From *n*-gram-based to CRF-based Translation Models

Lavergne - Crego - Allauzen - Yvon

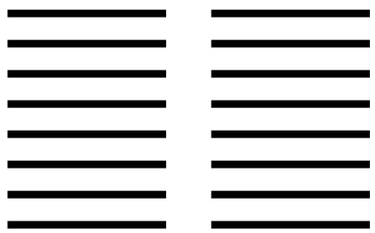
LIMSI


I CAN HAS SIMPLER TRAINING?

In principio erat
verbum: et ver
bum erat apud
deum: et deus
erat verbū. Hoc
erat in principio
ad deum. Omnia per ipsū
facta sunt: et sine ipso factum
est nihil: quod factum est. In

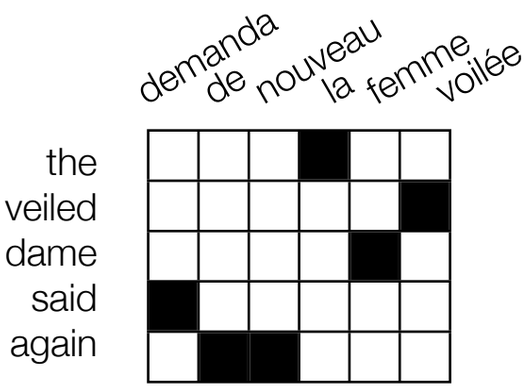
1-gram-based *noisy channel* MT

Bi-text



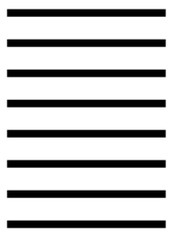
EM

Word alignments
& translation model
& reordering model



p(the|la)
p(veiled|voilée)
p(again|nouveau)
...

Monolingual
text



MLE

Language model

p(nouveau|de)
p(nouveau|demanda de)
...

Decoder

$$e^* = \underset{e}{\operatorname{argmax}}\, p(f|e)p(e)$$

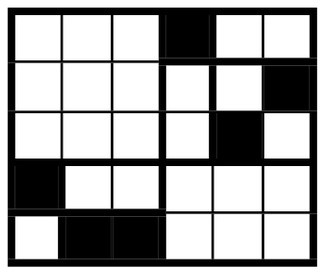
phrased-based *discriminative* MT

Bi-text

EM

the
veiled
dame
said
again

demanda de nouveau la femme voilée



p(the|la)
p(veiled|voilée)
p(again|nouveau)
...

MLE

p(veiled|voilée)
p(veiled dame|femme voilé)
p(said again|demanda de nouveau)
...

Phrase translation model
+ reordering model

Monolingual
text

MLE

Language model

p(nouveau|de)
p(nouveau|demanda de)
...

