \mathbf{e}, \mathbf{f}; \theta)$



Distortion θ_d

$$\begin{aligned} p(\text{↑↑}) &= 0.6 \\ p(\text{↑→}) &= 0.2 \\ p(\text{→X}) &= \mathbf{0.1} \\ \dots & \end{aligned}$$

Translation θ_t

$$\begin{aligned} p(\text{the} \rightarrow \text{le}) &= 0.53 \\ p(\text{the} \rightarrow \text{la}) &= 0.24 \\ p(\text{railroad} \rightarrow \text{ferroviaire}) &= \mathbf{0.19} \\ p(\text{NULL} \rightarrow \text{le}) &= 0.12 \\ \dots & \end{aligned}$$

EM training

Maximize $p(\mathbf{e}, \mathbf{f}; \theta)$

EM training

Maximize $p(\mathbf{e}, \mathbf{f}; \theta)$

Parameters: θ

Expectation over alignments: q

$$\boxed{q}$$

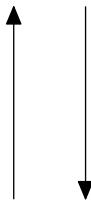
E-step:

$q(\mathbf{a} \mid \mathbf{e}, \mathbf{f}) := p(\mathbf{a} \mid \mathbf{e}, \mathbf{f}; \theta)$
(forward-backward)

M-step:

$\theta := \operatorname{argmax}_{\theta} \mathbb{E}_q \log p(\mathbf{a}, \mathbf{e}, \mathbf{f} \mid \theta)$
(normalizing counts)

$$\boxed{\theta}$$



EM training

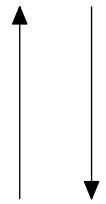
Maximize $p(\mathbf{e}, \mathbf{f}; \theta)$

Parameters: θ

Expectation over alignments: q

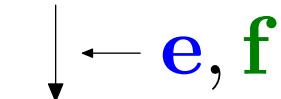
E-step:
 $q(\mathbf{a} | \mathbf{e}, \mathbf{f}) := p(\mathbf{a} | \mathbf{e}, \mathbf{f}; \theta)$
(forward-backward)

$$\boxed{q}$$



M-step:
 $\theta := \operatorname{argmax}_{\theta} \mathbb{E}_q \log p(\mathbf{a}, \mathbf{e}, \mathbf{f} | \theta)$
(normalizing counts)

$$\boxed{\theta}$$

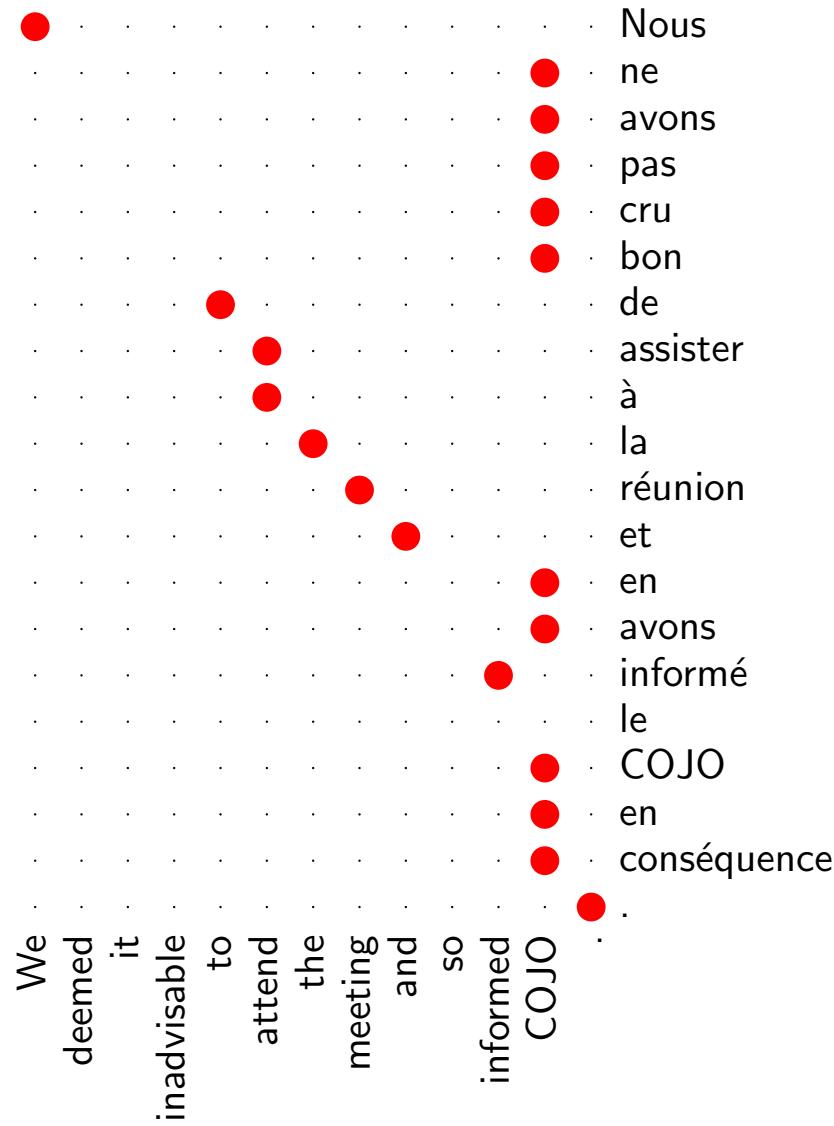


$$\mathbf{a}$$

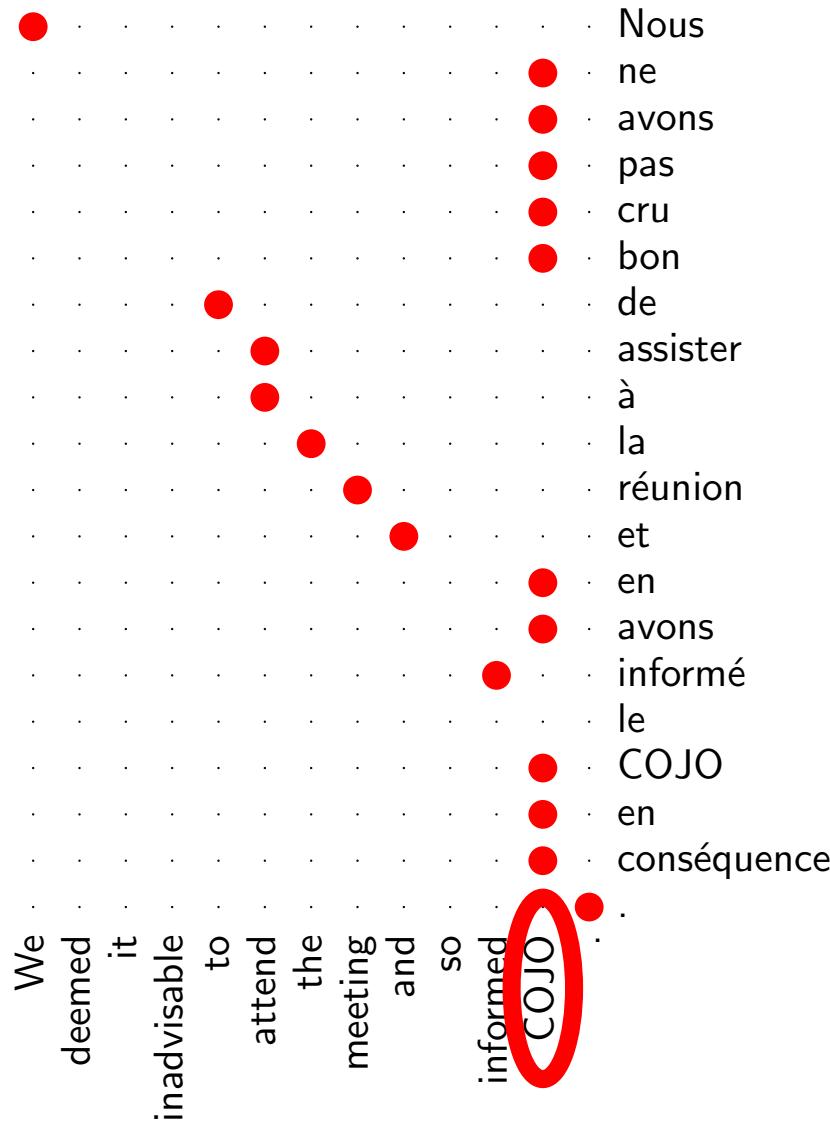
Output of one HMM model

We deemed it inadvisable to attend the meeting and so informed COJO
Nous ne avons pas cru bon de assister à la réunion et en avons informé le COJO en conséquence

Output of one HMM model

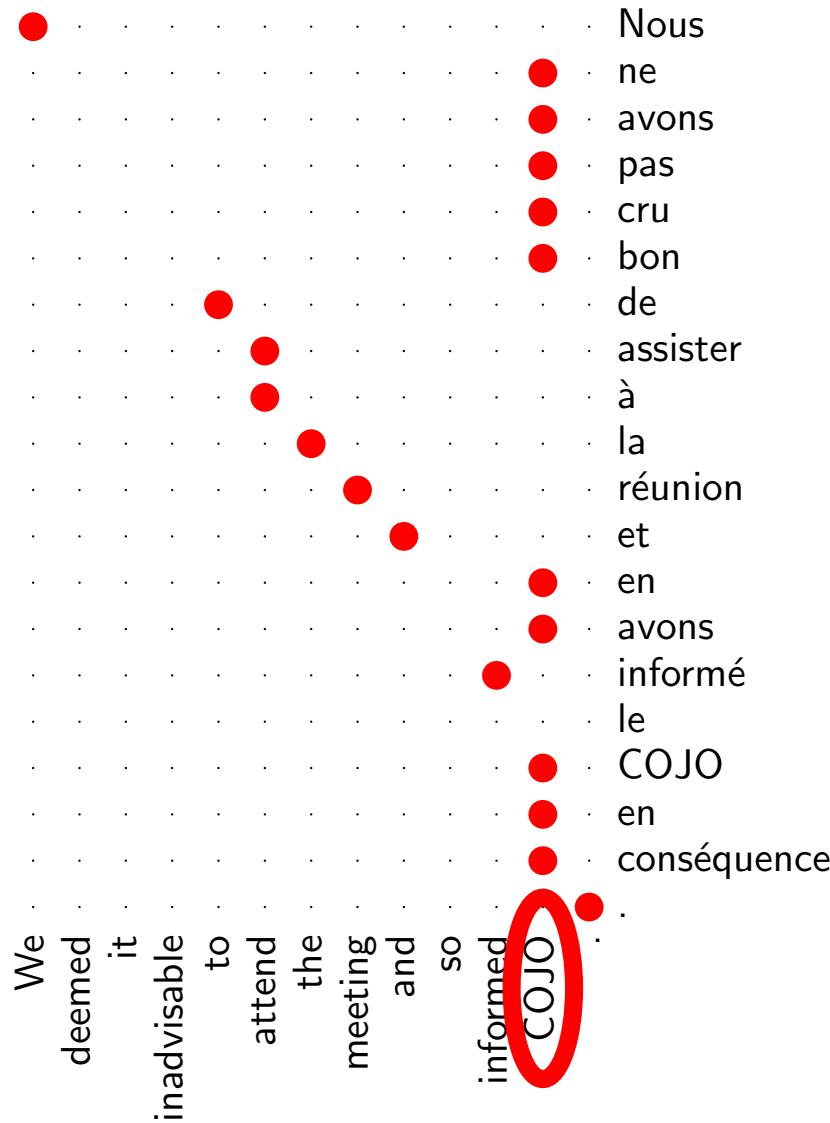


Output of one HMM model



- A problem:
 - Rare words garbage-collect alignments [Moore '05]

