Exploiting Monolingual Data at Scale for Neural Machine Translation

[1]Lijun Wu, [2]Yiren Wang, [3]Yingce Xia, [3]Tao Qin,
[1]Jianhuang Lai and [3]Tie-Yan Liu
[1]Sun Yat-sen University;
[2]University of Illinois at Urbana-Champaign;
[3]Microsoft Research Asia

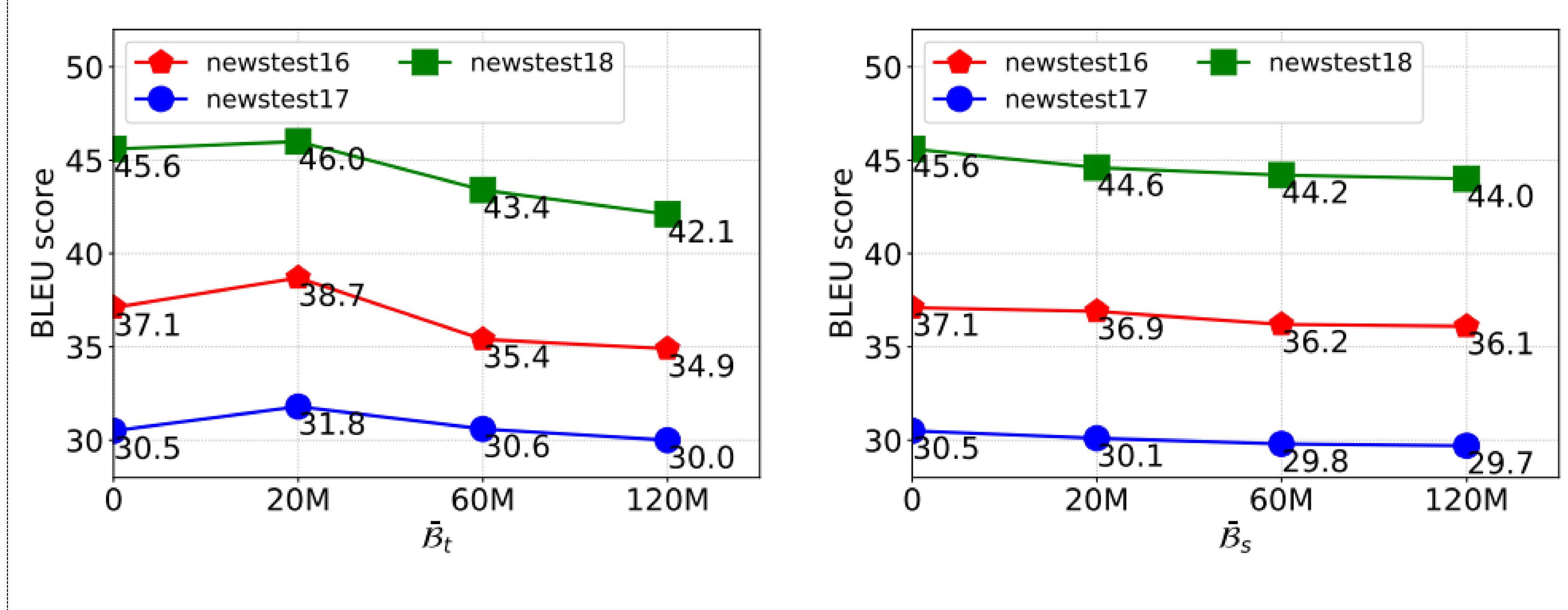
Conference on Empirical Methods in Natural Language Processing & International Joint Conference on Natural Language Processing 2019
November 3–November 7
Hong Kong

1. Introduction

- NMT consumes large amount of bilingual data. However, bilingual data is limited while monolingual data is unlimited.
- **Target-side** monolingual data has been proven to be effective through Back-Translation (BT) approach.
- **Source-side** monolingual data is not well studied. Using Forward-Translation only (FT) is not as effective as expected.
- *We propose an effective and simple strategy/pipeline to **leverage both** of the source-side and target-side monolingual data.*
- *We achieve SOTA results and make a comprehensive study on the effectiveness of source-side and target-side monolingual data with our approach.*

