Tsinghua University Neural Machine Translation Systems for CCMT 2020

Speaker: Xuancheng Huang

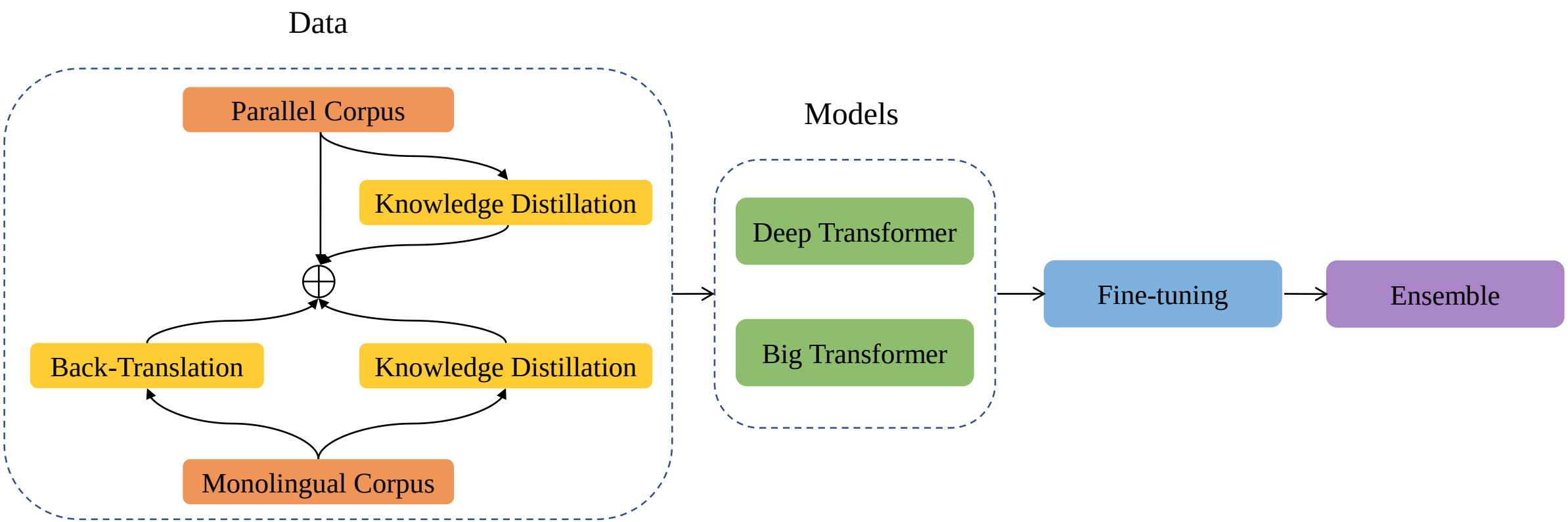
Authors: Gang Chen*, Shuo Wang*, Xuancheng Huang*, Zhixing Tan, Maosong Sun, and Yang Liu

2020-10-11

Contents

Overview

- We participated in English<->Chinese translation tasks.
- The overview of our methods is shown below.



Data Preprocessing

- We use CCMT and WMT data, which contains 26.7M sentence pairs.
- We apply the following procedures to preprocess:
  - Remove illegal UTF-8 characters.
  - Convert Traditional Chinese to Simplified Chinese.
  - Apply Unicode NFKC normalization.
  - Restore HTML/XML escape and normalize punctuation.

Data Filtering

- Furthermore, we utilize following rules to filter the parallel data:
  - Remove duplicate sentence pairs.
  - Sentence contains illegal characters (e.g., Chinese characters in English text) are removed.
  - We translate one side (e.g., Chinese) of bilingual data to another (e.g., English). Then, the sentence pairs whose BLEU score calculated between translated sentence and original sentence is lower than 5 are discarded.

