Robust Neural Machine Translation with Doubly Adversarial Inputs

Cheng et al.

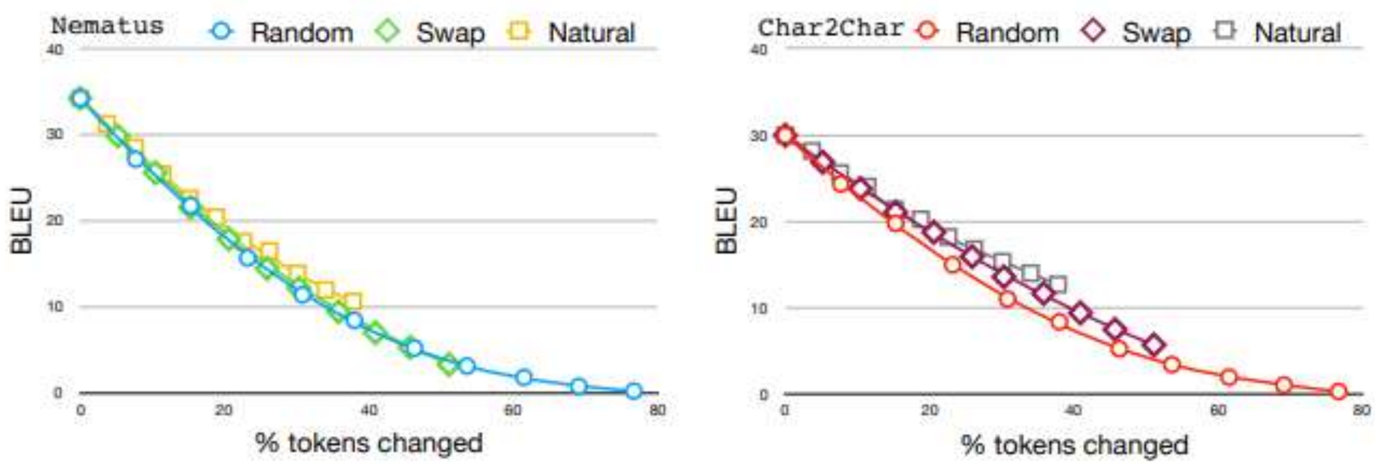
JungsooPark

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)

