BIT's system for AutoSimTrans 2021

Mengge Liu , Shuoying Chen , Minqin Li , Zhipeng Wang and **Yuhang Guo***
Beijing Institute of Technology, Beijing, China
lmg864355282@gmail.com guoyuhang@bit.edu.cn

Abstract

In this paper we introduce our Chinese-English simultaneous translation system participating in AutoSimTrans 2021. In simultaneous translation, translation quality and latency are both important. In order to reduce the translation latency, we cut the streaming-input source sentence into segments and translate the segments before the full sentence is received. In order to obtain high-quality translations, we pretrain a translation model with adequate corpus and fine-tune the model with domain adaptation and sentence length adaptation. The experimental results on the development dataset show that our system performs better than the baseline system.

1 Introduction

Machine translation greatly facilitates communication between people of different language, and the current neural machine translation model has achieved great success in machine translation field. However, for some occasions that have higher requirements for translation speed, such as in simultaneous interpretation dynamic subtitles and dynamic subtitles application fields. Machine translation models that use full sentences as translation units need to wait for the speaker to speak the full sentence before starting translation, in which the translation delay is unacceptable. In order to reduce the delay, translation must start before the complete sentence is received. But at the same time the incomplete sentence may have grammatical errors and semantic incompleteness, and the translation quality will decrease compared to the result obtained by full sentences. Further more, different languages may have different word order. There are also many reordering phenomenon when translating between Chinese and English which both belong to the same SVO sentence structure. Sentence reordering and different word-order expression habits bring a great difficult to simultaneous translation.

Since the latency of using a full sentence as translation unit is unacceptable, and the translation of incomplete sentences is difficult and not guaranteed to obtain reliable translations, we consider cutting long sentence into appropriate sub-sentences. And each sub-sentence is grammatically correct and semantically complete to get suitable translation result. By decomposing translating long sentences into translating shorter sub-sentences, the translation can be started before the complete long sentence is received. This strategy of achieving low-latency simultaneous translation can be summarized as segmentation strategy (Rangarajan Sridhar et al., 2013). At the same time, it is observed that a sentence can be divided into independent sub-sentences for translation. For the example in table 1, Chinese and English sentences can be cut, and the Chinese sub-sentences can be translated as a shorter translation unit. According to this example, we can also observe that there is no cross alignment between the two sub-sentences, that is, the English translation of the first Chinese sub-sentence has no semantic and word connections with the translation of second Chinese sub-sentence, and there is no cross word alignment between the two sub-sentences. This phenomenon indicates that it is feasible to divide the full sentence in the parallel corpus into shorter sub-sentences.

In the following of this paper, the second part will introduce the overall framework of the model, the third part will give a detailed description of the fine-tuning, finally will ex-

*Corresponding author

Proceedings of the 2nd Workshop on Automatic Simultaneous Translation, pages 12–18

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

Table 1: Segment example, first sub-sentence is in red and the second one is in black.

plain and analysis the experiment results.

2 System Architecture

This part mainly introduces the overall framework of our submission in AutoSimulTrans 2021 competition. The whole model uses typical segmentation strategy to achieve simultaneous translation. It consists of a sentence boundary detector and a machine translation module. The sentence boundary detector reads the streaming input text and obtains the appropriate segments. The segments are input to the downstream translation module, and the translation result of each segment is obtained and then spliced to obtain the full translation. The overall framework of the entire model is shown in the figure 1.

2.1 Sentence Boundary Detector

The sentence boundary detector can also be regarded as a text classifier. For the streaming-input sentence, detector needs to be able to judge whether the received part can be used as a suitable segment to be translated. The specific implementation of the boundary detector is based on a pre-trained Chinese BERT(Devlin et al. (2018)) model as a text representation, add a fully connected layer to form a classifier. In terms of data, long sentences are divided into segments according to punctuation marks, segments are regarded as sub-sentences. Positive and negative examples are constructed according to such rules to fine-tune the pre-trained model to obtain a classifier achieving an accuracy of 92.5%. According to the above processes, a boundary detector that can process streaming input text is constructed.

2.2 Translation Module

The translation module is implemented with the tensor2tensor framework, training the transformer-big model(Vaswani et al., 2017) as a machine translation module. We use the pre-training and fine-tuning method to get better performance on the target task.