Tuning bi-text

$\lambda_{TM}\ \lambda_{LM}\ (\lambda_i)_i$

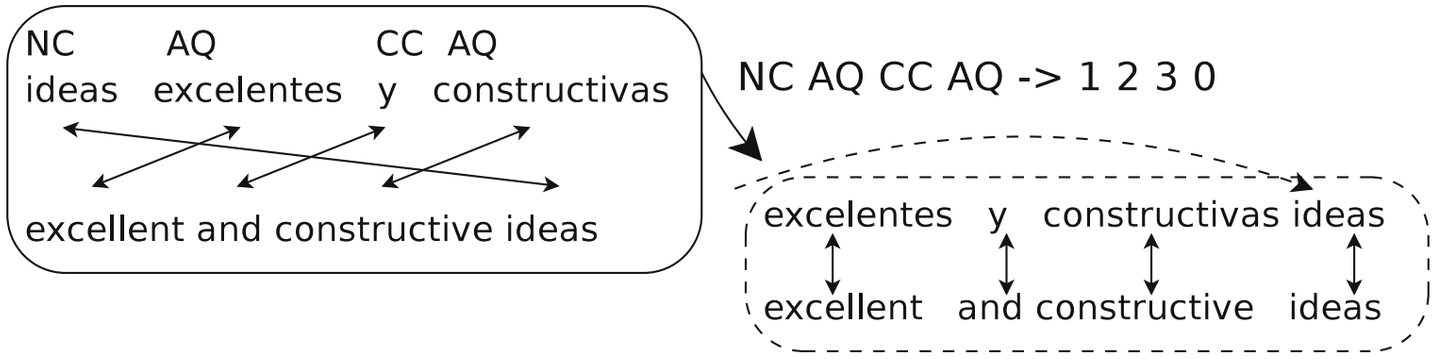
MERT

Decoder

$$e^* = \underset{e}{\operatorname{argmax}}\, \lambda_{TM} \log p(f|e) + \lambda_{LM} \log p(e) + \sum_i \lambda_i f_i(f, e)$$

Reordering for *n*-gram models

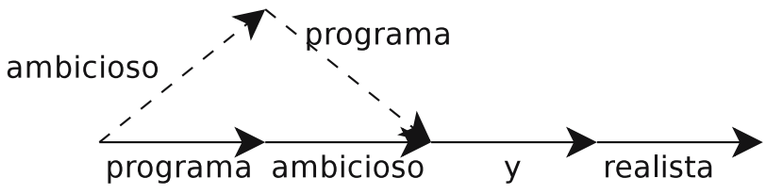
Training from word-aligned POS-tagged bi-text