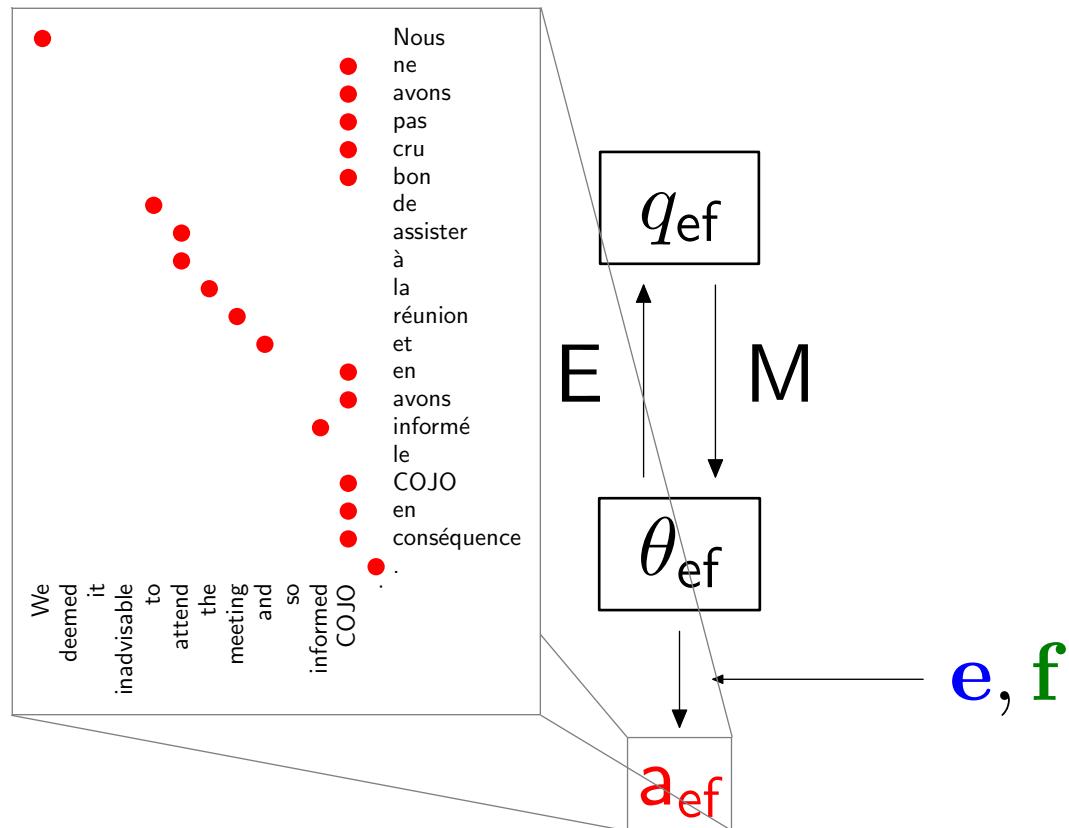
Output of one HMM model



- A problem:
 - Rare words garbage-collect alignments [Moore '05]
- One solution:
 - More complex models

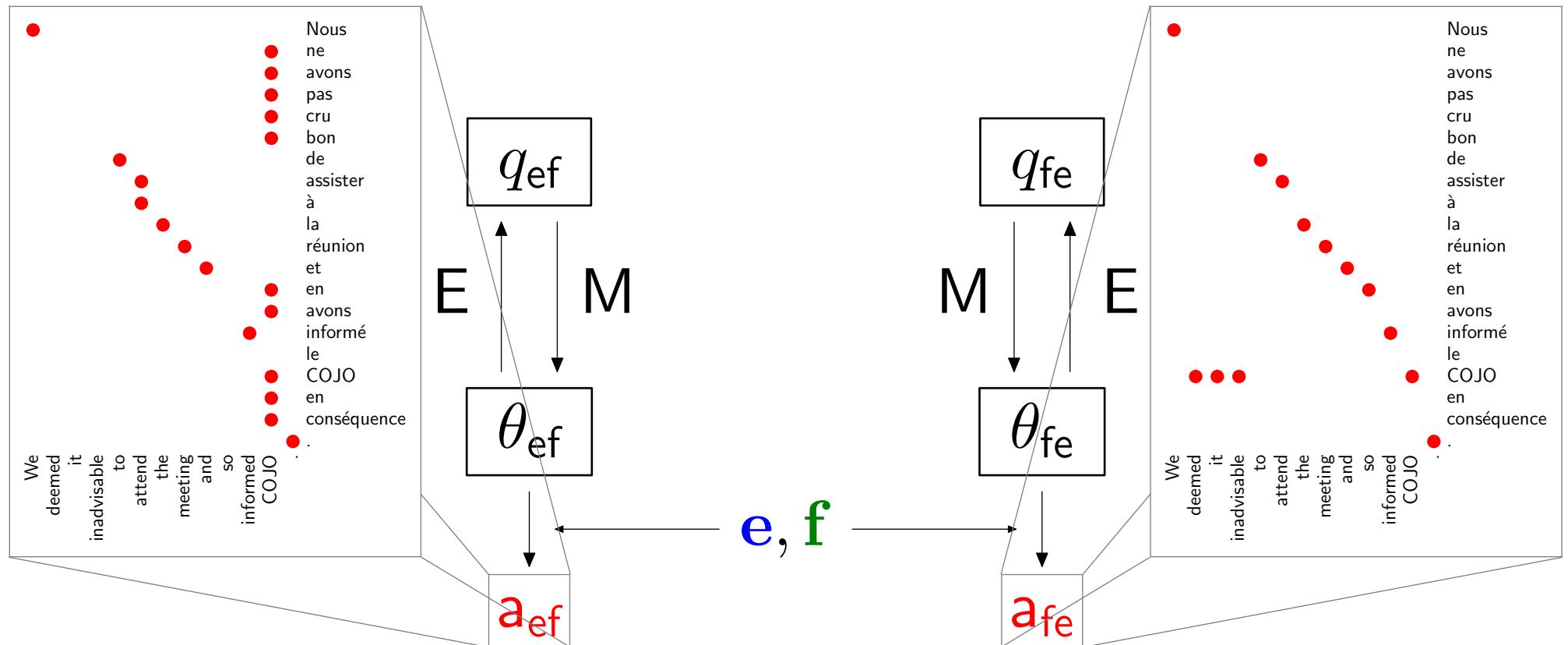
Two complementary models

One model is broken . . .

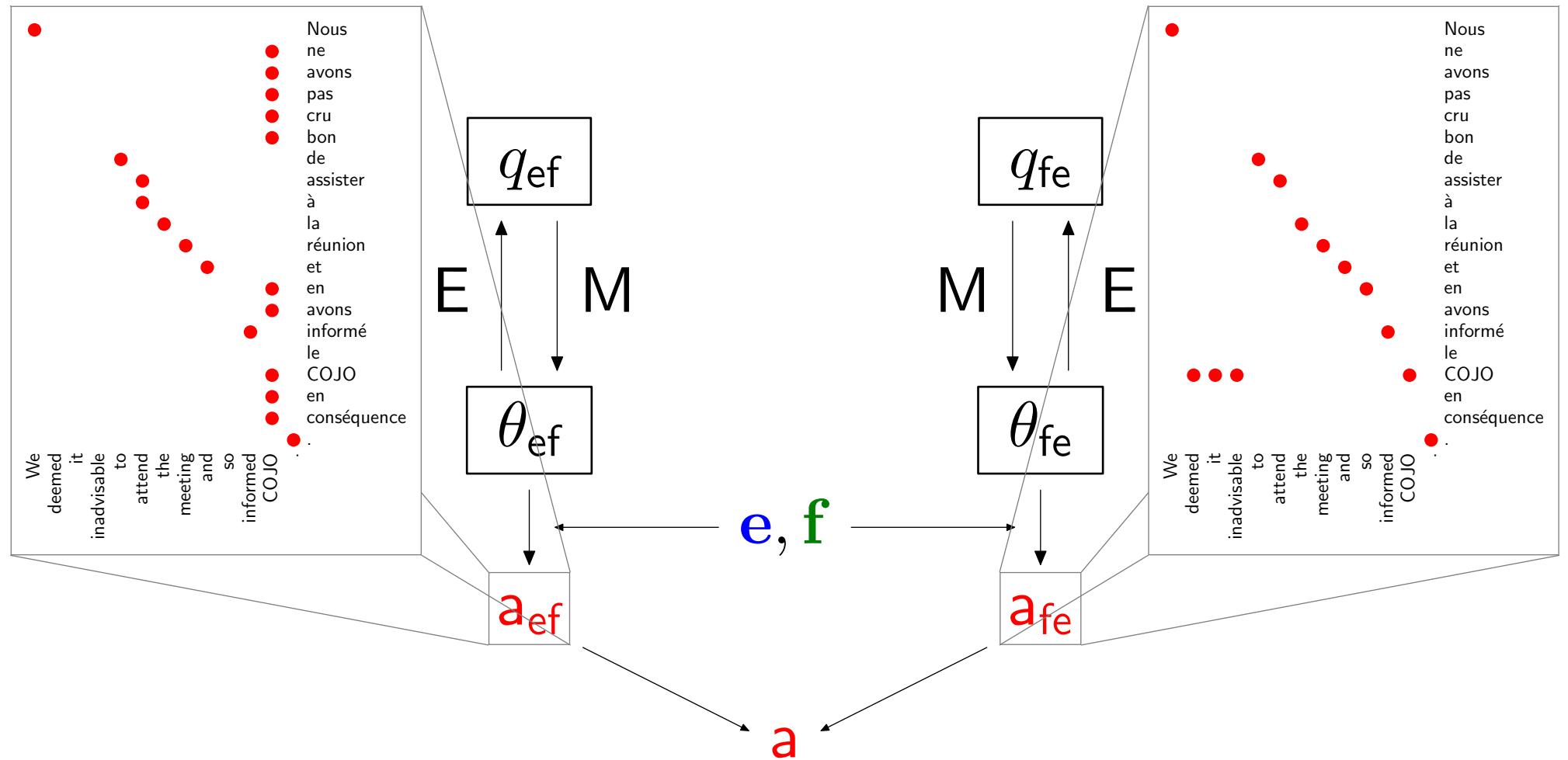


Two complementary models

But second model is not broken in the same way.

