Neural System Combination for Machine Translation

Long Zhou, Wenpeng Hu, Jiajun Zhang and Chengqing Zong
{long.zhou, wenpeng.hu, jjzhang, cqzong}@nlpr.ia.ac.cn

Background and Motivation

Neural machine translation (NMT) generates much more fluent results compared to statistical machine translation (SMT). However, SMT is usually better than NMT in translation adequacy. System combination is therefore a promising direction to unify the advantages of both NMT and SMT.

Our solution: **Neural System Combination (NSC)**
- **Step 1:** For a source sentence, generate translations with phrase-based SMT (PBMT), hierarchical phrase-based SMT (HPMT), and NMT.
- **Step 2:** Design a multi-source sequence-to-sequence model that takes as input the three translation results of PBMT, HPMT, and NMT, and produces the final target language translation.

Translation Example:

**Source:** 海珊 也与 恐怖组织建立了联系。
**Pinyin:** hǎishān yě yǔ tǒngbù zǔzhīwǎng jiànlì le liánxì。
**Ref.:** hussein has also established ties with terrorist networks.
**PBMT:** hussein also has established relations and terrorist group.
**HPMT:** hussein also and terrorist group established relations.
**NMT:** hussein also established relations with <UNK>.

Table 1: Translation examples of single system and our model.

Overview of our approach

**Inputs:** translation outputs of PBMT, HPMT and NMT.
**Outputs:** final target language translation results.
**Core idea:** hierarchical attention-based multi-source seq2seq model.

Layer I Attention:

\[ c_{jk} = \sum_{i=1}^{m} \alpha_{jk}^i h_i \quad \alpha_{jk}^i = \frac{\exp(e_{jk}^i)}{\sum_{j=1}^{n} \exp(e_{jk})} \]

where \( e_{jk}^i = v_j^T \tanh(W_j^s \hat{s}_{j-1} + U_a h_i) \) scores how well \( \hat{s}_{j-1} \) and \( h_i \) match.

Layer II Attention:

\[ c_j = \sum_{i=1}^{n} \beta_{j} h_i \quad \beta_j = \frac{\exp(s_{j-1} \cdot c_{jk})}{\sum_{j=1}^{n} \exp(s_{j-1} \cdot c_{jk})} \]

Training Data Simulation:

The neural system combination framework should be trained on the outputs of multiple translation systems and the gold target translations. In order to keep consistency in training and testing, we design a strategy to simulate the real scenario.

![Training data simulation diagram](image)

Experiments

**Data sets:** 2.08M sentence pairs of Chinese-English extracted from LDC corpus.

**Systems:** We compare our neural combination system with the best individual engines, and the state-of-the-art traditional combination system Jane.

<table>
<thead>
<tr>
<th>System</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT06</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBMT</td>
<td>37.47</td>
<td>41.20</td>
<td>36.41</td>
<td>36.03</td>
<td>37.78</td>
</tr>
<tr>
<td>HPMT</td>
<td>38.05</td>
<td>41.47</td>
<td>36.86</td>
<td>36.04</td>
<td>38.10</td>
</tr>
<tr>
<td>NMT</td>
<td>37.91</td>
<td>38.95</td>
<td>36.02</td>
<td>36.65</td>
<td>37.38</td>
</tr>
<tr>
<td>Jane</td>
<td>39.83</td>
<td>42.75</td>
<td>38.63</td>
<td>39.10</td>
<td>40.08</td>
</tr>
<tr>
<td>NSC</td>
<td>40.64</td>
<td>44.81</td>
<td>38.80</td>
<td>38.26</td>
<td>40.63</td>
</tr>
<tr>
<td>NSC+Source</td>
<td>42.16</td>
<td>45.51</td>
<td>40.28</td>
<td>39.02</td>
<td>41.75</td>
</tr>
<tr>
<td>NSC+Ensemble</td>
<td>41.67</td>
<td>45.95</td>
<td>40.37</td>
<td>39.02</td>
<td>41.75</td>
</tr>
<tr>
<td>NSC+Source+Ensemble</td>
<td>43.55</td>
<td>47.09</td>
<td>42.02</td>
<td>41.10</td>
<td>43.44</td>
</tr>
</tbody>
</table>

Table 2: Translation results (BLEU score) for different machine translation and system combination methods. Jane is an open source system combination toolkit that uses confusion network decoding.

Conclusions

- The proposed neural system combination method using hierarchical attentional seq2seq model can substantially improve the translation quality by combining the merits of SMT and NMT.
- Neural system combination architecture is simple and can be applied into other applications, such as summarization and text generation.