Unsupervised Neural Machine Translation with SMT as Posterior Regularization

Shuo Ren[1,2*], Zhirui Zhang[3], Shujie Liu[4], Ming Zhou[4] and Shuai Ma[1,2]

[1]SKLSDE Lab, Beihang University, China

[2]Beijing Advanced Innovation Center for Big Data and Brain Computing

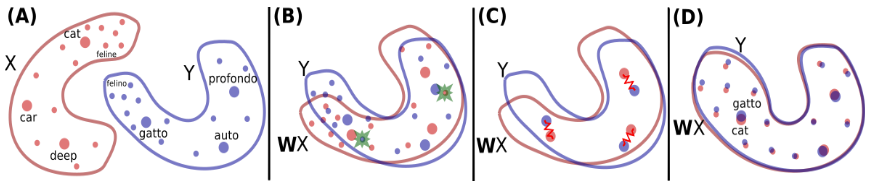
[3]University of Science and Technology of China, Hefei, China

[4]Microsoft Research Asia, Beijing, China

* Contribution during internship at Microsoft Research Asia.

Background

- Two main components of unsupervised NMT (Lample et al. 2018)
  - Model Initialization
  - Iterative Back-translation



Cross-lingual Word Embedding

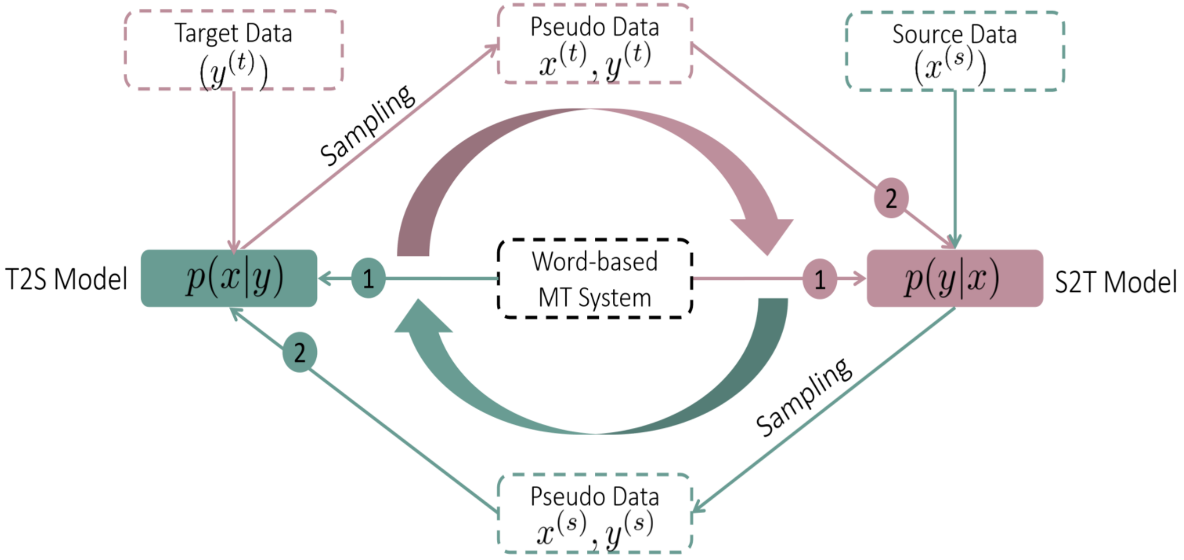
Translation Table

$P_{word}(y|x)$

$P_{word}(x|y)$

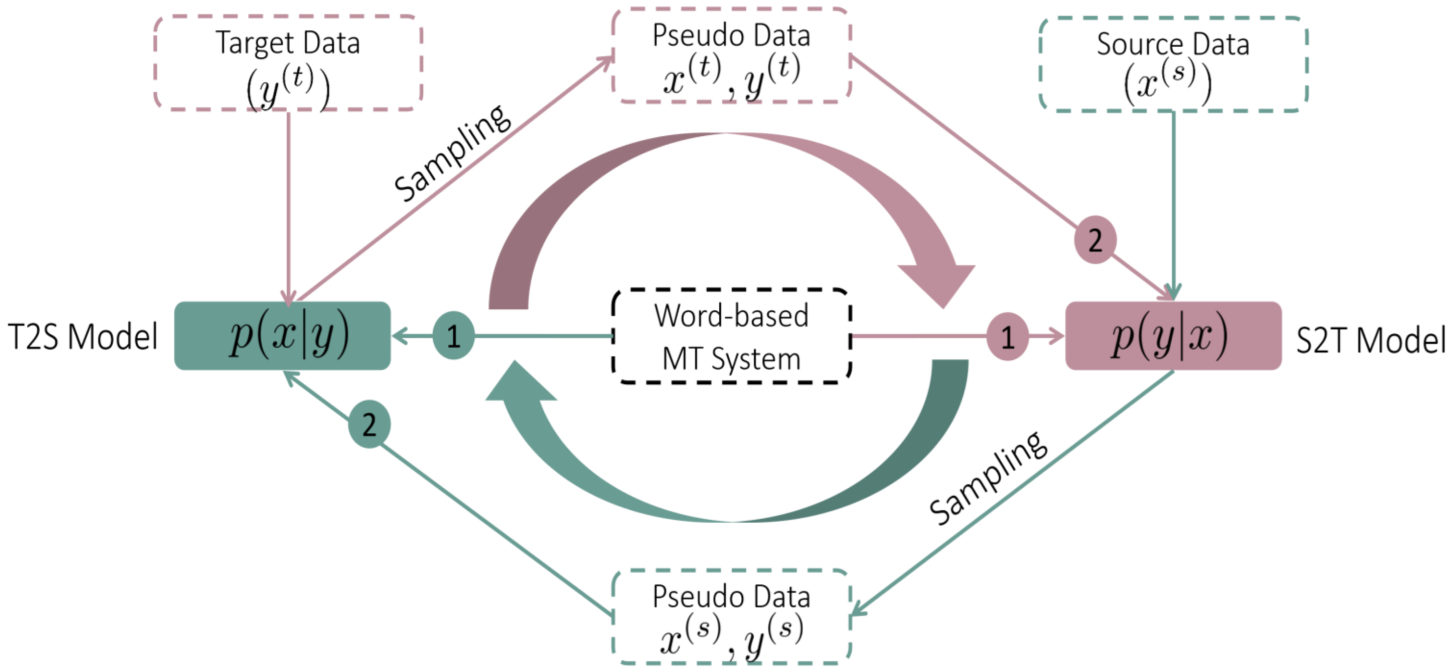
hello ↔ 你好
my ↔ 我的
good ↔ 好
bad ↔ 差
...

Word-based MT System



Motivation

- Noisy pseudo data generated with back-translation method



- Bad pseudo sentence pairs hurt the performance.
- Due to weak supervision, noises and errors will be accumulated and reinforced.
- SMT performs better than NMT in tackling noisy data (Khayrallah et al. 2018)

Unsupervised MT

- Unsupervised NMT (Artetxe et al. (2017), Lample et al. (2017), Yang et al. (2018))
  - Modifications of the enc-dec structure.
  - Weight sharing of both translation directions
  - Denoising auto-encoder and iterative back-translation are leveraged
- Unsupervised SMT (Artetxe et al. (2018), Lample et al. (2018))
  - Initialized by word-to-word translation tables.
  - Iterative back-translation performed by two SMT models

Combination NMT with SMT

- NMT as feature
  - He et al. (2016) integrate probability calculated by NMT as a feature into a log-linear model.
- Introducing phrase table into NMT