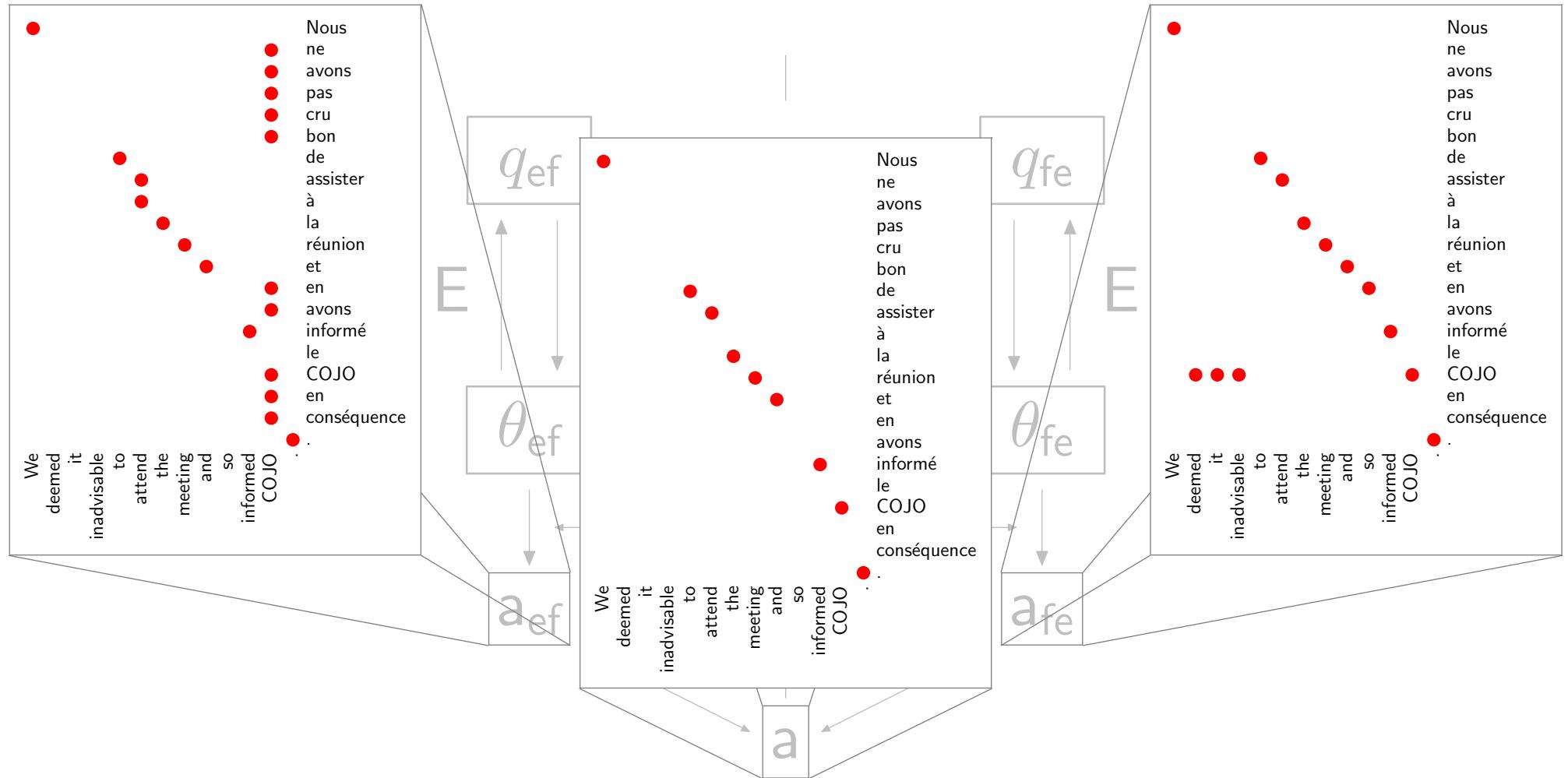


Two complementary models



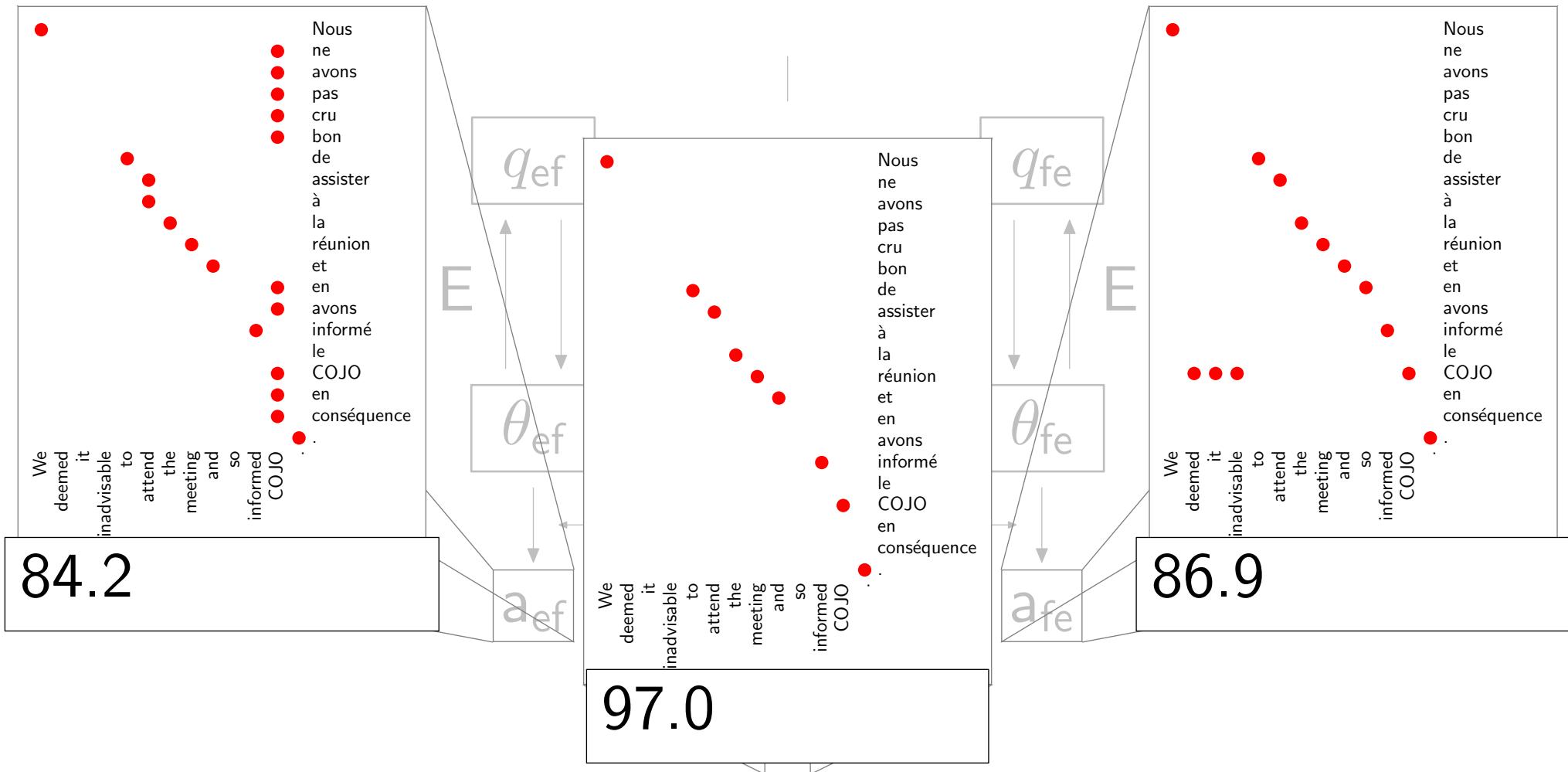
Two complementary models

Intersection kills many bad alignment edges.



