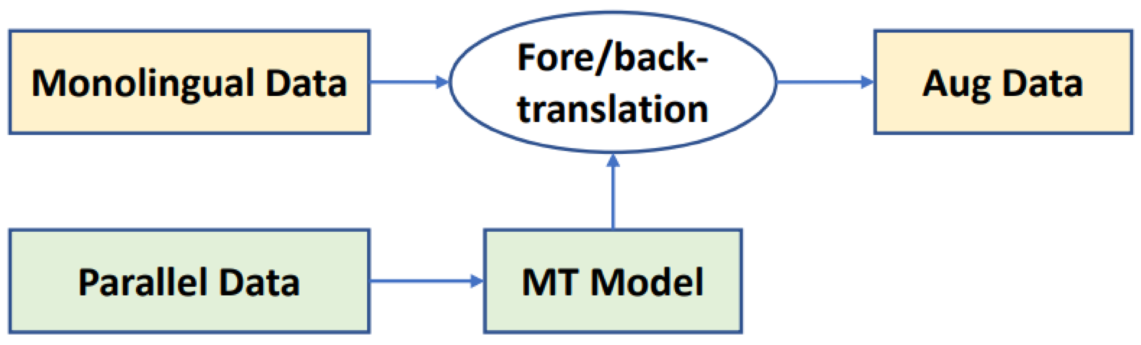
-Side Augmentation for Document-Level Machine Translation

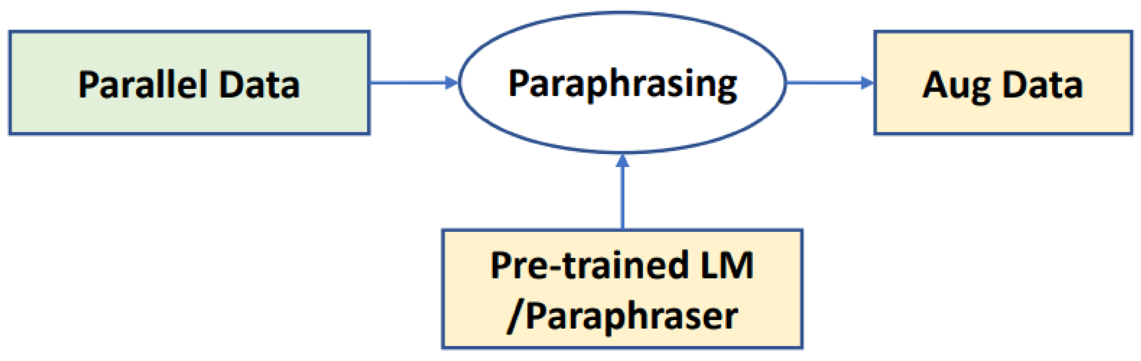
Guangsheng Bao, Zhiyang Teng, and Yue Zhang