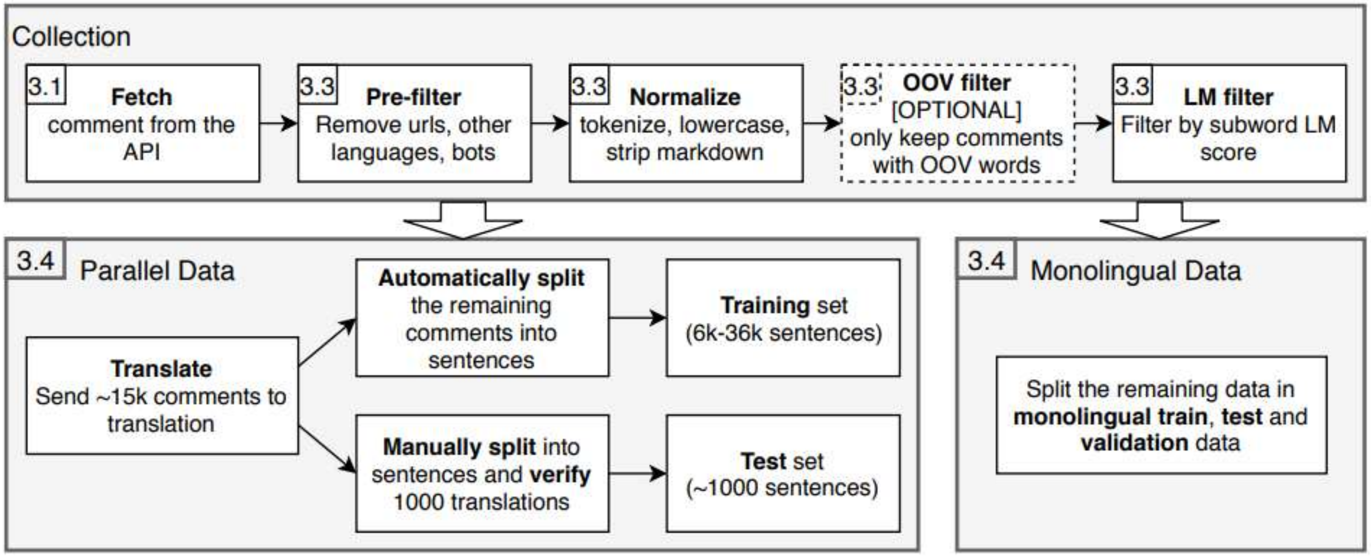


Current NMT models suffer from both synthetic and natural noise

1 Introduction

MTNT: A Testbed for Machine Translation of Noisy Text

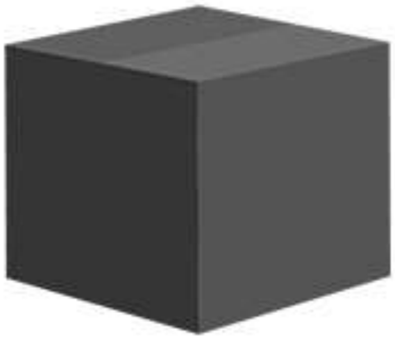
Michel et al. (EMNLP 2018)



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

2 Related Work

Research Trend



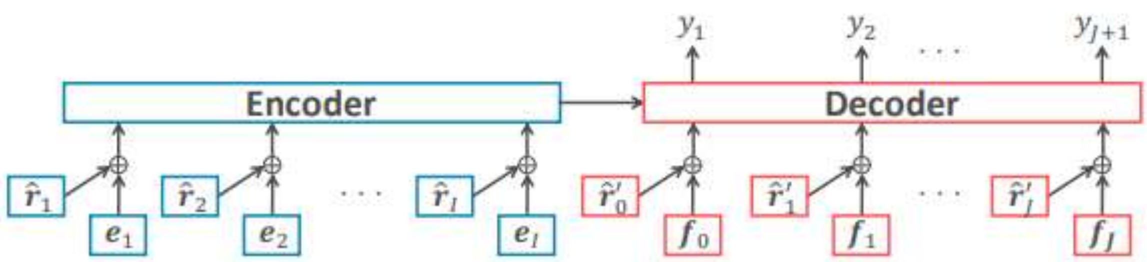
- Domain Adaptation
- Designing Synthetic and Natural Noise

- Adversarial Training

2 Related Work

Effective Adversarial Regularization for NMT

Sato et al. (ACL, 2018)



