UdS Submission for the WMT 19 Automatic Post-Editing Task

Hongfei Xu[1,2] *, Qiuhui Liu[3], Josef van Genabith[1,2]

[1]Saarland University, [2]DFKI, [3]China Mobile Online Services

*hongfei.xu@dfki.de

Introduction

The Automatic Post-Editing (APE) task is to automatically correct errors in machine translation outputs.

In our submission, we:

- **Utilize and adapt an NMT architecture originally developed for exploiting context information to APE;**
- **Explore joint training of the APE task with a de-noising encoder.**

Data

Use both both the training set provided by WMT and the synthetic eSCAPE corpus.

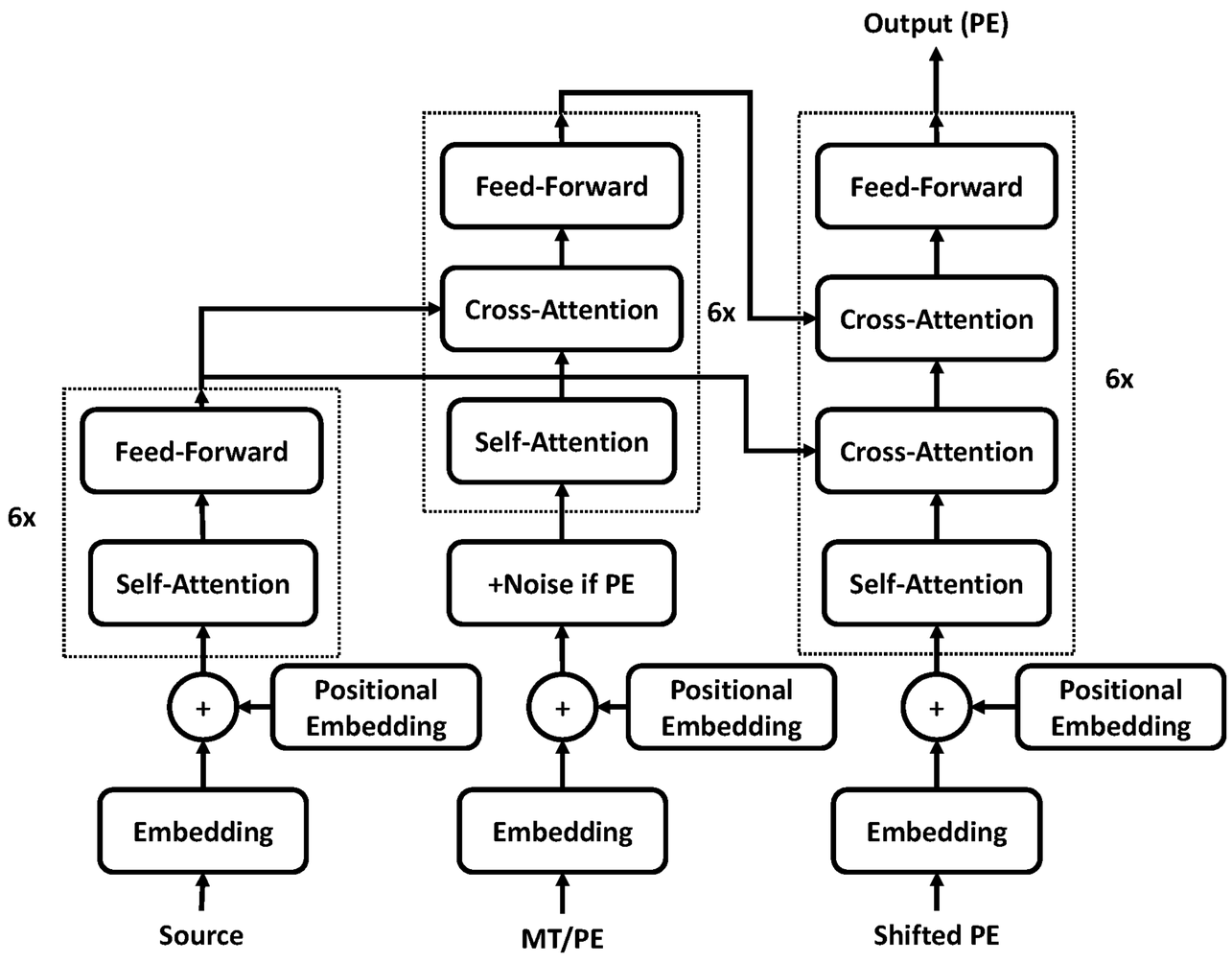
Pre-processing:

1. Re-tokenize (using arguments: -a -no-escape) and truecase with Moses;
2. Apply joint BPE with 40k merge operations and 50 as the vocabulary threshold;
3. Clean the data sets with scripts from the Neutron toolkit;
4. Up-sample the original training set 20 times.

Post-processing:

1. Recover BPE segmentation;
2. De-truecase and re-tokenize (without -a argument).

Our Model



Joint Training with De-noising Encoder

- Adaptive Gaussian / Uniform Noise

$$emb_{out} = emb + strength * \overline{abs(emb)} * N$$

- Combination of Objectives

$$loss = \lambda * loss_{ape} + (1 - \lambda) * loss_{de-noising}$$

Hyper Parameters

- Vocabulary Size: 42476;
- Dropout: 0.1;
- Embedding Dimension: 512;
- Hidden Units: 2048;
- Warm-up Steps: 8000;
- Batch Size: 25k tokens;
- Beam Size: 4;
- Checkpoint Saving: 1500 steps;
- Strength: 0.2;
- λ: 0.5.

Results

Development Set

Models	BLEU
MT as PE	76.76
Processed MT	76.61
Base Model	76.91~77.13
+Gaussian	76.94~77.08
+Uniform	77.01~77.10
Ensemble x5	77.22

Test Set

Models	TER	BLEU
MT as PE	16.84	74.73
Gaussian	16.79	75.03
Uniform	16.80	75.03
Ensemble x5	16.77	75.03

Analysis

- Additional pre-processing and post-processing introduced for training models hurts performance;
- The multi-source transformer (Base) model achieves the highest single model BLEU score without joint training;
- The performance gap between the best model and the worst model from joint training is smaller;
- Even the ensemble of 5 models does not result in significant differences especially in BLEU scores.

Acknowledgments

Hongfei Xu is supported by a doctoral grant from China Scholarship Council ([2018]3101, 201807040056).

This work is supported by the German Federal Ministry of Education and Research (BMBF) under the funding code 01IW17001 (Deeplee).

We thank the anonymous reviewers for their instructive comments.