First, we use the CWMT19 data set as a large-scale corpus to pre-train machine translation model. The CWMT19 corpus is a standard Chinese and English text corpus, but the target test set in the competition is the speech transcription and translation results, which have domain difference with the standard text. So it is necessary to use speech domain corpus to fine-tune the translation model. On the other hand, the translator needs to translate the sub-sentences when decoding. There is a mismatch between the length and the amount of information between the sub-sentence and the longer full sentences. So we further fine-tune the translation model to make it adapted to sub-sentences translation.

3 Fine-tuning Corpus

3.1 Domain fine-tuning

In order to make the machine translation model trained on the standard text corpus more suitable for translating the transcriptions in the speech field, the translation model needs to be fine-tuned with the corpus of the corresponding speech field. We use the manual transcription and translation text of the Chinese speech provided by the organizer as parallel corpus to fine-tune the pre-training translation model.

3.2 Sentence length fine-tuning

The pre-training and domain fine-tuning processes only train the translation model on the full sentence corpus. But when the model is used to perform the simultaneous translation and decoding process, the sub-sentences are needed to be translated, which causes mismatch between training and testing. In order to make the machine translation model adapt to the shorter sub-sentences translation sence, it is necessary to construct a sub-sentence corpus composed of Chinese and English sub-sentence pairs to further fine-tune the machine translation model. In order to meet the requirements of domain adaptation at the same time, sub-sentence corpus is constructed based

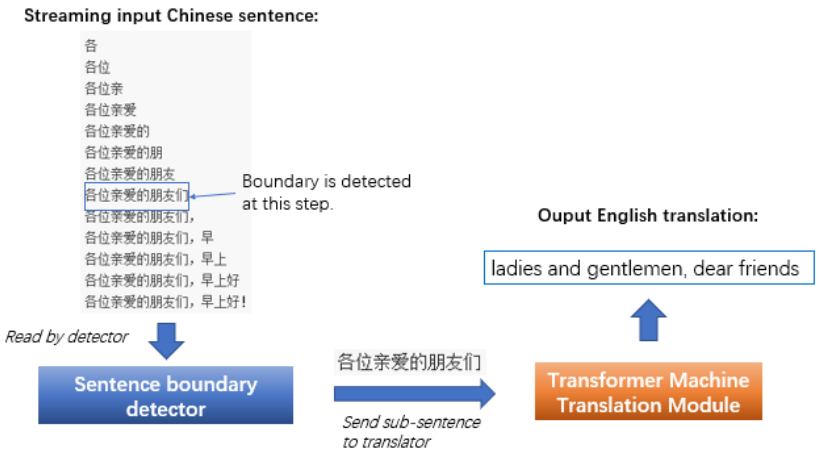

Figure 1: System Architecture

on the Chinese-English corpus provided by the organizer to fine-tune the machine translation model to adapt to the sub-sentence translation scenario. The following is a detailed description of the specific method of processing the full sentence into a sub-sentences.

The ideal sentence segmentation effect is that if the Chinese and English sentence pairs are divided into two or more sub-sentence pairs, Chinese sentence and the English sentence should be cut at the same time to obtain the same number of sub-sentences, and corresponding Chinese and English sub-sentences should contain same information. In another word, using Chinese sub-sentence can get enough information to translate the corresponding English sub-sentence. In order to meet the requirements of information integrity, we use the word alignment tool to obtain the word alignment information between Chinese and English sentence pairs, using the *fast_align*(Dyer et al., 2013) word alignment tool to obtain Chinese to English and English to Chinese alignments respectively, and merge them into symmetry alignments. The result of word alignment, such as the Chinese input sentence $X = \{x_1, x_2, ..., x_n\}$ and the target English sentence $Y = \{y_1, y_2, ..., y_m\}$, we can get a set of alignment results $A = \{< x_i, y_j > \mid x_i \in X, y_j \in Y\}$.

Then, the word alignment matrix is obtained according to the word alignment results. The segmentation of the Chinese and English full-sentence pairs is equivalent to the division of the word alignment matrix. The word alignment matrix can be divided into four blocks according to a division position, when the lower left and upper right matrices are both zero matrices, meaning that two sub-sentences do not have cross-word alignment. And sub-sentences can be obtained at the current segmentation position. Moreover, the traversal-based division algorithm can divide a sentence with multiple suitable methods, effectively increasing the number of sub-sentence pairs in the sub-sentence corpus.

An example of sentence segmentation using word alignment matrix is shown in the figure 2. According to the alignment results of Chinese and English words, an alignment matrix is constructed. The position is '1' means the Chinese word and English word have alignment and the remaining position have no alignment. Two dashed boxes are identified in the figure, corresponding to two reasonable division results. The dashed box is the first sub-sentence and remain part is second sub-sentence. We retain all reasonable fragmentation results when segmenting sentences, that is, both segmentation results in the figure will be retained.