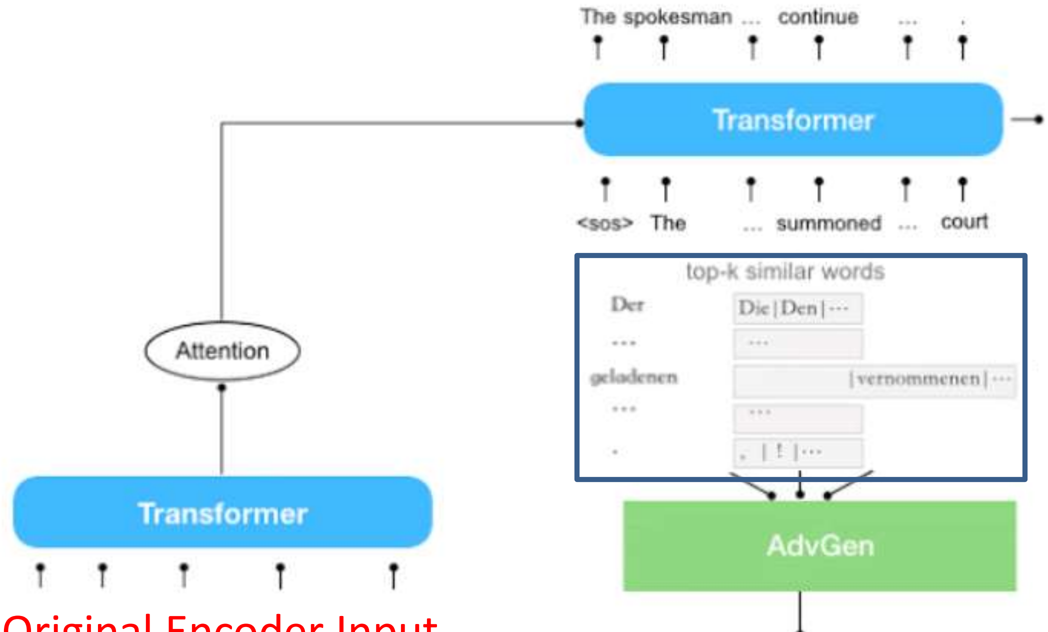
Inject Adversarial Perturbation(Noise) in Embedding Space

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

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

3 Method

AdvGen (Encoder)



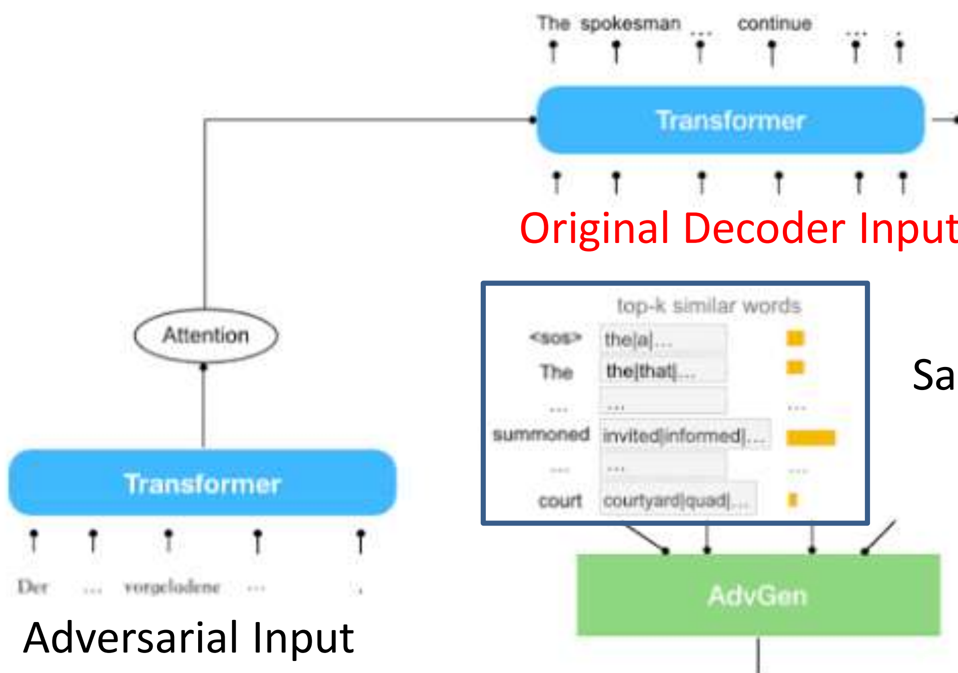
Sample words uniformly

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

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

Word Selection Distribution

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

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

Adversarial Attack on Word Composition

JungsooPark

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

Appendix

Limitation of Related Work

Subwords (_ means spaces)	Vocabulary id sequence
_Hell/o/_world	13586 137 255
_H/ello/_world	320 7363 255
_He/llo/_world	579 10115 255
_/He/l/l/o/_world	7 18085 356 356 137 255
_H/el/l/o/_/world	320 585 356 137 7 12295

- ✓ Using subword segmentation method, same word can be segmented in many ways.
- ✓ Thus if typo occurs, the error will be accumulated by word being segmented into entirely wrong segments

Appendix

Typos Make the Word Composition Entirely Wrong

"Actually"

"Actual" "ly"

"Actualy"

"Act" "ual" "ly"

Appendix

Subword Regularization

Kudo et al. (ACL 2018)

To resolve ambiguity in word segmentation and inform NMT model the composition of word, "Subword Regularization" was proposed.

