Document-Level Neural Machine Translation with Hierarchical Attention Networks
Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, James Henderson

Motivation

Why document-level NMT?
• Not considering the document context and discourse connections affects coherence and cohesion of a text.

Why hierarchical attention networks?
• Different abstraction levels: word-level and sentence-level.
• Allows dynamic access to the context for each predicted word.

Other advantages in our approach
• Joint optimization of multiple sentences.
• Shared hidden representations across sentence translations.
• Exploiting source and target context.
• Multi-head attention to capture different discourse phenomena.

Document-level NMT

\[
\max_{t} \sum_{n=1}^{N} \log(P_o(t^n|s^n)) \quad \rightarrow \quad \max_{t} \sum_{n=1}^{N} \log(P_o(t^n|s^n, D_s, D_t))
\]

Baseline NMT

- \( s^n \): source sentence, and \( D_s = (s^{n-k}, ..., s^{n-1}) \): source context
- \( t^n \): target sentence, and \( D_t = (t^{n-k}, ..., t^{n-1}) \): target context.
- Context \((k \) previous sentences) is modeled by HANs:

Hierarchical Attention Network (HAN)

Word-level attention:
\[
\hat{h}^t = \text{MultiHead}(q_{st}, h^t) \quad q_{st} = f_w(h^t)
\]

Sentence-level attention:
\[
d_t = \text{FFN}(\text{MultiHead}(q_{st}, \hat{h}^t)) \quad q_{st} = f_h(h^t)
\]

Multi-head attention

- “juego” can be translated as “game” or “set”
- “su” can be translated as “his”, “her”, or “its”

Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>TED Talks</th>
<th>Subtitles</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td>zh–en</td>
<td>35.44</td>
<td>35.20</td>
<td>21.36</td>
</tr>
<tr>
<td>es–en</td>
<td>34.66</td>
<td>35.49</td>
<td>22.36</td>
</tr>
<tr>
<td>zh–en</td>
<td>37.01</td>
<td>35.96</td>
<td>22.36</td>
</tr>
<tr>
<td>es–en</td>
<td>36.23</td>
<td>36.23</td>
<td>22.76</td>
</tr>
<tr>
<td>zh–en</td>
<td>36.46</td>
<td>36.46</td>
<td>22.76</td>
</tr>
<tr>
<td>es–en</td>
<td>36.59</td>
<td>36.59</td>
<td>22.80</td>
</tr>
<tr>
<td>zh–en</td>
<td>36.91</td>
<td>36.91</td>
<td>22.80</td>
</tr>
<tr>
<td>es–en</td>
<td>37.24</td>
<td>37.24</td>
<td>22.80</td>
</tr>
</tbody>
</table>

Discourse Evaluation

- HAN decoder helps in lexical cohesion and coherence.
- HAN encoder helps in pronoun and noun disambiguation.

Conclusion

- We proposed a hierarchical multi-head attention model for document-level context.
- It directly connects representations from previous sentence translations into the current sentence translation.
- It significantly outperforms two competitive baselines.
- It improves cohesion and coherence, and noun/pronoun translation.
- We show that target and source context is complementary.

Our multi-head HAN could be used to model context in other NLP tasks. Code available at https://github.com/idiap/HAN_NMT

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