BIT’s system for AutoSimTrans 2021

Beijing Institute of Technology, Beijing, China

Mengge Liu, Shuoying Chen, Minqin Li, Zhipeng Wang and Yuhang Guo

Reporter: Mengge Liu

2021.6.10
Introduction & Motivation

• Simultaneous machine translation is the task of generating partial translations before observing the entire source sentence, which is useful in simultaneous interpretation and dynamic subtitles. Here are two basic methods:

• Full-sentence translation (MT):
  • Good quality, unacceptable delay.

• Incomplete-sentence translation:
  • Begin translation before receiving the full sentence.
  • Good latency, may cause bad performance.
  • Use Sub-sentence as translation unit.
    • Sub-sentence is grammatically correct and semantically complete.
    • Better latency with acceptable quality.
Introduction & Motivation

• It is feasible to cut sentence into sub-sentences in translation, when:
  • Each sub-sentence is grammatically correct and semantically complete.
  • There is no cross alignment between sub-sentences.
  • A simple example:

<table>
<thead>
<tr>
<th>Source sentence</th>
<th>各位</th>
<th>亲爱</th>
<th>的</th>
<th>朋友</th>
<th>们</th>
<th>,</th>
<th>早上好</th>
<th>!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation by word</td>
<td>Everybody</td>
<td>dear</td>
<td>friend</td>
<td>s</td>
<td>,</td>
<td>good morning</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>Target sentence</td>
<td>Ladies and gentlemen,</td>
<td>dear</td>
<td>friend</td>
<td>s</td>
<td>,</td>
<td>good morning</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>
System Architecture

- Sentence boundary detector:
  - Read the streaming input text.
  - Detect sentence boundaries and cut segments.
  - Pass the segments to translator.

- Translation module:
  - Translate the segments as sub-sentences.
  - Concatenate translation results to form full translation.
Streaming input Chinese sentence:

各
各位
各位亲爱
各位亲爱的
各位亲爱的朋
各位亲爱的朋友
各位亲爱的朋友们
各位亲爱的朋友们，
各位亲爱的朋友们，早
各位亲爱的朋友们，早上好
各位亲爱的朋友们，早上好！

Boundary is detected at this step.

Output English translation:
ladies and gentlemen, dear friends

Read by detector
Sentence boundary detector
Send sub-sentence to translator
Transformer Machine Translation Module

Back to reading and translation.
Model Training

• Sentence boundary detector:
  • Text classification.
  • Use BERT to predict sentence boundaries in Chinese text.

• Translation module:
  • Pre-training and fine-tuning transformer translator.
  • Pre-trained on large parallel corpus.
  • Domain fine-tuning: adaptation for speech domain.
  • Sub-sentence fine-tuning: adaptation for shorter sub-sentences.

(Short of sub-sentence corpus)
Sub-sentence corpus

• Parallel sub-sentence corpus:
  • Cut a Zh-En sentence pair into several sub-sentence pairs.
  • Basic rule: no cross alignment between sub-sentences.

• Use word-alignment tool for cutting:
  • Obtain alignment informations of sentence pairs.
    • $A = \{< x_i, y_j > | x_i \in X, y_j \in Y \}$
  • Cut sentence pairs based on alignment matrix.
    • No cross alignment: matrix blocks in the bottom left and top right corners are all zeroes.
<table>
<thead>
<tr>
<th>各位</th>
<th>亲爱</th>
<th>的</th>
<th>朋友</th>
<th>们</th>
<th>,</th>
<th>早上好</th>
<th>!</th>
</tr>
</thead>
<tbody>
<tr>
<td>ladies</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>!</td>
</tr>
<tr>
<td>and</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gentlemen</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dear</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>friends</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>good</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>morning</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>!</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments

• Settings
  • Boundary detector: BERT, *chinese_L-12_H-768_A-12*.
  • Translator: Tensor2tensor, *transformer-big*.

• Datasets
  • Pre-training:
    • CWMT19.
  • Fine-tuning:
    • Domain fine-tuning: parallel text of BSTC(给出全称) corpus.
    • Sub-sentence fine-tuning: the sub-sentences pairs constructed from BSTC corpus.
Results

- The pre-trained model is domain fine-tuned firstly and sub-sentence fine-tuned secondly.
- The results show negative influence of sub-sentence fine-tuning.
- Filtering good-quality sub-sentence pairs makes some positive influence.
- (After fix some mistake, we see BLEU 19.60 for sub-sentence fine-tuned model.)
Thanks