Exploiting Monolingual Data at Scale for Neural Machine Translation

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, Tieyan Liu (EMNLP 2019)

Presenter: Jiao, Wenxiang

2020-11-09

Abstract

- How to effectively leverage monolingual data is an important research topic for NMT.
    - There are plenty of works studying this problem.

- Back-Translation (BT) is one of the most cited approach.
    - Sennrich et al. (2016) and Edunov et al. (2018) leverage the target-side monolingual data.

- On the other hand, the investigation on source-side monolingual data is very limited.
    - Zhang and Zong (2016) and Ueffing et al. (2007) use the source-side data to make the synthetic target data.

- This paper study how to leverage both source-side and target-side monolingual data to boost the accuracy of NMT.

Improving NMT by Monolingual Data

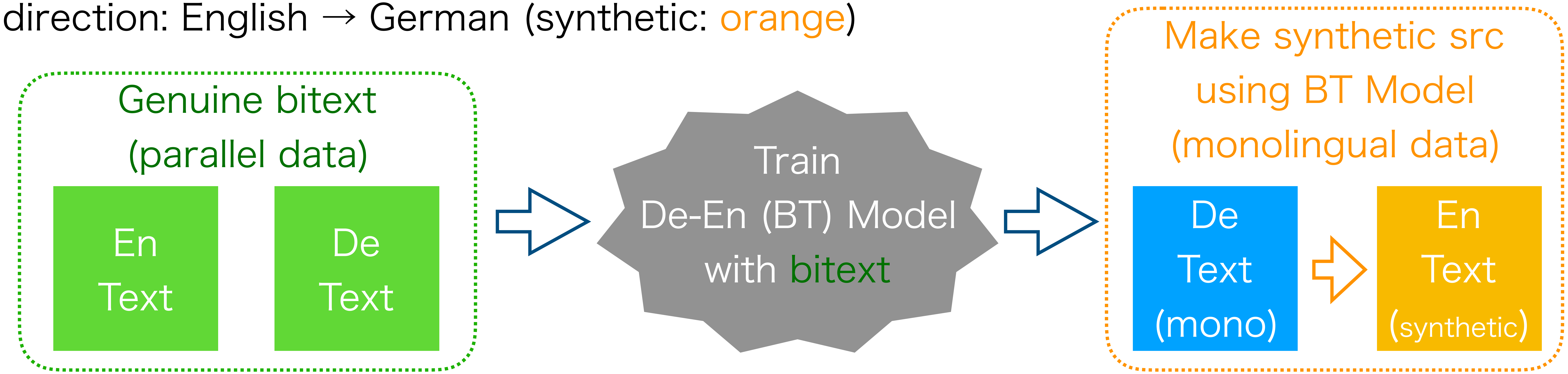
- Target-side monolingual data.
  - Fusion: NMT + Language model.
  - BT: Make synthetic source-side data from target-side monolingual data.

- Source-side monolingual data.
  - Self-learning: Make synthetic target-side data from source-side monolingual data.

Improving NMT by Monolingual Data

- BT requires training an additional target-to-source NMT model given the bilingual dataset.
- The translation output and the target-side monolingual data then paired as synthetic parallel corpus to augment the original bilingual dataset.

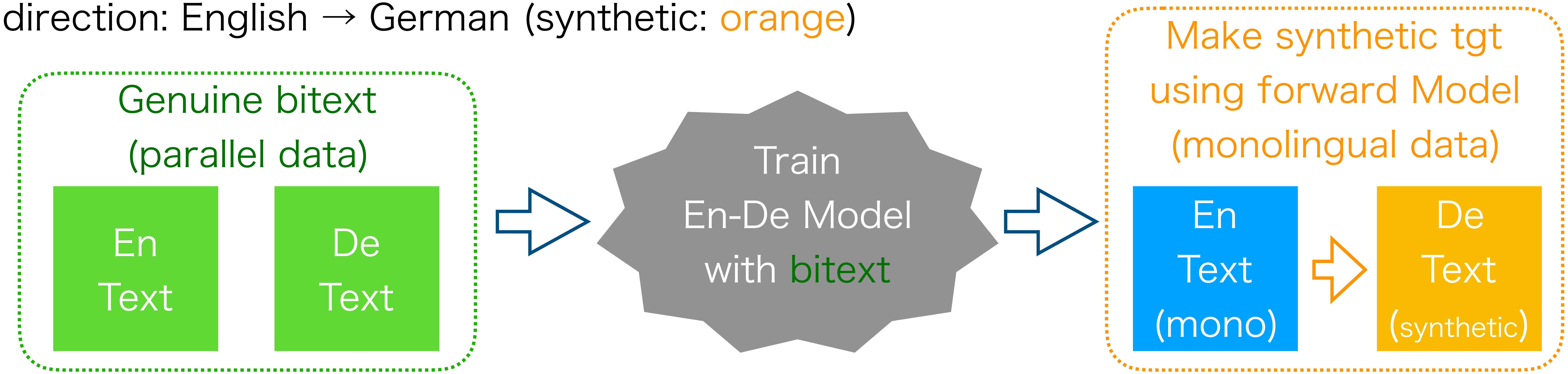
direction: English → German (synthetic: orange)



