Context-aware Decoder for Neural Machine Translation Using a Target-side Document-Level Language Model

Amane Sugiyama
The University of Tokyo

Naoki Yoshinaga
Institute of Industrial Science
The University of Tokyo
Document-level MT and the Standard Approach

Directly **optimize using document-level parallel data**

![Diagram showing end-to-end neural network for document-level MT](image)

\[ \hat{y} = \arg \max_y p(y|x, c(x), c(y)) \]

**Problem: Lack of document-level parallel data**

Most of existing parallel data are built from only reliable sentence alignments in parallel/comparable documents.

*Can we perform document-level translation without using document-level parallel data?*
Decoding with a Document-level Language Model

Approximate the objective function by sentence-level translation model, document-level language model, and sentence-level language model scores.

\[
\hat{y} = \arg\max_y \log p(y|x, c^{(y)}) \approx \arg\max_y \left[ \log p(y|x) + \log p(y|c^{(y)}) - \log p(y) \right]
\]

C-Score

• Document-level parallel is not required for training
• \( PMI(c^{(y)}, y) = \log p(y|c^{(y)}) - \log p(y) \): association between \( y \) and \( c^{(y)} \)
Decoding Strategy

**Reranking with C-Score (§ 2.2.1)**

Generate n-best hypotheses by sentence-level decoding and select the one that maximizes C-Score

**Context-aware Beam Search (§ 2.2.2)**

Decompose C-Score into token-wise C-Score and perform beam search

\[
C\text{-Score}(y; x, c(y)) = \sum_i \left[ \log p_{S-TM}(y_i|x, y_{<i}) + \log p_{D-LM}(y_i|c(y), y_{<i}) - \log p_{S-LM}(y_i|y_{<i}) \right]
\]

(φ) 君を探しいたよ

Encoder → Decoder

log \( p(y_i|x, y_{<i}) \)

log \( p(y|c(y), y_{<i}) - \log p(y|y_{<i}) \) (PMI)
Experiments

Overall translation performance measured by BLEU score

<table>
<thead>
<tr>
<th>Model</th>
<th>para only</th>
<th>+30M mono</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer w/ BT</td>
<td>32.36</td>
<td>32.40</td>
</tr>
<tr>
<td>DocTransformer</td>
<td>32.50</td>
<td>31.59</td>
</tr>
<tr>
<td>DocRepair</td>
<td>n/a</td>
<td>32.35</td>
</tr>
<tr>
<td>Bayes DocReranker</td>
<td>n/a</td>
<td>33.75**</td>
</tr>
<tr>
<td>w/o context</td>
<td>n/a</td>
<td>33.67**</td>
</tr>
<tr>
<td>Ours (Context-aware beam search)</td>
<td>n/a</td>
<td>32.27</td>
</tr>
<tr>
<td>Ours (Reranking with C-Score)</td>
<td>n/a</td>
<td>32.93*</td>
</tr>
</tbody>
</table>

• Bayes DocReranker and ours (rerank) achieved significant improvements the baseline
• Bayes DocReranker performed almost as well without context.

Evaluation of the ability to capture context [Voita+ 2019]

<table>
<thead>
<tr>
<th>Model</th>
<th>deixis</th>
<th>lex.c</th>
<th>ell.infl</th>
<th>ell vp</th>
</tr>
</thead>
<tbody>
<tr>
<td>DocTransformer</td>
<td>50.0</td>
<td>45.9</td>
<td>56.0</td>
<td>57.2</td>
</tr>
<tr>
<td>DocRepair</td>
<td>89.1</td>
<td>75.8</td>
<td>82.2</td>
<td>67.2</td>
</tr>
<tr>
<td>Bayes DocReranker</td>
<td>65.2</td>
<td>72.2</td>
<td>59.6</td>
<td>44.6</td>
</tr>
<tr>
<td>C-Score (ours)</td>
<td>86.9</td>
<td>94.9</td>
<td>78.2</td>
<td>77.0</td>
</tr>
<tr>
<td>PMI</td>
<td>96.8</td>
<td>97.8</td>
<td>75.8</td>
<td>90.6</td>
</tr>
</tbody>
</table>

C-Score achieves higher scores than DocRepair in two test sets
Conclusion

• We proposed an approach to document-level MT, trainable without document-level parallel data

• We confirmed the effectiveness of our methods in terms of BLEU and the contrastive test
Appendix
BLEU vs #context sents.