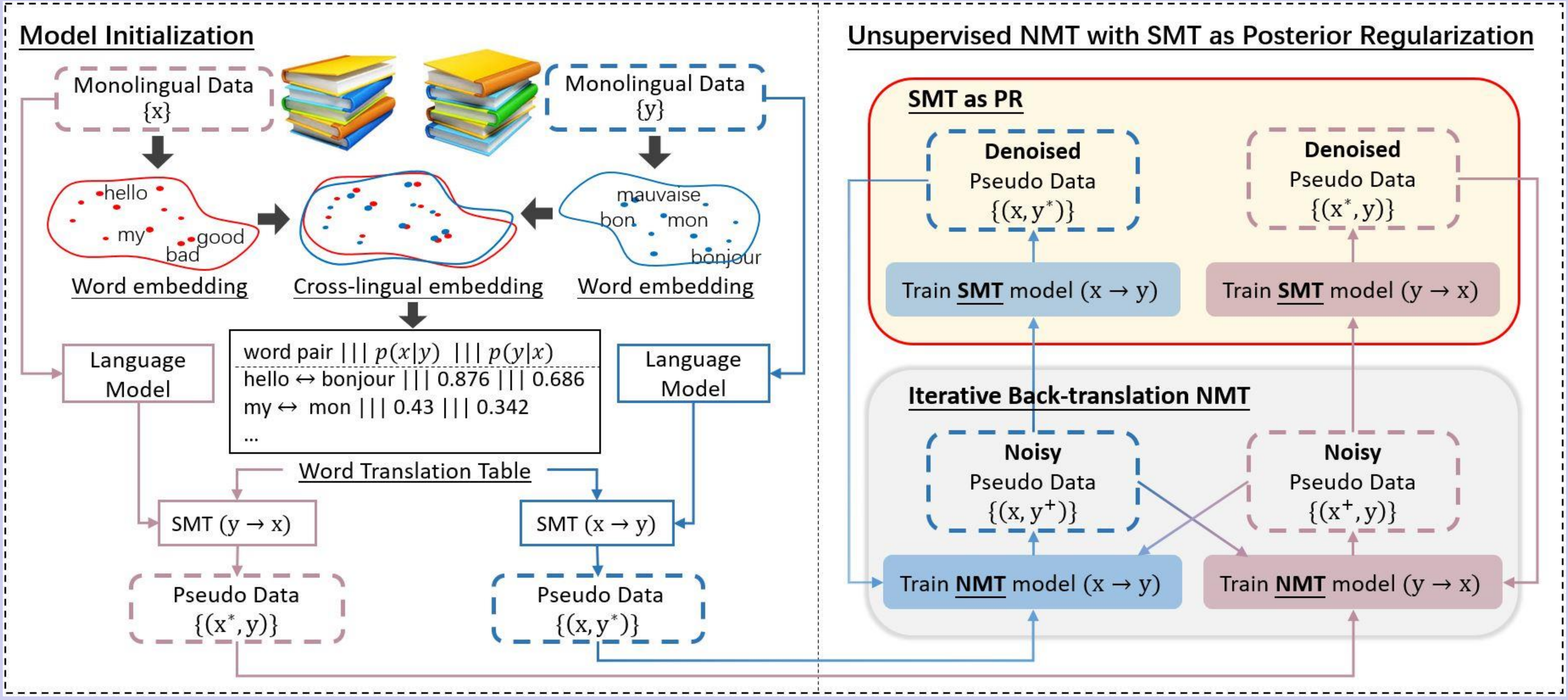
Unsupervised Neural Machine Translation with SMT as Posterior Regularization

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, Shuai Ma

Overview

The whole procedure consists of two parts, i.e., model initialization and unsupervised training.


Model Initialization
Monolingual Data {x}
Monolingual Data {y}
hello
my
good
bad
mauvaise
bon
mon
bonjour
Word embedding
Cross-lingual embedding
Word embedding
Language Model
word pair ||| $p(x|y)$ ||| $p(y|x)$
hello ↔ bonjour ||| 0.876 ||| 0.686
my ↔ mon ||| 0.43 ||| 0.342
…
Language Model
Word Translation Table
SMT (y → x)
SMT (x → y)
Pseudo Data $\{(x^*, y)\}$
Pseudo Data $\{(x, y^*)\}$
Unsupervised NMT with SMT as Posterior Regularization
SMT as PR
Denoised Pseudo Data $\{(x, y^*)\}$
Denoised Pseudo Data $\{(x^*, y)\}$
Train SMT model (x → y)
Train SMT model (y → x)
Iterative Back-translation NMT
Noisy Pseudo Data $\{(x, y^+)\}$
Noisy Pseudo Data $\{(x^+, y)\}$
Train NMT model (x → y)
Train NMT model (y → x)

- **Model initialization**: Build two SMT models using language models and word translation tables inferred from unsupervised cross-lingual embeddings.
- **Iterative back-translation**: Traditional methods to create pseudo parallel data for unsupervised NMT.
- **SMT as PS**: Feed the noisy data generated by NMT models to SMT models to filter out noises by constructing strong phrase tables with good and frequent translation patterns.

Motivation

Unsupervised Neural Machine Translation (NMT) typically requires pseudo parallel data generated with the back-translation method for the model training. However, due to weak supervision, the pseudo data inevitably contains noises and errors that will be accumulated and reinforced in the subsequent training process.

Pseudo training data (fr-en)	[src1]: … de sa grand - mère blanche gravement malade , … [trg1]: … of his severely ill - fated white grandmother … [src2]: … au chevet de son épouse gravement malade … [trg2]: … at his severely ill wife 's bedside … [src3]: … ils sont malades ou couverts de moisissures ! [trg3]: … they are sick or covered with mold ! …
Test Result	[src]: … avec lui jusqu'à ce qu'il devienne vieux, malade. [sys]: … with him until he's become old, ill-fated. [ref]: … with him until he became old and ill.

