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
• 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.
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.
• We use **Back Translation** (Sennrich et al., 2016) to augment parallel data.
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.
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.
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.
We use **Ensemble**, which calculates the word-level averaged log-probability among different models during decoding, to combine different models for stronger performance.
• 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 newestst2019 Chinese-English test set.

<table>
<thead>
<tr>
<th>Settings</th>
<th>Transformer Big</th>
<th>Deep Transformer Base</th>
<th>Deep Transformer Big</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>27.94</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+Data Augmentation</td>
<td>28.59</td>
<td>29.74</td>
<td>29.85</td>
</tr>
<tr>
<td>+Finetuning</td>
<td>35.97</td>
<td>37.48</td>
<td>37.74</td>
</tr>
<tr>
<td>Ensemble</td>
<td></td>
<td></td>
<td><strong>38.95</strong></td>
</tr>
</tbody>
</table>

• 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>+BT</th>
<th>+Noise BT</th>
<th>+BT&amp;EKD</th>
<th>+BT&amp;R2LKD</th>
</tr>
</thead>
<tbody>
<tr>
<td>newstest2019</td>
<td>34.85</td>
<td>34.53</td>
<td>35.05</td>
<td>35.65</td>
<td>35.02</td>
</tr>
</tbody>
</table>

• Back translation does not work well.

• Knowledge distillation is more effective.
Experiments – EN->ZH


<table>
<thead>
<tr>
<th>Dataset</th>
<th>Data Augmentation</th>
<th>Deep model</th>
<th>Finetuning</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>newstest2019</td>
<td>36.94</td>
<td>36.91</td>
<td>38.33</td>
<td>39.56</td>
</tr>
</tbody>
</table>

- 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:

  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