Improving NMT by Monolingual Data

- Self-learning approach generates the synthetic data for the source-side monolingual data, which is a semi-supervised method.

direction: English → German (synthetic: orange)


Genuine bitext
(parallel data)
En Text
De Text
Train En-De Model with bitext
Make synthetic tgt using forward Model (monolingual data)
En Text (mono)
De Text (synthetic)

Notation

- X, Y: languages
- X, Y: the collection of all sentences for each language
- $B = \{(x_i, y_i)\}_{i=1}^{N}$: the bilingual training pairs, $x_i \in X, y_i \in Y$
- $M_x = \{x_j\}_{j=1}^{M_x}, M_y = \{y_j\}_{j=1}^{M_y}$: the collection of monolingual sentences, $x_i \in X, y_i \in Y$
- $f : X \rightarrow Y$: translation model

Training Strategy: Data Preparation

- Train two translation models on the bilingual data B
  - Forward-translation model: $f_b : X \rightarrow Y$
  - Backward-translation model: $g_b : Y \rightarrow X$

- Generate the following two synthetic datasets
  - Forward-translation: $\bar{B}_s = \{(x, f_b(x)) \,|\, x \in M_x\}$
  - Backward-translation: $\bar{B}_t = \{(g_b(y), y) \,|\, y \in M_y\}$

- This paper mainly adopt beam search to generate the sentences.

Training Strategy: Large-Scale Noised Training

- They add noise to the source-side data of both $\bar{B}_s$ and $\bar{B}_t$ for training instead of directly using them to train models.
  - Two noised datasets:

$$\bar{B}_s^n = \{(\sigma(x), y) \,|\, (x, y) \in \bar{B}_s\} \qquad \bar{B}_t^n = \{(\sigma(x), y) \,|\, (x, y) \in \bar{B}_t\}$$

- $\sigma(\,\cdot\,)$: Noise function.
  - Randomly replace a word to be a special tokens with probability 0.1.
  - Randomly drop the words with probability 0.1.
  - Randomly shuffle (swap) the words with constraint that the words will not be shuffled further than three positions distance.

- They then train an NMT model f_n for X-to-Y translation on $B \cup \bar{B}_s^n \cup \bar{B}_t^n$.

Training Strategy: Clean Data Tuning

- They further fine-tune the noised training model on the clean version of the synthetic data without adding noise manually.

- Train another f_b for X-to-Y translation and another g_b for Y-to-X translation.
  - Use them to build new synthetic data $\bar{B}_s$ and $\bar{B}_t$.
  - Subsample sentences to form $\bar{B}_s^s$ and $\bar{B}_t^s$.

- Fine-tune the translation model f_n (noised training model) on the new data.

$$\min \sum_{(x,y)\in B\cup\bar{B}_s^s\cup\bar{B}_t^s} -\log P(y\,|\,x;f)$$

Experimental Setup

- Task: WMT19 En-De, De-En, De-Fr, Fr-De
- Data: En-De, De2En
  - Parallel corpus
    - WMT (5M pairs); WMTPC (18M pairs).
  - Monolingual corpus
    - En, De News Crawl 2016-2018 => 120M sentences.
- Data: De-Fr, Fr-De
  - Parallel corpus
    - WMT (2M pairs); WMTPC (4.8M pairs).
  - Monolingual corpus
    - News Crawl: 60M sentences.

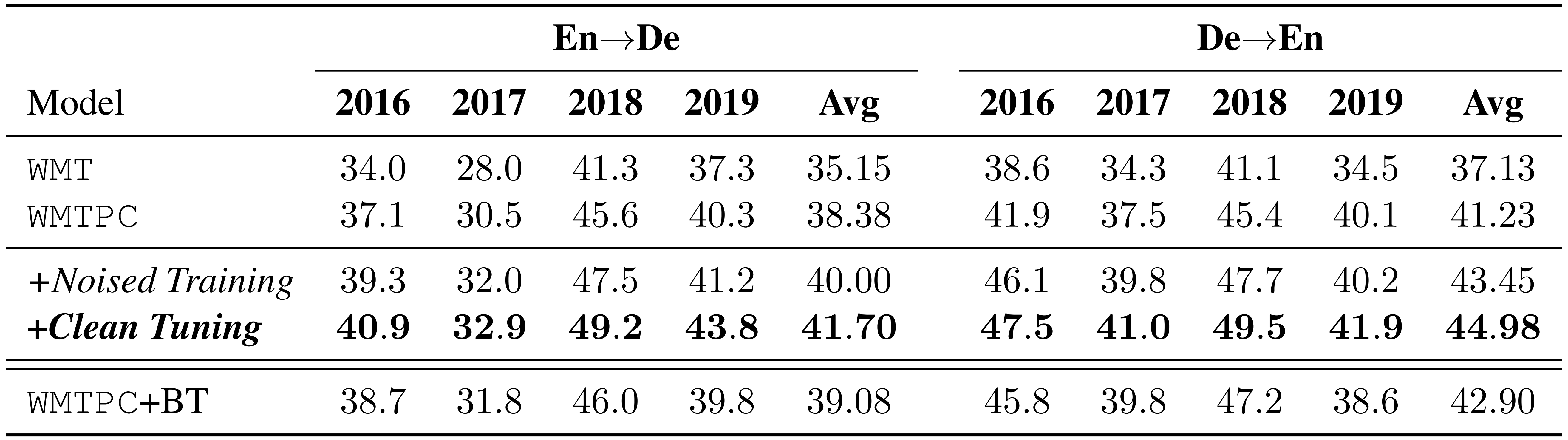
- 35K BPE; Transformer-big; Batch 4096 x 4; Update every 16 batches.

