Improved Neural Machine Translation with a Syntax-Aware Encoder and Decoder (2017)

Huadong Chen, Shujian Huang, David Chiang, Jiajun Chen

Presenter: Ziqiao Ding
Motivation [Shi et al. (2016)]

The current encoder-decoder model relies solely on the implicit structure it learns during training and can suffer from syntactic errors.
Proposal

Incorporate source-side syntactic (binarized) tree with both encoder and decoder
What we have (Before Spring Break) [Bahdanau et al. (2015)]

A Bidirectional Sequential Encoder

\[ \overrightarrow{h}_i = \text{GRU}(\overrightarrow{h}_{i-1}, s_i) \]

A Forward GRU Decoder (With Attention)

\[
P(y_j | y_{<j}, x; \theta) = \text{softmax}(t_{j-1}, d_j, c_j)
\]

\[
\alpha_{j,i} = \frac{\exp(e_{j,i})}{\sum_{i'=1}^{I} \exp(e_{j,i'})}
\]

\[
d_j = \text{GRU}(d_{j-1}, t_{j-1}, c_j)
\]

\[
c_j = \sum_{i=1}^{I} \alpha_{j,i} \overrightarrow{h}_i
\]

\[
e_{j,i} = v_a^T \tanh(W_a d_{j-1} + U_a \overrightarrow{h}_i)
\]

A Bidirectional RNN

Bidirectional RNN
GRU (quick reminder)

Reset Gate (How much info to forget):

\[ r_t = \sigma_g (W_r x_t + U_r h_{t-1} + b_r) \]

Update Gate (How much info to pass on):

\[ z_t = \sigma_g (W_z x_t + U_z h_{t-1} + b_z) \]

Updated Hidden State:

\[ h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \sigma_h (W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \]
Tree-GRU Encoder [Eriguchi et al. (2016)]

Leaf Node: $h^\uparrow_k = h_k$

Internal Node: $h^\uparrow_k = f(h^\uparrow_{L(k)}, h^\uparrow_{R(k)})$

Bottom Up Tree!
Tree-GRU Encoder [Eriguchi et al. (2016)]

$\hat{h}_k^\uparrow = f(h_{L(k)}^\uparrow, h_{R(k)}^\uparrow)$

Children Nodes:

\[
\begin{align*}
 r_L &= \sigma(U_L^{(rL)} h_{L(k)}^\uparrow + U_R^{(rL)} h_{R(k)}^\uparrow + b^{(rL)}) \\
 r_R &= \sigma(U_L^{(rR)} h_{L(k)}^\uparrow + U_R^{(rR)} h_{R(k)}^\uparrow + b^{(rR)}) \\
 z_L &= \sigma(U_L^{(zL)} h_{L(k)}^\uparrow + U_R^{(zL)} h_{R(k)}^\uparrow + b^{(zL)}) \\
 z_R &= \sigma(U_L^{(zR)} h_{L(k)}^\uparrow + U_R^{(zR)} h_{R(k)}^\uparrow + b^{(zR)})
\end{align*}
\]

Current Node:

\[
\begin{align*}
 z &= \sigma(U_L^{(z)} h_{L(k)}^\uparrow + U_R^{(z)} h_{R(k)}^\uparrow + b^{(z)}) \\
 \tilde{h}_k^\uparrow &= \tanh(U_L(r_L \odot h_{L(k)}^\uparrow) + U_R(r_R \odot h_{R(k)}^\uparrow)) \\
 h_k^\uparrow &= z_L \odot \hat{h}_{L(k)}^\uparrow + z_R \odot \hat{h}_{R(k)}^\uparrow + z \odot \tilde{h}_k^\uparrow
\end{align*}
\]
Bidirectional Tree-GRU Encoder (This Paper)

Root:

\[ h_\rho^\downarrow = \tanh (W h_\rho^\uparrow + b) \]

Non-root:

\[ h_k^\downarrow = \text{GRU}(h_{p(k)}^\downarrow, h_k^\uparrow) \]

Add Top Down Layer!
Decoder (Coverage Model) [Tu et al. (2016)]

A Forward GRU Decoder (With Attention)

\[
P(y_j \mid y_{<j}, x; \theta) = \text{softmax}(t_{j-1}, d_j, c_j)
\]

\[
d_j = \text{GRU}(d_{j-1}, t_{j-1}, c_j)
\]

\[
c_j = \sum_{i=1}^{l} \alpha_{j,i} \overrightarrow{h_i}
\]

\[
\alpha_{j,i} = \frac{\exp(e_{j,i})}{\sum_{i'=1}^{l} \exp(e_{j,i'})}
\]

\[
e_{j,i} = v_a^T \tanh(W_d d_{j-1} + U_a \overrightarrow{h_i})
\]

\[
C_{j,i} = \text{GRU}(C_{j-1,i}, \alpha_{j,i}, d_{j-1}, h_i)
\]

\[
e_{j,i} = v_a^T \tanh(W_d d_{j-1} + U_a h_i + V_a C_{j-1,i})
\]
Coverage? What?

More intuitive in SMT:
(Which word is translated already?)

\[ x = \{ x_1, x_2, x_3, x_4 \} \]

\[ C = \{ 0, 0, 0, 0 \} \rightarrow C = \{ 0, 1, 1, 0 \} \rightarrow C = \{ 1, 1, 1, 1 \} \]

Initial \hspace{4cm} Goal!