  - Tang et al. (2016) and Wang et al. (2017) leverage gate mechanisms to introduce a phrase table or candidates provided by SMT into NMT models.
- SMT as posterior regularization
  - Zhang et al. (2017) integrate more prior knowledge defined by some SMT features into NMT with the framework of posterior regularization.

Combination NMT with SMT

- NMT as feature
  - He et al. (2016) integrate probability calculated by NMT as a feature into a log-linear model.
- Introducing phrase table into NMT
  - Tang et al. (2016) and Wang et al. (2017) leverage gate mechanisms to introduce a phrase table or candidates provided by SMT into NMT models.
- SMT as **posterior regularization**
  - Zhang et al. (2017) integrate more prior knowledge defined by some SMT features into NMT with the framework of posterior regularization.
  - **√ leaving the architecture of NMT unchanged**

Posterior Regularization

- Posterior regularization (Ganchev et al. 2010) can incorporates indirect supervision from a desired distribution $q(y)$ via constraints on posterior distrib

$$F(q;\theta) = \mathcal{L}(\theta) - \sum_{n=1}^{N} \min_{q \in Q} \mathbf{KL}(q(\mathbf{y})||p(\mathbf{y}|\mathbf{x}_n;\theta))$$

Q is a constraint posterior set satisfying: $Q = \{q(\mathbf{y}) : \mathbf{E}_q[\phi(\mathbf{x},\mathbf{y})] \leq \mathbf{b}\}$

- Update models via EM framework:

$$E : q^{t+1} = \arg\min_{q \in Q} \mathbf{KL}(q(\mathbf{y})||p(\mathbf{y}|\mathbf{x}_n;\theta^t))$$

$$M : \theta^{t+1} = \arg\max_{\theta} \mathcal{L}(\theta) + \mathbf{E}_{q^{t+1}}[\log p(\mathbf{y}|\mathbf{x}_n;\theta)]$$

SMT as Posterior Regularization

- Leverage SMT to denoise and guide the training of unsupervised NMT models in the iterative back-translation process

