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 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 sample example, first sub-sentence is in red and the second one is in black.

Source sentence	各位	亲爱	的	朋友	们	,	早上好	!
Translation by word	Everybody	dear		friend	s	,	good morning	.
Target sentence	Ladies and gentlemen,	dear		friend	s	,	good morning	.

System Architecture

- Sentence boundary detector:
  - Read the streaming input text.
  - Detect sentence boundarys and cut segments.
  - Pass the segments to translator.
- Tranlation module:
  - Translate segments as sub-sentences.
  - Concatenate translation results to form full translation.

Streaming input Chinese sentence:

各
各位
各位亲
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
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.

Read by detector

Sentence boundary detector

各位亲爱的朋友们

Send sub-sentence to translator

Transformer Machine Translation Module

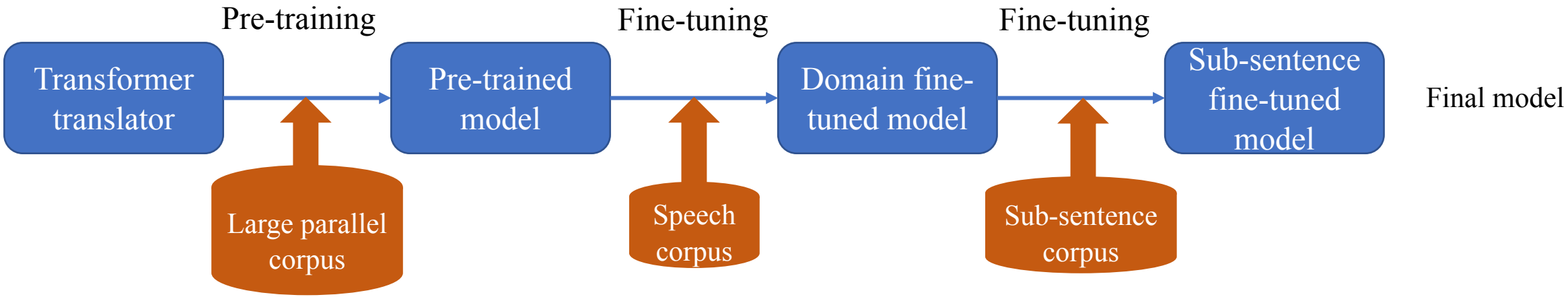
Ouput English translation:

ladies and gentlemen, dear friends

Model Training

- Sentence boundary detector:
  - Text classification problem.
  - Use BERT to predict punctuation marks in Chinese text.
- Tranlation 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 Zh-En sentence pair into 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.

	各位	亲爱	的	朋友	们	，	早上好	！
ladies		1						
and			1					
gentlemen	1	1						
,			1					
dear		1						
friend				1	1			
,					1	1		
good							1	
morning							1	
!								1

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 corpus.
    - Sub-sentence fine-tuning: sub-sentences pairs constructed from BSTC corpus.

Results

Model	AL	BLEU
domain fine-tuned	7.467	19.45
sub-sentence fine-tuned(golden+ASR)	7.478	19.02
sub-sentence fine-tuned(golden)	7.823	16.28
sub-sentence fine-tuned(filtered golden)	7.795	16.67

- 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