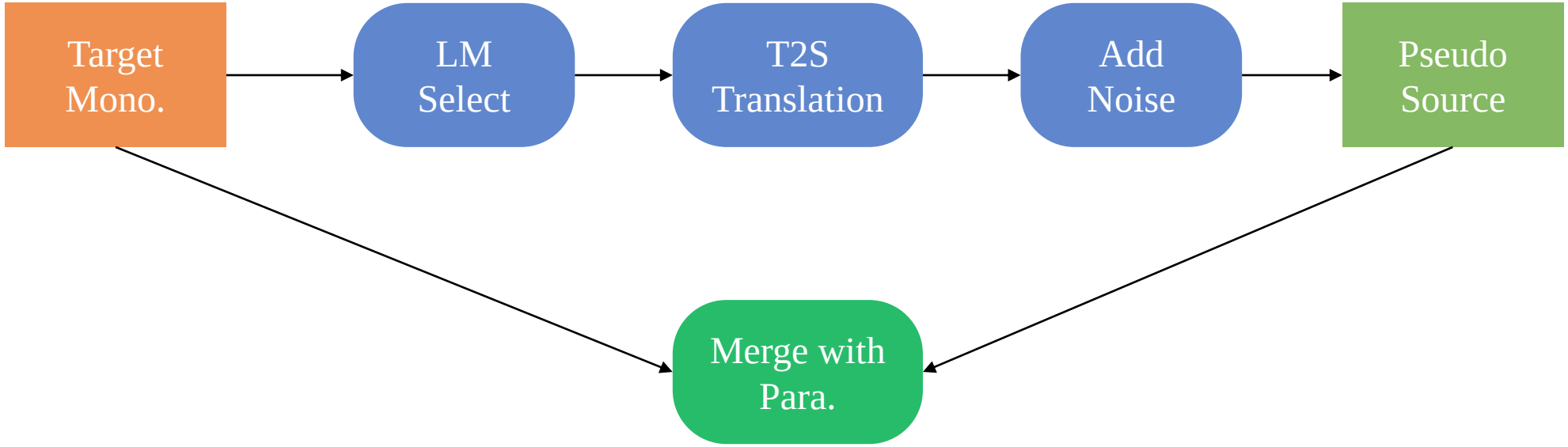
Models

- We use **Big** and **Deep** Transformer (Vaswani et al., 2017).
  - Big Transformer [hidden size=1024, filter size=4096, num of heads=16]
  - Deep Big Transformer (Big Transformer with 15 encoder layers)
  - Deeper Transformer (Base Transformer with 50 encoder layers)

- For Deep(er) Transformer, pre-layer normalizations (Wang et al., 2019) are used.

Data Augmentation

- We use **Back Translation** (Sennrich et al., 2016) to augment parallel data.