For noised training, the models used for translating monolingual data are trained on WMT.

For fine-tuning, the models used for translating monolingual data are trained on WMTPC.

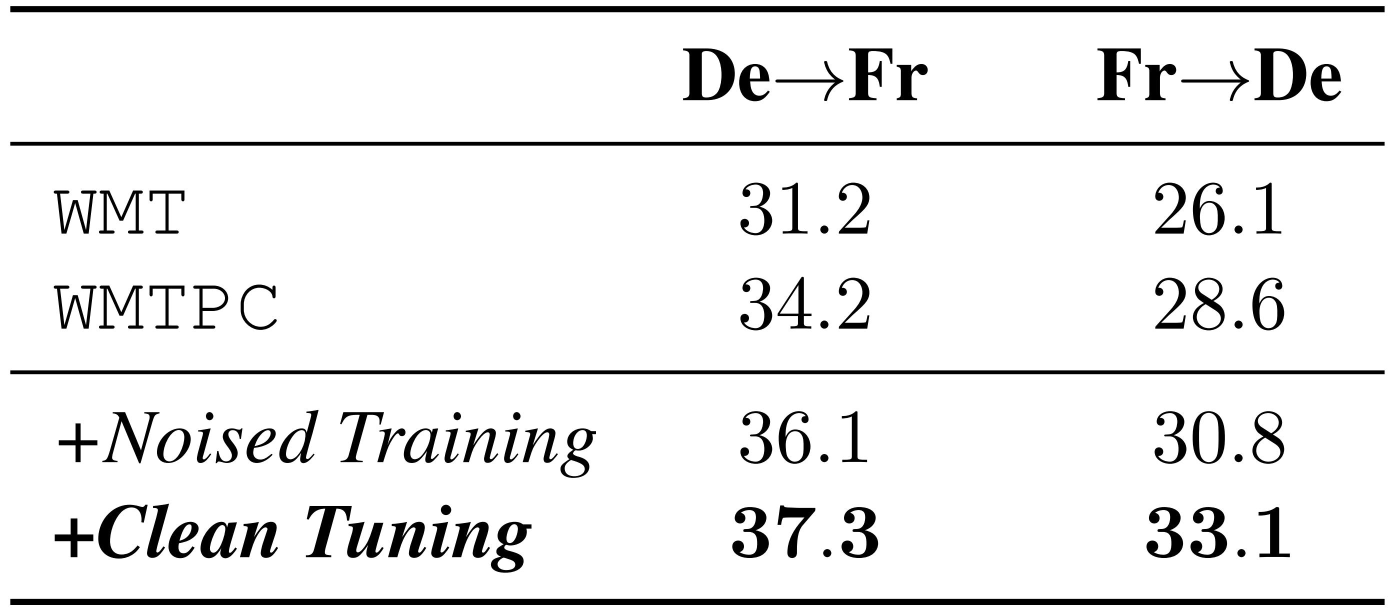
Experiments: Results

- This paper compares their strategy to the vanilla BT, which consists of 20M synthetic data and WMTPC (WMTPC+BT).



	En→De					De→En				
Model	**2016**	**2017**	**2018**	**2019**	**Avg**	**2016**	**2017**	**2018**	**2019**	**Avg**
WMT	34.0	28.0	41.3	37.3	35.15	38.6	34.3	41.1	34.5	37.13
WMTPC	37.1	30.5	45.6	40.3	38.38	41.9	37.5	45.4	40.1	41.23
+*Noised Training*	39.3	32.0	47.5	41.2	40.00	46.1	39.8	47.7	40.2	43.45
+Clean Tuning	**40.9**	**32.9**	**49.2**	**43.8**	**41.70**	**47.5**	**41.0**	**49.5**	**41.9**	**44.98**
WMTPC+BT	38.7	31.8	46.0	39.8	39.08	45.8	39.8	47.2	38.6	42.90

Table 1: De-tokenized case-sensitive SacreBLEU on WMT En↔De newstest2016, newstest2017, newstest2018, newstest2019 and the average score. "Avg" means the average BLEU score. "+" is conducted upon WMTPC dataset.

