System Description on Automatic Simultaneous Translation Workshop

Zecheng Li[1*], Yue Sun[2], and Haoze Li[3]

[1]Zhejiang University, Hangzhou, China
[2]Xiamen University, Xiamen, China
[3]North China Institute of Aerospace Engineering, Langfang, China
[1]*lizechng@zju.edu.cn*
[2]*njauyuesun@qq.com*
[3]*lohanz@foxmail.com*

Abstract

This paper describes our system submitted on the third automatic simultaneous translation workshop at NAACL2022. We participate in the Chinese audio→English text direction of Chinese-to-English translation. Our speech-to-text system is a pipeline system, in which we resort to rhymological features for audio split, ASRT model for speech recoginition, STACL model for streaming text translation. To translate streaming text, we use wait-*k* policy *trained* to generate the target sentence concurrently with the source sentence, but always *k* words behind. We propose a competitive simultaneous translation system and rank 3rd in the audio input track. The code will release soon.

1 Introduction

Simultaneous translation refers to translating the message from the speaker to the audience in real-time without interrupting the speakers, which is a challenging task and has become an increasingly popular research field in recent years.

In this paper, we describe our system submitted at the 3rd automatic simultaneous translation workshop, which consists of a rhymeological features based audio split model, an end to end speech recognition model and a wait-*k*(Ma et al., 2019) based streaming text translation model. The system input is Chinese audio file and the output is English translation text. A temporary Streaming transcription is obtained by audio split and speech recognition model, then transmitted into machine translation model to get the target system output.

For automatic audio split model, we calculate the rhythmological features(Weninger et al., 2013) of the audio input, resort to adaptive policy to set short-term energy threshold and zero crossing rate threhold for speech split. For automatic speech recognition model, we use ASRT model[1], which is based DCNN model and CTC decoder(Graves et al., 2006). Whilst, we expand the training data set by adding Aishell-1(Bu et al., 2017) and Thchs-30(Wang and Zhang, 2015) datasets. For streaming text translation, our model is based on STACL(Ma et al., 2019). We use some human rules and the pre-trained language model to filter the parallel corpus. At the step of inference, we apply the wait-*k* words policy. Both the pre-processing and post-precessing are applied to improve the terminology translation and deal with the word error produced by the ASR system.

Since our submission is a pipeline system, the rest of this paper describes separately regards to audio split, automatic speech recognition and matchine translation sub-modules. We firstly describe the training and development datasets we used, then the data precessing methods is introduces. Secondly, we describe our system architecture and experiment results. Lastly, we draw a conclusion of our system by analyzing the experiments.

2 Dataset

For audio data of ASR, we use qianyan audio datasets provided by NAACL workshop(Zhang et al., 2021), Aishell-1(Bu et al., 2017), Thchs-30(Wang and Zhang, 2015). For text data of MT, we use CWMT19[2] and the simultaneous translation corpus provided by the organizer of the workshop.

2.1 Audio data

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.

For other 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 each audio.

[1]https://github.com/nl8590687/ASRT_SpeechRecognition

[2]http://mteval.cipsc.org.cn:81/agreement/description

Proceedings of the Third Workshop on Automatic Simultaneous Translation, pages 18 - 21
July 15-16, 2022 ©2022 Association for Computational Linguistics

Data Source	Duration	Size
Qianyan(NAACL)	65h	5.4G
Aishell-1	178h	14.51G
Thchs-30	40h	6.01G

Table 1: Zh-En audio training datasets.

In order to mitigate the matching issues between audio file and transcription text, we use pre-trained ASRT model to produce pronunciation results from audio input, and then obtain streaming text from pronumciation models. Table 1 shows the number of training data.

2.2 Text data

For CWMT19 and Baidu Speech Translation Corpus(BSTC)(Zhang et al., 2021) datasets, we firstly filter out the sentence whose English sentence is longer than 120 words. Meanwhile, there are a few Chinese characters in the data which are traditional characters. We convert them to simplified ones. Then all Chineses sentences are segmented with Jieba Chinese Segmentation Tool[3] and all English sentences are tokenized and truecased with Moses[4]. Lastly, Both Chinese and English data are encoded by BPE(Sennrich et al., 2015) with Subword-NMT[5] to train a bytes pairs encoding model.

3 System description

Our system consist of a rhymeological features based audio split model, an end to end speech recognition model, and a wait-*k* based streaming text translation model. The model training process for speech recognition and machine translation model are implemented on a device with four GPUs of Nvidia 1080ti.

3.1 Audio split

For automatic audio split model, we use the traditional acoustic methods. We firstly calculate the rhythmological features of the audio input based on Librosa audio processing library[6] and the openSMILE toolkit(Eyben et al., 2010). According to short-term energy and zero crossing rate of the rhythmological features, we can detect the endpoint of voice. This can detect all valid speech parts of a section of speech. The endpoint detection consists of two steps. The first step is the overall endpoint detection used to segment the long audio file, the second step is the fine-tune of the splited audio. The audio split process is shown in Figure 1. The super-parameters we use are shown in the Table 2.

