Robust Neural Machine Translation with Doubly Adversarial Inputs

Cheng et al.

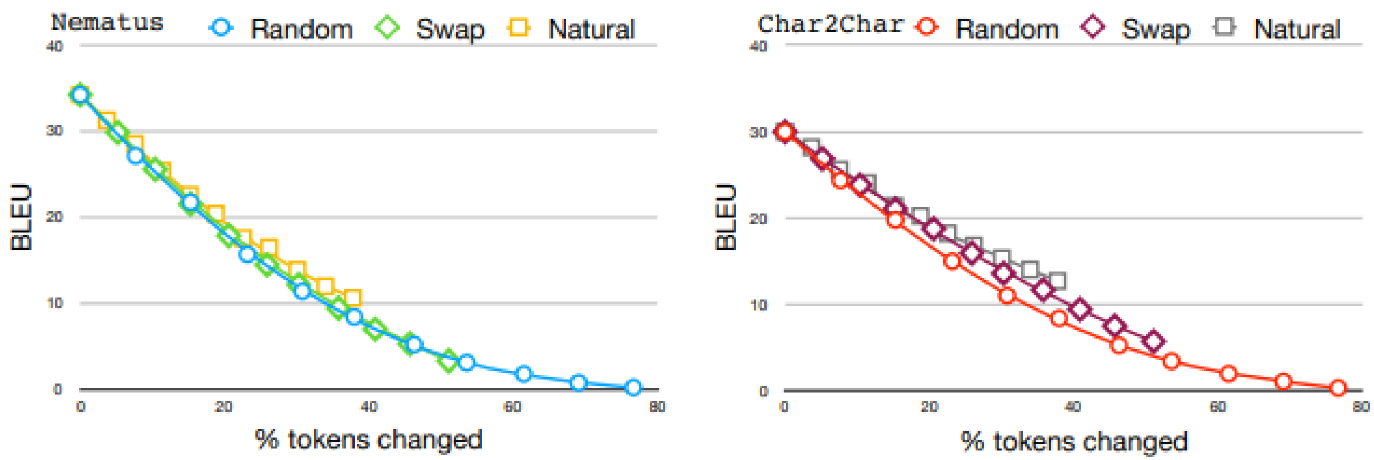
JungsooPark

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

1 Introduction

Synthetic and Natural Noise Both Break NMT

Belinkov et al. (ICLR 2018)

