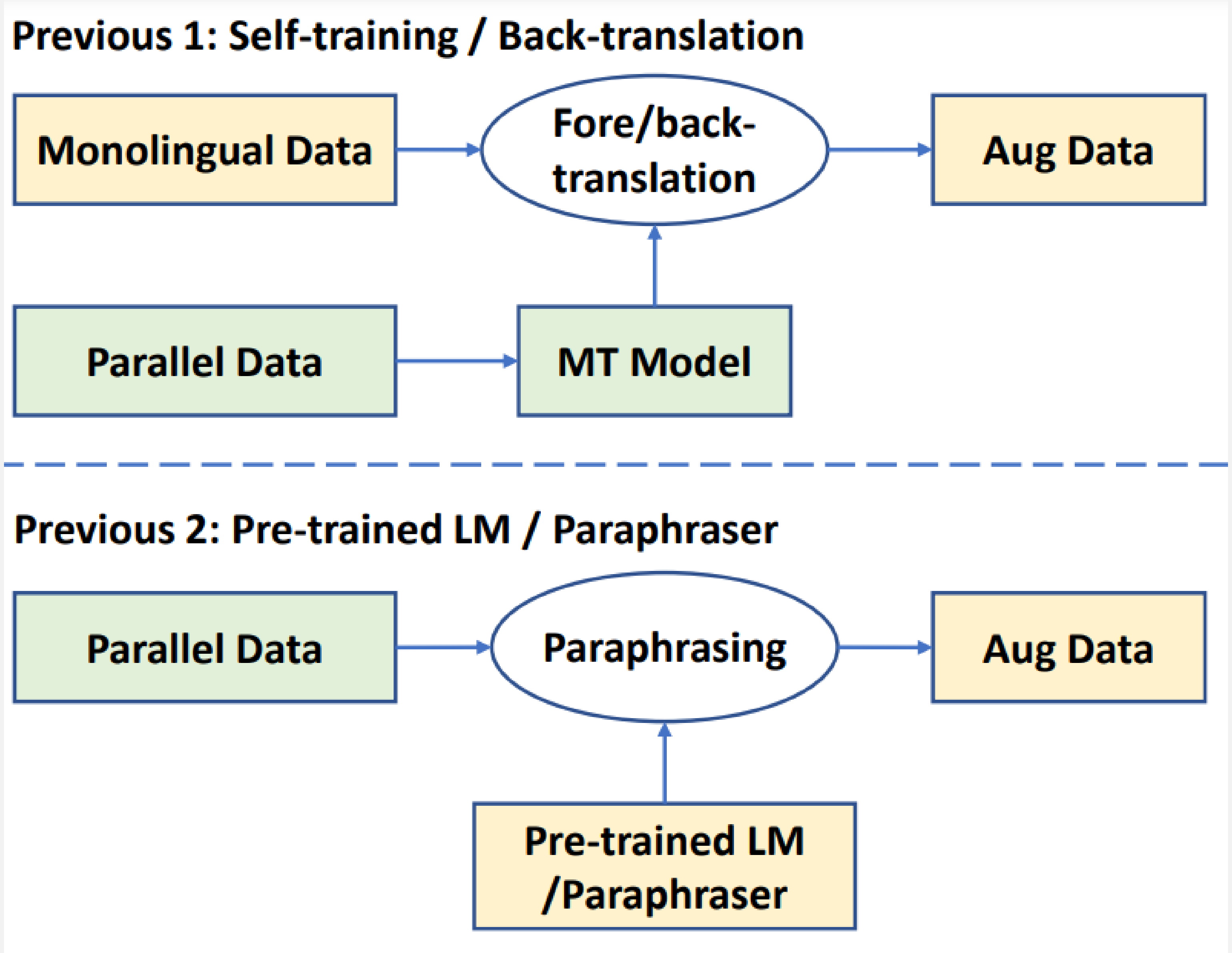
target-Side Augmentation for Document-level Machine Translation