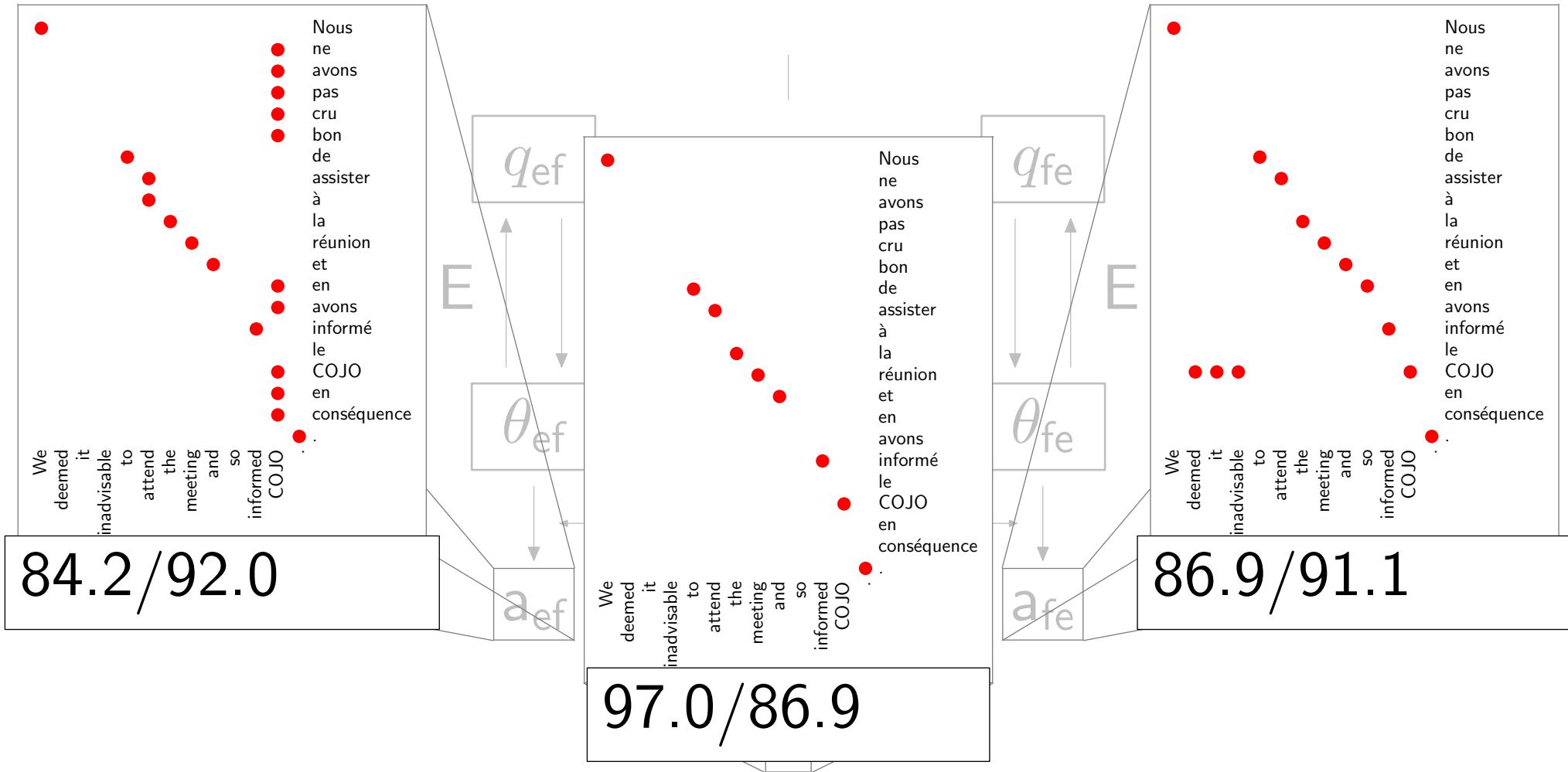
Two complementary models

Precision improves . . .



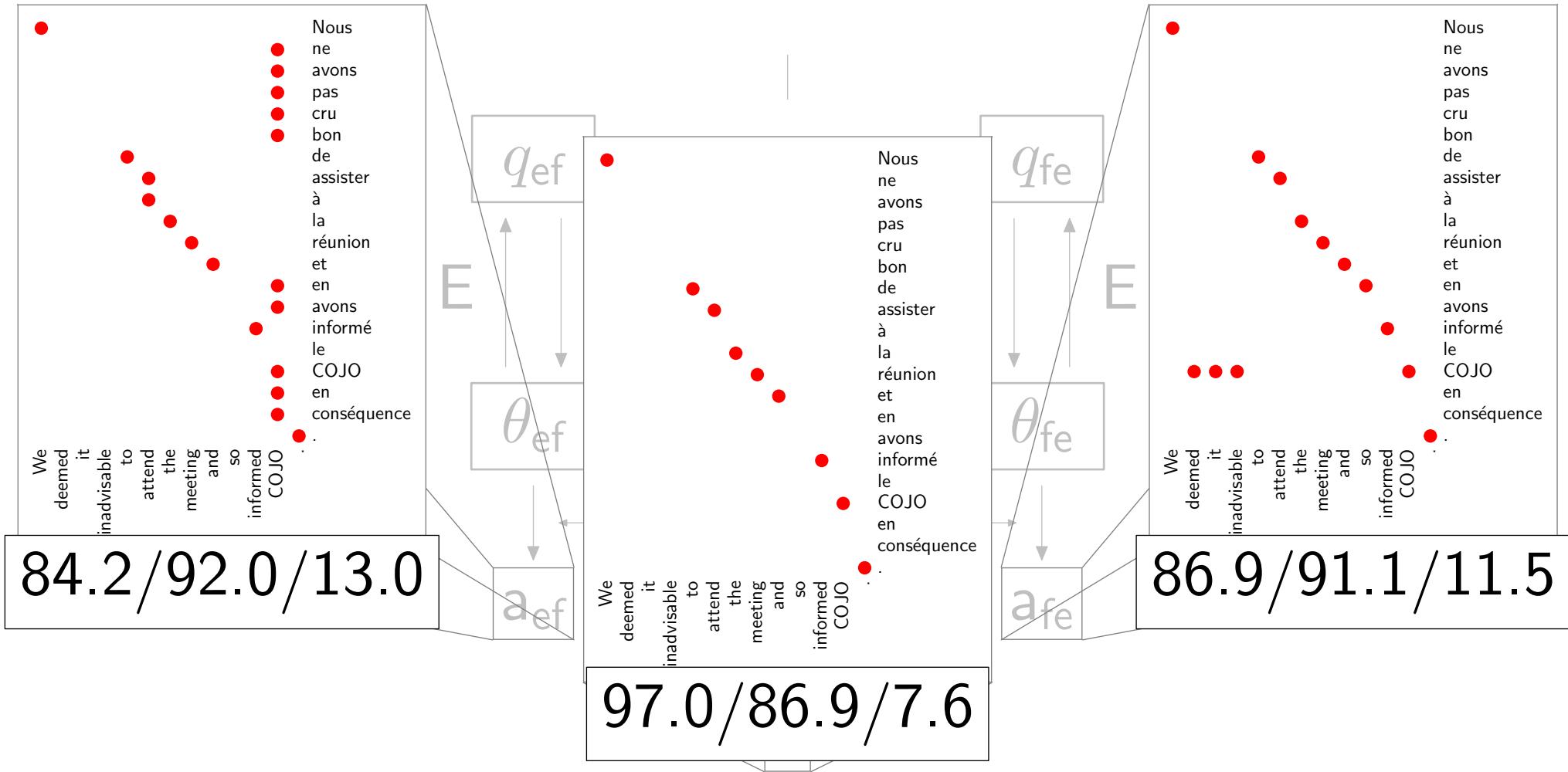
Two complementary models

Precision improves . . . Recall suffers . . .



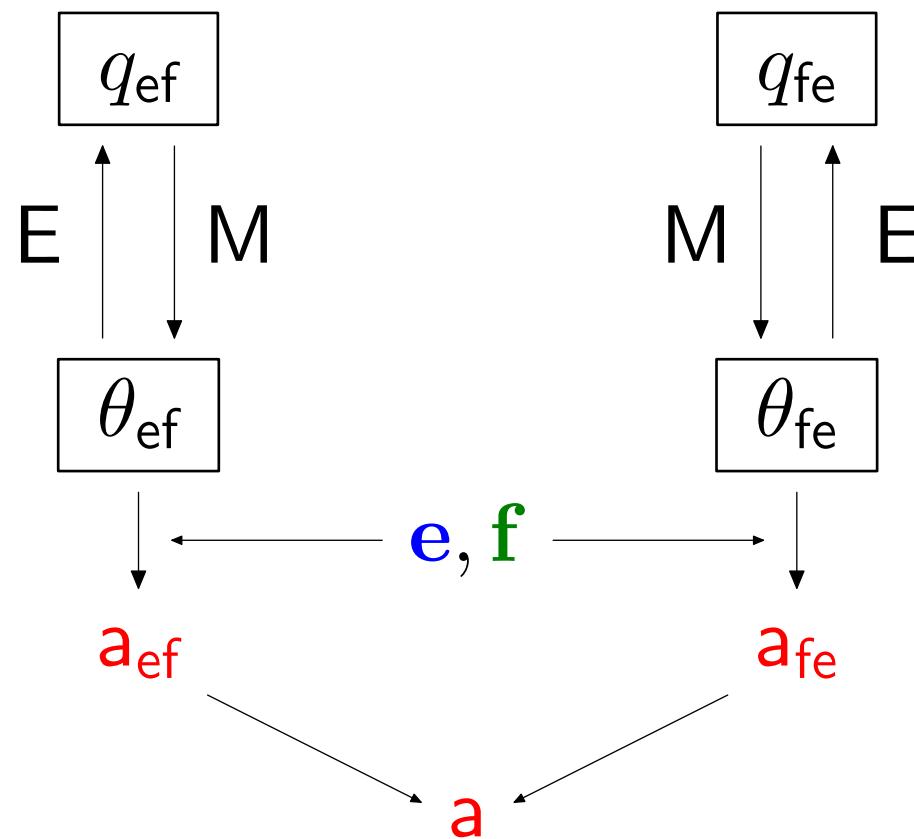
Two complementary models

Precision improves . . . Recall suffers . . . AER improves.



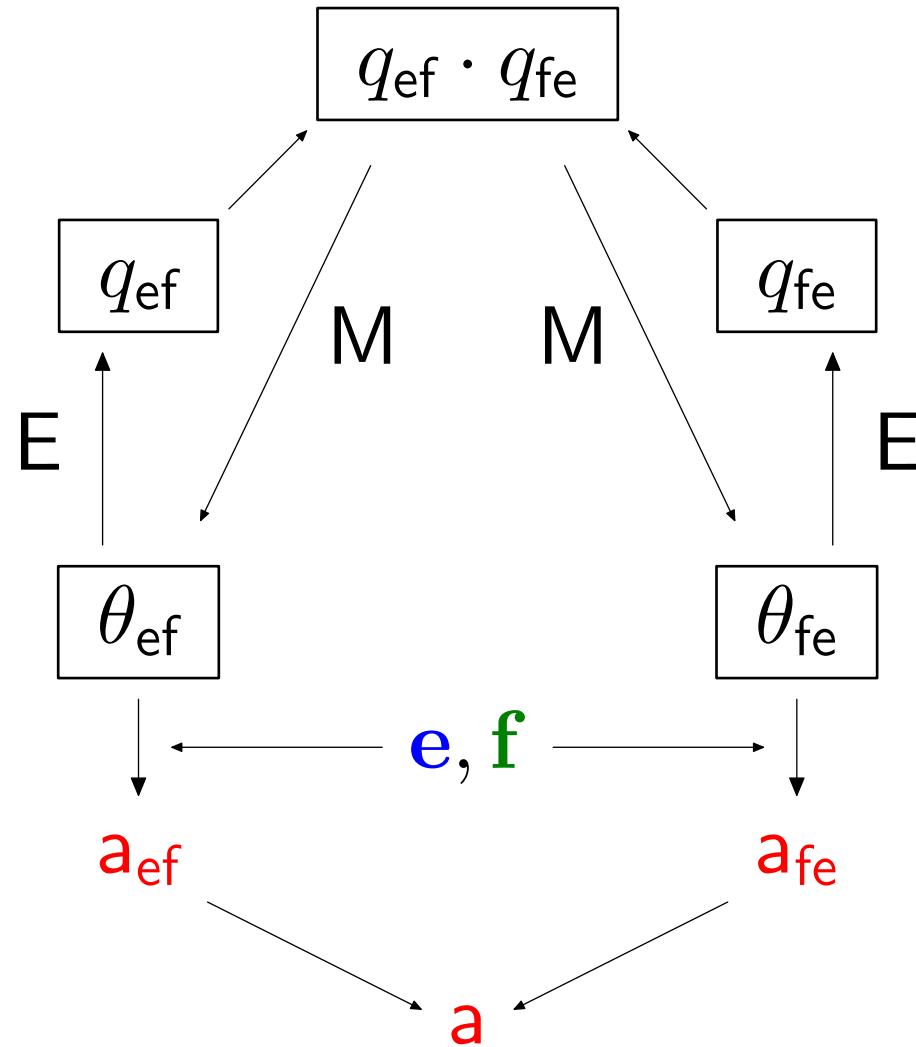
Two complementary models

Can we extend the agreement idea?