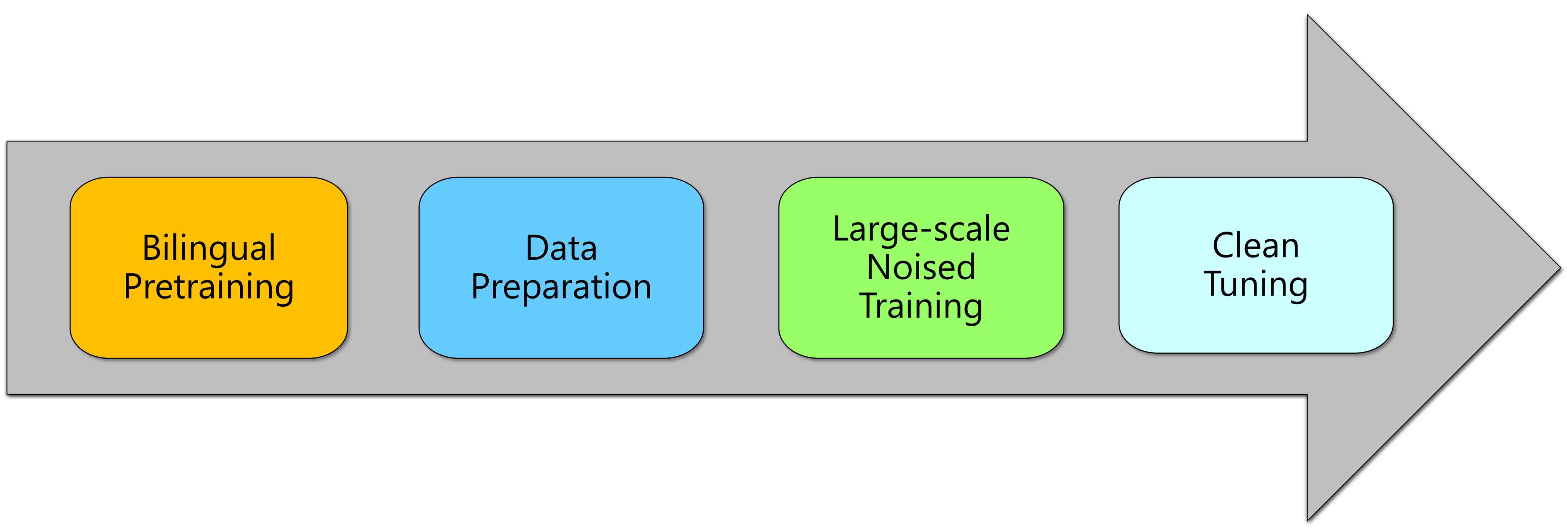
2. Monolingual Data

- The effectiveness of the source-side and target-side monolingual data under **different data scales**
  - ➢ Target-side monolingual data [$\mathcal{B}_t$]: Back-Translation (BT)
  - ➢ Source-side monolingual data [$\mathcal{B}_s$]: Forward-Translation (FT)
  - ➢ Data scales: 20M, 60M, 120M monolingual sentences
- Observations: ***one-side usage is not effective***
  - ➢ Back-Translation: first improve the performance, then drop quickly
  - ➢ Forward-Translation: performance drop little by little



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

3. Training Strategies



- *Stage-1*: **Bilingual** model pretrain on bitext.
- *Stage-2*: BT and FT **data generation** with pretrained bilingual model.
- *Stage-3*: **Combine** BT and FT data and **Add Noise** to source sentence; train on the large-scale noised data.
  - – Randomly replace words to be <unk> $\mathcal{B}\cup\bar{\mathcal{B}}_s^n\cup\bar{\mathcal{B}}_t^n$
  - – Randomly drop words
  - – Randomly shuffle words
- *Stage-4*: **Resample** BT and FT data, and train on the clean resampled data. $\mathcal{B}\cup\bar{\mathcal{B}}_s^s\cup\bar{\mathcal{B}}_t^s$