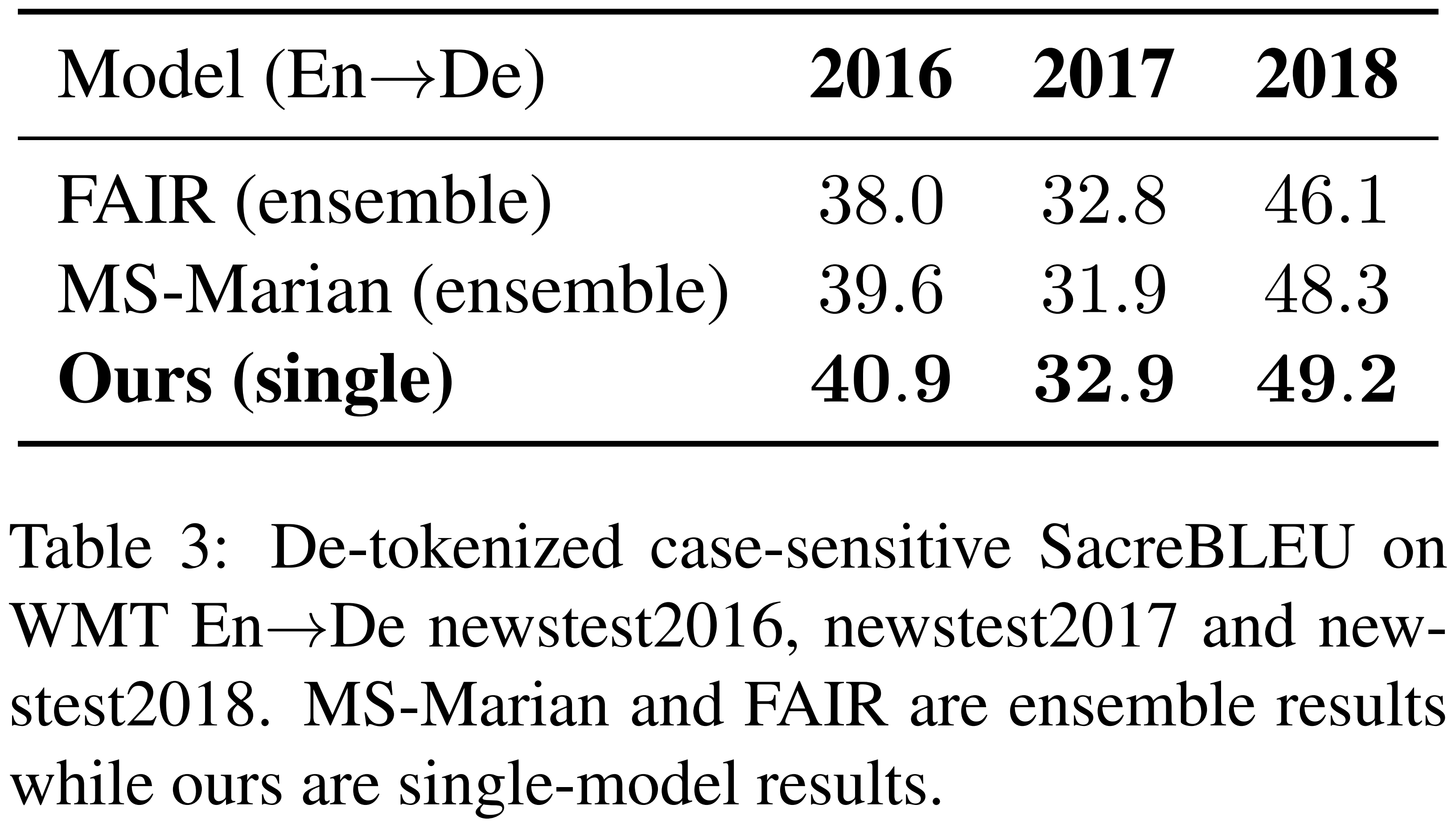


	De→Fr	**Fr→De**
WMT	31.2	26.1
WMTPC	34.2	28.6
+*Noised Training*	36.1	30.8
+Clean Tuning	**37.3**	**33.1**

Table 2: De-tokenized case-sensitive SacreBLEU on WMT De↔Fr newstest2019. "+" is conducted upon WMTPC dataset.

