Cross-model Back-translated Distillation for Unsupervised Machine Translation

Xuan-Phi Nguyen[1,2], **Shafiq Joty**[1,3],
Thanh-Tung Nguyen[1,2], **Wu Kui**[2], **Ai Ti Aw**[2]
[1]Nanyang Technological University
[2]Institute for Infocomm Research (I^{2}R), A*STAR
[3]Salesforce Research
Singapore

Data Diversity

- Recent Unsupervised MT (UMT) models [Lample et al., 2018] generate and feed diversified synthetic data during training.
  - Denoising-Auto-Encoding: randomly noises the input sentence.

$$\mathbb{X} \dashrightarrow x \xrightarrow{noise} \hat{x} \xRightarrow{(\hat{x},x)} \theta$$

  - Iterative Back-translation: translates monolingual data to obtain pseudo-parallel data and train it via back-translation.

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow{\theta} y_t \xRightarrow{(y_t,x_s)} \theta$$

$$\mathbb{X}_t \dashrightarrow x_t \xrightarrow{\theta} y_s \xRightarrow{(y_s,x_t)} \theta$$

Data Diversity

- Recent Unsupervised MT (UMT) models [Lample et al., 2018] generate and feed diversified synthetic data during training.
  - Denoising-Auto-Encoding: randomly noises the input sentence.

$$\mathbb{X} \dashrightarrow x \xrightarrow{noise} \hat{x} \xRightarrow{(\hat{x},x)} \theta$$

  - Iterative Back-translation: translates monolingual data to obtain pseudo-parallel data and train it via back-translation.

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow{\theta} y_t \xRightarrow{(y_t,x_s)} \theta$$

$$\mathbb{X}_t \dashrightarrow x_t \xrightarrow{\theta} y_s \xRightarrow{(y_s,x_t)} \theta$$

- As training matures, the MT model may have covered the data distribution these strategies can diversely generate.
- Motivation: increasing synthetic data diversity can help improve unsupervised MT.

Cross-model Back-translated Distillation (CBD)

- A method aim to artificially increase data diversity
- Uses 2 *distinct* UMT models instead: θ_1 and θ_2 to generate synthetic data in a cross-model fashion.
- Uses the data to train a final *supervised bidirectional model* θ.

Cross-model Back-translation Process

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow[s \to t]{\theta_\alpha} y_t$$

Figure: The sampling process of $x_s, y_t, z_s, x_t, y_s, z_t$. The variable ordered set $(\theta_\alpha, \theta_\beta)$ is replaced with (θ_1, θ_2) and (θ_2, θ_1) iteratively in during training. All synthetic parallel pairs are used to train θ in a supervised way.

Cross-model Back-translation Process

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow[s \to t]{\theta_\alpha} y_t \xrightarrow[t \to s]{\theta_\beta} z_s$$

Figure: The sampling process of $x_s, y_t, z_s, x_t, y_s, z_t$. The variable ordered set $(\theta_\alpha, \theta_\beta)$ is replaced with (θ_1, θ_2) and (θ_2, θ_1) iteratively in during training. All synthetic parallel pairs are used to train θ in a supervised way.

Cross-model Back-translation Process

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow[s \to t]{\theta_\alpha} y_t \xrightarrow[t \to s]{\theta_\beta} z_s \xRightarrow[(y_t, z_s), (z_s, y_t)]{(x_s, y_t), (y_t, x_s)} \theta$$

Figure: The sampling process of $x_s, y_t, z_s, x_t, y_s, z_t$. The variable ordered set $(\theta_\alpha, \theta_\beta)$ is replaced with (θ_1, θ_2) and (θ_2, θ_1) iteratively in during training. All synthetic parallel pairs are used to train θ in a supervised way.

Cross-model Back-translation Process

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow[s\to t]{\theta_\alpha} y_t \xrightarrow[t\to s]{\theta_\beta} z_s \xRightarrow[(y_t,z_s),(z_s,y_t)]{(x_s,y_t),(y_t,x_s)} \theta$$

$$\mathbb{X}_t \dashrightarrow x_t \xrightarrow[t\to s]{\theta_\alpha} y_s \xrightarrow[s\to t]{\theta_\beta} z_t \xRightarrow[(y_s,z_t),(z_t,y_s)]{(x_t,y_s),(y_s,x_t)} \theta$$

Figure: The sampling process of $x_s, y_t, z_s, x_t, y_s, z_t$. The variable ordered set $(\theta_\alpha, \theta_\beta)$ is replaced with (θ_1, θ_2) and (θ_2, θ_1) iteratively in during training. All synthetic parallel pairs are used to train θ in a supervised way.