Statistical Machine Translation (SMT) performs better than NMT in tackling noisy data by constructing a strong phrase table with good and frequent translation patterns and filtering out infrequent errors and noises. → Incorporating SMT into unsupervised NMT.

Method

Adopt the framework of posterior regularization (PR) to incorporate SMT into unsupervised NMT. The objective function is:

$$J\left(\theta_{x\to y}, \theta_{y\to x}, \overrightarrow{p_s}, \overleftarrow{p_s}\right) = \bar{\mathcal{L}}\left(\theta_{x\to y}, \theta_{y\to x}\right) \underbrace{- \sum_{i=1}^{M} \min_{\overrightarrow{p_s}} \mathrm{KL}\left(\overrightarrow{p_s}(y|x_i) || \overrightarrow{p_n}(y|x_i; \theta_{x\to y})\right) - \sum_{j=1}^{N} \min_{\overleftarrow{p_s}} \mathrm{KL}\left(\overleftarrow{p_s}(x|y_j) || \overleftarrow{p_n}(x|y_j; \theta_{y\to x})\right)}_{\text{PR terms}}$$

SMT models

NMT models

where $\bar{\mathcal{L}}\left(\theta_{x\to y}, \theta_{y\to x}\right)$ corresponds to the training objective of iterative back-translation for NMT models. Then we use an EM training algorithm to optimize the above objective.

Experiments

Comparison

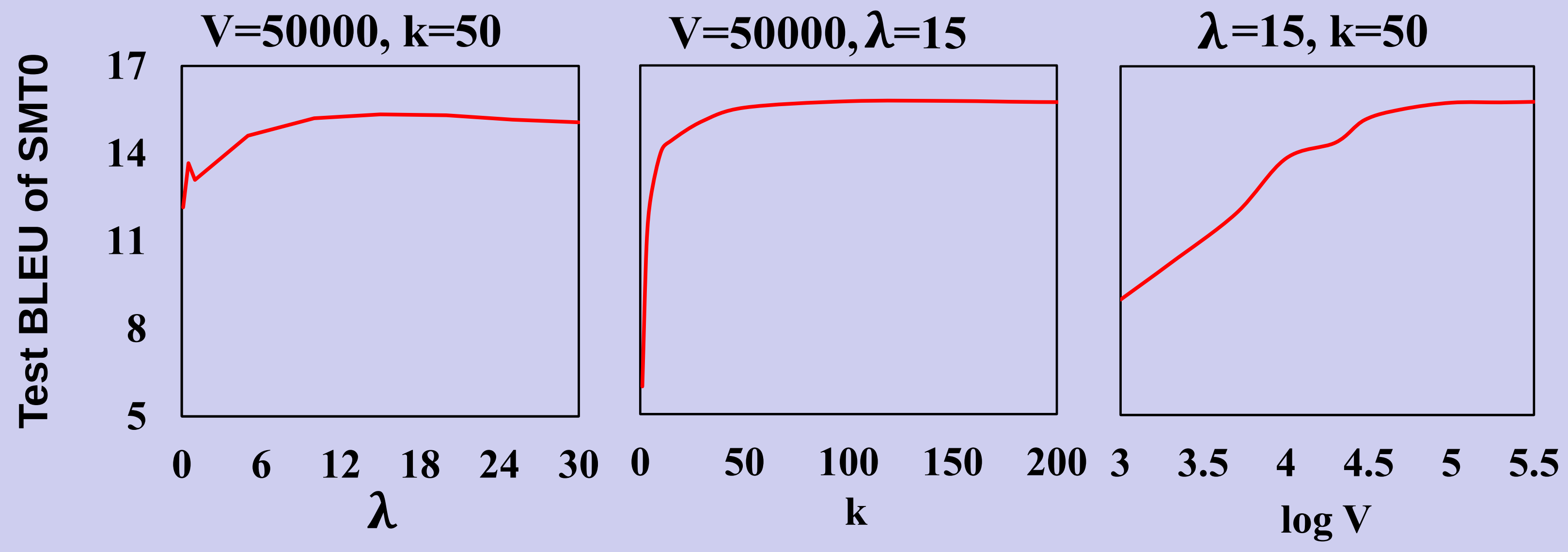
	fr-en	en-fr	de-en	en-de
(Artetxe et al. 2017)	15.56	15.13	-	-
(Lample et al. 2017)	14.31	15.05	13.33	9.64
(Yang et al. 2018)	15.58	16.97	14.62	10.86
(Lample et al. 2018), best sys	27.68	28.11	25.19	20.23
Our method	28.79	29.21	25.92	21.07
(+R2L regularization)	**28.92**	**29.53**	**26.32**	**21.65**

Discussion on model initialization

Given cross-lingual embeddings of both languages, i.e., $\{e_{x_i}\}_{i=1}^{V}$ and, $\{e_{y_j}\}_{j=1}^{V}$ the word translation probability from x_i to y_j is :

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

where λ is the distribution peakiness controller, e_x is the cross-lingual embedding of a certain word x, and we choose k translation candidates for each word. The performance of initial SMT models with various hyper-params (λ, k, V) is:



Conclusion

We introduce SMT models as posterior regularization to denoise and guide unsupervised NMT models with the ability of eliminating bad patterns generated in the back-translation iterations of NMT. We unify SMT and NMT models within the EM training algorithm where they can be trained jointly and benefit from each other incrementally. In the future, we may delve into the initialization stage, which is crucial to the final performance of the proposed method.