- We replace the posterior regularization term $q(y)$ with the SMT models ($\overrightarrow{p_s}(y|x;\theta_{x\to y})$ and $\overleftarrow{p_s}(x|y;\theta_{y\to x})$):

$$\mathcal{J}(\theta_{\mathbf{x}\to\mathbf{y}}, \theta_{\mathbf{x}\leftarrow\mathbf{y}}, \overrightarrow{p_s}, \overleftarrow{p_s}) = \bar{\mathcal{L}}(\theta_{\mathbf{x}\to\mathbf{y}}, \theta_{\mathbf{x}\leftarrow\mathbf{y}}) - \sum_{i=1}^{M} \min_{\overrightarrow{p_s}} \mathbf{KL}(\overrightarrow{p_s}(\mathbf{y}|\mathbf{x}_i) || \overrightarrow{p_n}(\mathbf{y}|\mathbf{x}_i;\theta_{\mathbf{x}\to\mathbf{y}})) - \sum_{j=1}^{N} \min_{\overleftarrow{p_s}} \mathbf{KL}(\overleftarrow{p_s}(\mathbf{x}|\mathbf{y}_j) || \overleftarrow{p_n}(\mathbf{x}|\mathbf{y}_j;\theta_{\mathbf{x}\leftarrow\mathbf{y}}))$$

$$\bar{\mathcal{L}}(\theta_{\mathbf{x}\to\mathbf{y}}, \theta_{\mathbf{x}\leftarrow\mathbf{y}}) = \sum_{i=1}^{M} \mathbf{E}_{\mathbf{y}\sim\overrightarrow{p_n}(\mathbf{y}|\mathbf{x}_i;\theta_{\mathbf{x}\to\mathbf{y}})}[\log \overleftarrow{p_n}(\mathbf{x}_i|\mathbf{y};\theta_{\mathbf{x}\leftarrow\mathbf{y}})] + \sum_{j=1}^{N} \mathbf{E}_{\mathbf{x}\sim\overleftarrow{p_n}(\mathbf{x}|\mathbf{y}_j;\theta_{\mathbf{x}\leftarrow\mathbf{y}})}[\log \overrightarrow{p_n}(\mathbf{y}_j|\mathbf{x};\theta_{\mathbf{x}\to\mathbf{y}})]$$

EM Training Algorithm

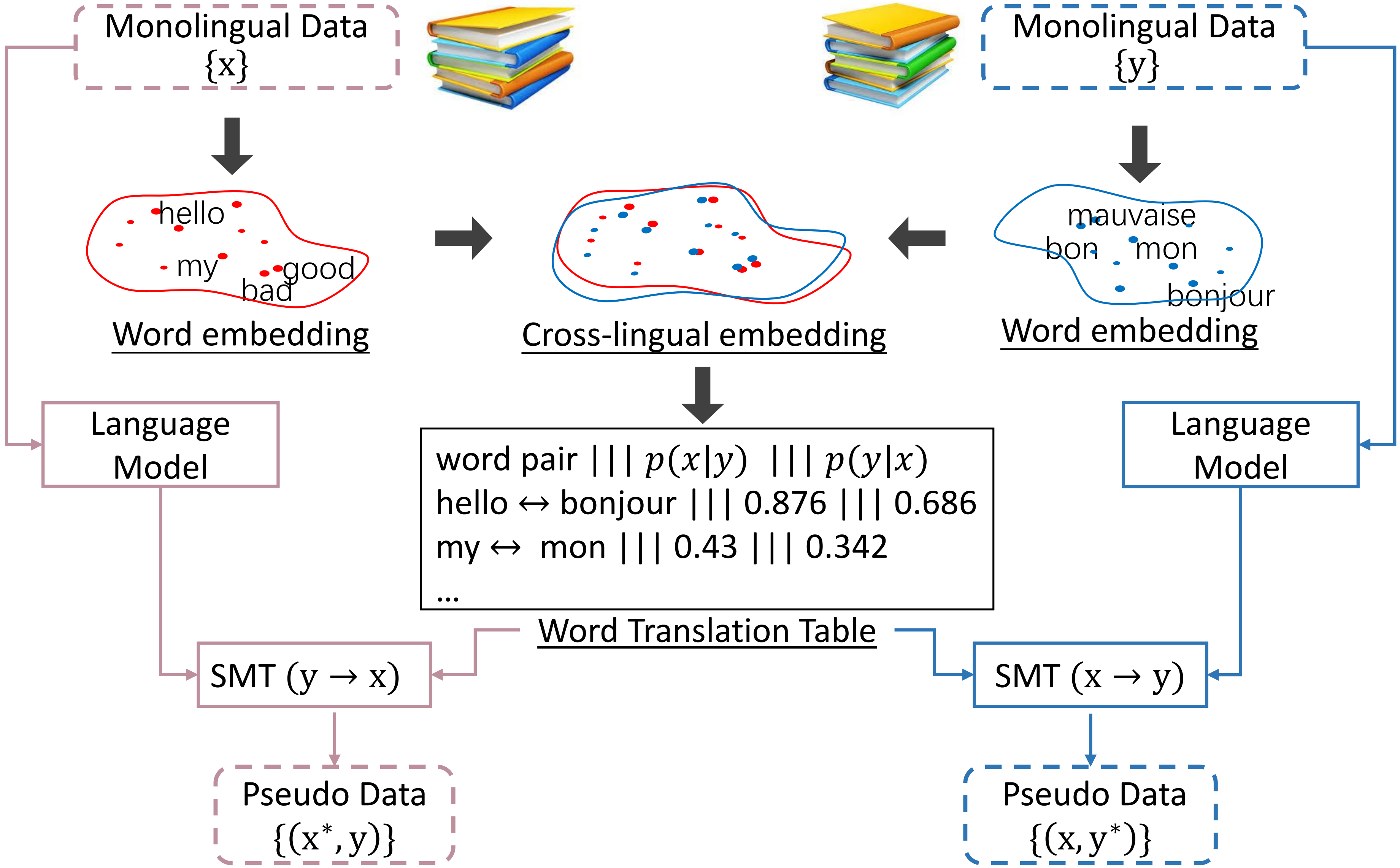
- E-Step: Optimize SMT models to minimize the KL distance between SMT models and NMT models