Guangsheng Bao, Zhiyang Teng , Yue Zhang, ACL 2023

Introduction

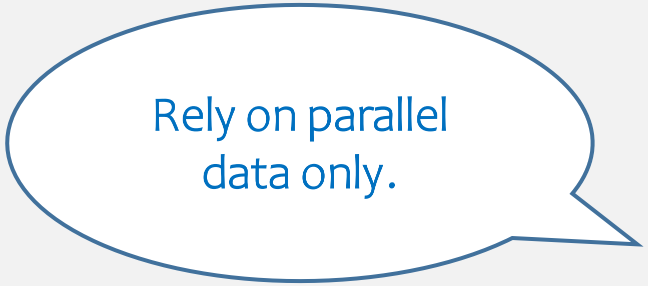
Data Augmentation in MT

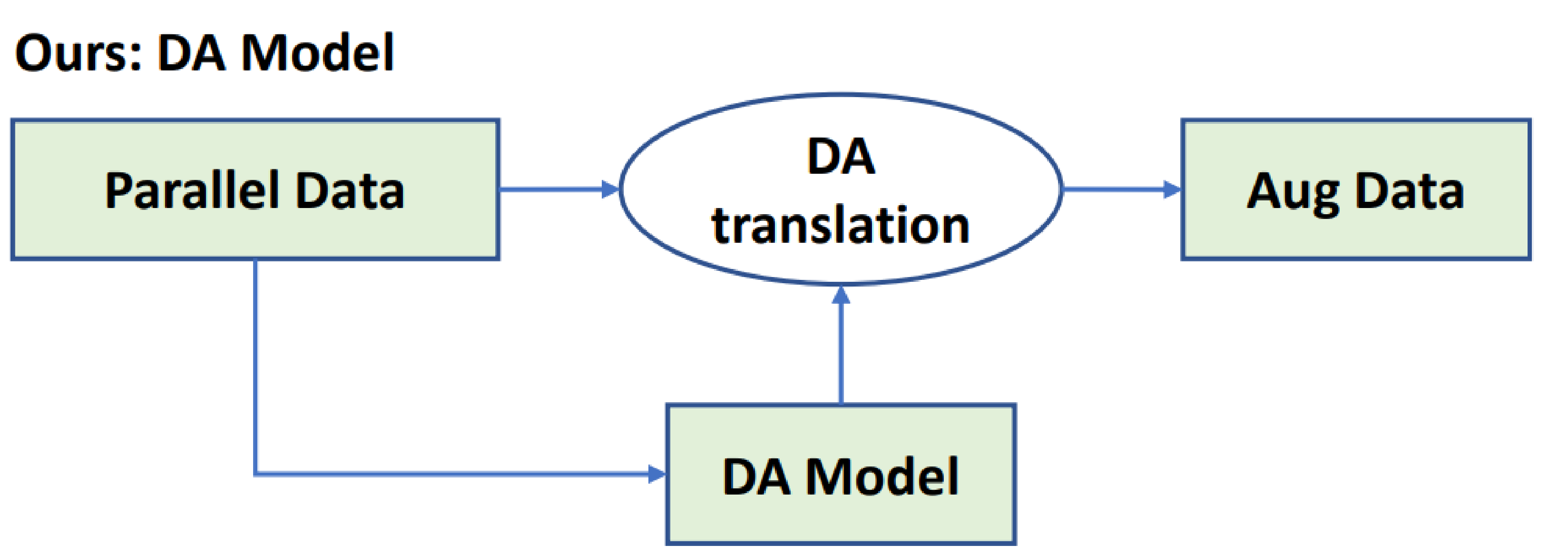
Rely on external data or models.



Introduction

Data Augmentation in MT





Target-Side Data Augmentation

Samples from data distribution for training:

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

Step 1. DA model training:

die meisten freien Gesellschaften halten diese Einschränkungen für sinnvoll , aber die Gesetze wurden in letzter Zeit verschärft . x_i

DA Model
$P_{da}(y|x_i, y_i)$

y_i

y_i

One reference:

most free societies accept such limits as reasonable , but the law has recently become more restrictive .

Step 2. Target-side data augmentation:

die meisten freien Gesellschaften halten diese Einschränkungen für sinnvoll , aber die Gesetze wurden in letzter Zeit verschärft . x_i

DA Model
$P_{da}(y|x_i, y_i)$

y_i

$\hat{y}_j$

Sample from DA model:

$\hat{y}_1$: *while most free societies consider these restrictions useful , the law has recently been tightened .*

$\hat{y}_2$: *most free societies regard such restrictions as reasonable , but the law has been strengthened lately .*

$\hat{y}_3$: ...