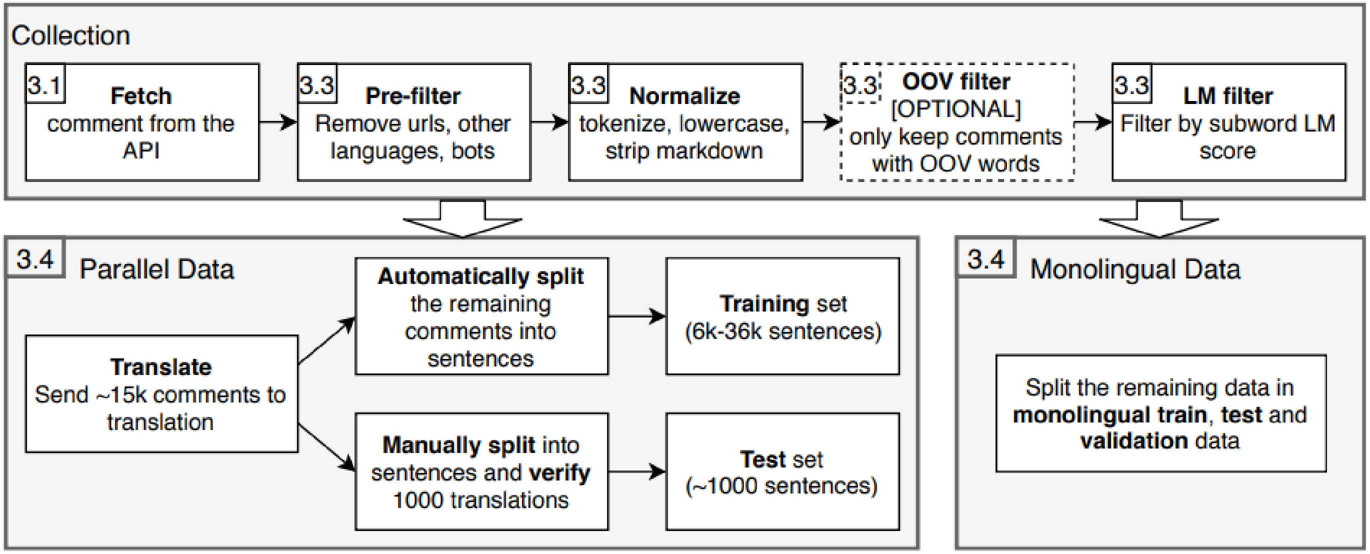


Current NMT models suffer from both synthetic and natural noise

1 Introduction

MTNT: A Testbed for Machine Translation of Noisy Text

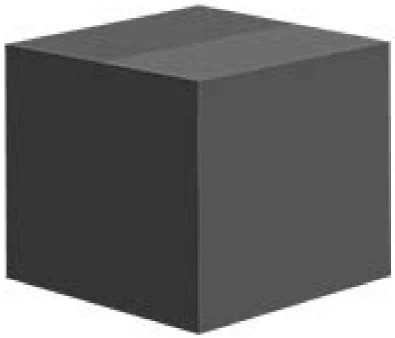
Michel et al. (EMNLP 2018)



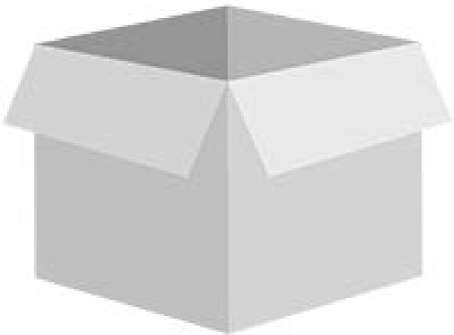
A Surge of Interest Towards Building Robust NMT Models to Noisy Text

2 Related Work

Research Trend



- Domain Adaptation
- Designing Synthetic and Natural Noise

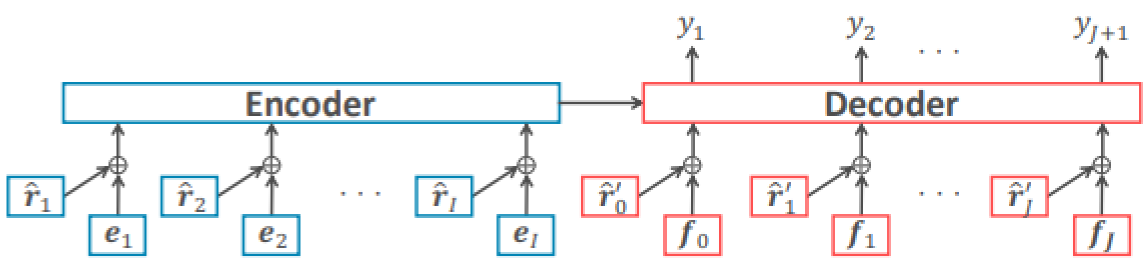


- Adversarial Training

2 Related Work

Effective Adversarial Regularization for NMT

Sato et al. (ACL, 2018)