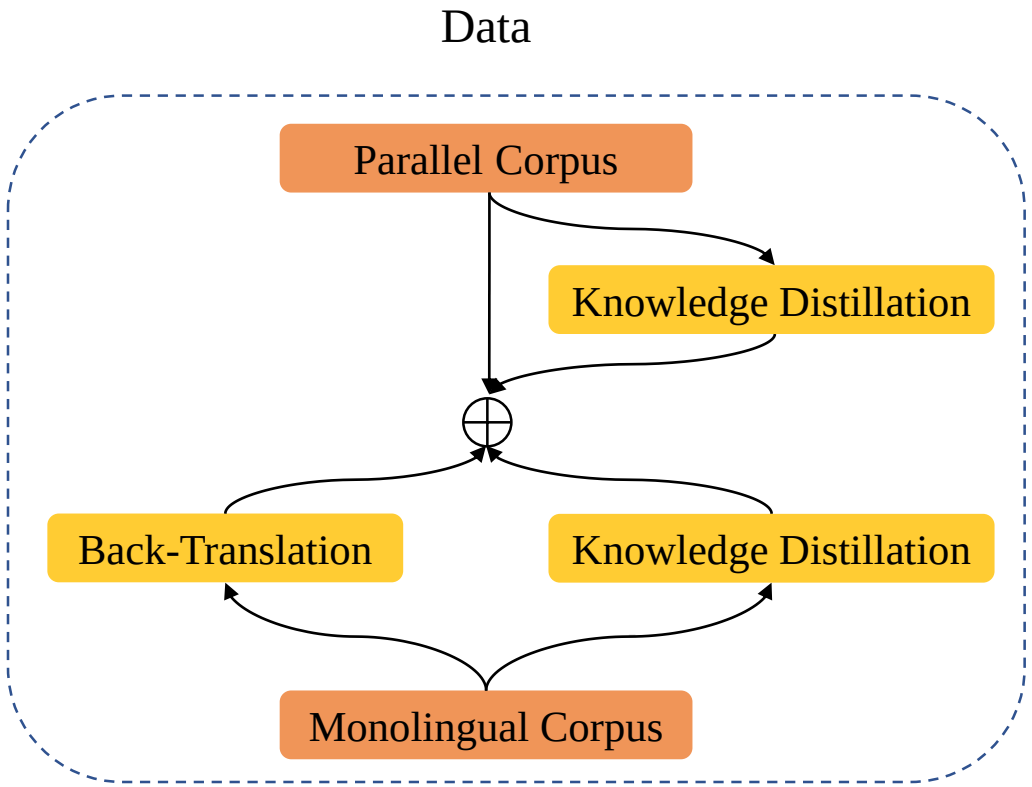


Data Augmentation (continued)

- We use **Knowledge Distillation** (Kim et al., 2016) to augment and regenerate parallel data.
  - Augmentation
    - Monolingual data

      -> pseudo parallel data
  - Regeneration
    - Source side of parallel data

      -> pseudo parallel data



Data Augmentation (continued)

- Different models are used for **Knowledge Distillation**:
  - Right2Left models.
  - Ensemble models.

Finetuning

- Previous work (Sun et al., 2019) found that finetuning with in-domain data can bring huge improvements.
- For ZH->EN, we use CWMT2008, CWMT2009 and original-Chinese part of newsdev2017, newstest2017, newstest2018 as in-domain data.
- For EN->ZH, we original-English part of newsdev2017, newstest2017, newstest2018 as in-domain data.

Finetuning (continued)

- We use in-domain data to finetune all models (strongest model, R2L model, etc.)
- During finetuning, we use a larger dropout rate and a smaller constant learning rate than those in the training process. The model parameters are updated after each epoch.

Ensemble

- We use **Ensemble**, which calculates the word-level averaged log-probability among different models during decoding, to combine different models for stronger performance.

Ensemble (continued)

- We also try to use **Transductive Ensemble Learning (TEL)** (Wang et al., 2020) , utilizing the synthetic test data (consists of original source sentences and translations of target-side sentences) of different models to finetune a single model.
- TEL and ensemble are not orthogonal.
- 5 L2R and 2 R2L models are used in our primary submission.

Experiments – ZH->EN

- Effectiveness of each method on the newstest2019 Chinese-English test set.

Settings	Transformer Big	Deep Transformer Base	Deep Transformer Big
Baseline	27.94	-	-
+Data Augmentation	28.59	29.74	29.85
+Finetuning	35.97	37.48	37.74
Ensemble		**38.95**	

- Our final submission with TEL achieves 48.12 BLEU-SBP on the ccmt2020 test set, which gains 1.1 BLEU-SBP improvements over the submission with standard model ensemble.

Experiments – ZH->EN (continued)

- Effectiveness of different data augmentation methods on Chinese-English task.

Dataset	Baseline	+BT	+Noise BT	+BT&EKD	+BT&R2LKD
newstest2019	34.85	34.53	35.05	35.65	35.02

- Back translation does not work well.
- Knowledge distillation is more effective.

Experiments – EN->ZH

- Evaluation results on the newstest2019 English-Chinese test set.

Dataset	Data Augmentation		Deep model	Finetuning	Ensemble
	BT	Noise BT			
newstest2019	36.94	37.02	36.91	38.33	39.56

- Our final submission achieves 63.43 BLEU-SBP on ccmt2020 English-Chinese test set.

Conclusion

- We participated in ZH <-> EN translation task.
- Larger Transformer model, data augmentation, finetuning, and ensemble improve translation performance substantially.
- Our open source NMT toolkit:

New! https://github.com/THUNLP-MT/THUMT (**Pytorch** version is available!)

- Our MT Reading List:
  - https://github.com/THUNLP-MT/MT-Reading-List

Thanks for listening