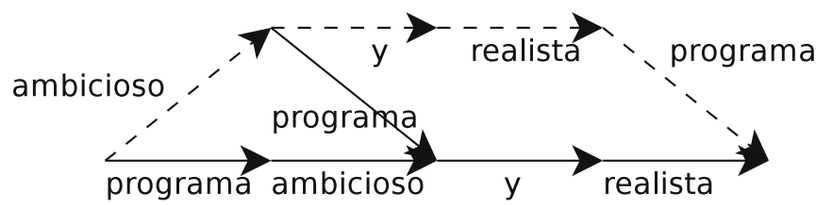
Input search graph

programa ambicioso y realista
NC AQ CC AQ

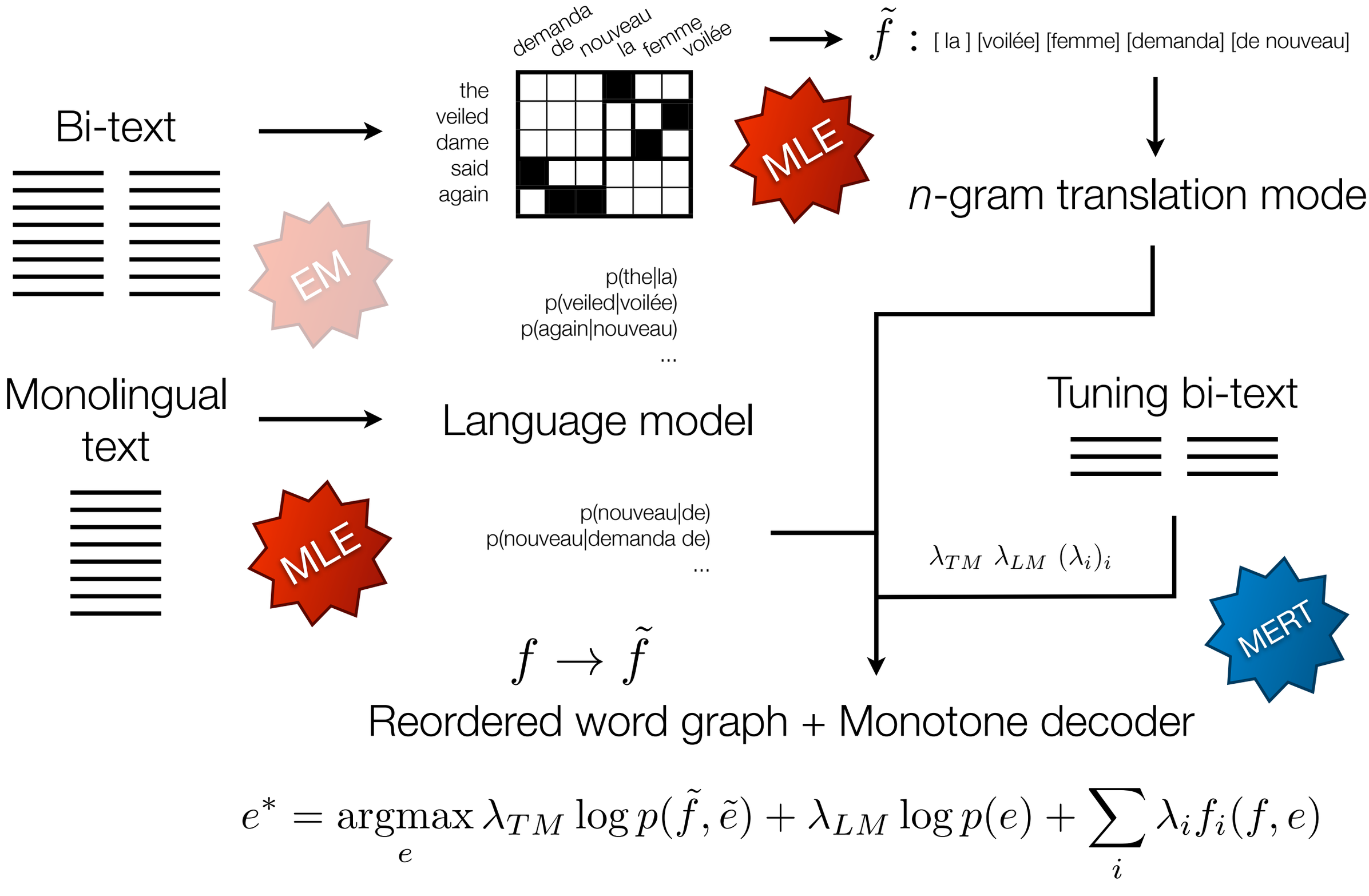
NC AQ -> 1 0



NC AQ CC AQ -> 1 2 3 0



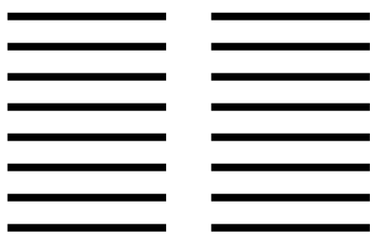
n-gram-based *discriminative* MT



$$e^* = \underset{e}{\operatorname{argmax}}\, \lambda_{TM} \log p(\tilde{f}, \tilde{e}) + \lambda_{LM} \log p(e) + \sum_i \lambda_i f_i(f, e)$$

CRF-based MT

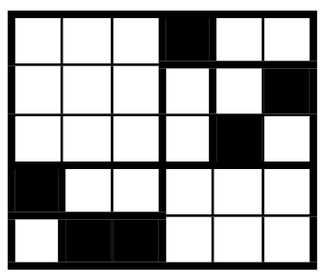
Bi-text



EM

demanda de nouveau la femme voilée

the veiled dame said again



MLE

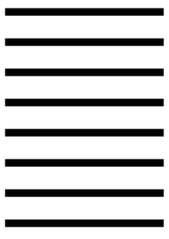
p(the|la)
p(veiled|voilée)
p(again|nouveau)

$\tilde{f}$: [la] [voilée] [femme] [demanda] [de nouveau]