Introduction: Data Augmentation in MT

Previous 1: Self-training / Back-translation



Previous 2: Pre-trained LM / Paraphraser

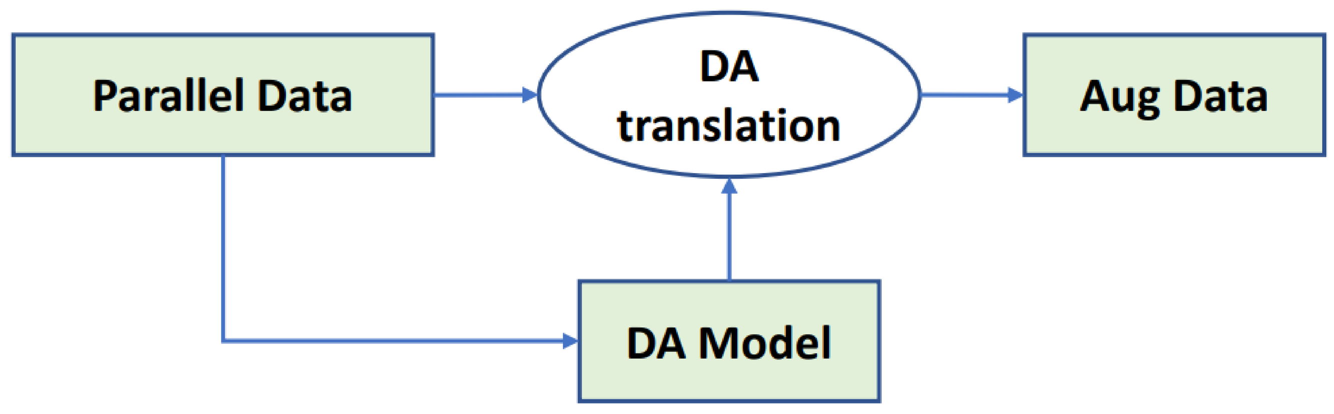


Rely on external data or models.

VS

Rely on parallel data only.

Ours: DA Model



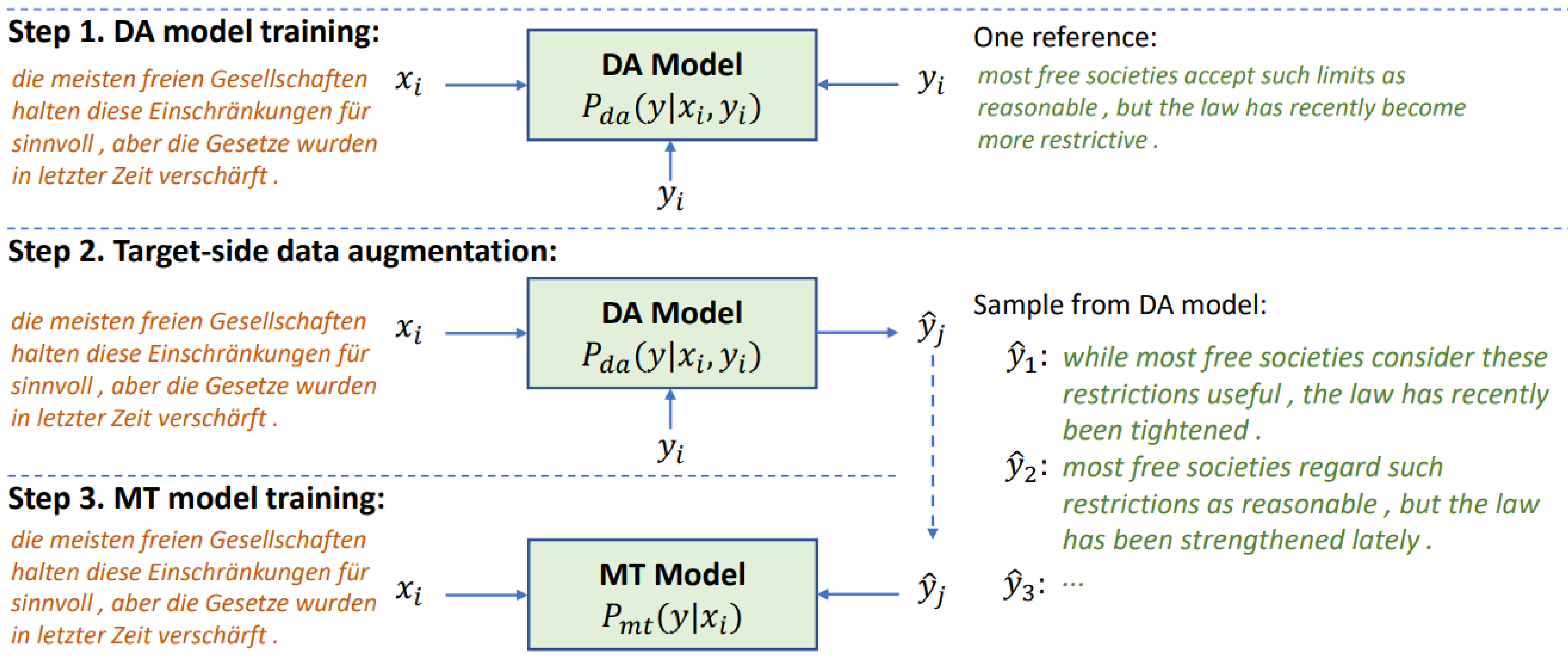
- The DA model estimates the posterior distribution of translations given observed x_i and y_i.
- The posterior distribution balances the Diversity and Deviation of generated translations, providing better estimation (lower PPL).