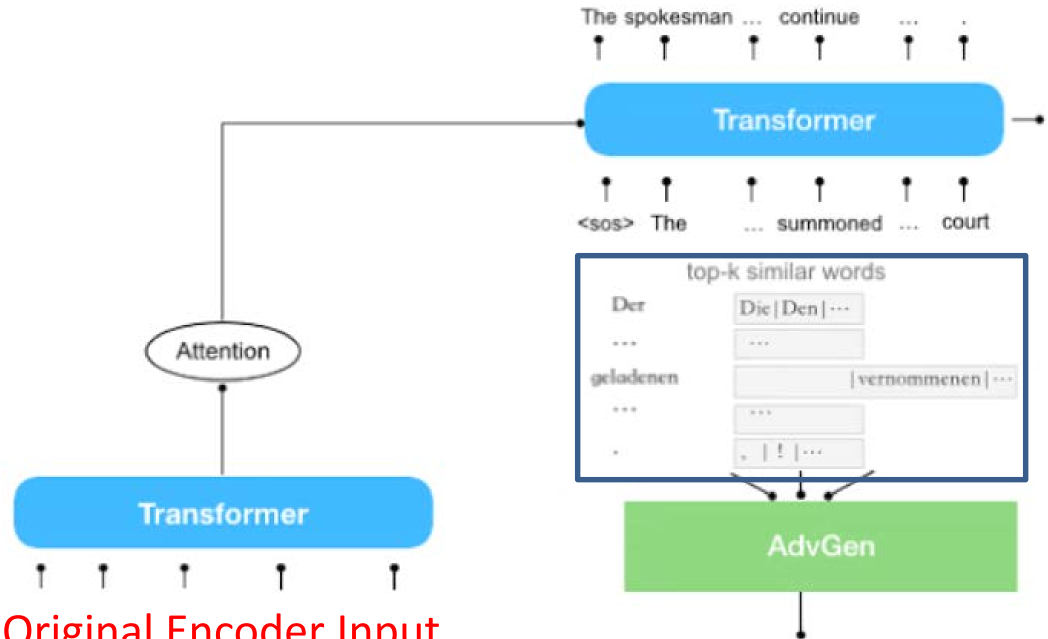
Inject Adversarial Perturbation(Noise) in Embedding Space

$$e'_i = \boldsymbol{E}\boldsymbol{x}_i + \hat{\boldsymbol{r}}_i.$$

$$\hat{\boldsymbol{r}} = \underset{\boldsymbol{r}, ||\boldsymbol{r}|| \leq \epsilon}{\operatorname{argmax}} \left\{ \ell(\boldsymbol{X}, \boldsymbol{r}, \boldsymbol{Y}, \boldsymbol{\Theta}) \right\},$$

3 Method

AdvGen (Encoder)



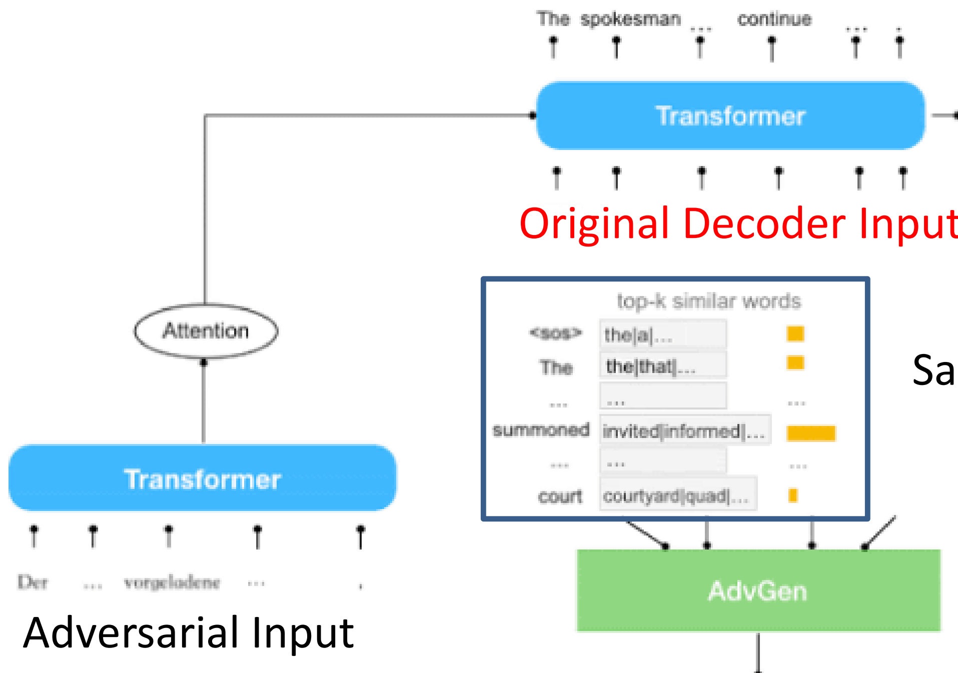
Sample words uniformly

Original Encoder Input

Replace the selected words in encoder input with adversarial one.

3 Method

AdvGen (Decoder)



Sample words according to attention score

Replace the selected words in decoder input with adversarial one.

