USST's System for AutoSimTrans 2022

Jiahui Zhu [1], **Jun Yu** [2]
[1] University of Shanghai for science and technology
[2] East China University of Science and Technology

Outline

- Tracks
- Data and preprocessing
- Text-to-text system
- Conclusion

Tracks

- We participated in two streaming transcription track:
  - Zh→En Translation. (text-to-text)
  - En→Es Translation. (text-to-text)

Streaming transcription	Translation
我	
我下	I
我下面	
我下面来	’m
我下面来讲	
我下面来讲我	going
我下面来讲我们	
我下面来讲我们这	to
我下面来讲我们这段	talk
我下面来讲我们这段故	
我下面来讲我们这段故事	about
我下面来讲我们这段故事。	this story.

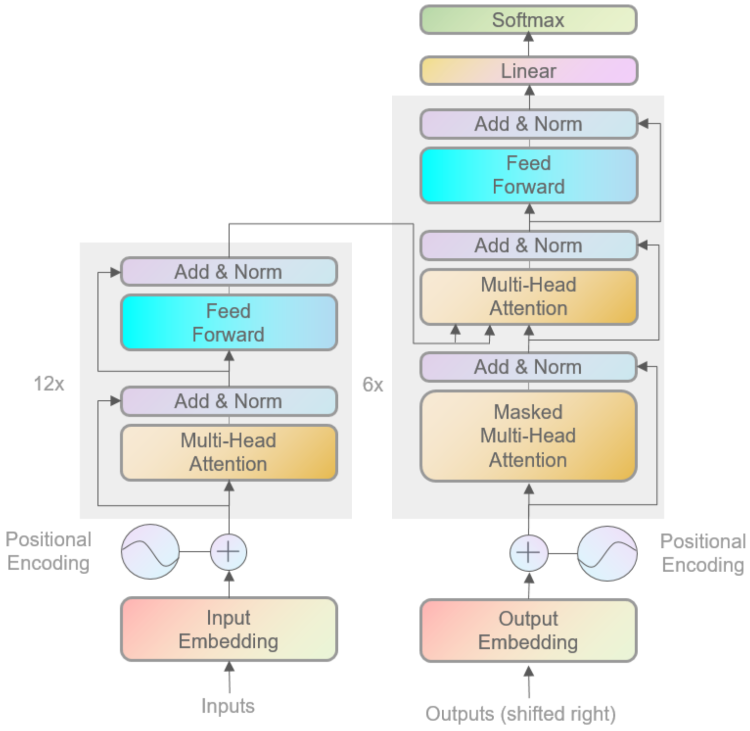
Data and preprocessing

- Datasets
  - Zh→En: We pretrained our MT model on CWMT21 (9.1M) and fine-tined on Baidu Speech Translation Corpus (39K).
  - En→Es: We trained our MT model on UN Parallel Coupus (21M).
- Preprocessing
  - Word Segmentation.
  - Length filter.
  - Langage identification.
  - Deduplication.
  - Byte-pair-encoding.

Text-to-text system

- 1. For abtaining rich source representations, we train a Transformer base with 12 layer encoder.

- 2. For stabilize training , we initialize the model with the method mentioned in DeepNet.




```
def deepnorm(x):
    return LayerNorm(x * α + f(x))

def deepnorm_init(w):
    if w is ['ffn', 'v_proj', 'out_proj']:
        nn.init.xavier_normal_(w, gain=β)
    elif w is ['q_proj', 'k_proj']:
        nn.init.xavier_normal_(w, gain=1)
```

Architectures	Encoder α	Encoder β	Decoder α	Decoder β
Encoder-only (e.g., BERT)	$(2N)^{\frac{1}{4}}$	$(8N)^{-\frac{1}{4}}$	-	-
Decoder-only (e.g., GPT)	-	-	$(2M)^{\frac{1}{4}}$	$(8M)^{-\frac{1}{4}}$
Encoder-decoder (e.g., NMT, T5)	$0.81(N^4M)^{\frac{1}{16}}$	$0.87(N^4M)^{-\frac{1}{16}}$	$(3M)^{\frac{1}{4}}$	$(12M)^{-\frac{1}{4}}$

Figure 2: (a) Pseudocode for DEEPNORM. We take Xavier initialization (Glorot and Bengio, 2010) as an example, and it can be replaced with other standard initialization. Notice that α is a constant. (b) Parameters of DEEPNORM for different architectures (N-layer encoder, M-layer decoder).

Text-to-text system

- 3.For domain adaption, we use in-domain data filtering to mine data from CWMT that is most similar to BSTC, and then mix it with BSTC for fine-tuning.

$$Score(p)_{abs} = |P_I(p) - P_N(p)| \qquad (5)$$

- 4. To reduce the latency, we adopt fixed read/write policy wait-k.

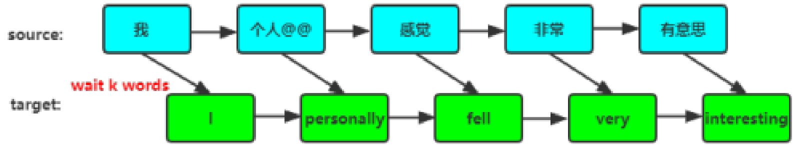


Figure 1: An example of prefix-to-prefix (wait 1).

Conclusion

- 1. This paper describes our text-to-text simultaneous translation system, which uses a deep Transformer to improve translation quality and adopt wait-k policy to reduce latency.

- 3. We plan to research on some dynamic read-write policies in order to better balance quality and latency.

Thank you !