- Using probabilistic model(unigram language model) for generating segmentation candidates for a given sequence.
- During training, each sequence can be segmented in to multiple candidates, thus informing the model word composition.
- Main method for "Sentencepiece"

Appendix

Example

```
>>> import sentencepiece as spm
>>> s = spm.SentencePieceProcessor()
>>> s.Load('spm.model')
>>> for n in range(5):
...     s.SampleEncodeAsPieces('New York', -1, 0.1)
...
['▁', 'N', 'e', 'w', '▁York']
['▁', 'New', '▁York']
['▁', 'New', '▁Y', 'o', 'r', 'k']
['▁', 'New', '▁York']
['▁', 'New', '▁York']
```

Appendix

Adversarial Attack on Word Composition

Informing model the word composition is quite critical in making NMT models robust,

- How about sampling some words and make those words segmented into other compositions in the direction of making the model most vulnerable
- Candidates will be listed by unigram language model by probability
- As a result, we expect the model to get information of word composition.

Appendix

Adversarial Attack on Word Composition

```
-0.08174121379852295    The U.S. government.
-0.1639 -0.0978 -0.0686 -0.0263 -0.0600 -0.0571 -0.1228 -0.0575
Der US-Regierung.
-0.1015702486038208     The U.S. government.
-0.1631 -0.0244 -0.2612 -0.0186 -0.0384 -0.1305 -0.1166 -0.0597
Der US-Regierung.
-0.39662614464759827    Der U.S. government.
-0.7268 -0.1862 -2.2555 -0.0698 -0.0260 -0.0583 -0.0661 -0.1197 -0.0613
Der US-Regierung.
-0.3923000693321228     Your U.S. government.
-2.1688 -0.5472 -0.0666 -0.0384 -0.0589 -0.0478 -0.1468 -0.0639
Der US-Regierung.
-0.306985080242157      The U.S. government.
-1.6898 -0.3488 -0.0727 -0.0303 -0.0572 -0.0498 -0.1520 -0.0553
Der US-Regierung.
-0.30663415789604187    Der U.S. government.
-0.5064 -0.1880 -0.0298 -1.9115 -0.0607 -0.0293 -0.0567 -0.1106 -0.1131 -0.0602
Der US-Regierung.
-0.3574221432209015     It's the U.S. government.
-2.2462 -0.3567 -0.0535 -0.7455 -0.1489 -0.0587 -0.0176 -0.0552 -0.0333 -0.1479 -0.0681
Der US-Regierung.
-0.23668985068798065    Der US government.
-0.5999 -0.1138 -0.1847 -0.4039 -0.1677 -0.1222 -0.0645
Der US-Regierung.
-0.09902060031890869    The U.S. government.
-0.2121 -0.1229 -0.0903 -0.0193 -0.0426 -0.1315 -0.1164 -0.0570
```

```
-0.4872387945652008     And so the investigators, without their consent, got his phone call straight.
-0.3938 -0.1372 -0.3753 -0.7674 -0.0508 -0.0315 -0.0967 -0.3734 -0.0991 -1.2012 -0.2110 -0.0625 -1.526
Und so haben sich die Ermittler, ohne sein Einverständnis, seine Telefonnachweise geheim besorgt,
-0.5079879760742188     And so the Ermittlers, without consent, got his phone requests straight.
```