3 Method

AdvGen (Encoder)

Adversarial Objective

$$\left\{\mathbf{x}' \,|\, \mathcal{R}(\mathbf{x}', \mathbf{x}) \leq \epsilon, \operatorname*{argmax}_{\mathbf{x}'} -\log P(\mathbf{y}|\mathbf{x}'; \boldsymbol{\theta}_{mt})\right\}$$

Replacing

$$\begin{aligned} x_i' &= \operatorname*{argmax}_{x \in \mathcal{V}_x} \operatorname{sim}\left(e(x) - e(x_i), \mathbf{g}_{x_i}\right) \\ \mathbf{g}_{x_i} &= \nabla_{e(x_i)} - \log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) \end{aligned}$$

Candidate Minimization

$$Q_{src}(x_i, \mathbf{x}) = P_{lm}(x|\mathbf{x}_{<i}, \mathbf{x}_{>i}; \boldsymbol{\theta}_{lm}^{x})$$

$$\mathcal{V}_{x_i} = top\_n(Q(x_i, \mathbf{x}))$$

AdvGen (Decoder)

Adversarial Objective

$$\mathbf{z}' = AdvGen(\mathbf{z}, Q_{trg}, D_{trg}, -\log P(\mathbf{y}|\mathbf{x}'))$$

Substitution Candidate Reduction

$$\begin{aligned} Q_{trg}(z_i, \mathbf{z}) = \quad & \lambda P(z|\mathbf{z}_{<i}, \mathbf{z}_{>i}; \boldsymbol{\theta}_{lm}^{y}) \\ & +(1-\lambda)P(z|\mathbf{z}_{<i}, \mathbf{x}'; \boldsymbol{\theta}_{mt}) \end{aligned}$$

Word Selection Distribution

$$P(j) = \frac{\sum_i \mathcal{M}_{ij}\delta(x_i, x_i')}{\sum_k \sum_i \mathcal{M}_{ik}\delta(x_i, x_i')}, j \in \{1, .., |\mathbf{y}|\}$$

4 Experiment

Method	Model	MT06	MT02	MT03	MT04	MT05	MT08
Vaswani et al. (2017)	Trans.-Base	44.59	44.82	43.68	45.60	44.57	35.07
Miyato et al. (2017)	Trans.-Base	45.11	45.95	44.68	45.99	45.32	35.84
Sennrich et al. (2016a)	Trans.-Base	44.96	46.03	44.81	46.01	45.69	35.32
Wang et al. (2018)	Trans.-Base	45.47	46.31	45.30	46.45	45.62	35.66
Cheng et al. (2018)	$\text{RNMT}_{lex.}$	43.57	44.82	42.95	45.05	43.45	34.85
	$\text{RNMT}_{feat.}$	44.44	46.10	44.07	45.61	44.06	34.94
Cheng et al. (2018)	$\text{Trans.-Base}_{feat.}$	45.37	46.16	44.41	46.32	45.30	35.85
	$\text{Trans.-Base}_{lex.}$	45.78	45.96	45.51	46.49	45.73	36.08
Sennrich et al. (2016b)*	Trans.-Base	46.39	47.31	47.10	47.81	45.69	36.43
Ours	Trans.-Base	46.95	47.06	46.48	47.39	46.58	37.38
Ours + BackTranslation*	Trans.-Base	**47.74**	**48.13**	**47.83**	**49.13**	**49.04**	**38.61**

Evaluation on NIST Test Dataset

4 Experiment

Method	0.00	0.05	0.10	0.15
Vaswani et al.	44.59	41.54	38.84	35.71
Miyato et al.	45.11	42.11	39.39	36.44
Cheng et al.	45.78	42.90	40.58	38.46
Ours	**46.95**	**44.20**	**41.71**	**39.89**

Evaluation on Noisy Dataset

$\mathcal{L}_{clean}$	$\mathcal{L}_{robust}$		$\mathcal{L}_{lm}$	BLEU
	$\mathbf{x}' \neq \mathbf{x}$	$\mathbf{z}' \neq \mathbf{z}$		
✓				44.59
✓			✓	45.08
✓	✓		✓	45.23
✓		✓	✓	46.26
✓	✓	✓		46.61
✓	✓	✓	✓	**46.95**

Ablation Study