System Description on Automatic Simultaneous Translation Workshop

Linjie Chen, Jianzong Wang,[*] Zhangcheng Huang, Xiongbin Ding, Jing Xiao
Ping An Technology (Shenzhen) Co., Ltd.
chenlinjie887@pingan.com.cn,
jzwang@188.com,
huangzhangcheng624@pingan.com.cn,
dingxiongbin106@pingan.com.cn,
xiaojing661@pingan.com.cn

Abstract

This paper shows our submission on the second automatic simultaneous translation workshop at NAACL2021. We participate in all the two directions of Chinese-to-English translation, Chinese audio→English text and Chinese text→English text. We do data filtering and model training techniques to get the best BLEU score and reduce the average lagging. We propose a two-stage simultaneous translation pipeline system which is composed of Quartznet and BPE-based transformer. We propose a competitive simultaneous translation system and achieves a BLEU score of 24.39 in the audio input track.

1 Introduction

Our submitted system consists of an end to end speech recognition model and a neural machine translation model which follows the traditional pipeline framework in simultaneous translation task. The system input is Chinese audio file and the output is English translation text. A temporary Streaming transcription is obtained by speech recognition model and transmitted into machine translation model to get the target system output.

For automatic speech recognition(ASR) model, we use the QuartzNet model (Kriman et al., 2019) of Nvidia Jarvis. At the moment, we expand the train data set by adding Aishell-1 and data that collected, then using plenty of rules to filter audio data and deal with parallel transcription. Compared to the Jasper model (Li et al., 2019), it can reduce number of parameters quickly by using separable 1D convolutions including time channel.

Our neural machine translation model is Transformer (Vaswani et al., 2017). We use some human rules and the pre-trained language model to filter the parallel corpus. The method of back translation (Sennrich et al., 2016) is also applied to generate synthetic Chinese sentences.

At the step of inference, we apply the wait-k words method (Ma et al., 2018). Both the preprocessing and post-processing are applied to improve the terminology translation and deal with the word error produced by the ASR system.

Since our submission is a two-stage system, the rest of this paper describes separately regards to the Automatic speech recognition(ASR and Machine translation(MT) sub-modules. We firstly describe the training and development datasets we used, then the data filtering methods we applied is introduced. Secondly, the system architecture is discussed and it is verified by the experiments. Lastly, we draw a conclusion of our system by analyzing the experiments.

2 Datasets

For audio data of ASR, we use qianyan audio datasets provided by NAACL workshop (Zhang et al., 2021), Aishell-1 (Hui Bu, 2017) and lip sentences we collect by smartphone(16kHz, 16-bit).

2.1 Audio Data

We invite 20 volunteers in data collection. Each volunteer performed two hours of Mandarin Chinese audio about 1000 sentences in the quiet room. To keep data diversity, different domains of audios were collected, including artificial intelligent, industrial production, business conversation and medical. Finally, we get a total of 19800 sentences (audio and transcription) in this way.

For qianyan audio datasets, we split each audio into sentences according to the sentence-level transcription. After processing, the blank part of all entire audio files was removed, and duration time of audios was reduced from the original 68 hours to about 52 hours.

For AIshell-1 datasets, we firstly deal with transcription files by using rules to get path and filename of every transcription. Then using wave library to read audio files to get the duration time of

Proceedings of the 2nd Workshop on Automatic Simultaneous Translation, pages 24–27
June 10, 2021. ©2021 Association for Computational Linguistics

each audio.

Noting that the audio data and the transcription may not exactly match. In order to improve the accuracy of the data, we use a pre-trained ASR model to transcript audio data to produce text result. Then using similarity matching algorithm to filter audio and original transcription data of lower similarity. Table1 shows the number of train data after filtering.

Table 1: ZH-EN audio train datasets

Data Source	Duration	Total Samples
Qianyan(NAACL)	70hours	36,140
Aishell-1	178hours	120,098
Collection	40hours	19,800

2.2 Text Data

The corpus we use to build our machine translation system is CWMT 19 corpus [1]. It includes both the bilingual and monolingual data.

For the bilingual data, we apply data filtering techniques. The main process is described as follows. Firstly, we set the punctuation ratio and sentence length ratio of the sentence pairs to abandon the sentences higher than the ratio. Secondly, we calculate the cross entropy of each English sentence by a pre-trained language model and removed the sentence pair exceed the threshold. Thirdly, we construct a terminology table using the methods of name entity resolution and word alignment. The terminology such as companies, organizations and human names are replaced with specific words.

For the monolingual data, we follow the method proposed by (Sennrich et al., 2016). We firstly train an English to Chinese machine translation model. Then the monolingual English sentences are translated to generate synthetic Chinese translation. All the synthetic parallel data are filtered with the same strategies applied in bilingual data.