4. Experiments

- **Overall Results**
  - – WMT16,17,18,19 En↔De

Model	En→De 2016	En→De 2017	En→De 2018	En→De 2019	En→De Avg	De→En 2016	De→En 2017	De→En 2018	De→En 2019	De→En 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

- **Comparison with SOTA systems**
  - – WMT16,17,18,19 En↔De

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**

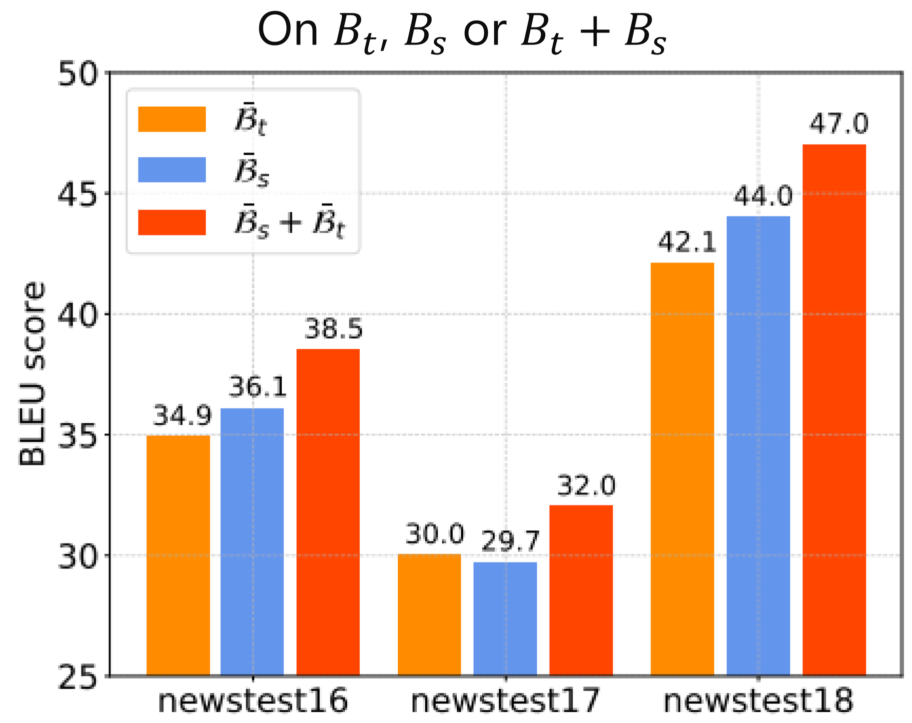
Model (De→En)	2016	2017	2018
UCAM (ensemble)	45.1	38.7	48.0
RWTH (ensemble)	46.0	39.9	48.4
Ours (single)	**47.5**	**41.0**	**49.5**

