The trees are obtained from the attention weights of the self-attentive
proposed approach:
- The residual connections facilitate the flow of information.
- The self-attention allows selective use of previously predicted words.

**Other Self-Attentive Networks**

**Memory RNN**
RNN with memory cells of previous representations
[Cheng et al., EMNLP 2016]

**Self-Attentive RNN**
RNN with a summary vector from past predictions
[Daniluk et al., ICLR 2016]

- The baseline NMT decoder uses a residual connection to the previously
  predicted word \(y_{t-1}\).
- We propose to use residual connections from all previously translated
  words \(y_1, \ldots, y_{t-1}\) with a summary vector \(d_t\).

**Experimental Setup**

**Datasets**: En-ZH UN Corpus 0.5M, Es-En WMT 2.1M, En-De WMT 4.5M

**Architecture**: Attention-based NMT with GRUs of dimension 1024, 500 for
word embeddings, and vocabulary of 50K.

**Results**

<table>
<thead>
<tr>
<th>Models</th>
<th>(\Theta)</th>
<th>En–Zh</th>
<th>En–En</th>
<th>En–De</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMT baseline</td>
<td>–</td>
<td>21.6</td>
<td>25.2</td>
<td>23.2</td>
</tr>
<tr>
<td>NMT transformer (comparable model)</td>
<td>109.0M</td>
<td>22.0</td>
<td>25.9</td>
<td>24.1</td>
</tr>
<tr>
<td>NMT baseline</td>
<td>108.7M</td>
<td>22.6</td>
<td>25.4</td>
<td>24.8</td>
</tr>
<tr>
<td>+ Memory RNN</td>
<td>109.7M</td>
<td>22.5</td>
<td>25.3</td>
<td>24.9</td>
</tr>
<tr>
<td>+ Self-attentive RNN</td>
<td>110.2M</td>
<td>22.0</td>
<td>25.1</td>
<td>24.3</td>
</tr>
<tr>
<td>+ Mean residual connections</td>
<td>108.7M</td>
<td>23.6</td>
<td>26.7</td>
<td>24.9</td>
</tr>
<tr>
<td>+ Self-attentive residual connections</td>
<td>108.9M</td>
<td>24.0</td>
<td>26.3</td>
<td>25.5</td>
</tr>
</tbody>
</table>

**Conclusion**

- We proposed self-attentive residual learning framework.
- Improvements over a standard baseline, and two variants of
  self-attention.
- Analysis of the attention shows syntactic-like structures.
- It can be applied to other tasks based on RNNs.

**Acknowledgements**

Supported by the European Union Horizon 2020 SUMMA project (grant 688139).
www.summa-project.eu