But harder to model single word in NN, so we model the attention:

\[ C_{j,i} = \text{GRU}(C_{j-1,i}, \alpha_{j,i}, d_{j-1}, h_i) \]

Hidden State! Attention Weight, decoder state, encoder state

Attention adjusted according to coverage:

\[ e_{j,i} = v_a^T \tanh (W_{ad_j-1} + U_a h_i + V_a C_{j-1,i}) \]
Tree Coverage Decoder (This paper)

\[ C_{j,i} = \text{GRU}(C_{j-1,i}, \alpha_{j,i}, d_{j-1}, h_i) \]

\[ C_{j,i} = \text{GRU}(C_{j-1,i}, \alpha_{j,i}, d_{j-1}, h_i, \]
\[ C_{j-1,L(i), \alpha_{j,L(i)}, \]
\[ C_{j-1,R(i), \alpha_{j,R(i)}}. \]

\[ e_{j,i} = \nu_{a}^{T} \tanh (W_{ad}d_{j-1} + U_{ah}h_{i} + V_{a}C_{j-1,i}) \]

Tree Coverage Vector:
And there’s a point in using Coverage Model? Yes...
Note: Extending GRU parameters

\[ C_{j,i} = \text{GRU}(C_{j-1,i}, \alpha_{j,i}, d_{j-1}, h_i, C_{j-1,L(i)}, \alpha_{j,L(i)}, C_{j-1,R(i)}, \alpha_{j,R(i)}). \]
# Performance (Encoder Only)

<table>
<thead>
<tr>
<th>#</th>
<th>Encoder</th>
<th>Coverage</th>
<th>MT02</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT06</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sequential</td>
<td>no</td>
<td>33.76</td>
<td>31.88</td>
<td>33.15</td>
<td>30.55</td>
<td>27.47</td>
<td>30.76</td>
</tr>
<tr>
<td>2</td>
<td>Tree-LSTM</td>
<td>no</td>
<td>33.83</td>
<td>33.15</td>
<td>33.81</td>
<td>31.22</td>
<td>27.86</td>
<td>31.51(+0.75)</td>
</tr>
<tr>
<td>3</td>
<td>Tree-GRU</td>
<td>no</td>
<td>35.39</td>
<td>33.62</td>
<td>35.1</td>
<td>32.55</td>
<td>28.26</td>
<td>32.38(+1.62)</td>
</tr>
<tr>
<td>4</td>
<td>Bidirectional</td>
<td>no</td>
<td>35.52</td>
<td>33.91</td>
<td>35.51</td>
<td>33.34</td>
<td>29.91</td>
<td><strong>33.17(+2.41)</strong></td>
</tr>
</tbody>
</table>

**BLEU scores** against:

- **NMT**: the standard attentional NMT model (Bahdanau et al., 2015).
- **Tree-LSTM**: the attentional NMT model extended with the Tree-LSTM encoder (Eriguchi et al., 2016).
## Performance (w. Word / Tree Coverage)

<table>
<thead>
<tr>
<th>#</th>
<th>Encoder</th>
<th>Coverage</th>
<th>MT02</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT06</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Sequential</td>
<td>word</td>
<td>34.21</td>
<td>32.73</td>
<td>34.17</td>
<td>31.64</td>
<td>28.29</td>
<td>31.71(+0.95)</td>
</tr>
<tr>
<td>6</td>
<td>Tree-LSTM</td>
<td>word</td>
<td>35.81</td>
<td>33.62</td>
<td>34.84</td>
<td>32.6</td>
<td>28.52</td>
<td>32.40(+1.64)</td>
</tr>
<tr>
<td>7</td>
<td>Tree-GRU</td>
<td>word</td>
<td>35.91</td>
<td>33.71</td>
<td>35.46</td>
<td>33.02</td>
<td>29.14</td>
<td>32.84(+2.08)</td>
</tr>
<tr>
<td>8</td>
<td>Bidirectional</td>
<td>word</td>
<td>36.14</td>
<td>35.00</td>
<td>36.07</td>
<td>33.74</td>
<td>30.40</td>
<td><strong>33.80(+3.04)</strong></td>
</tr>
<tr>
<td>9</td>
<td>Tree-LSTM</td>
<td>tree</td>
<td>34.97</td>
<td>33.91</td>
<td>35.21</td>
<td>33.08</td>
<td>29.38</td>
<td>32.90(+2.14)</td>
</tr>
<tr>
<td>10</td>
<td>Tree-GRU</td>
<td>tree</td>
<td>35.67</td>
<td>34.25</td>
<td>35.72</td>
<td>33.47</td>
<td>29.95</td>
<td>33.35(+2.59)</td>
</tr>
<tr>
<td>11</td>
<td>Bidirectional</td>
<td>tree</td>
<td><strong>36.57</strong></td>
<td><strong>35.64</strong></td>
<td><strong>36.63</strong></td>
<td><strong>34.35</strong></td>
<td><strong>30.57</strong></td>
<td><strong>34.30(+3.54)</strong></td>
</tr>
</tbody>
</table>

- **Coverage**: the attentional NMT model extended with word coverage (Tu et al., 2016).
Performance against sentence length
Potential Further Improvements

- More training complexity analysis

- Make use of node labels from the tree

- Using syntactic information on the target side