CRF translation model

$$P(y_1^L|x_1^L) = \frac{1}{Z(x_1^L;\theta)} \exp(\theta^T G(x_1^L, y_1^L))$$

Monolingual text



MLE

Language model

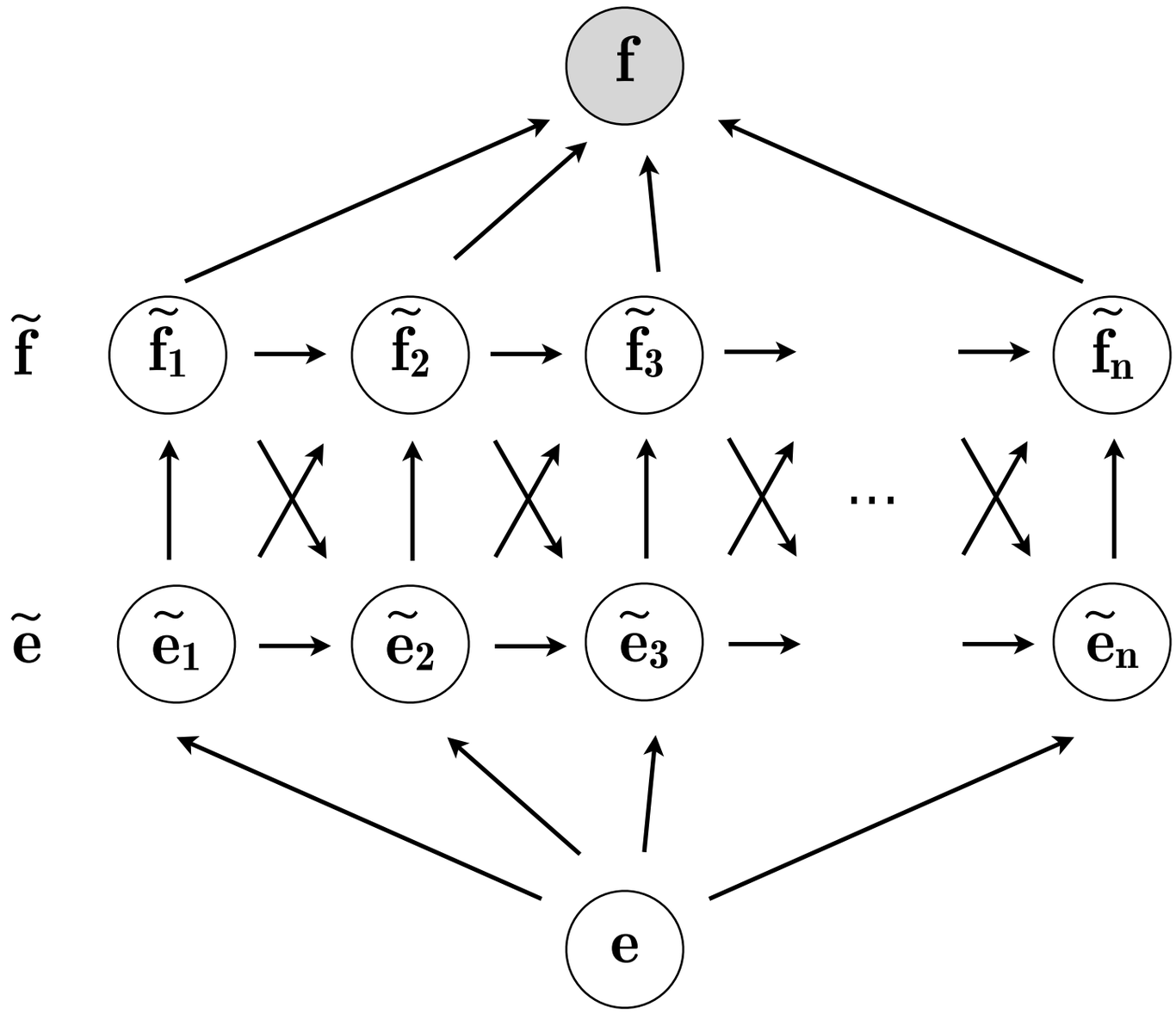
p(nouveau|de)
p(nouveau|demanda de)
...