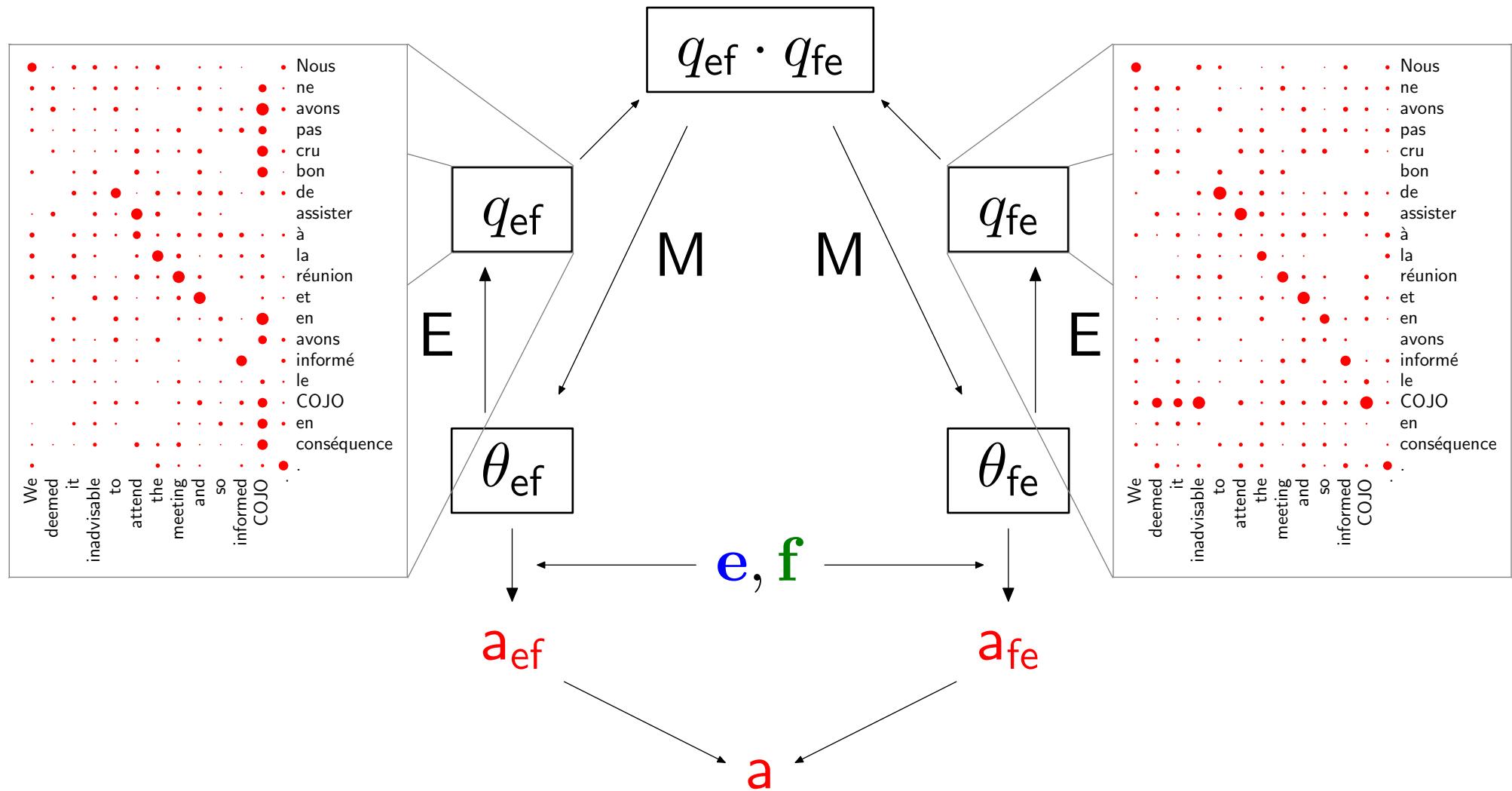
Two complementary models

Key: intersect alignments at training time



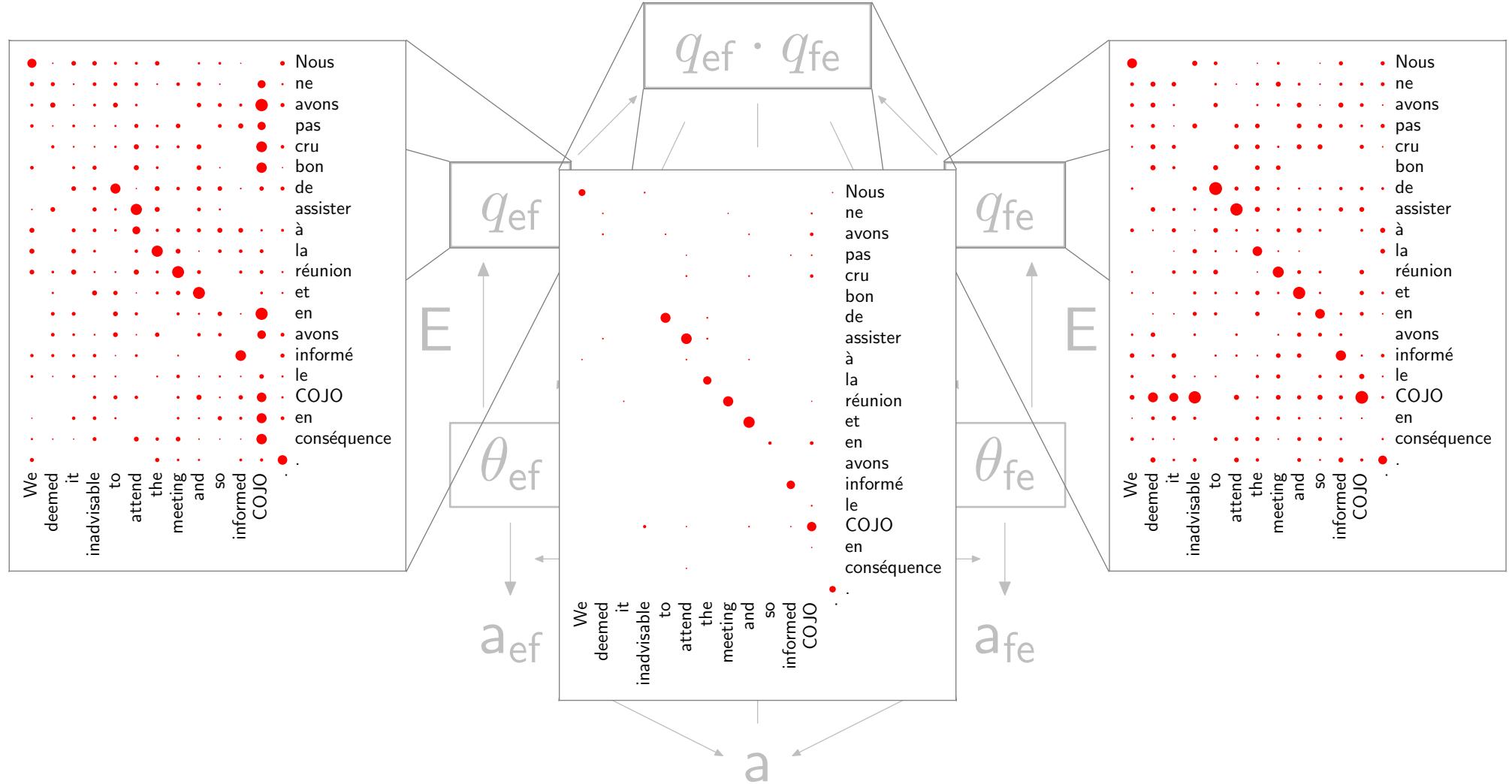
Two complementary models

Fractional alignments



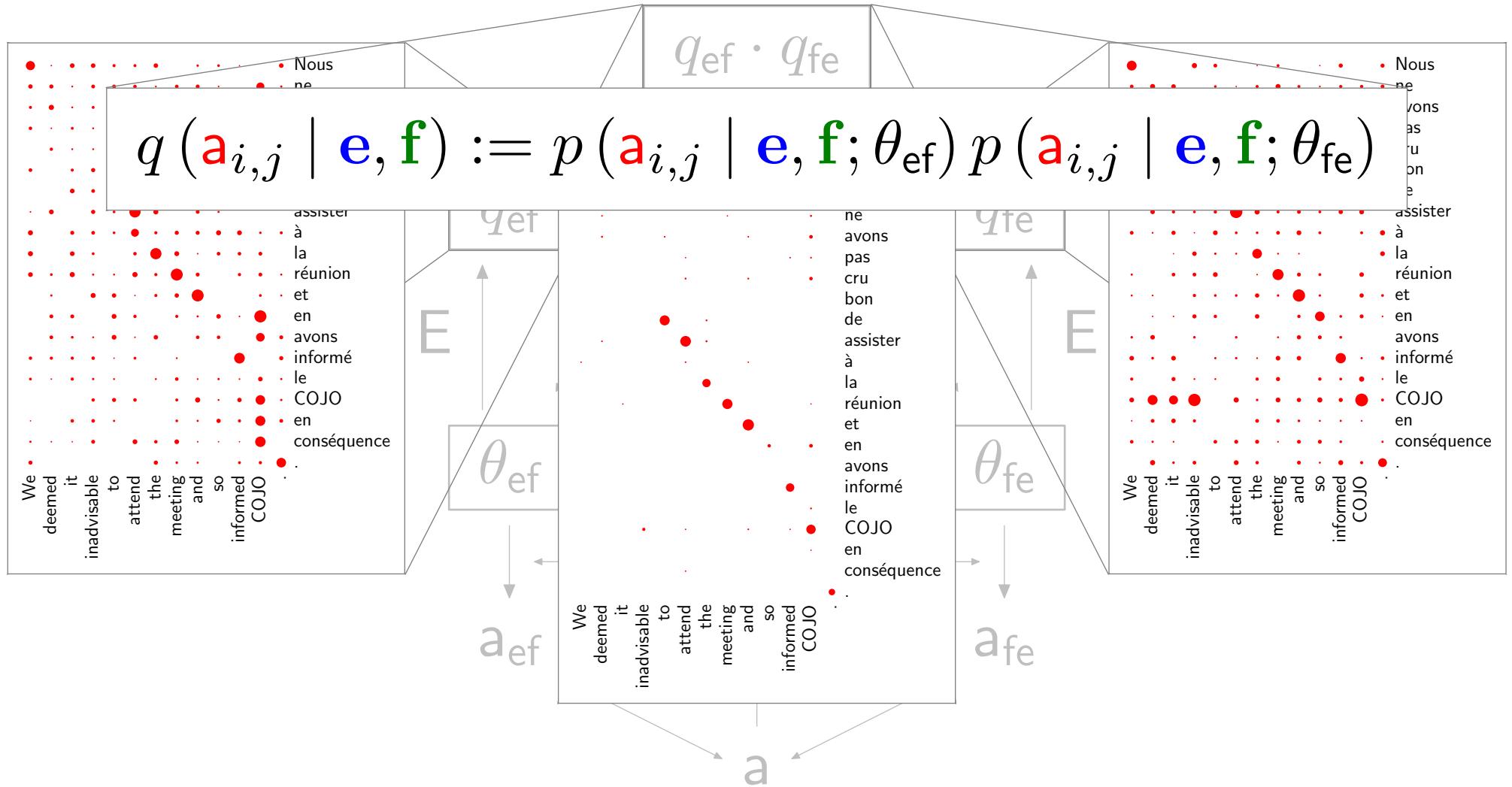
Two complementary models