Step 3. MT model training:

die meisten freien Gesellschaften halten diese Einschränkungen für sinnvoll , aber die Gesetze wurden in letzter Zeit verschärft . x_i

MT Model
$P_{mt}(y|x_i)$

$\hat{y}_j$

Target-Side Data Augmentation

The DA Model

$$P_{da}(y|x_i, y_i) = \sum_{z \in \mathcal{Z}_i} P_{\varphi}(y|x_i, z) P_{\alpha}(z|y_i), \quad (1)$$

$$P_{da}(y|x_i, y_i) \approx \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} P_{\varphi}(y|x_i, z), \quad (2)$$

$$\begin{aligned} \mathcal{L}_{da} &= -\sum_{i=1}^{N} \log P_{da}(y = y_i|x_i, y_i) \\ &\approx -\sum_{i=1}^{N} \log \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} P_{\varphi}(y = y_i|x_i, z) \quad (3) \\ &\leq -\sum_{i=1}^{N} \frac{1}{|\hat{\mathcal{Z}}_i|} \sum_{z \in \hat{\mathcal{Z}}_i} \log P_{\varphi}(y = y_i|x_i, z), \end{aligned}$$

Target-Side Data Augmentation

The MT Model

$$\mathcal{L}_{mt} = -\sum_{i=1}^{N} \sum_{y \in \mathcal{Y}_i} P_{da}(y|x_i, y_i) \log P_{mt}(y|x_i), \tag{5}$$

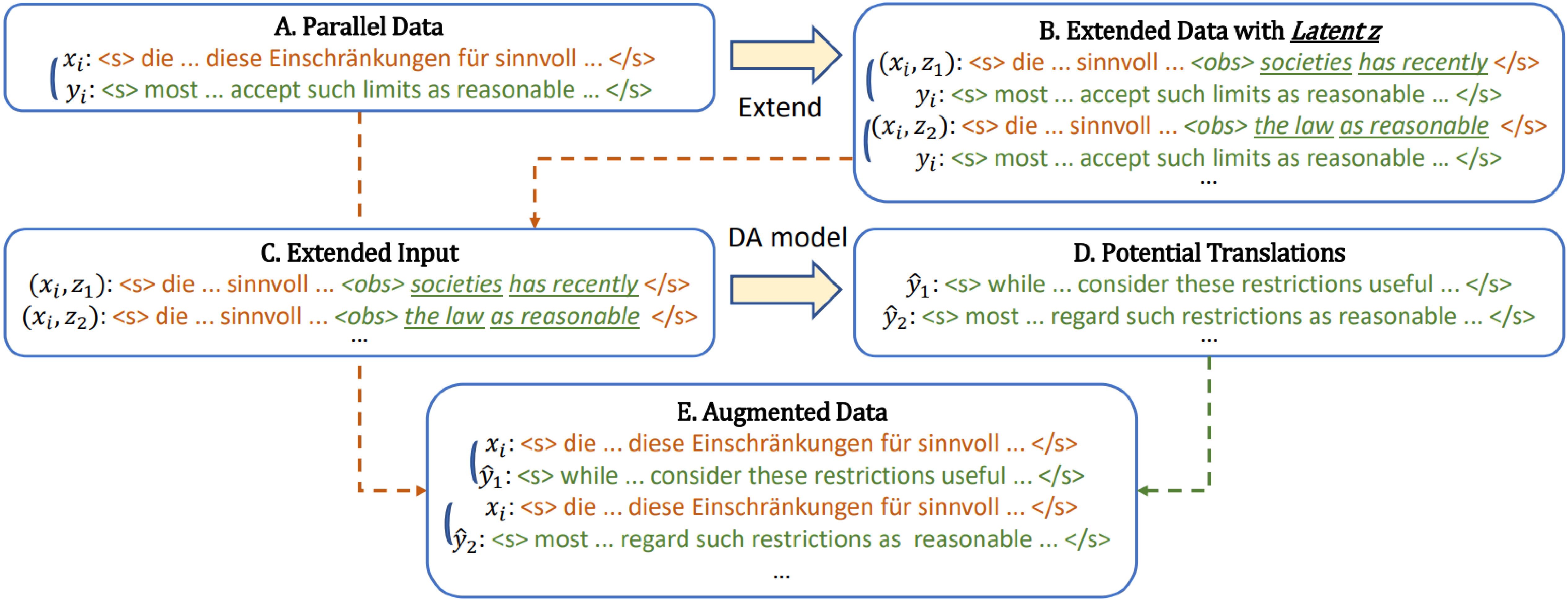
$$\hat{\mathcal{Y}}_i = \{\arg\max_{y} P_{\varphi}(y|x_i, z_j) | z_j \sim P_{\alpha}(z|y_i)\}_{j=1}^{M}, \tag{6}$$

$$\mathcal{L}_{mt} \approx -\sum_{i=1}^{N} \frac{1}{|\hat{\mathcal{Y}}_i|} \sum_{y \in \hat{\mathcal{Y}}_i} \log P_{\theta}(y|x_i), \tag{7}$$