CBD Training Algorithm

Algorithm Cross-model Back-translated Distillation: Given monolingual data $\mathbb{X}_s$ and $\mathbb{X}_t$ of languages s and t, return a UMT model with parameters θ.

```
Train the 1st UMT agent with parameters θ_1
Train the 2nd UMT agent with parameters θ_2
Initialize model θ (randomly or with pretrained model)
while until convergence do
    θ ← θ − η∇_θ L_θ(θ_α = θ_1, θ_β = θ_2)
    θ ← θ − η∇_θ L_θ(θ_α = θ_2, θ_β = θ_1)
return θ
```

$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_\theta(\theta_\alpha = \theta_1, \theta_\beta = \theta_2)$

$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_\theta(\theta_\alpha = \theta_2, \theta_\beta = \theta_1)$

WMT Unsupervised Machine Translation

Method / Data	En-Fr	Fr-En	En-De	De-En	En-Ro	Ro-En
NMT [Lample et al., 2018]	25.1	24.2	17.2	21.0	21.1	19.4
PBSMT [Lample et al., 2018]	27.8	27.2	17.7	22.6	21.3	23.0
XLM [Conneau and Lample, 2019]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [Song et al., 2019]	37.5	34.9	28.3	35.2	35.2	33.1
CBD	**38.2**	**35.5**	**30.1**	**36.3**	**36.3**	**33.8**

Table: BLEU scores on the *large scale* WMT'14 English-French (En-Fr), WMT'16 English-German (En-De) and WMT'16 English-Romanian (En-Ro) unsupervised translation tasks.

Benefit of Cross-model Back-translation

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow[s \to t]{\theta_\alpha} y_t \xrightarrow[t \to s]{\theta_\alpha} z_s \xrightarrow[(y_t, z_s), (z_s, y_t)]{(x_s, y_t), (y_t, x_s)} \theta$$

Figure: No cross-model Back-translated Distillation - BD(2/2), where only 1 model involves in the two-stage translation processes.

Method	**En-Fr**	**Fr-En**	**En-De**	**De-En**	**En-Ro**	**Ro-En**
NMT	24.7	24.5	14.5	18.2	16.7	16.3
BD(1/1)	24.5	24.5	14.0	17.5	16.1	15.9
BD(1/2)	24.6	24.6	14.1	17.8	16.4	16.2
BD(2/2)	24.8	24.7	14.4	18.1	16.9	16.4
CBD	**26.6**	**25.7**	**16.6**	**20.5**	**18.1**	**17.8**

Table: BLEU comparison of CBD vs. no cross-model variants in the *base* WMT'14 English-French (En-Fr), WMT'16 English-German (En-De) and English-Romanian (En-Ro) tasks.

CBD Creates Data Diversity

$$\mathbb{X}_s \dashrightarrow x_s \xrightarrow[s \to t]{\theta_\alpha} y_t \xrightarrow[t \to s]{\theta_\beta} z_s \Rightarrow BLEU_{recon}(x_s, z_s)$$

Figure: How the reconstruction BLEU score is computed.

Method	**En-Fr**	**Fr-En**	**En-De**	**De-En**	**En-Ro**	**Ro-En**
BD	76.0	72.4	75.3	63.7	73.2	71.5
CBD	63.1	59.7	60.3	50.5	61.1	56.9

Table: Reconstruction BLEU scores of BD and CBD in different languages for the *base* WMT unsupervised translation tasks. Lower BLEU means more diverse.

CBD vs Other Diversity-related Methods

WMT	**En-Fr**	**Fr-En**	**En-De**	**De-En**
XLM	33.0	31.5	23.9	29.3
Sampling (temp=0.3)	33.5	32.2	24.3	30.2
Top-k sampling	33.18	32.26	24.0	29.9
Top-p sampling		Diverge		
Target noising	32.8	30.7	24.0	29.6
Multi-agent dual learning	33.5	31.7	24.6	29.9
CBD	35.4	33.0	26.1	31.5

Table: Comparison with other alternatives on the *base* WMT En-Fr, Fr-En, En-De and De-En, with XLM as the base model.

- Thank you.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc., 2019.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1549. URL https://www.aclweb.org/anthology/D18-1549.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.