FST: - segment & reorder
- monotone decoder

$$\widetilde{\mathbf{e}}^* = \arg\max_{\widetilde{\mathbf{e}}} P(\widetilde{\mathbf{e}}|\mathbf{f})$$

$$\approx \arg\max_{\tilde{\mathbf{f}} \in \mathcal{L}(\mathbf{f}), \widetilde{\mathbf{e}}} P(\widetilde{\mathbf{e}}, |\tilde{\mathbf{f}}, \mathbf{f}) P(\tilde{\mathbf{f}}|\mathbf{f})$$

n-gram generative story



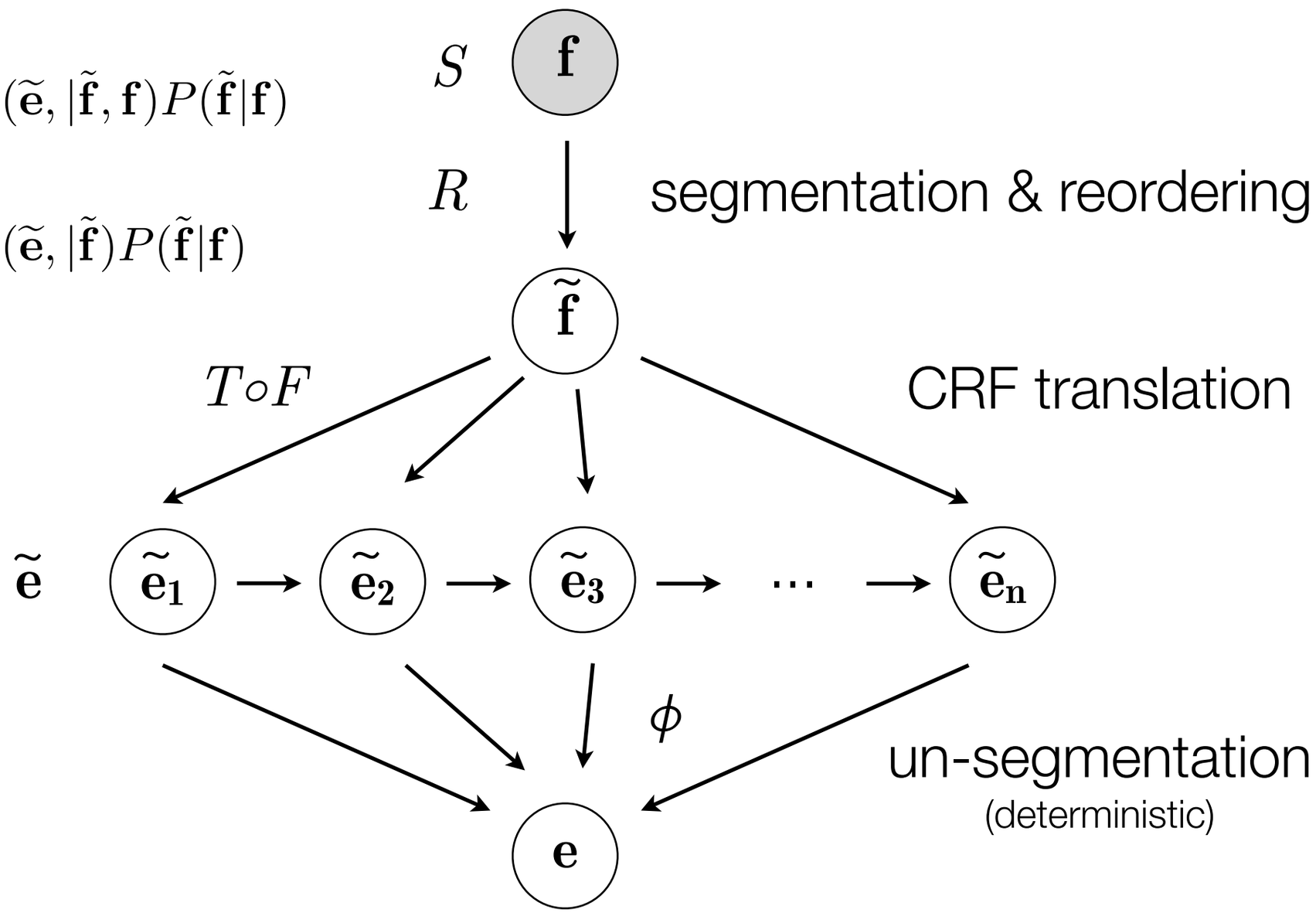
reordering (POS)

