Syntax-Enhanced Neural Machine Translation with Syntax-Aware Word Representations (SAWR)

Meishan Zhang (Tianjin University, China), Zhenghua Li, Guohong Fu and Min Zhang (Soochow University, China)
Outlines

- Research progress on parsing (dependency)
- Seq2seq NMT
- Approaches for exploiting source-language syntax
- Experiments
- Conclusions
The goal is to build an acyclic, directed tree.
Recent progress

- The deep learning techniques achieve great success.
  - Transition-based: Chen and Manning, 2014; Dyer et al., 2015; Zhou et al., 2015; Andor et al., 2016;
  - Graph-based: Kiperwasser and Goldberg, 2016; Dozat and Manning, 2017

Labeled Attachment Score (LAS)
- English PTB: 92 (traditional) -> 94 (Biaffine) -> 95 (+Bert)
- Chinese: 78 (traditional) -> 85 (Biaffine) -> 89 (+Bert)
The Biaffine Parser (Dozat and Manning, 2017)

- Multi-layer BiLSTMs for global sentence encoding
- Reasonable design of scoring architecture
- Good settings of hyper-parameters, such as dropouts
CoNLL-2009 Chinese Test Data

- CoNLL best (2009)
- Graph CRF (-2014)
- Transition (-2014)
- Google (2016)
- Standford Biaffine (2017)
- Biaffine +ELMo
- Biaffine +BERT
Motivations

- Inspired by the success of syntax-based SMT, two questions remain for NMT.
  - Can state-of-the-art syntactic parsing help NMT?
  - How? What is the most effective way?
Why utilize Syntax for NMT?

Pros:
- Long-distance dependencies
- Structural correspondences, reordering

Cons:
- Error propagation: imperfect performance of syntactic parsing
  - Especially on open-domain texts
- Annotation inconsistencies
- Non-optimal choices made in annotation guideline
Going-on Works

- **Data annotation** [hlt.suda.edu.cn/index.php/SUCDT](hlt.suda.edu.cn/index.php/SUCDT)
  - 70-page guideline (systematic, scientific)
  - High-quality (strict double annotation -> inconsistency handling)
  - ~100K sentences from different sources and genres
  - Li+ ACL19: data and semi-supervised domain adaptation

- **Data conversion** (Jiang et al., ACL-2018)
  - Utilizing heterogeneous treebanks (~100-200K sentences)

Seq2Seq NMT (Luong+15)
Biaffine Parser & Seq2Seq NMT
NMT w/ SAWR (Syntax-aware Word Representations)

Word repr from parser encoder, free of error propagation

Similar to ELMo
$$s_i = W{o_i} + b$$

$$h = \text{Encoder}(x_1 \oplus s_1, \ldots, x_n \oplus s_n)$$
NMT w/ SAWR (Syntax-aware Word Representations)

Offline-trained Biaffine Parser

Whether to fine-tune parser encoder?
Baseline Approach 1: Tree-GRU (Chen+ ACL17)

Step 1: Tree-GRU on embs
Step 2: Use Tree-GRU word repr as extra inputs

Efficient issue: careful batching
Baseline Approach 2: Tree-Linearization (Li+ ACL17)

教育 is 现代 civilization 的 基石
education is modern civilization of cornerstone

\[ x_1: \text{SH(教育)} \quad x_2: \text{SH(是(is))} \quad x_3: \text{RL(top)} \]
\[ x_4: \text{SH(现代)} \quad x_5: \text{SH(文明)} \quad x_6: \text{RL(amod)} \]
\[ x_7: \text{SH(的)} \quad x_8: \text{RR(assm)} \quad x_9: \text{SH(基石)} \]
\[ x_{10}: \text{RL(assmod)} \quad x_{11}: \text{RR(attr)} \quad x_{12}: \text{PR} \]
Chinese-English
LDC(1.25M, max-length 50) and NIST 02-06
Parser: CTB7, 81.02% (LAS)

English-Vietnamese
IWSLT 2015 (133K, max-length 150)
Parser: Penn Treebank, 93.84% (LAS)

Vocabulary
Source: 50K
Target: 32K BPE

Other details (beam 5, hidden 512+1024, …)
## Experiments (Speed comparison)

### Chinese-English Translation

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>105 min</td>
</tr>
<tr>
<td>SAWR</td>
<td>142 min</td>
</tr>
<tr>
<td>Tree-RNN</td>
<td>498 min</td>
</tr>
<tr>
<td>Tree-Linearization</td>
<td>137 min</td>
</tr>
</tbody>
</table>
## Single-model Results