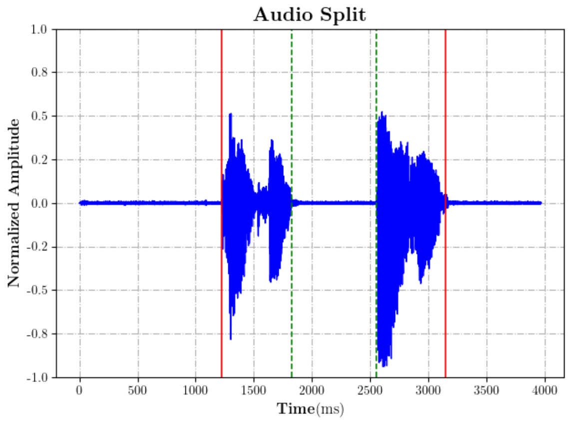

Figure 1: Audio Split Process. The solid red line is the reult of Step-1, and the dashed green line is the result of Step-2

Parameter	Step-1	Step-2
Frame length	400	240
Min. turbid interval	25	20
Short-term energy threhold	1.0	0.4
Zero crossing rate threhold	0.8	1.2

Table 2: Audio split model parameters.

3.2 Speech recognition

The speech recognition model we use is ASRT model, based on deep convolutional neural network and long-short memory neural network, attention mechanism and CTC to implement.

We firstly limit the maximum length of splited audio to 16 seconds, as the input of ASRT model. The speech recognition model will output the corresponding pronunciation sequence. Then we resort to probability map based maximum entropy Markov model to convert the pronunciation sequence to recogized text. To improve the recognition accuracy, we use the model pre-trained on AiShell-1 and Thchs-30 datasets and fine-tune on audio dataset provided by NAACL workshop. We list the model configuration in Table 3

3.3 Machine translation

We use STACL as our machine translation model. We train the model for over two days,

[3]https://github.com/fxsjy/jieba
[4]https://github.com/moses-smt/mosesdecoder
[5]https://github.com/rsennrich/subword-nmt
[6]https://github.com/librosa/librosa

Configuration	Value
Audio length	1600
Feature length	200
Label length	64
Channels	1
Output size	1428
Optimizer	Adam

Table 3: Speech recognition model configuration

Configuration	Value
Encoder/Decoder depth	6
Attention heads	8
Word Embedding	512
Chinese Vacabulary size	10000
English Vacabulary size	10000
Optimizer	Adam

Table 4: Machine translation model configuration

the BLEU(Papineni et al., 2002) score increased rapidly at the beginning and the growth slowed after 20 hours. After the loss converged, we save the last checkpoint as the final model. We list the model configuration in Table 4 and training parameters in Table 5.

The simultaneous policy we use is wait-k, which first wait k source words, and then translates concurrently with the reset of source sentence, i.e., the output is always k words behind the input.

We implement fine-tuning on the STACL model using the BSTC dataset to improve the translation quality on simultaneous translation task. Since fine-tuning is effective to build a domain-adaptive model.

4 Experiment

In this section, we evaluate our system on the development set of the Baidu Speech Translation Corpus. The two used metrics are case-sensitive detokenized BLEU(Papineni et al., 2002) and Consecutive Wait(CW)(Gu et al., 2016), for translation quality and latency respectively. CW considers on

Parameter	Value
Label smoothing	0.1
Learning rate	2.0
Warmup steps	8000
Maximum sentence length	120

Table 5: Machine translation model training parameters

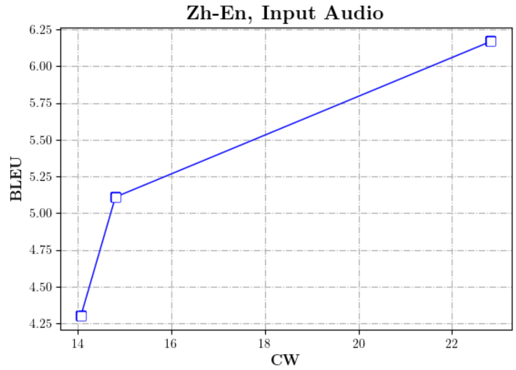

Figure 2: Experimental Result of speech-to-text track

how many source words are waited for consecutively between two target words, and thus larget CW means longer latency.

We set the threshold k in the wait-k policy to various values and get multiple results, as shown in Figure 2. Due to the speech in the development set is difficult for ASR model trained ourselves, resulting in a high character error rate. The errors caused by ASR are brought to MT, and thus the BLEU is much lower than that in the text-to-text track.

5 Conclusion

This paper describe our submission to the 3rd automatic simultaneous workshop at NAACL2022. We detail our process of data filtering and model training. The Consecutive Wait(CW) of the best point reached to 14.06, while we get the BLEU value of 6.17 in the audio input track. In future work, we will continue to research on end-to-end speech translation model.

References

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5. IEEE.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. 2016. Learning to translate in real-time with neural machine translation. *arXiv preprint arXiv:1610.00388*.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

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*, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Dong Wang and Xuewei Zhang. 2015. Thchs-30: A free chinese speech corpus. *arXiv preprint arXiv:1512.01882*.

Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer. 2013. On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology*, 4:292.

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*.