$$
\begin{aligned}
E: \overleftarrow{p_s}^{t+1} &= \arg\max_{\overleftarrow{p_s}} \mathcal{J}(\theta_{\mathbf{x}\rightarrow\mathbf{y}}, \theta_{\mathbf{x}\leftarrow\mathbf{y}}, \overrightarrow{p_s}, \overleftarrow{p_s}) \\
&= \arg\min_{\overleftarrow{p_s}} \mathbf{KL}(\overleftarrow{p_s}(\mathbf{x}|\mathbf{y}_j)||\overleftarrow{p_n}(\mathbf{x}|\mathbf{y}_j;\theta^t_{\mathbf{x}\leftarrow\mathbf{y}})) \\
\overrightarrow{p_s}^{t+1} &= \arg\max_{\overrightarrow{p_s}} \mathcal{J}(\theta_{\mathbf{x}\rightarrow\mathbf{y}}, \theta_{\mathbf{x}\leftarrow\mathbf{y}}, \overrightarrow{p_s}, \overleftarrow{p_s}) \\
&= \arg\min_{\overrightarrow{p_s}} \mathbf{KL}(\overrightarrow{p_s}(\mathbf{y}|\mathbf{x}_i)||\overrightarrow{p_n}(\mathbf{y}|\mathbf{x}_i;\theta^t_{\mathbf{x}\rightarrow\mathbf{y}}))
\end{aligned}
$$

- M-Step: Optimize NMT models using the pseudo data generated by SMT models and the corresponding reverse NMT models

