ZXN's Simultaneous Translation System at NAACL 2022

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

[1] Zhejiang University, Hangzhou, China
[2] Xiamen University, Xiamen, China
[3] North China Institute of Aerospace Engineering, Langfang, China

Tasks

- We participated in two tasks.
  - Zh->En Translation, input: streaming transcription. (text-to-text)
  - Zh->En Translation, input: audio file. (speech-to-text)

Data and preprocessing

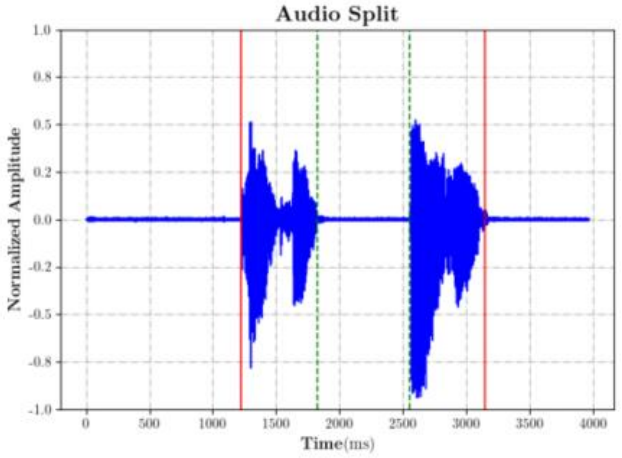
- Dataset.
  - For audio data of ASR, we use QianYan audio datasets provided by NAACL workshop, Aishell-1.
  - For text data of MT, we use CWMT19 and the simultaneous translation corpus provided by the organizer.
- Preprocessing
  - Filter out long sentence pairs.
  - Convert full-width characters into half-width characters.
  - Segment Chinese text and tokenize English text.
  - Apply byte-pair-encoding to all sentences.

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.
    - Audio Split.
    - Speech Recognition.
    - Machine Translation.

System Description

- 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 and the openSMILE toolkit. According to short-term energy and zero crossing rate of the rhythmological features, we can detect the endpoint of voice.
  - 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.



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

Illustration of Audio Split process. (a) The solid red line is the reult of Step-1, and the dashed green line is the result of Step-2 (b) Audio split model parameters.

System Description

- 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.

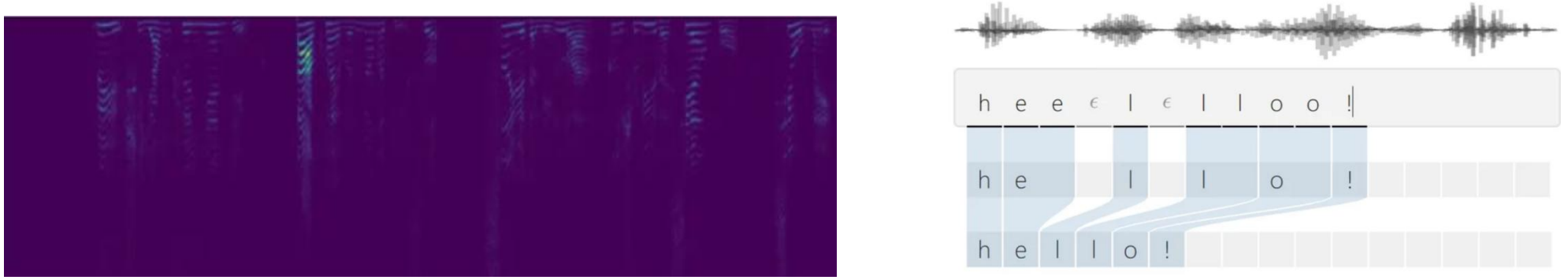


Illustration of Speech Recognition process. (a)The ordinary WAV speech signal is transformed into a two-dimensional spectral image signal through frame segmentation and window addition (b) Merge successive identical symbols into the same symbol, and then remove the mute delimiter

System Description

- Machine translation
  - We use STACL as our machine translation model.
  - The simultaneous policy we use is waik-*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.