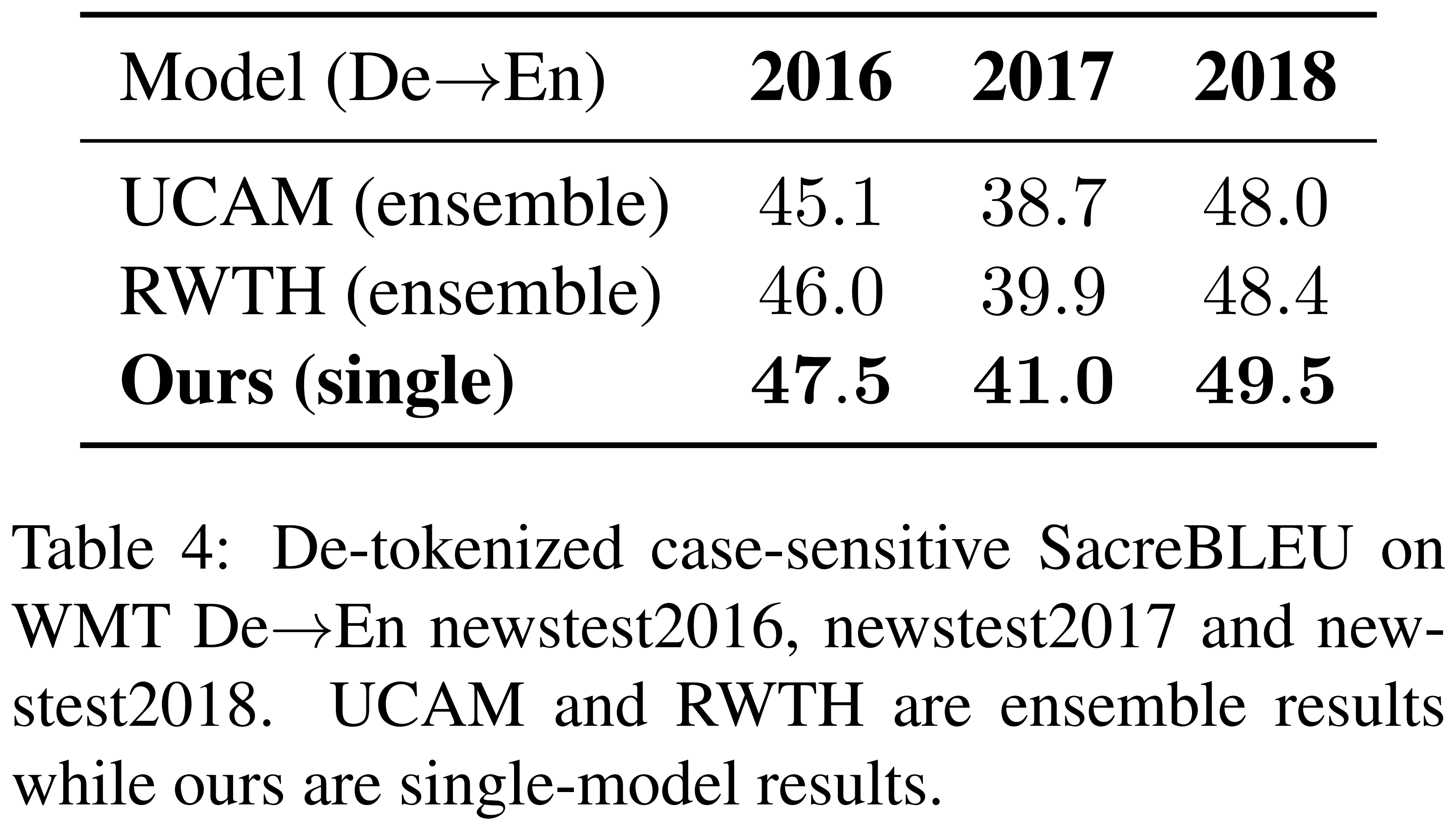
Experiments: Results

- Comparison with SOTA.



Model (En→De)	**2016**	**2017**	**2018**
FAIR (ensemble)	38.0	32.8	46.1
MS-Marian (ensemble)	39.6	31.9	48.3
Ours (single)	**40.9**	**32.9**	**49.2**

Table 3: De-tokenized case-sensitive SacreBLEU on WMT En→De newstest2016, newstest2017 and newstest2018. MS-Marian and FAIR are ensemble results while ours are single-model results.



Model (De→En)	**2016**	**2017**	**2018**
UCAM (ensemble)	45.1	38.7	48.0
RWTH (ensemble)	46.0	39.9	48.4
Ours (single)	**47.5**	**41.0**	**49.5**

Table 4: De-tokenized case-sensitive SacreBLEU on WMT De→En newstest2016, newstest2017 and newstest2018. UCAM and RWTH are ensemble results while ours are single-model results.

Analysis: Source or Target Monolingual Data

- They compare three different ways to use monolingual data, including leveraging:
  - Source-side monolingual data only ($\bar{B}_s$).
  - Target-side monolingual data only ($\bar{B}_t$).
  - Monolingual data from both sides ($\bar{B}_s + \bar{B}_t$).

- Total monolingual data size: 120M.

- Without noise.