Soft intersection: multiply fractional alignment



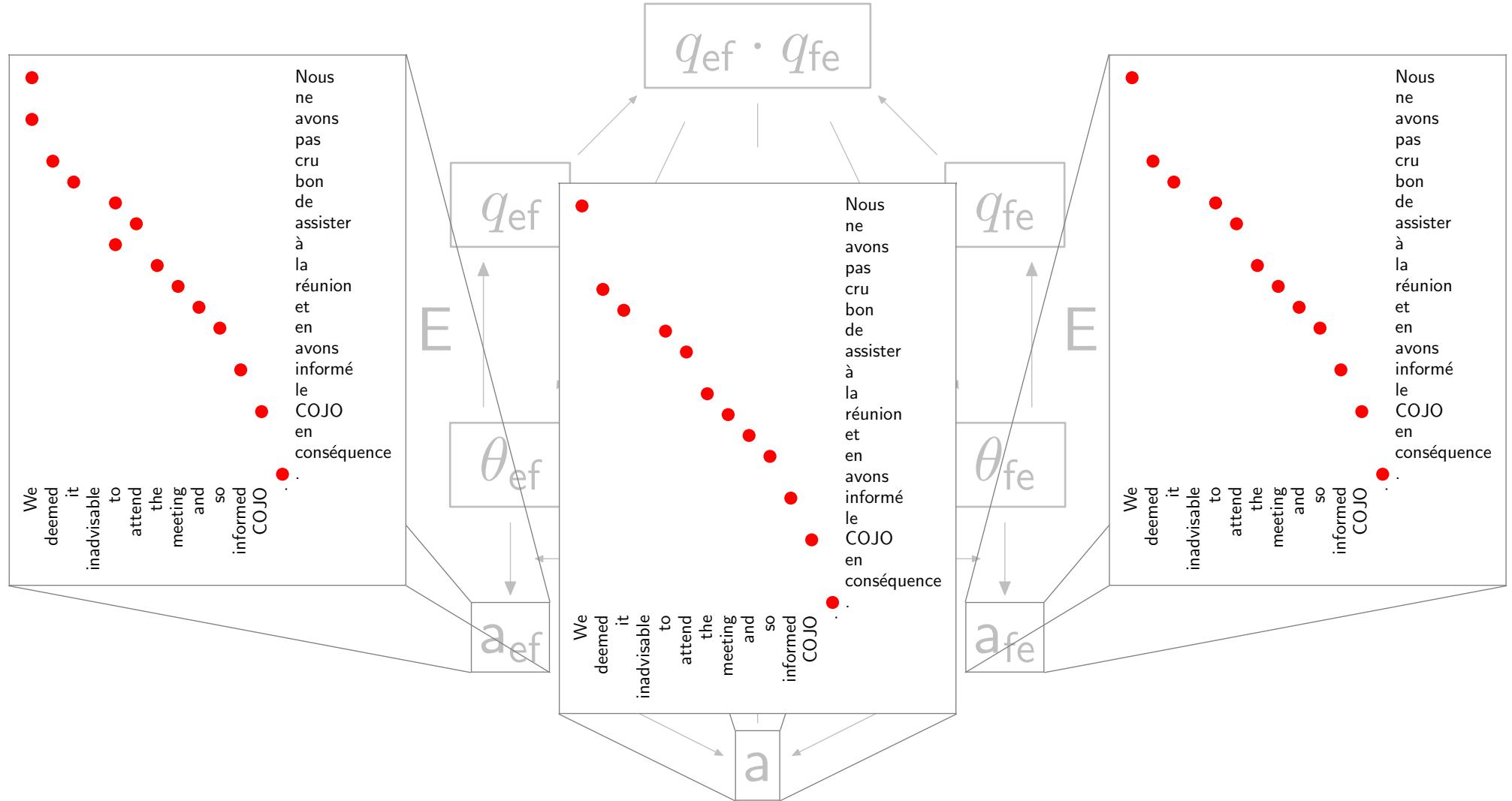
Two complementary models

Soft intersection: multiply fractional alignment



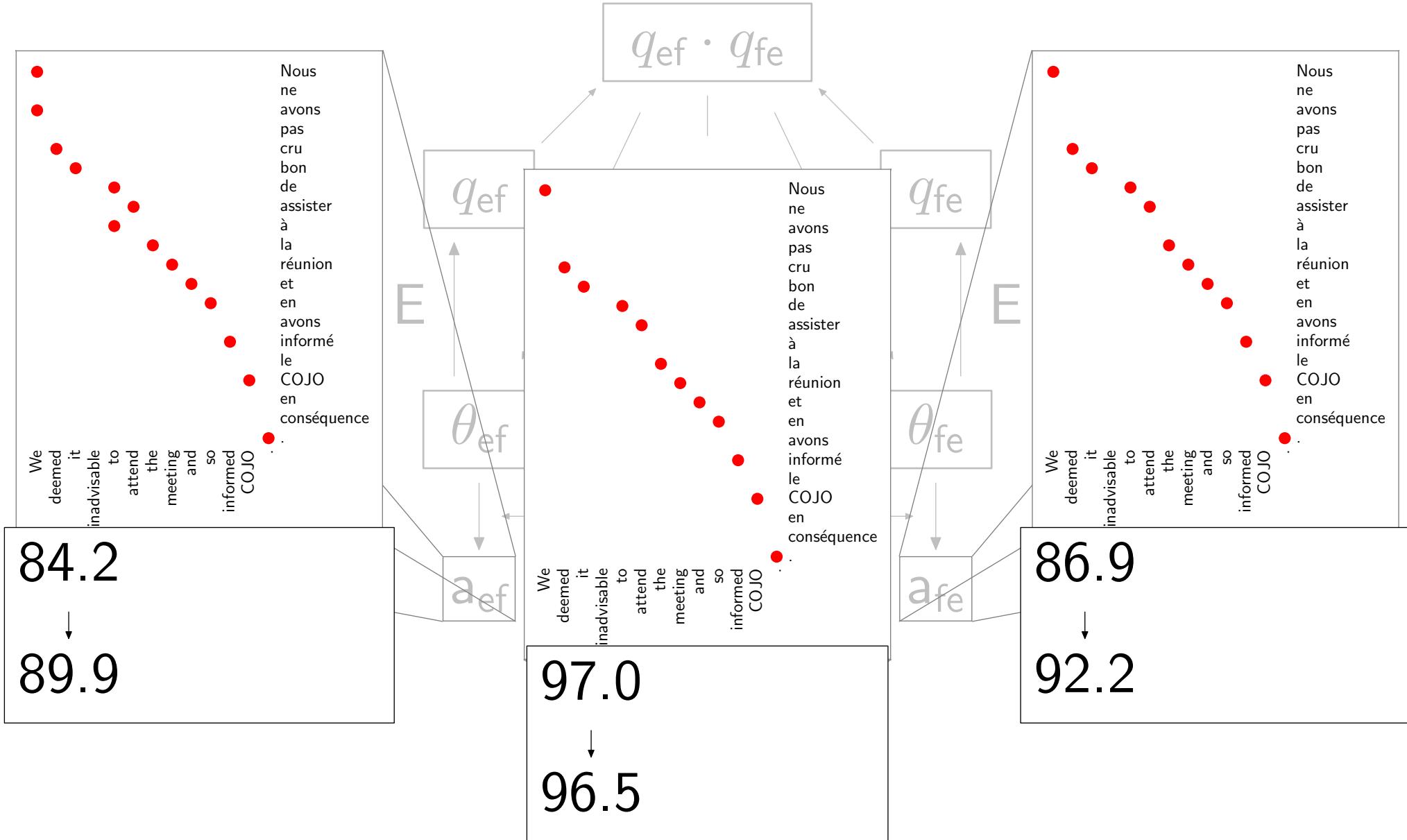
Two complementary models

Models that are trained to agree predict better.