n-gram translation

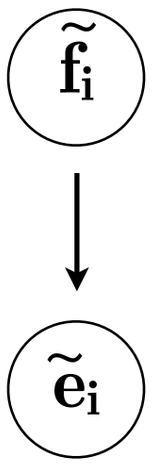
segmentation

Discriminative story

$$
\begin{aligned}
P(\mathbf{e}|\mathbf{f}) &= \sum_{\tilde{\mathbf{f}},\tilde{\mathbf{e}}|\phi(\tilde{\mathbf{e}})=\mathbf{e}} P(\tilde{\mathbf{e}},\tilde{\mathbf{f}}|\mathbf{f}) \\
&= \sum_{\tilde{\mathbf{f}},\tilde{\mathbf{e}}|\phi(\tilde{\mathbf{e}})=\mathbf{e}} P(\tilde{\mathbf{e}},|\tilde{\mathbf{f}},\mathbf{f})P(\tilde{\mathbf{f}}|\mathbf{f}) \\
&= \sum_{\tilde{\mathbf{f}},\tilde{\mathbf{e}}|\phi(\tilde{\mathbf{e}})=\mathbf{e}} P(\tilde{\mathbf{e}},|\tilde{\mathbf{f}})P(\tilde{\mathbf{f}}|\mathbf{f})
\end{aligned}
$$



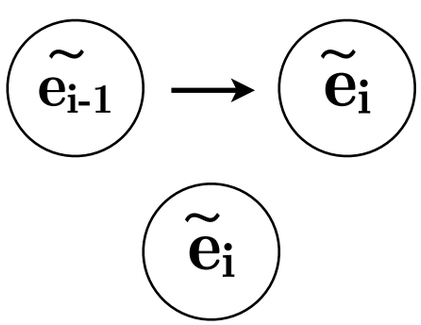
CRF Features



translation features [**trs**]

+context features [**ctx**]

+suffix/prefix features [**ix**]
+segmentation (length) features [**seg**]
+distortion (**Δ**+lex) features [**ord**]



target bigram features [**trg**]

+LM (non-local)

Negative results

	dev	test	# feat.
Moses (3g)	21.2	20.5	
n-gram (2g,3g)	20.6	20.2	755K
n-gram (3g,3g)	21.5	21.2	755K
CRF (**trs,trg**)	-	18.3	660K
CRF +**ctx**	-	18.8	1.5M
CRF +**ix,ord,seg**	-	19.1	1.5M
CRF +**ix,ord,seg**+3g	-	19.1	1.5M

+Language model does not help

Weaknesses:

- scoring of reordering & segmentation
- target LM

Oracle segmentation / reordering

	dev	test	# feat.
decoding with optimal segmentation/reordering			
CRF (**trs,trg**)	23.8	25.1	660K
CRF +**ctx**	24.1	25.4	1.5M
CRF +**ix,ord,seg**	24.3	25.6	1.5M
decoding with optimal reordering			
n-gram (2g,3g)	20.6	24.1	755K
n-gram (3g,3g)	21.5	25.2	755K
CRF **trs,trg**	-	22.8	660K
CRF +**ctx**	-	23.1	1.5M
CRF +**ix,ord,seg**	-	23.5	1.5M
regular decoding			
Moses (3g)	21.2	20.5	
n-gram (2g,3g)	20.6	20.2	755K
n-gram (3g,3g)	21.5	21.2	755K
CRF (**trs,trg**)	-	18.3	660K
CRF +**ctx**	-	18.8	1.5M
CRF +**ix,ord,seg**	-	19.1	1.5M
CRF +**ix,ord,seg**+3g	-	19.1	1.5M