Method	Diversity ↑	Deviation ↓	PPL ↓
Prior distribution	**78.68**	76.55	8.68
Posterior distribution	45.42	**47.14**	**7.00**

Method: Target-Side Data Augmentation

Samples from data distribution for training:

$$x_i \sim P_{data}(x),\ \ y_i \sim P_{data}(y|x_i)$$



DA Model:

$$P_{da}(y|x_i, y_i) = \sum_{z \in \mathcal{Z}_i} P_\varphi(y|x_i, z) P_\alpha(z|y_i),$$

$$P_{da}(y|x_i, y_i) \approx \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} P_\varphi(y|x_i, z),$$

$$\begin{aligned}
\mathcal{L}_{da} &= -\sum_{i=1}^{N} \log P_{da}(y = y_i|x_i, y_i) \\
&\approx -\sum_{i=1}^{N} \log \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} P_\varphi(y = y_i|x_i, z) \\
&\leq -\sum_{i=1}^{N} \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} \log P_\varphi(y = y_i|x_i, z),
\end{aligned}$$

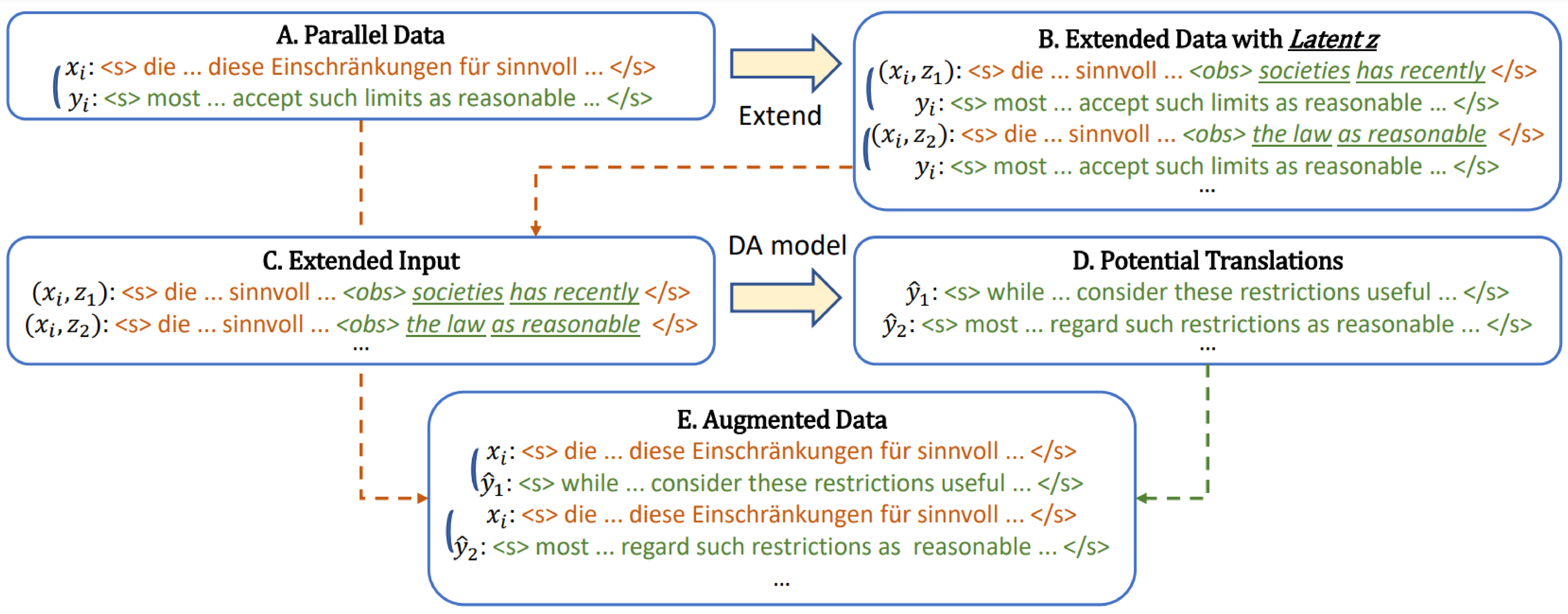
MT Model:

$$\hat{\mathcal{Y}}_i = \{\arg\max_y P_\varphi(y|x_i, z_j) | z_j \sim P_\alpha(z|y_i)\}_{j=1}^{M},$$

$$\mathcal{L}_{mt} = -\sum_{i=1}^{N} \sum_{y \in \mathcal{Y}_i} P_{da}(y|x_i, y_i) \log P_{mt}(y|x_i),$$

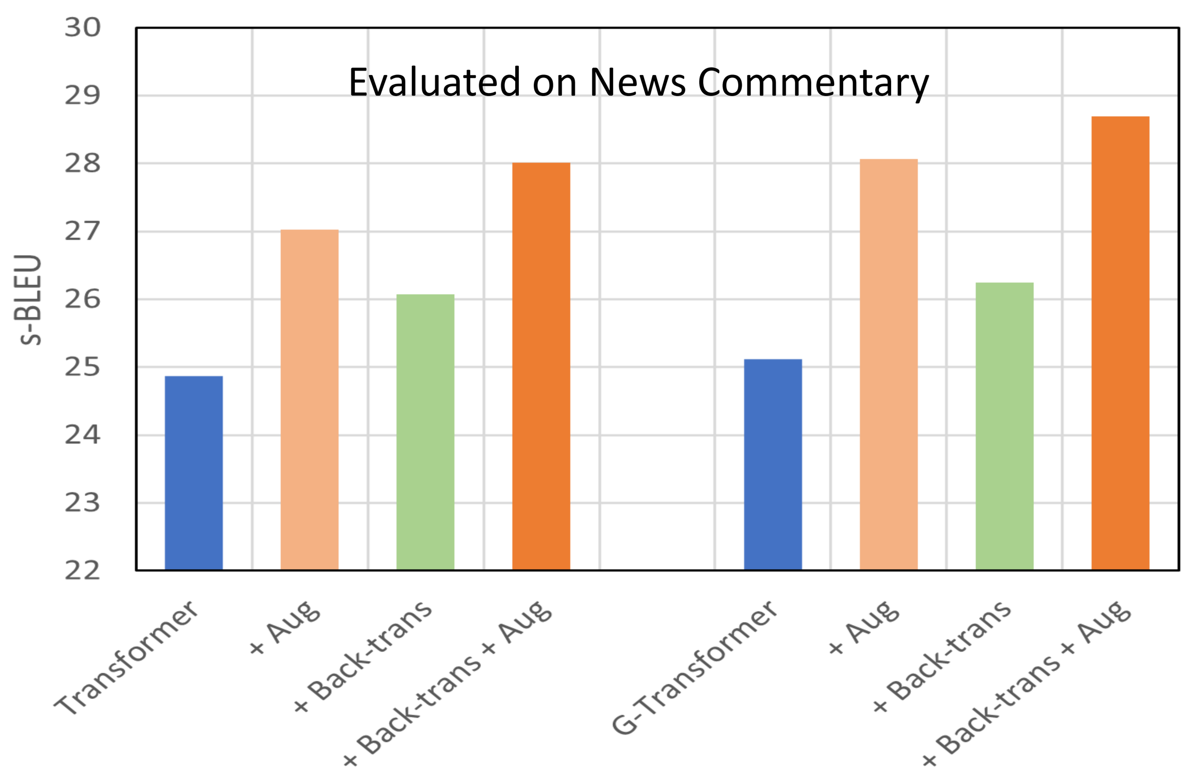
$$\mathcal{L}_{mt} \approx -\sum_{i=1}^{N} \frac{1}{|\hat{\mathcal{Y}}_i|} \sum_{y \in \hat{\mathcal{Y}}_i} \log P_\theta(y|x_i),$$

Data Augmentation Process:



Results & Analysis

Compare to Baselines and Back-translations:



Compare to Paraphraser:

Method	Dev	Test
Transformer (base)	34.85	33.87
+ T5 paraphraser ♢	34.01	33.10
+ Target-side augmentation	**36.42**	**35.42**

Table 6: Target-side augmentation vs paraphraser on sentence-level MT, evaluated on IWSLT14 German-English (De-En). ♢ – nucleus sampling with $p = 0.95$.