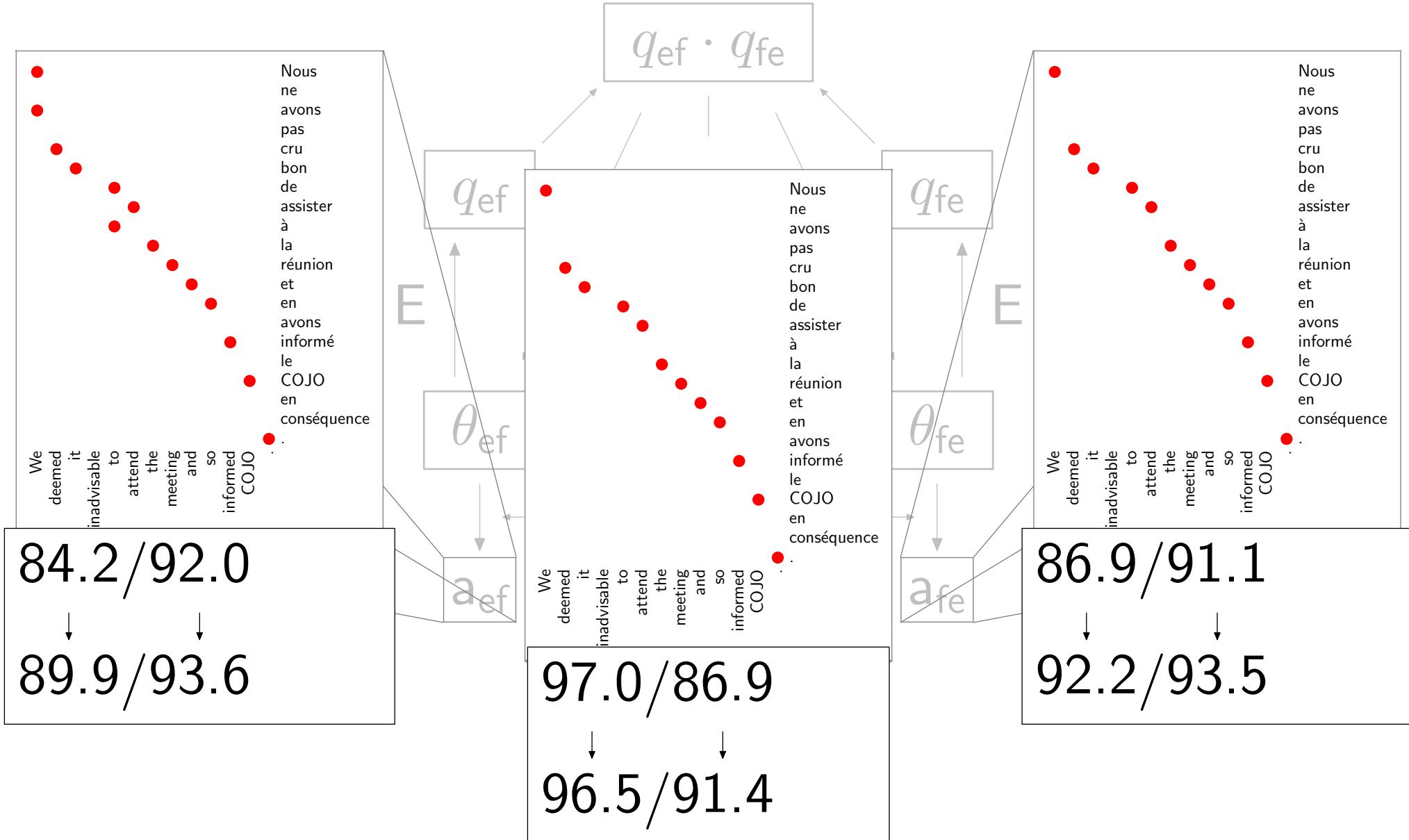
Two complementary models

Models that are trained to agree predict better.



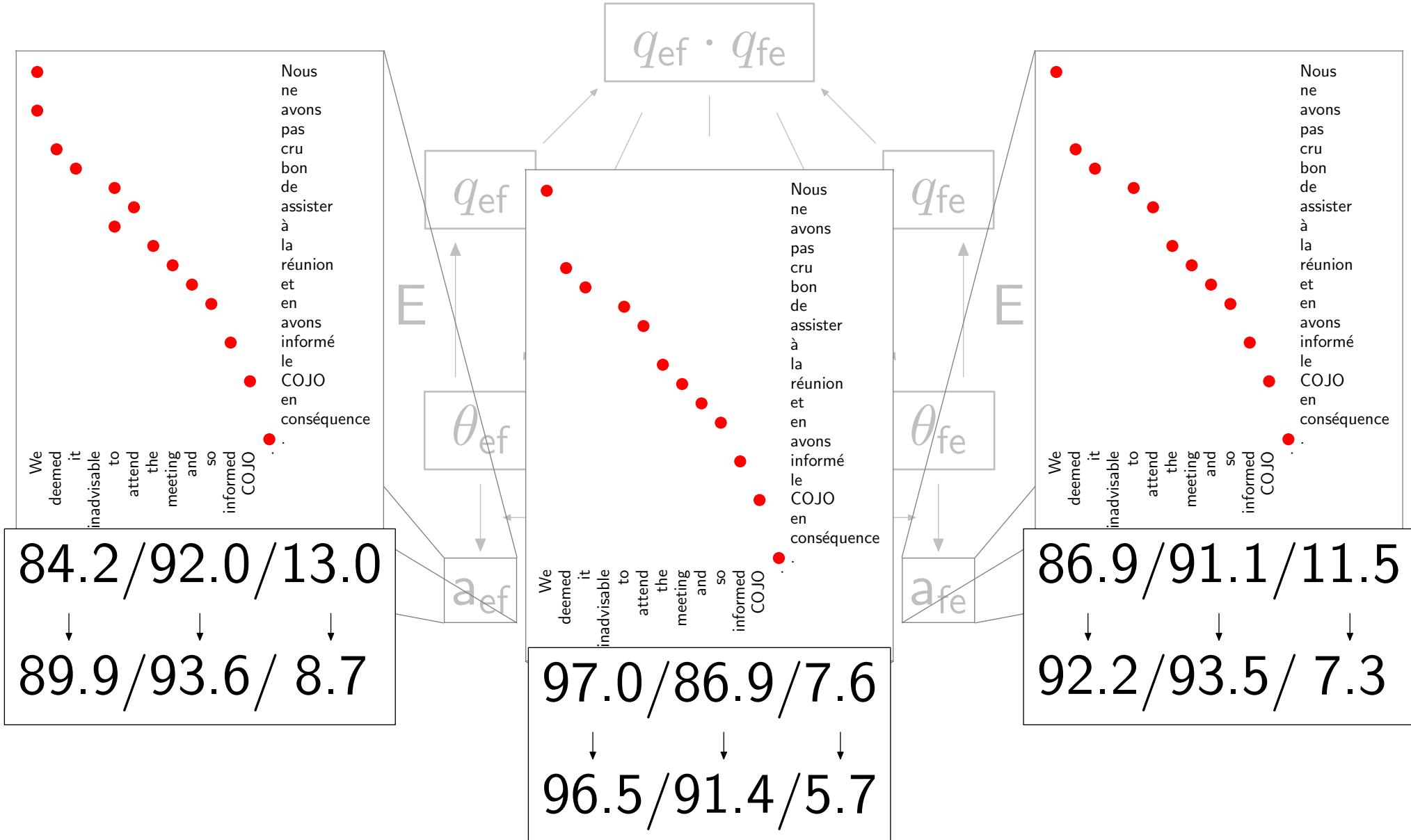
Two complementary models

Models that are trained to agree predict better.



Two complementary models

Models that are trained to agree predict better.



Initialization

Jointly-trained models less sensitive to initialization

| Initialization | Indep. HMMs |
|----------------|-------------|
| Uniform | AER>50 |

Initialization

Jointly-trained models less sensitive to initialization

| Initialization | Indep. HMMs |
|----------------|-------------|
| Uniform | AER>50 |
| Model 1 | 6.6 |

Initialization

Jointly-trained models less sensitive to initialization

| Initialization | Indep. HMMs | Joint HMMs |
|----------------|-------------|------------|
| Uniform | AER>50 | 5.7 |
| Model 1 | 6.6 | 5.2 |

Initialization

Jointly-trained models less sensitive to initialization

| Initialization | Indep. HMMs | Joint HMMs |
|----------------|-------------|------------|
| Uniform | AER>50 | 5.7 |
| Model 1 | 6.6 | 5.2 |

- Two models have somewhat disjoint capacities for producing bad alignments
- Agreement biases parameters away from troublesome areas

Agreement provides staged training

E-step:

$$q(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}) := p(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}; \theta_{\text{ef}}) p(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}; \theta_{\text{fe}})$$

Agreement provides staged training

E-step:

$$q(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}) := p(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}; \theta_{\text{ef}}) p(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}; \theta_{\text{fe}})$$

M-step:

$$\theta_t(to \rightarrow de) \propto \sum_{\mathbf{e}_i = to, \mathbf{f}_j = de} q(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f})$$

- Magnitude of fractional q = influence in M-step

Agreement provides staged training

E-step:

$$q(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}) := p(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}; \theta_{\text{ef}}) p(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f}; \theta_{\text{fe}})$$

M-step:

$$\theta_t(to \rightarrow de) \propto \sum_{\mathbf{e}_i = to, \mathbf{f}_j = de} q(\mathbf{a}_{i,j} \mid \mathbf{e}, \mathbf{f})$$

- Magnitude of fractional q = influence in M-step
- Downweight hard cases where two models disagree
- As models get better, harder examples contribute

General unsupervised approach

- Input $\mathbf{x} = (\mathbf{e}, \mathbf{f})$, output $\mathbf{z} = \mathbf{a}$
- Two complementary models $p_1(\mathbf{x}, \mathbf{z}; \theta_1), p_2(\mathbf{x}, \mathbf{z}; \theta_2)$

General unsupervised approach

- Input $\mathbf{x} = (\mathbf{e}, \mathbf{f})$, output $\mathbf{z} = \mathbf{a}$
- Two complementary models $p_1(\mathbf{x}, \mathbf{z}; \theta_1), p_2(\mathbf{x}, \mathbf{z}; \theta_2)$

$$\overbrace{\max_{\theta_1, \theta_2} \log p_1(\mathbf{x}; \theta_1) + \log p_2(\mathbf{x}; \theta_2)}^{\text{Independent training}}$$

General unsupervised approach

- Input $\mathbf{x} = (\mathbf{e}, \mathbf{f})$, output $\mathbf{z} = \mathbf{a}$
- Two complementary models $p_1(\mathbf{x}, \mathbf{z}; \theta_1), p_2(\mathbf{x}, \mathbf{z}; \theta_2)$
- Joint training objective:

$$\underbrace{\max_{\theta_1, \theta_2} \log p_1(\mathbf{x}; \theta_1) + \log p_2(\mathbf{x}; \theta_2)}_{\text{Independent training}} + \log \underbrace{\sum_{\mathbf{z}} p_1(\mathbf{z} \mid \mathbf{x}; \theta_1) p_2(\mathbf{z} \mid \mathbf{x}; \theta_2)}_{\text{agreement}}$$