After the filtering process, we normalize the punctuation for both Chinese and English sentences. We apply Chinese word segmentation using LAC toolkit[2] (Jiao et al., 2018) for Chinese sentences. For the English sentences, we apply the Tokenizer and Truecaser toolkit provided by Moses scripts (Koehn et al., 2007). Finally, we train a bytes pairs encoding model and applied it for both Chinese and English sentences.

3 System description

The model training process for both the speech recognition and machine translation model are implemented on a device with eight GPUs of Nvidia TESLA V100.

3.1 Automatic speech recognition

The QuartzNet15x5 model is as our based model on ASR, we also use Memory-Self-Attention(MSA) (Luo et al., 2021) modules in the model structure of CTC and RNN-T.

3.1.1 Training Scheme

After data pre-processing, we use the file of json structure to train quartznet 15x5 model. We list the model configuration and train parameters in Table2. When the model was trained, the size of each sample audio should be controlled to less than 16.7 s. To do this, it can improve the accuracy of model and accelerate the training speed. The ASR model was trained over three days and reached to the best WER. After the loss value converged, we use the last saved model to try to transform test datasets and get average score. We use WER-BEAT (Sheshadri et al., 2021) to evaluate our model. And we get closed to 1.0 WER.

Table 2: Model Configuration

Configuration	**Value**
Sample rate	16,000
Repeat	5
n fft	512
activation	relu
Chinese Vocabulary size	5,270
Optimizer	Adam
residual	true
filters	256/512
batch size	64

To increase the accuracy of model recognition, we use MSA modules in the model structure of CTC and RNN-T. The operation complexity of the model maintains a linear relationship with the length of the input speech, which greatly improves the efficiency of the model, and there will be no serious decline in efficiency as the input increases.

3.1.2 Model Usage

Before we use the model, in order to improve the accuracy of recognition, we need to process the input voice file.

[1] http://mteval.cipsc.org.cn:81/agreement/AutoSimTrans

[2] https://github.com/baidu/lac

In the end, we only submit one point in the competition. This point is to directly use the previously segmented audio transcription text as the input of the translation model, thereby obtaining a more accurate English text output.

3.2 Machine translation

We use Transformer as our based model on machine translation, the attention mechanism is strength-able at capturing the Semantic relationship on a sentence. The development toolkit we used in machine translation is Marian (Junczys-Dowmunt et al., 2018).

3.2.1 Training Scheme

After completing the data preprocessing on both the bilingual data and back-translated data, we train our baseline model by evaluating BLEU. The language tool for evaluation is uncased 4-gram BLEU (Papineni et al., 2002). We list the model configuration in Table3 and training parameters in Table 4.

We train the model for over three days, the BLEU score increased rapidly at the beginning and the growth slowed after 30 hours. After the loss converged, we collect the last 20 checkpoints of the model in the time interval of one hour and applied checkpoint average to get the final model.

Table 3: Model Configuration

Configuration	Value
Encoder/Decoder depth	6
Attention heads	16
Word Embedding	1024
FFN size	4096
Chinese Vocabulary size	50,000
English Vocabulary size	50,000
Optimizer	Adam

Table 4: Training Parameters

Parameter	Value
Label smoothing	0.1
Learning rate	16
Warmup rates	15,000
Maximum sentence length	120
Clip normalization	5

3.2.2 Fine-tuning

We implement fine-tuning on the Transformer model using the development set of qianyan audio datasets (956 sentence pairs) to improve the translation quality on simultaneous translation task. Since fine-tuning is effective to build a domain-adaptive model.

4 Conclusion

This paper describes our submission to the second automatic simultaneous translation workshop at NAACL2021. We detail our process of data filtering and model training. The Consecutive Wait(CW) (Klein et al., 2017) of the best point reached to 18.4 while we get the BLEU value of 24.39 in the audio input track. In future work, we will continue to research on end-to-end speech translation model from Chinese speech input to English text output.

References

Xingyu Na Bengu Wu Hao Zheng Hui Bu, Jiayu Du. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, page Submitted.

Zhenyu Jiao, Shuqi Sun, and Ke Sun. 2018. Chinese lexical analysis with deep bi-gru-crf network. *arXiv preprint arXiv:1807.01882*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 116–121. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2019. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions.

Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M. Cohen, Huyen

Nguyen, and Ravi Teja Gadde. 2019. Jasper: An end-to-end convolutional neural acoustic model.

Jian Luo, Jianzong Wang, Ning Cheng, and Jing Xiao. 2021. Unidirectional memory-self-attention transducer for online speech recognition.

Mingbo Ma, Liang Huang, Hao Xiong, Kaibo Liu, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, and Haifeng Wang. 2018. STACL: simultaneous translation with integrated anticipation and controllable latency. *CoRR*, abs/1810.08398.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Akshay Krishna Sheshadri, Anvesh Rao Vijjini, and Sukhdeep Kharbanda. 2021. Wer-bert: Automatic wer estimation with bert in a balanced ordinal classification paradigm.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. Bstc: A large-scale chinese-english speech translation dataset. *arXiv preprint arXiv:2104.03575*.