Impact of Aug Scale:

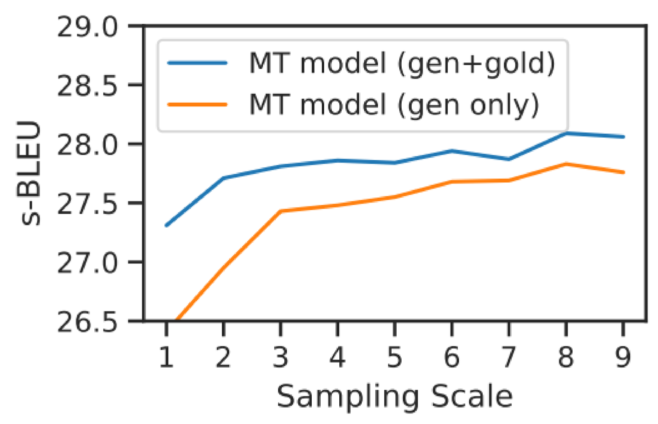
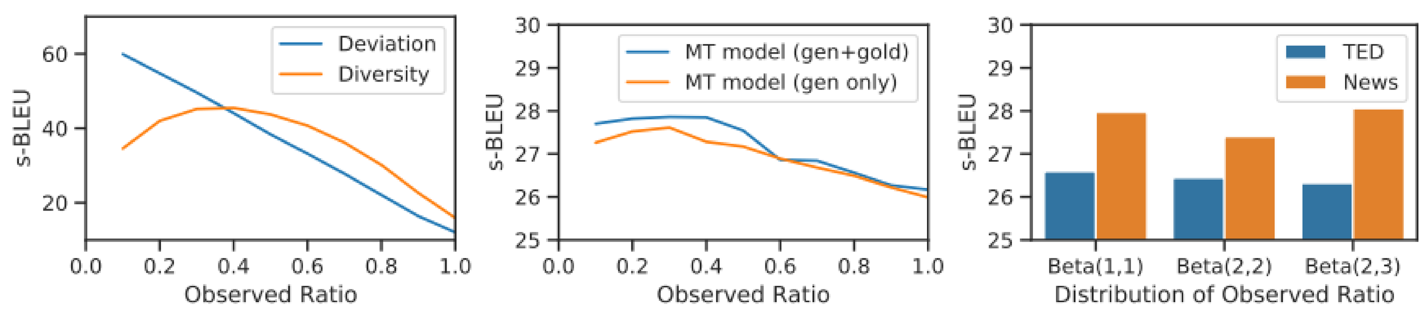


Figure 3: Impact of the sampling scale for z, trained on G-Transformer (fnt.) and evaluated in *s-BLEU* on *News*.

Impact of Latent Variable:



(a) Quality of translations generated by the DA model, evaluated on *News*

(b) Performance of MT model on augmented data, evaluated on *News*

(c) Performance of MT models trained using mixed observed ratios

Figure 4: Impact of the observed ratio for z, trained on G-Transformer (fnt.) and evaluated in *s-BLEU*. Beta(a,b) – The function curves are shown in Appendix B.3.

Conclusion

- Target-side data augmentation mitigates data sparsity effectively.
- Balancing Diversity and Deviation is the key for the DA model to obtain the best effect.
- Poster distribution can approximate the data distribution better than prior distribution.
- Given single translation parallel data, we model poster distribution by introducing an intermediate latent variable.

Contact: baoguangsheng @westlake.edu.cn

Zhejiang University

Westlake University

Nanyang Technological University