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

\[ \max \sum_{n=1}^{N} \log(P(t^n|s^n)) \quad \rightarrow \quad \max \sum_{n=1}^{N} \log(P(t^n|s^n, D_{\phi}, D_{\gamma})) \]

Baseline NMT

Document-level NMT

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

Hierarchical Attention Network (HAN)

Word-level attention:
\[ h^j = \text{MultiHead}(q_{ot}, h^j) \quad q_{ot} = f_q(h_i) \]

Sentence-level attention:
\[ d_e = \text{FFN}(\text{MultiHead}(q_{st}, h^j)) \quad q_{st} = f_s(h_i) \]

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>TED Talks</th>
<th>Subtitles</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zh–En</td>
<td>Es–En</td>
<td>Zh–En</td>
</tr>
<tr>
<td>NMT transformer</td>
<td>16.87</td>
<td>35.44</td>
</tr>
<tr>
<td>+ cache</td>
<td>17.32***</td>
<td>36.46***</td>
</tr>
<tr>
<td>+ HAN encoder</td>
<td>17.81***</td>
<td>36.91***</td>
</tr>
<tr>
<td>+ HAN decoder</td>
<td>17.39***</td>
<td>37.03***</td>
</tr>
<tr>
<td>+ HAN joint</td>
<td>17.79***</td>
<td>37.24***</td>
</tr>
</tbody>
</table>

BLEU scores. Significance with respect to NMT \(*\), and to cache model \(*\). P-values: \( \dagger \ < .05, \dagger\dagger \ < .01, \dagger\dagger\dagger \ < .001.\)

- **Significant improvement over strong baselines on multiple data sets.**
- **Context from source and target sides are complementary.**

Discourse Evaluation

<table>
<thead>
<tr>
<th>Coherence</th>
<th>Lexical Cohesion</th>
<th>Pronouns</th>
<th>Nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT transformer</td>
<td>28.42</td>
<td>47.98</td>
<td>62.84</td>
</tr>
<tr>
<td>+ HAN encoder</td>
<td>28.60</td>
<td>48.35</td>
<td>64.48</td>
</tr>
<tr>
<td>+ HAN decoder</td>
<td>28.78</td>
<td>48.51</td>
<td>64.04</td>
</tr>
<tr>
<td>+ HAN joint</td>
<td>28.82</td>
<td>48.61</td>
<td>64.32</td>
</tr>
<tr>
<td>Human reference</td>
<td>29.79</td>
<td>52.94</td>
<td>100.0</td>
</tr>
</tbody>
</table>

- **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.

---

Reference
Code available at https://github.com/idiap/HAN_NMT