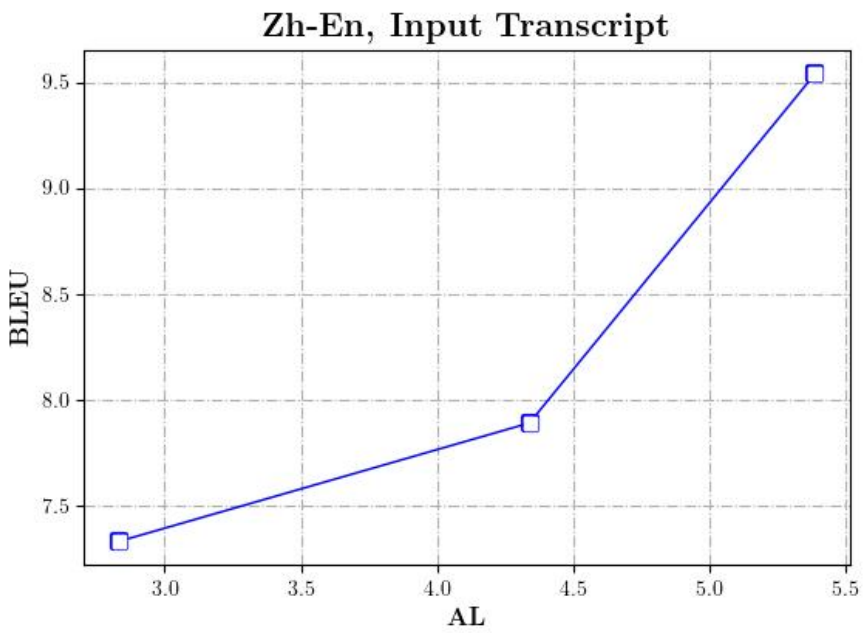
Configuration	Value
Audio length	1600
Feature length	200
Label length	64
Channels	1
Output size	1428
Optimizer	Adam

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

Illustration of Machine Translation process. (a)Speech recognition model configuration (b) Machine translation model configuration

Experiments

- The two systems are evaluated on the development set of the Baidu Speech Translation Corpus.



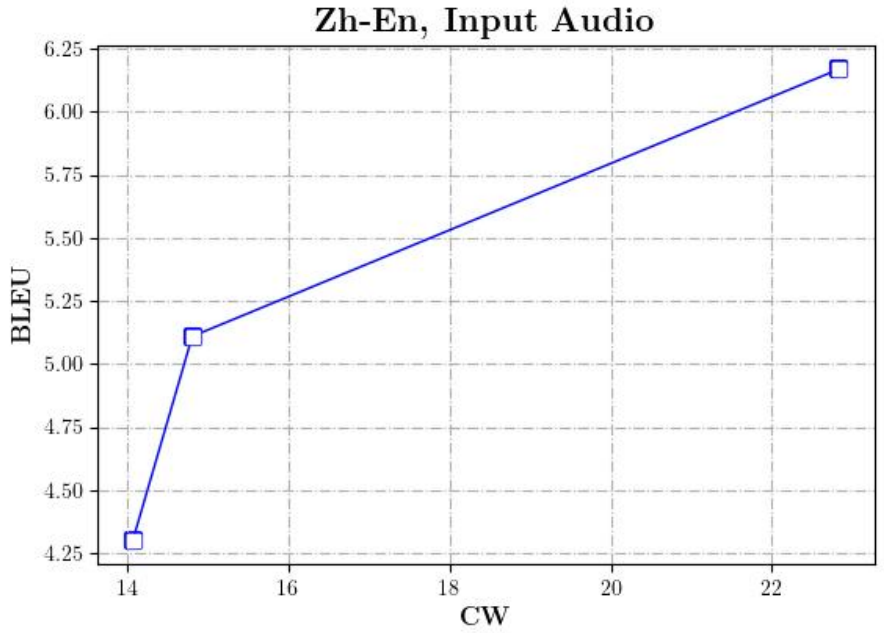


Illustration of Experiment Results. (a)The text-to-text system. (b) The speech-to-text system.

Thanks