### Chinese-English Translation

<table>
<thead>
<tr>
<th>System</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT06</th>
<th>Average/Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>36.44</td>
<td>39.35</td>
<td>36.26</td>
<td>36.32</td>
<td>37.09</td>
</tr>
<tr>
<td>SAWR</td>
<td>38.42</td>
<td>40.60</td>
<td>38.27</td>
<td>38.04</td>
<td>38.83/+1.74</td>
</tr>
<tr>
<td>Tree-RNN</td>
<td>38.12</td>
<td>40.35</td>
<td>37.86</td>
<td>37.32</td>
<td>38.41/+1.32</td>
</tr>
<tr>
<td>Tree-Linearization</td>
<td>37.95</td>
<td>40.24</td>
<td>37.64</td>
<td>37.44</td>
<td>38.32/+1.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Previous Work</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al. (2017a)</td>
<td>35.64</td>
<td>36.63</td>
<td>34.35</td>
<td>30.57</td>
<td>34.30/+2.59</td>
</tr>
<tr>
<td>Li et al. (2017)</td>
<td>34.9</td>
<td>38.6</td>
<td>35.5</td>
<td>35.6</td>
<td>36.15/+1.45</td>
</tr>
<tr>
<td>Chen et al. (2017b)</td>
<td><strong>35.91</strong></td>
<td><strong>38.73</strong></td>
<td>34.18</td>
<td>33.76</td>
<td>35.65/+1.52</td>
</tr>
</tbody>
</table>
## English-Vietnamese Translation

<table>
<thead>
<tr>
<th>System</th>
<th>tst 2013 / Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>28.29</td>
</tr>
<tr>
<td><strong>SAWR</strong></td>
<td><strong>29.09/+0.80</strong></td>
</tr>
<tr>
<td>Tree-RNN</td>
<td>28.51/+0.22</td>
</tr>
<tr>
<td>Tree-Linearization</td>
<td>28.93/+0.64</td>
</tr>
</tbody>
</table>
# Ensemble Results

## Chinese-English Translation

<table>
<thead>
<tr>
<th>System</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT06</th>
<th>Average/Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline × 3</td>
<td>40.90</td>
<td>43.25</td>
<td>40.64</td>
<td>40.16</td>
<td>41.24</td>
</tr>
<tr>
<td>SAWR × 3</td>
<td>41.94</td>
<td>44.59</td>
<td>41.91</td>
<td>41.97</td>
<td>42.60/+1.36</td>
</tr>
<tr>
<td>Tree-RNN × 3</td>
<td>42.03</td>
<td>44.15</td>
<td>41.50</td>
<td>41.41</td>
<td>42.27/+1.03</td>
</tr>
<tr>
<td>Tree-Linearization × 3</td>
<td>41.74</td>
<td>44.23</td>
<td>41.32</td>
<td>41.44</td>
<td>42.18/+0.94</td>
</tr>
<tr>
<td>Hybrid</td>
<td>42.72</td>
<td>45.14</td>
<td>42.38</td>
<td>42.15</td>
<td>43.10/+1.86</td>
</tr>
</tbody>
</table>
Whether to fine-tune parser encoder

Chinese-English Translation

<table>
<thead>
<tr>
<th>Parser</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT06</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>no Tune</td>
<td>38.42</td>
<td>40.60</td>
<td>38.27</td>
<td>38.04</td>
<td>38.83</td>
</tr>
<tr>
<td>Tune</td>
<td>37.33</td>
<td>39.45</td>
<td>36.93</td>
<td>37.03</td>
<td>37.69</td>
</tr>
</tbody>
</table>

Translation output

Diagram:
- Encoder
- GRU
- Projection
- Embedding
- Input
Effect of parser performance

Chinese-English Translation

![Graph showing BLEU scores for different parser performance levels.](image)
Effect regarding to source sentence length

Chinese-English Translation

![Graph showing BLEU scores for different sentence length categories]
## Results w/ Transformer as baseline

### Chinese-English Translation

<table>
<thead>
<tr>
<th>System</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT06</th>
<th>Average/Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer</td>
<td>40.45</td>
<td>42.76</td>
<td>40.09</td>
<td>39.67</td>
<td>40.74</td>
</tr>
<tr>
<td>SAWR</td>
<td>41.63</td>
<td>43.60</td>
<td>41.68</td>
<td>40.21</td>
<td>41.78/+1.04</td>
</tr>
<tr>
<td>Tree-RNN</td>
<td>41.24</td>
<td>43.38</td>
<td>41.04</td>
<td>40.02</td>
<td>41.42/+0.68</td>
</tr>
<tr>
<td>Tree-Linearization</td>
<td>41.12</td>
<td>43.02</td>
<td>41.04</td>
<td>39.86</td>
<td>41.26/+0.52</td>
</tr>
</tbody>
</table>
Conclusions

- We propose a simple approach for syntax-enhanced NMT with SAWR, surpassing previous approaches (Tree-GRU & Tree-Linear)
  - Free of error propagation
  - Efficient
- Future directions
  - Use heterogeneous parsing results
  - Exploration of more model architectures
Q/A?
Thanks!
## English-Germany Results

<table>
<thead>
<tr>
<th>System</th>
<th>newstest2014 / Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>19.47</td>
</tr>
<tr>
<td>SAWR</td>
<td>20.04/+0.57</td>
</tr>
<tr>
<td>Tree-RNN</td>
<td>20.20/+0.73</td>
</tr>
<tr>
<td>Tree-Linearization</td>
<td>20.11/+0.64</td>
</tr>
</tbody>
</table>