5. Studies

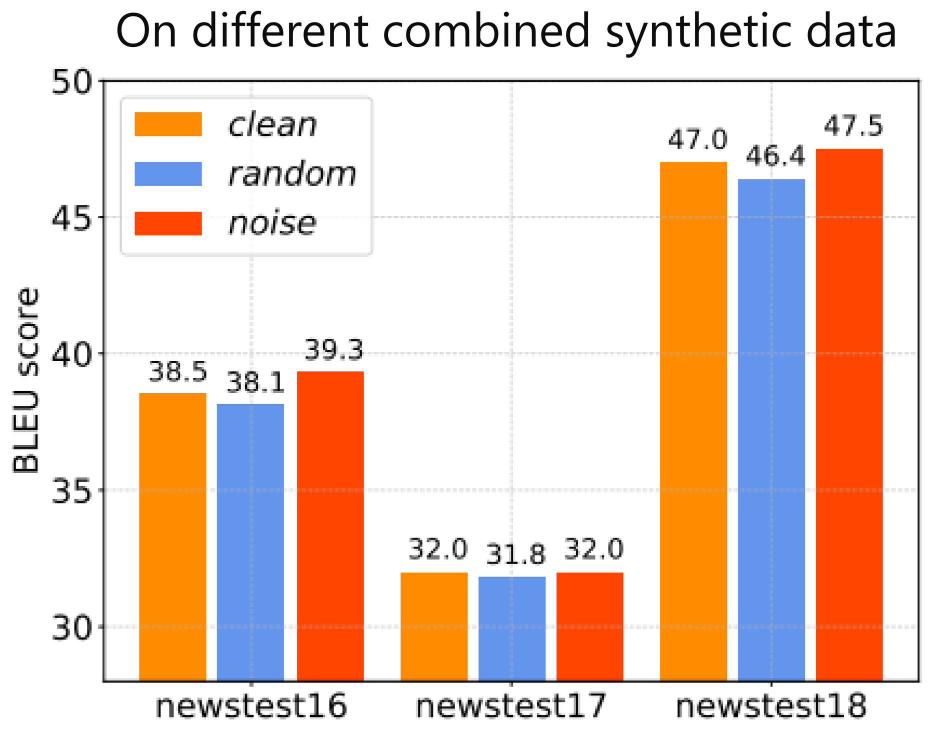
Contact
- wulijun3@mail2.sysu.edu.cn

- **Analysis**

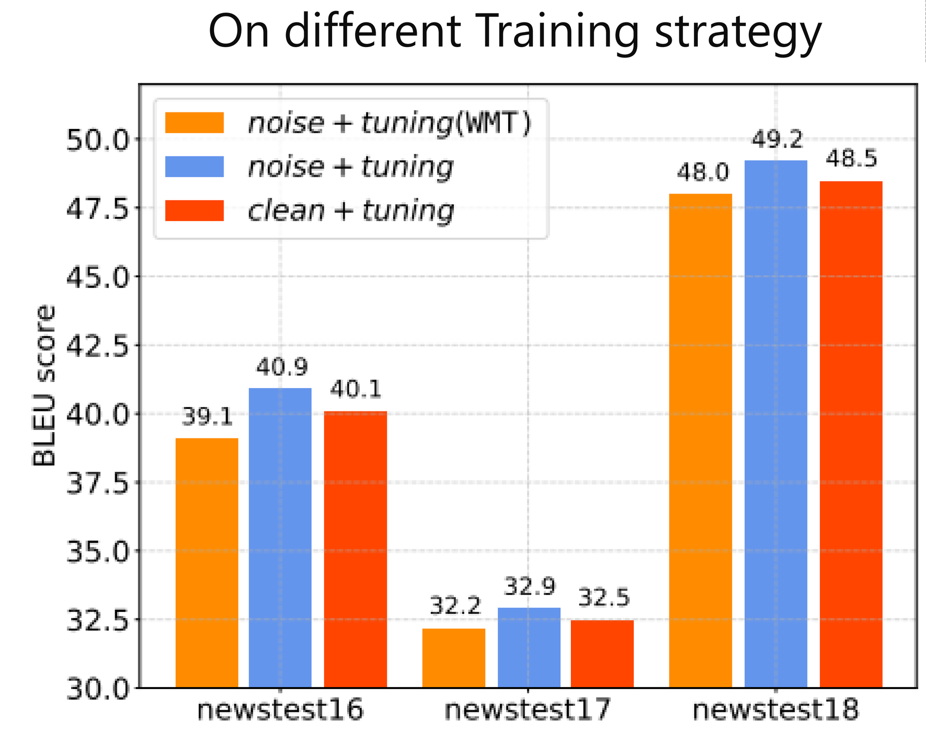
On B_t, B_s or $B_t + B_s$



On different combined synthetic data



On different Training strategy



Summary

✓ Combine both sides monolingual data

✓ Add noise to large-scale synthetic data

✓ Tune on the clean synthetic data