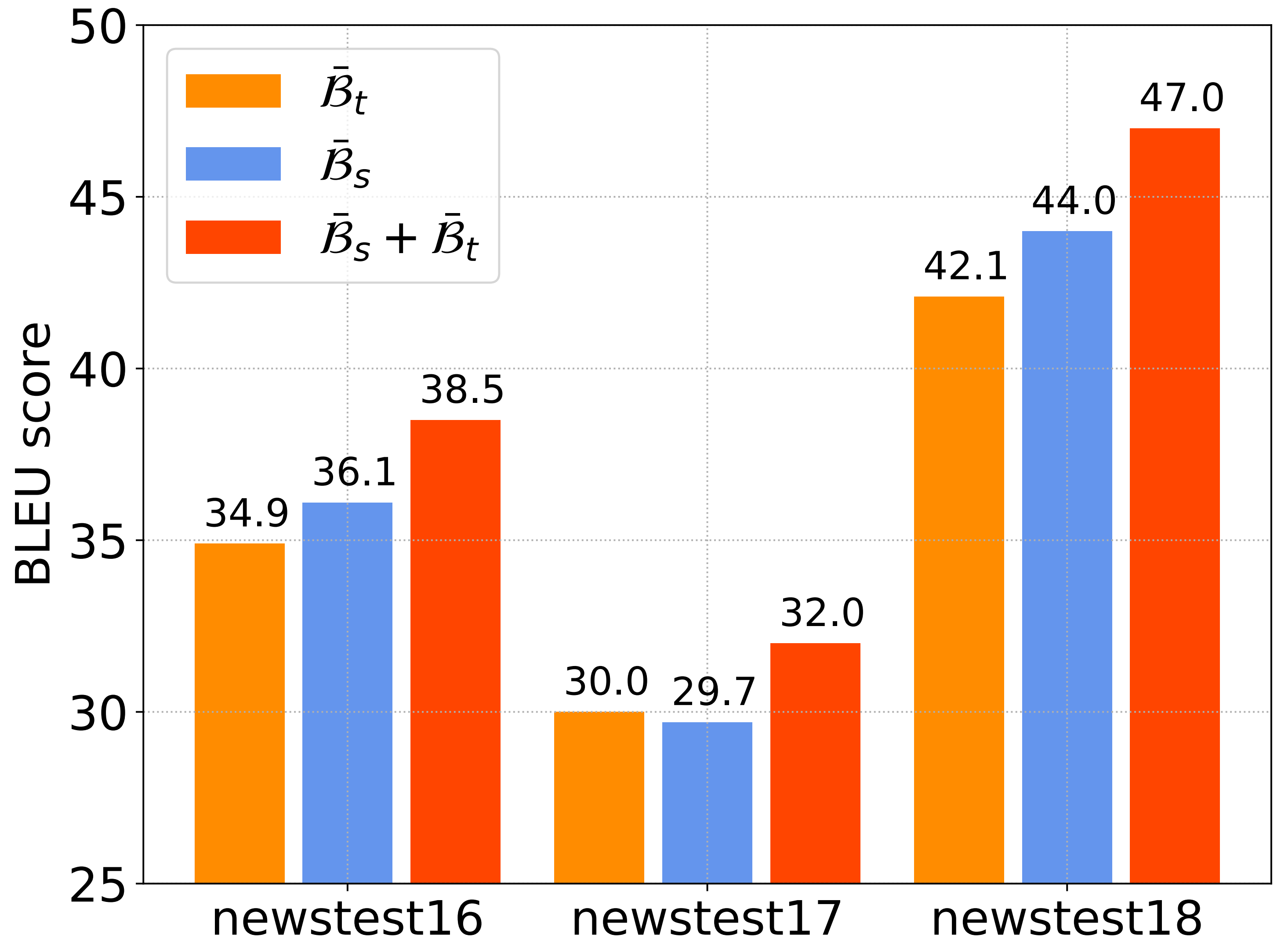


Figure 1: The de-tokenized SacreBLEU scores on En→De newstest2016, newstest2017 and newstest2018 of the models trained by different synthetic data: (1) $\bar{\mathcal{B}}_s$ from source-side monolingual data only, (2) $\bar{\mathcal{B}}_t$ from target-side monolingual data only and (3) the combination of $\bar{\mathcal{B}}_s$ and $\bar{\mathcal{B}}_t$.

Analysis: Synthetic Data Generation

- They conduct experiments on different generated synthetic data, to verify whether adding noise id essential.

- They compare their noised training data $\bar{B}_s^n$ and $\bar{B}_t^n$ with another two baselines:
  - $\bar{B}_s$ and $\bar{B}_t$ without any transformation.
  - $\bar{B}_s$ and BT data by sampling $\bar{B}_t^r$.

- Only noised training.

