When and Why is Document-level Context Useful in Neural Machine Translation?

Yunsu Kim, Duc Thanh Tran, Hermann Ney
RWTH Aachen University, Aachen, Germany
kim@cs.rwth-aachen.de

Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)
November 3, 2019
Document-level NMT Hype

Dozens of methods for integrating document-level context in NMT:

• Jean & Lauly+ 17 [Tiedemann & Scherrer 17] [Tu & Liu+ 18] [Cao & Xiong 18]
  • [Zhang & Luan+ 18] [Voita & Serdyukov+ 18] [Maruf & Haffari 18] [Kuang & Xiong+ 18b]
  • [Kuang & Xiong 18a] [Bawden & Sennrich+ 18] [Miculicich & Ram+ 18] [Stojanovski & Fraser 18]
  • [Voita & Sennrich+ 19] [Maruf & Martins+ 19] [Junczys-Dowmunt 19]

• 14 out of 15 papers: trained **only with (restricted) document-level data**
  ▶ TED, News commentary, Europarl, OpenSubtitles, ...
  ▶ For both sentence-level and document-level training

• In average: $\sim+1.0$ **BLEU [%]** over a sentence-level baseline

**Question:** Where do all improvements come from? Coreference? Coherence?
Causes of Improvements

How often is document-level context utilized for coreference/coherence?

<table>
<thead>
<tr>
<th></th>
<th>#sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>en-it</td>
</tr>
<tr>
<td>Total</td>
<td>1,147</td>
</tr>
<tr>
<td>Total ( \text{TER} ) improved</td>
<td>379</td>
</tr>
<tr>
<td>Coreference</td>
<td>21</td>
</tr>
<tr>
<td>Topic-aware lexical choice</td>
<td>66</td>
</tr>
<tr>
<td>Others</td>
<td>292</td>
</tr>
</tbody>
</table>

- **92.5%** of improvements ("Others"): irrelevant to document-level context
- Our hypothesis: context causes the regularization effect
Stronger Sentence-level NMT

Document-level training on top of **well-regularized sentence-level models**

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Dropout</th>
<th>System</th>
<th><strong>BLEU [%]</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>en-it</strong></td>
</tr>
<tr>
<td>Small</td>
<td>0.1</td>
<td>Sentence-level</td>
<td>31.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Document-level</td>
<td>32.5</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>Sentence-level</td>
<td>33.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Document-level</td>
<td>33.5</td>
</tr>
<tr>
<td>Large</td>
<td>0.1</td>
<td>Sentence-level</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Document-level</td>
<td>-</td>
</tr>
</tbody>
</table>

- No improvements in **BLEU** over strong sentence-level systems
- Targeted test sets: improvements might not carry over to real scenarios
Remarks

1. Do **not** sell the improvements in **BLEU** by document-level context
2. Check the **real** document-level improvements **manually**
3. Build the sentence-level system **as strong as possible** first

Come to our poster to also check:

- How much **modeling power** is needed for context encoding?
- Is the **entire context** sentence needed?
- Is a **long-range context** useful?

When and Why is Document-level Context Useful in Neural Machine Translation?

Yunsu Kim, Duc Thanh Tran, Hermann Ney
References


[Cao & Xiong 18] Q. Cao, D. Xiong.
Encoding gated translation memory into neural machine translation.

Does neural machine translation benefit from larger context?


Fusing recency into neural machine translation with an inter-sentence gate model.

Modeling coherence for neural machine translation with dynamic and topic caches.

References

Selective attention for context-aware neural machine translation.


[Stojanovski & Fraser 18] D. Stojanovski, A. Fraser.
Coreference and coherence in neural machine translation: A study using oracle experiments.

Neural machine translation with extended context.

Learning to remember translation history with a continuous cache.

When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion.

Context-aware neural machine translation learns anaphora resolution.

Improving the transformer translation model with document-level context.