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.

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-trucase and re-tokenize (without -a argument).

Our Model

- Adaptive Gaussian / Uniform Noise
  \[ \text{emb}_{\text{out}} = \text{emb} + \text{strength} \times \text{abs}(\text{emb}) \times N \]
- Combination of Objectives
  \[ \text{loss} = \lambda \times \text{loss}_{\text{ape}} + (1 - \lambda) \times \text{loss}_{\text{de-noising}} \]

Hyper Parameters

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

Results

<table>
<thead>
<tr>
<th>Development Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Models</td>
<td>BLEU</td>
</tr>
<tr>
<td>MT as PE</td>
<td>76.76</td>
</tr>
<tr>
<td>Processed MT</td>
<td>76.61</td>
</tr>
<tr>
<td>Base Model</td>
<td>76.91–77.13</td>
</tr>
<tr>
<td>+Gaussian</td>
<td>76.94–77.08</td>
</tr>
<tr>
<td>+Uniform</td>
<td>77.01–77.10</td>
</tr>
<tr>
<td>Ensemble x5</td>
<td>77.22</td>
</tr>
</tbody>
</table>

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.