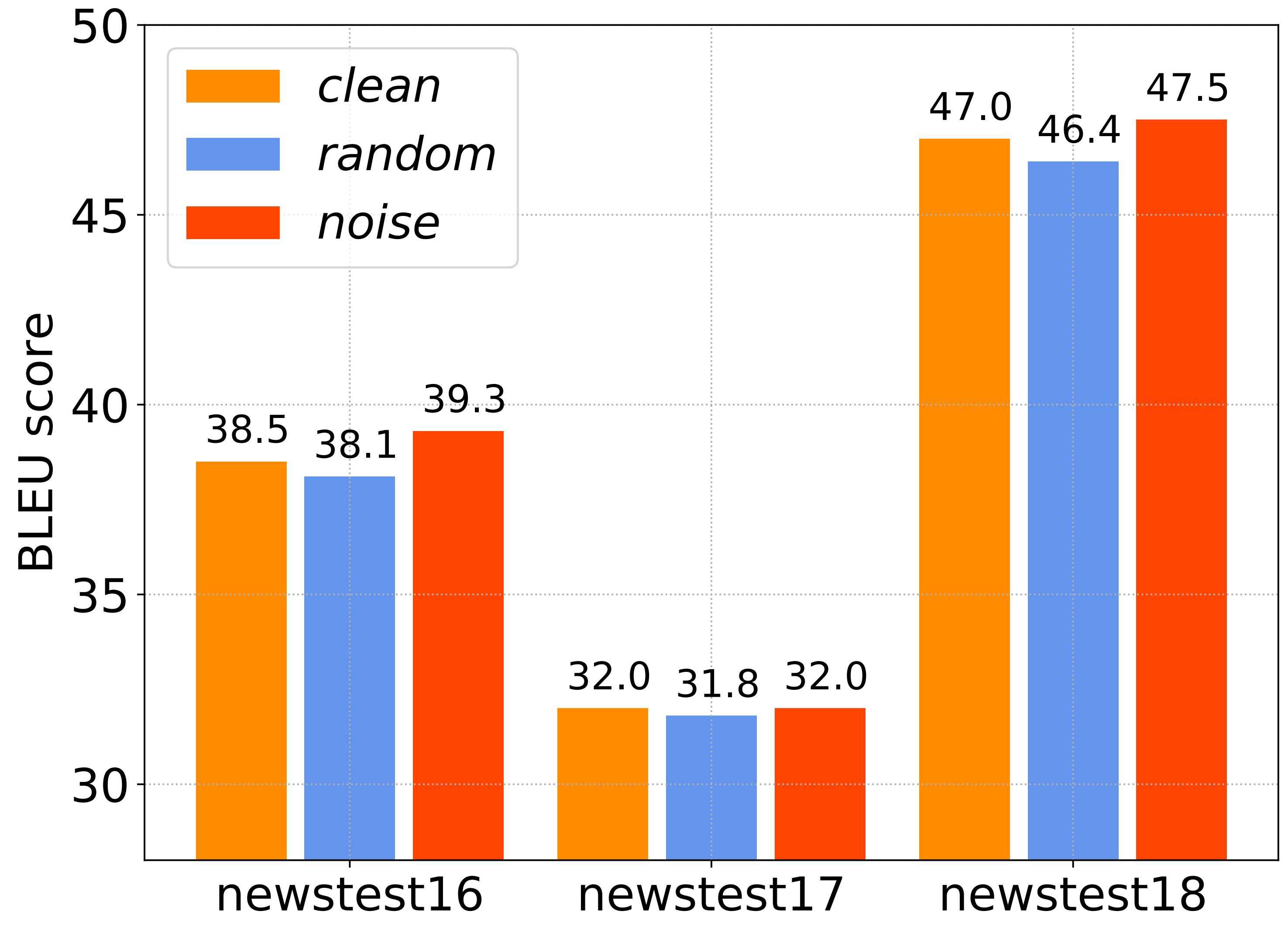
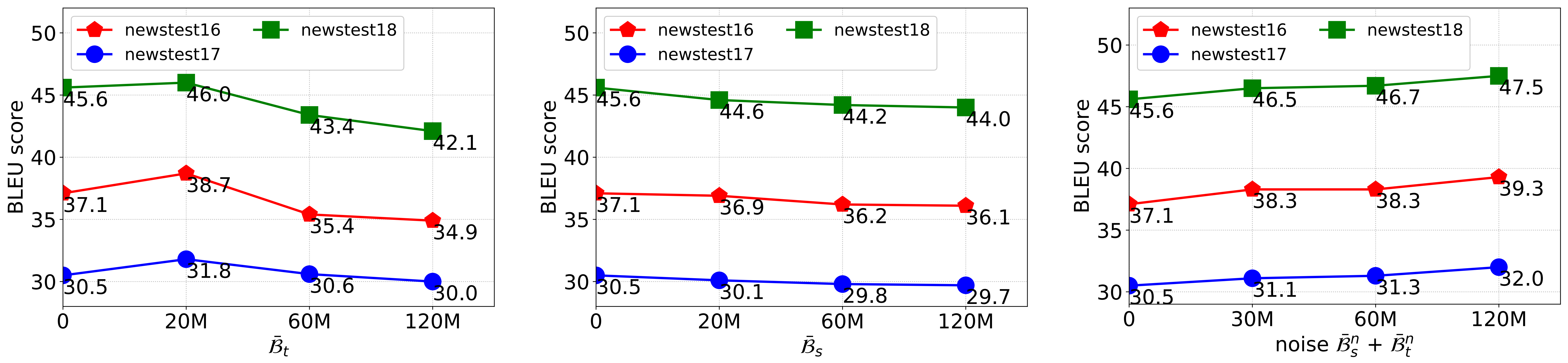


Figure 2: The de-tokenized SacreBLEU scores on En→De newstest2016, newstest2017 and newstest2018 of the models trained by synthetic data generated in different ways: (1) clean $\bar{\mathcal{B}}_s$ and $\bar{\mathcal{B}}_t$ data, (2) $\bar{\mathcal{B}}_s^r$ and randomly sampled $\bar{\mathcal{B}}_t^r$ data, and (3) noised $\bar{\mathcal{B}}_s^n$ and $\bar{\mathcal{B}}_t^n$ data.

Analysis: Scale of Monolingual Data

- They give a comparison of different data scales for each kind of synthetic data.
  - Source-side monolingual data is helpful and the best way to use it is to combine with target-side monolingual data.
  - They show that adding noise to synthetic data outperforms that without noise.



(a) Different scales of $\bar{\mathcal{B}}_t$ data.

(b) Different scales of $\bar{\mathcal{B}}_s$ data.

(c) Different scales of noised $\bar{\mathcal{B}}_s + \bar{\mathcal{B}}_t$ data.

Figure 3: The de-tokenized SacreBLEU scores on newstest2016, newstest2017, newstest2018 of the models trained with varied data scales of (a) $\bar{\mathcal{B}}_t$ data, (b) $\bar{\mathcal{B}}_s$ data, and (c) combined $\bar{\mathcal{B}}_s$ and $\bar{\mathcal{B}}_t$ data.