General unsupervised approach

- Input $\mathbf{x} = (\mathbf{e}, \mathbf{f})$, output $\mathbf{z} = \mathbf{a}$
- Two complementary models $p_1(\mathbf{x}, \mathbf{z}; \theta_1), p_2(\mathbf{x}, \mathbf{z}; \theta_2)$
- Joint training objective:

$$\underbrace{\max_{\theta_1, \theta_2} \log p_1(\mathbf{x}; \theta_1) + \log p_2(\mathbf{x}; \theta_2)}_{\text{Independent training}} + \log \underbrace{\sum_{\mathbf{z}} p_1(\mathbf{z} \mid \mathbf{x}; \theta_1) p_2(\mathbf{z} \mid \mathbf{x}; \theta_2)}_{\text{agreement}}$$

E-step: $q(\mathbf{z} \mid \mathbf{x}) \propto p_1(\mathbf{z} \mid \mathbf{x}; \theta_1) p_2(\mathbf{z} \mid \mathbf{x}; \theta_2)$

General unsupervised approach

- Input $\mathbf{x} = (\mathbf{e}, \mathbf{f})$, output $\mathbf{z} = \mathbf{a}$
- Two complementary models $p_1(\mathbf{x}, \mathbf{z}; \theta_1), p_2(\mathbf{x}, \mathbf{z}; \theta_2)$
- Joint training objective:

$$\underbrace{\max_{\theta_1, \theta_2} \log p_1(\mathbf{x}; \theta_1) + \log p_2(\mathbf{x}; \theta_2)}_{\text{Independent training}} + \log \underbrace{\sum_{\mathbf{z}} p_1(\mathbf{z} \mid \mathbf{x}; \theta_1) p_2(\mathbf{z} \mid \mathbf{x}; \theta_2)}_{\text{agreement}}$$

$$\text{E-step: } q(\mathbf{z} \mid \mathbf{x}) \propto p_1(\mathbf{z} \mid \mathbf{x}; \theta_1) p_2(\mathbf{z} \mid \mathbf{x}; \theta_2)$$

Useful in grammar induction [Klein, Manning '04]

General unsupervised approach

- Input $\mathbf{x} = (\mathbf{e}, \mathbf{f})$, output $\mathbf{z} = \mathbf{a}$
- Two complementary models $p_1(\mathbf{x}, \mathbf{z}; \theta_1), p_2(\mathbf{x}, \mathbf{z}; \theta_2)$
- Joint training objective:

$$\underbrace{\max_{\theta_1, \theta_2} \log p_1(\mathbf{x}; \theta_1) + \log p_2(\mathbf{x}; \theta_2)}_{\text{Independent training}} + \log \underbrace{\sum_{\mathbf{z}} p_1(\mathbf{z} \mid \mathbf{x}; \theta_1) p_2(\mathbf{z} \mid \mathbf{x}; \theta_2)}_{\text{agreement}}$$

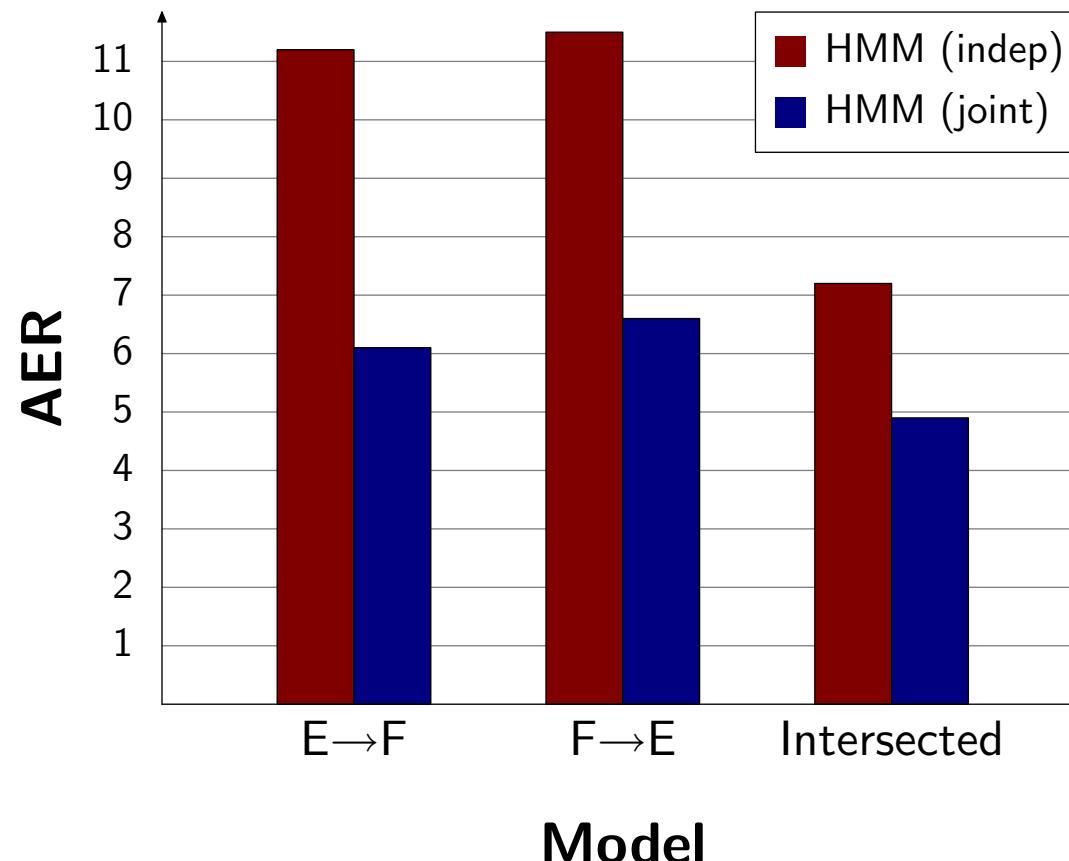
E-step: $q(\mathbf{z} \mid \mathbf{x}) \propto p_1(\mathbf{z} \mid \mathbf{x}; \theta_1) p_2(\mathbf{z} \mid \mathbf{x}; \theta_2)$

Useful in grammar induction [Klein, Manning '04]

Related work: co-training [Blum, Mitchell '98]
CoBoost [Collins, Singer '99]

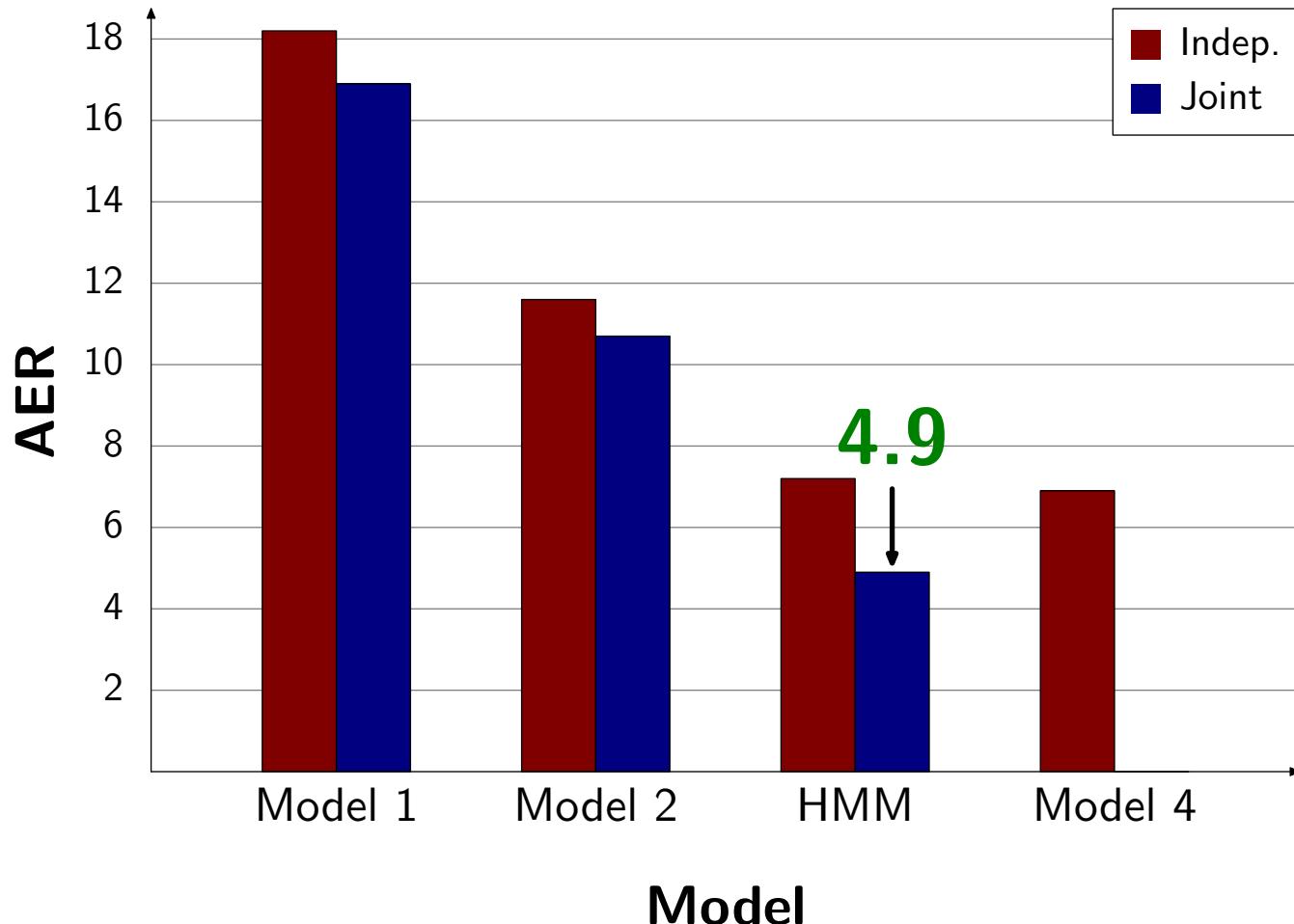
Final results

Hansards (1.1M training sentences, 347 test sentences)



Joint training improves both
directional and intersected models

Final results



Significant error reduction for various models

29% reduction in AER over model 4

Conclusion

- Simple and efficient procedure → 4.9% AER

Conclusion

- Simple and efficient procedure → 4.9% AER
- Suggests a general approach for unsupervised learning

Conclusion

- Simple and efficient procedure → 4.9% AER
- Suggests a general approach for unsupervised learning
- Achieves insignificantly better BLEU score

Conclusion

- Simple and efficient procedure → 4.9% AER
- Suggests a general approach for unsupervised learning
- Achieves insignificantly better BLEU score
- Provides features for discriminative methods