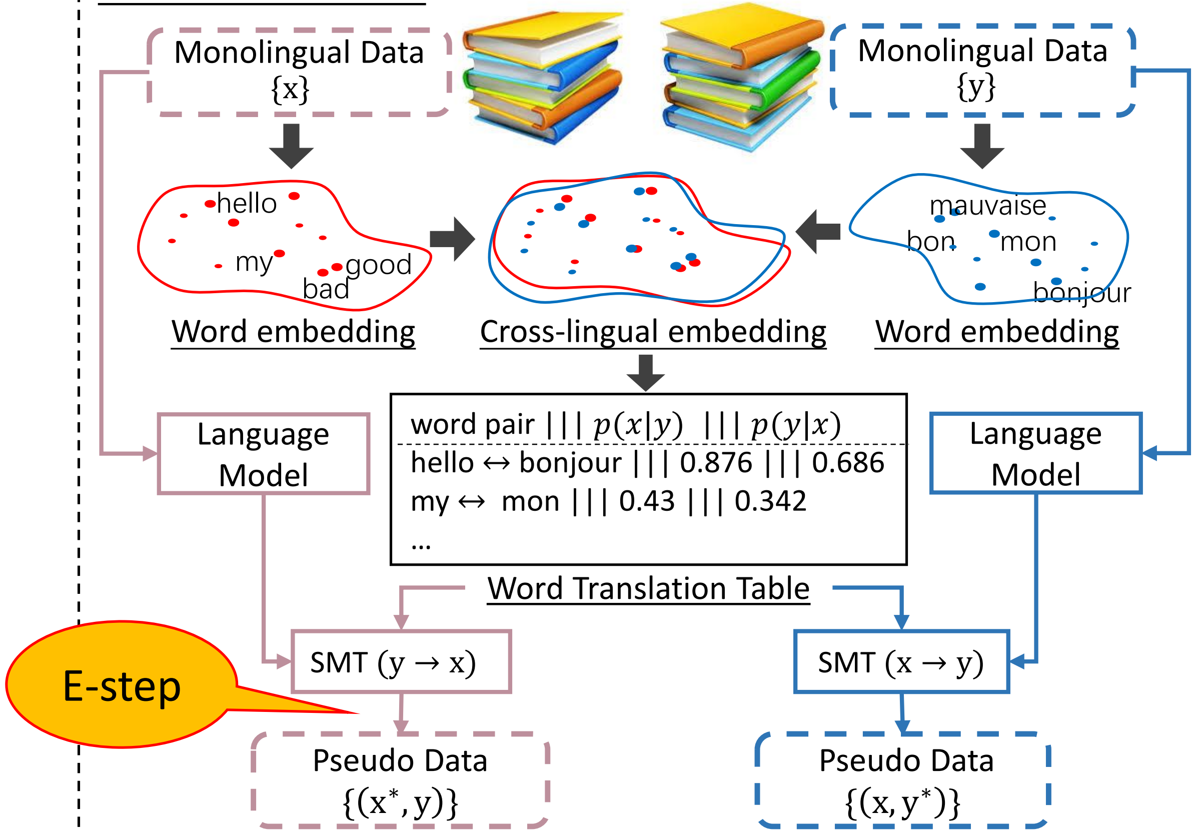
$$
\begin{aligned}
M: \theta^{t+1}_{\mathbf{x}\leftarrow\mathbf{y}} &= \arg\max_{\theta_{\mathbf{x}\leftarrow\mathbf{y}}} \mathcal{J}(\theta_{\mathbf{x}\rightarrow\mathbf{y}}, \theta_{\mathbf{x}\leftarrow\mathbf{y}}, \overrightarrow{p_s}, \overleftarrow{p_s}) \\
&= \arg\max_{\theta_{\mathbf{x}\leftarrow\mathbf{y}}} \{\mathbf{E}_{\overleftarrow{p_s}^{t+1}}[\log \overleftarrow{p_n}(\mathbf{x}|\mathbf{y}_j;\theta_{\mathbf{x}\leftarrow\mathbf{y}})] \\
&\quad + \mathbf{E}_{\overrightarrow{p_n}(\mathbf{y}|\mathbf{x}_i;\theta^t_{\mathbf{x}\rightarrow\mathbf{y}})}[\log \overleftarrow{p_n}(\mathbf{x}_i|\mathbf{y};\theta_{\mathbf{x}\leftarrow\mathbf{y}})]\} \\
\theta^{t+1}_{\mathbf{x}\rightarrow\mathbf{y}} &= \arg\max_{\theta_{\mathbf{x}\rightarrow\mathbf{y}}} \mathcal{J}(\theta_{\mathbf{x}\rightarrow\mathbf{y}}, \theta_{\mathbf{x}\leftarrow\mathbf{y}}, \overrightarrow{p_s}, \overleftarrow{p_s}) \\
&= \arg\max_{\theta_{\mathbf{x}\rightarrow\mathbf{y}}} \mathbf{E}_{\overrightarrow{p_s}^{t+1}}[\log \overrightarrow{p_n}(\mathbf{y}|\mathbf{x}_i;\theta_{\mathbf{x}\rightarrow\mathbf{y}})] \\
&\quad + \mathbf{E}_{\overleftarrow{p_n}(\mathbf{x}|\mathbf{y}_j;\theta^t_{\mathbf{x}\leftarrow\mathbf{y}})}[\log \overrightarrow{p_n}(\mathbf{y}_j|\mathbf{x};\theta_{\mathbf{x}\rightarrow\mathbf{y}})]
\end{aligned}
$$