Analysis: Synthetic Tuning

- Is it helpful to use two groups of models of building synthetic data for noised training and clean tuning?
  - At fine-tuning step, use a subsample of synthetic data of noised training step.

- Is it helpful to use noised training first regarding future BLEU score achieved by the fine-tune step? (With clean tuning)
  - Training without noise first and then fine-tuning using clean data.

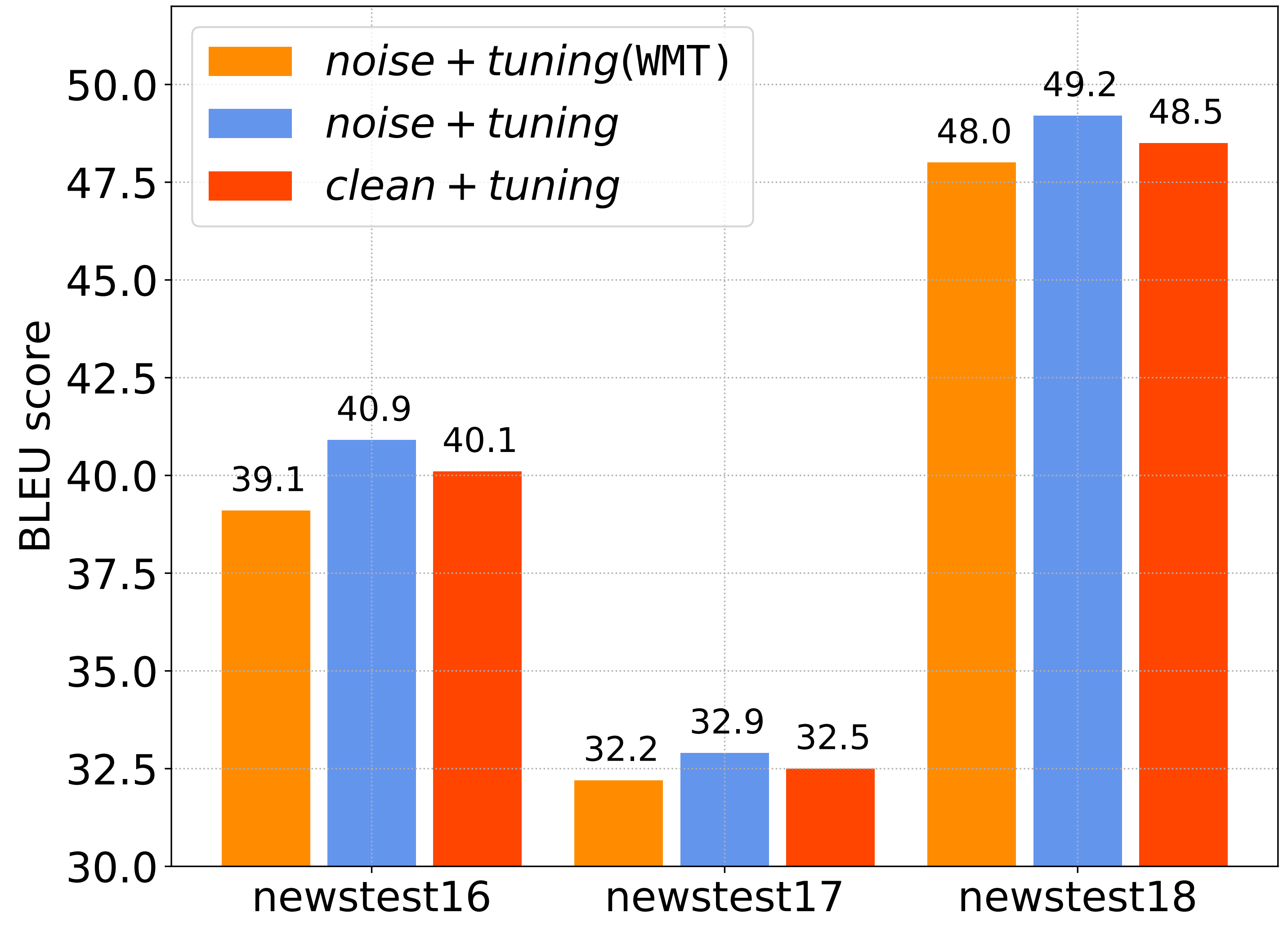


Figure 4: The de-tokenized SacreBLEU scores on En→De newstest2016, newstest2017 and newstest2018 of the models tuned by different synthetic data and pretrained models: (1) noised training model tuned on the subset of $\bar{\mathcal{B}}_s$ and $\bar{\mathcal{B}}_t$ (WMT in the figure), (2) noised training model tuned on the synthetic data as introduced in Section 4.1, and (3) clean training model tuned on the same synthetic data generated as (2).

Summary

- Exploited the monolingual data at scale for the neural machine translation task.
- Proposed an effective training strategy to boost the NMT performance by leveraging both source-side and target-side monolingual data.
- Future directions:
    - More language pairs.
    - Other sequence-to-sequence task.
    - With other data augmentation approaches.