1. Summary

Abstract

- This paper describes the NICT’s neural machine translation systems for Chinese↔English directions in the CCMT-2019 shared news translation task.
- Our system makes use of techniques that have been proven to be most effective to improve the performance of NMT model, and thereby generates the primary submissions of Chinese↔English translation tasks.

Key points

- Using Marian Toolkit (v1.7.6) to build baseline translation model for Chinese↔English translation tasks.
- Back-translation technique is used to performe data augmentation for providing a large quantity of pseudo parallel data.
- Fine-tuning technique to further optimize the trained NMT model on small development data.
- Ensemble technique focuses on making full use of diverse NMT models to generate better translations.

2. Datasets and Preprocessing

Datasets: pre-processed parallel data (left table) and pre-processed monolingual data (right table)

<table>
<thead>
<tr>
<th>Language pair</th>
<th>#Sentence pairs</th>
<th>#Tokens</th>
<th>Language</th>
<th>#Sentences</th>
<th>#Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese↔English</td>
<td>24.8M</td>
<td>509.9M</td>
<td>English</td>
<td>338.7M</td>
<td>7.5B</td>
</tr>
<tr>
<td></td>
<td>576.2M</td>
<td></td>
<td>Chinese</td>
<td>130.5M</td>
<td>2.3B</td>
</tr>
</tbody>
</table>

Pre-processing steps

- For Chinese and English monolingual data:
  - Empirical selecting the first ten million lines of the News Crawl 2019 English corpus
  - English: tokenizer and truecaser in Moses
  - Chinese: Jieba
  - Filtering out sentences longer than 80 tokens in the training data
  - Replacing characters forbidden by Moses

3. Experiments

Marian Toolkit (v1.7.6) for Chinese↔English

- 50,000 sub-word vocabularies
- Batch sizes of 4096 words
- The number of dimensions of input and output layers was set to 512
- The inner feed-forward neural network layer was set to 2048
- The number of attention heads was set to eight
- Warm-up steps of 16,000
- The label smoothing and attention dropout were set to 0.1

Main results (BLEU-cased)

<table>
<thead>
<tr>
<th>#</th>
<th>System</th>
<th>ZH→EN</th>
<th>EN→ZH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Single Transformer (base) model (w/o back-translation)</td>
<td>23.3</td>
<td>30.3</td>
</tr>
<tr>
<td>2.</td>
<td>Single Transformer (base) model (w/ back-translation)</td>
<td>25.3</td>
<td>31.8</td>
</tr>
<tr>
<td>3.</td>
<td>Single Transformer (base) model (w/ fine-tuning)</td>
<td>27.5</td>
<td>33.1</td>
</tr>
<tr>
<td>4.</td>
<td>ensembling five single models (w/ back-translation and fine-tuning)</td>
<td>31.0</td>
<td>34.5</td>
</tr>
</tbody>
</table>

Observation

- Back-translation obtains improvements by 2.0 BLEU scores (#2) over single Transformer (base) model (w/o back-translation) (#1)
- Large improvements with Single Transformer (base) model (w/ fine-tuning)
- Best results with ensemble five single models (w/ back-translation and fine-tuning)
- This results conformed that these three methods can incrementally improve translation performance.