Model Initialization



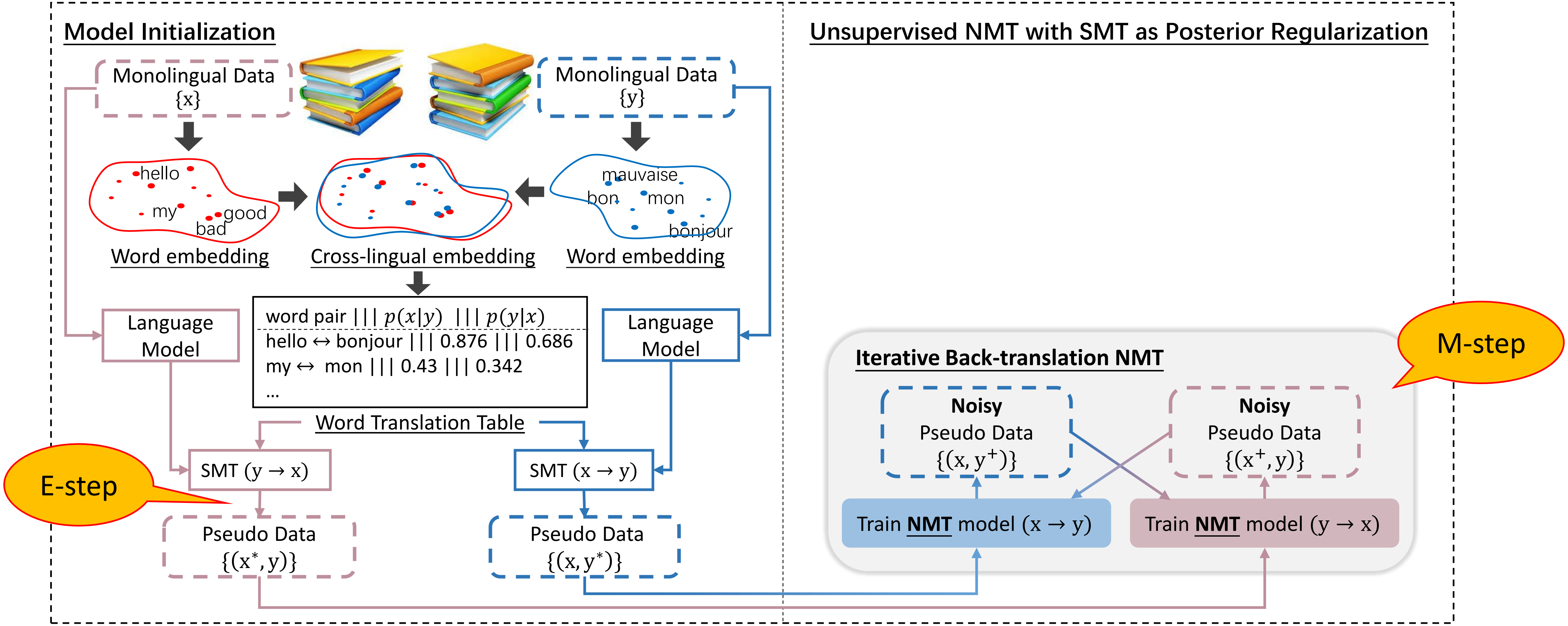
SMT as Posterior Regularization

Model Initialization

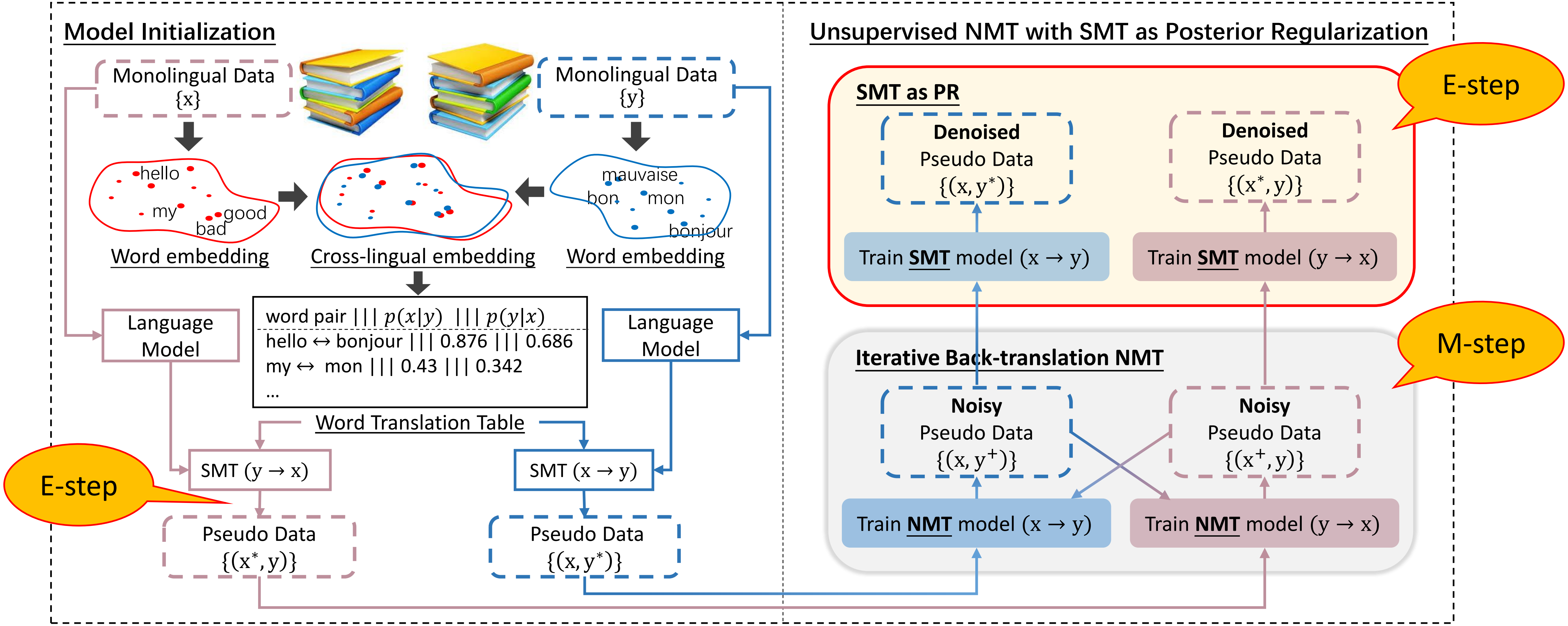


Unsupervised NMT with SMT as Posterior Regularization

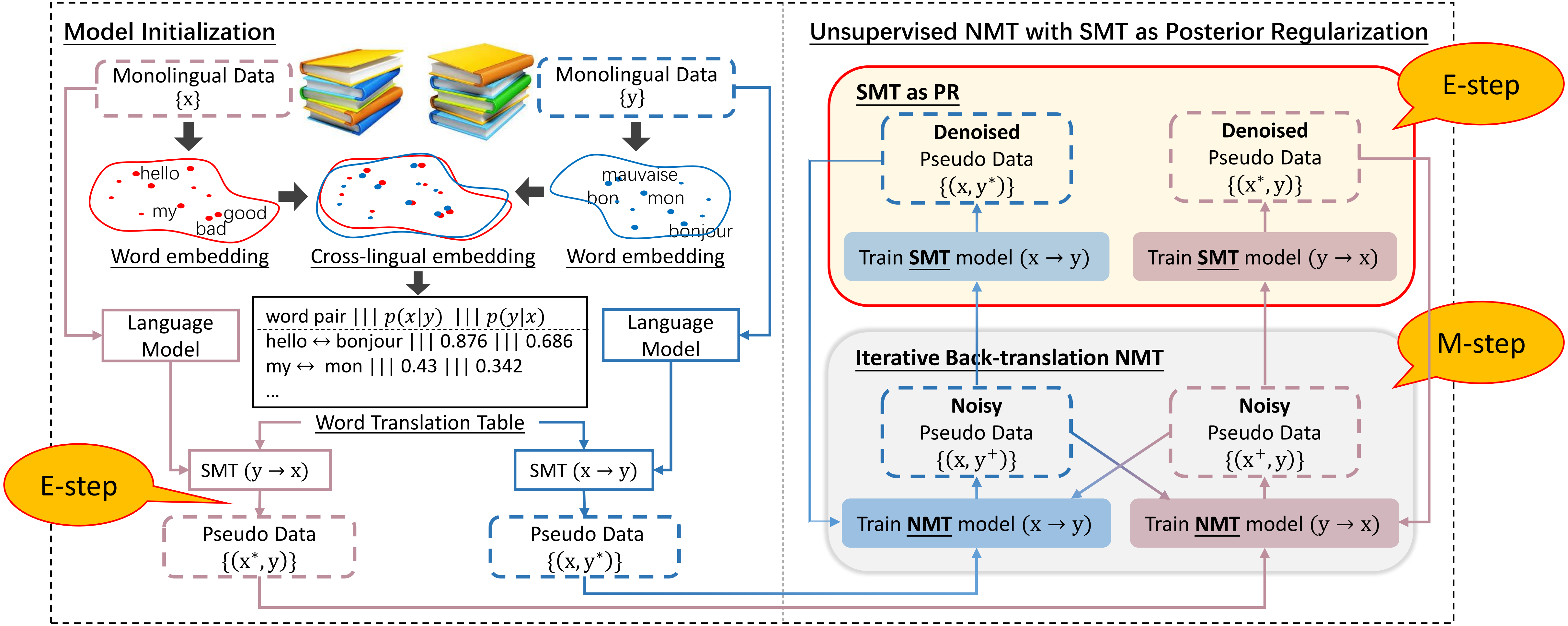
SMT as Posterior Regularization



SMT as Posterior Regularization



SMT as Posterior Regularization



Comparison Results

- Dataset
  - Monolingual data: Following the setting in (Lample et al. 2018), select 50M English, French and German sentences in NewsCrawl
  - Test data
    - English-French translation task: news-test 2014
    - English-German translation task: news-test 2014 and 2016
- Result