SECTION 2

Target-Side Data Augmentation

Data Augmentation Process



SECTION 3

Main Results

Method	TED		News		Europarl		Average
	s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU
HAN (Miculicich et al., 2018)	24.58	-	25.03	-	28.60	-	26.07
SAN (Maruf et al., 2019)	24.42	-	24.84	-	29.75	-	26.34
Hybrid Context (Zheng et al., 2020)	25.10	-	24.91	-	30.40	-	26.80
Flat-Transformer (Ma et al., 2020)	24.87	-	23.55	-	30.09	-	26.17
G-Transformer (rnd.) (Bao et al., 2021)	23.53	25.84	23.55	25.23	32.18	33.87	26.42
G-Transformer (fnt.) (Bao et al., 2021)	25.12	27.17	25.52	27.11	32.39	34.08	27.68
MultiResolution (Sun et al., 2022)	25.24	29.27	25.00	26.71	32.11	34.48	27.45
RecurrentMem (Feng et al., 2022)	25.62	**29.47**	25.73	27.78	31.41	33.50	27.59
SMDT (Zhang et al., 2022)	25.12	-	25.76	-	32.42	-	27.77
Transformer (sent baseline) ◇	24.91	-	24.82	-	31.22	-	26.98
+ Target-side data augmentation (ours)	26.14*	-	27.03*	-	31.75*	-	28.31
G-Transformer (fnt.) (doc baseline) ◇	25.20	27.94	25.12	27.02	31.93	33.88	27.42
+ Target-side augmentation (ours)	**26.59***	29.20*	28.06*	29.83*	**32.85***	**34.76***	**29.17**
Transformer + Back-translation (sent) ♡	25.03	-	26.07	-	31.12	-	27.41
Target-side augmentation (ours)	26.13	-	28.01	-	31.27	-	28.47
G-Transformer + Back-translation (doc) ♡	25.45	28.06	26.25	28.21	32.00	33.94	27.90
Target-side augmentation (ours)	26.21	28.58	**28.69**	**30.41**	32.52	34.50	29.14
Pre-training Setting for Comparison							
Flat-Transformer+BERT (Ma et al., 2020)	26.61	-	24.52	-	31.99	-	27.71
G-Transformer+BERT (Bao et al., 2021)	26.81	-	26.14	-	32.46	-	28.47
G-Transformer+mBART (Bao et al., 2021)	28.06	30.03	30.34	31.71	32.74	34.31	30.38

Analysis

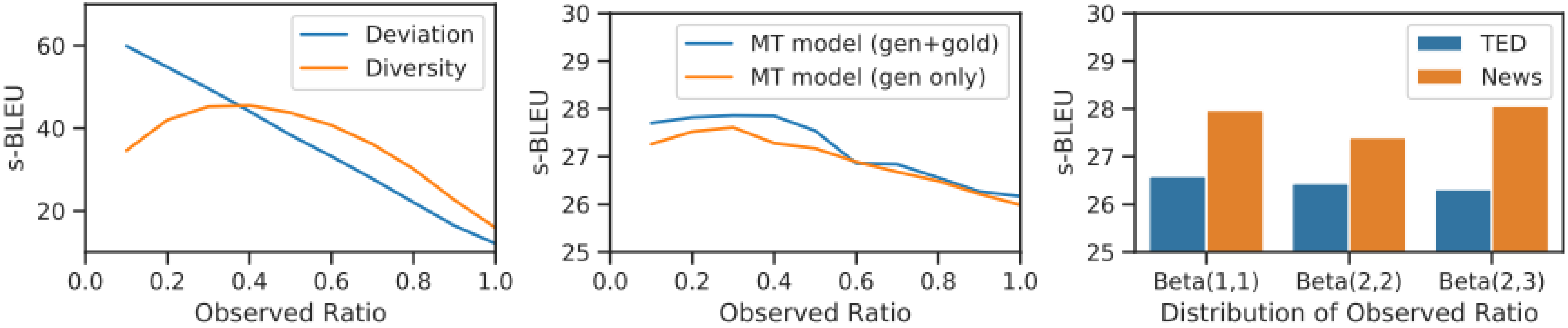
Posterior vs Prior Distribution

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

Table 4: Quality of generated translations and accuracy of the estimated distributions from the DA model, evaluated on *News*.

Analysis

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.

Q & A

Thank you!

Github:
https://github.com/baoguangsheng/target-side-augmentation

Zhejiang University

Westlake University

Nanyang Technological University