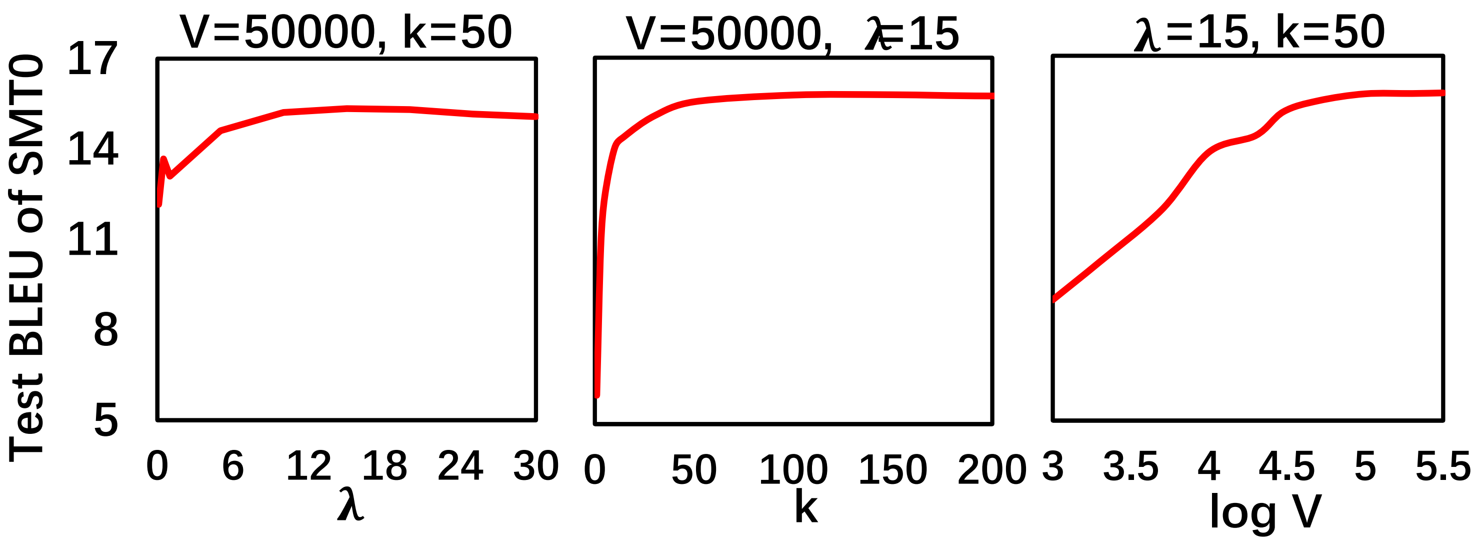
Method	fr-en	en-fr	de-en (2014)	en-de (2014)	de-en (2016)	en-de (2016)
(Artetxe et al. 2017)	15.56	15.13	10.21	6.89	-	-
(Lample, Denoyer, and Ranzato 2017)	14.31	15.05	-	-	13.33	9.64
(Yang et al. 2018)	15.58	16.97	-	-	14.62	10.86
(Lample et al. 2018), NMT	24.18	25.41	-	-	21.00	17.16
(Lample et al. 2018), PBSMT	27.16	28.11	-	-	22.68	17.77
(Lample et al. 2018), NMT+PBSMT	26.29	27.12	-	-	22.06	17.52
(Lample et al. 2018), PBSMT+NMT	27.68	27.60	-	-	25.19	20.23
Our Method	**28.92**	**29.53**	**20.43**	**16.97**	**26.32**	**21.65**

Table 1: Comparison with previous methods.

Supervised MT:
en-fr: 41.8
en-de(2014): 28.4

Discussion on Initialization

- Test of initial models with various hyper-parameters



$$p(y_j|x_i) = \frac{\exp\left[\lambda \cos(e_{x_i}, e_{y_j})\right]}{\sum_k \exp\left[\lambda \cos(e_{x_i}, e_{y_k})\right]}$$

λ: the peakiness controller
K: tok-k candidates in the word-to-word translation table
V**:** vocabulary size of both languages

- The effect of SMT0 (the first SMT models) on NMT0 (the first NMT models)

Initialization Method	fr-en	en-fr	de-en	en-de
NMT0 without SMT0	12.29	12.46	7.32	4.81
NMT0 with SMT0	24.06	24.82	16.29	12.88

Using word-to-word translation to generate pseudo data rather than SMT0 models

Example

Source	J'ai eu des relations difficiles avec lui jusqu'à ce qu'il devienne vieux, malade.
SMT0	I've gotten of difficult relations with him until he will become old, sick.
NMT0	I've had difficult relations with him until he's become old, ill-fated.
SMT1	I've had difficult relationships with him until he became old, sick.
NMT1	I had difficult relations with him until he became old and sick.
Reference	I had a difficult relationship with him until he became old and ill.
Source	Le fonds d'investissement qui était propriétaire de cette bâtisse-là avait des choix à faire.
SMT0	The owner of this building, so had to make a choice of which was an investment fund.
NMT0	The investment fund that was an owner of that canopy-back business had plenty of choice to do.
SMT1	The investment fund that was the owner of this building just had to make choices.
NMT1	The investment fund that was the owner of this building had choices to make.
Reference	The investment fund that owned the building had to make a choice.
Source	M. Dutton a rendu visite à Mme Plibersek pour garantir qu'aucun dollar du plan de sauvetage ne sera dépensé en bureaucratie supplémentaire.
SMT0	Mr Dutton paid a visit to Ms Plibersek to guarantee that the greenback no rescue plan of not be spent in extra bureaucracy.
NMT0	Mr Dutton said Ms Plibersek'visit to guarantee any dollar from the rescue plan will be spent in extra bureaucracy.
SMT1	Mr Dutton was visiting Ms Plibersek to guarantee that no dollar rescue plan will be spent on additional bureaucracy.
NMT1	Mr Dutton paid a visit to Ms Plibersek to guarantee that no dollar from the rescue plan will be spent on extra bureaucracy.
Reference	Mr Dutton called on Ms Plibersek to guarantee that not one dollar out of the rescue package would be spent on additional bureaucracy.

Table 4: Cases of translation results from French to English in *newstest* 2014. The models of SMT0, NMT0, SMT1 and NMT1 are corresponding to the steps in Table 2.

Thanks!
Q & A