4 Experiment

4.1 Experiment settings

The boundary detector is based on the pre-training BERT of *chinese_L-12_H-768_A-12* as the pre-training model, the hidden size of fully connected layer is the same of BERT. Using the simultaneous interpretation corpus provided by the organizer, cutting into sub-

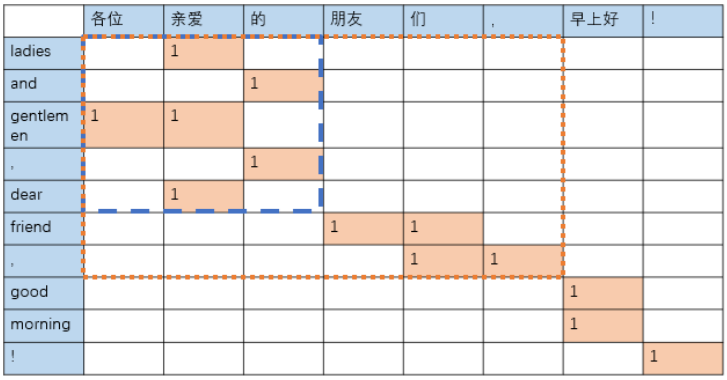

Figure 2: Segment sentence by word alignment matrix.

sentences based on punctuation, constructing positive and negative examples for fine-tuning training. Then we obtain a sentence boundary recognizer that can recognize sentence boundaries and realize real-time segmentation of streaming input.

Our translation model is based on the *tensor2tensor* framework. We set the parameters of the model as *transformer_big*. And we set the parameter problem as *translate_enzh_wmt32k_rev*. We train the model on 6 GPUs for 9 days.

In experiment, we pre-train translator on CWMT19 dataset, fine-tune translator on BSTC(Zhang et al., 2021) dataset, and evaluate model on BSTC development dataset containing transcription and translation of 16 speeches. CWMT19 is a standard text translation corpus. BSTC contains 68h Chinese speech and corresponding Chinese transcription and English translation text. In this article, we only use Chinese and English texts in the speech field.

4.2 Sub-sentence fine-tuning

In terms of domain adaptability, we use golden transcribed text as fine-tuning corpus. In terms of sentence length adaptability, we use corpus containing only golden transcriptions and corpus containing ASR and golden transcriptions to construct sub-sentence corpus, and use boundary detector as a filter to remove some unsuitable sub-sentence. The situation of fine-tuning corpus is shown in the table 2. The same sentence boundary detector is used by all model, and different machine translation modules are as follows:

- domain fine-tuned: pre-trained on CWMT19 corpus, and fine-tuned on golden transcription.

- sub-sentence fine-tuned(golden+ASR): based on domain fine-tuned model, fine-tuned by segmented golden&ASR transcription corpus.

- sub-sentence fine-tuned(golden): based on domain fine-tuned model, fine-tuned by segmented golden transcription corpus.

- sub-sentence fine-tuned(filtered golden): based on domain fine-tuned model, fine-tuned by filtered segmented golden transcription corpus.

Learning rate is set as 2e-5 in fine-tuning, domain fine-tuning is carried out for 2000 steps and segmentation fine-tuning is carried out for 4000 steps.

4.3 Latency metric

Here is the definition of AL latency metric as used in (Ma et al., 2018). t is decoding step, τ is cut-off decoding step where source sentence is finished, $g(t)$ denote the number of source words read by encoder at decoding step t, and $r = |x|/|y|$ is target-to-source length ratio. The lower AL value means lower latency and better real-time simultaneous system.

$$AL = \frac{1}{\tau}\sum_{t=1}^{\tau} g(t) - \frac{t-1}{\gamma}$$

$$\tau = \arg\min_{t}[g(t) = |x|]$$

4.4 Results and analysis

The performance of each model on the development set is list in table 3.According to the

Fine-tuning corpus	Type	Sentence Pairs
golden transcription	full-sentence	37k
segmented golden&ASR transcription	sub-sentence	2555k
segmented golden transcription	sub-sentence	668k
segmented golden(filtered) transcription	sub-sentence	246k

Table 2: First full-sentence corpus is provided by organizer. Three sub-sentence corpus constructed by word alignment, constructed from golden and ASR transcription corpus provided by organizer. The third line is the filtered segmentation corpus.

experimental results, the performance of the fine-tuning model did not meet expectations. Using only the corpus made by golden transcription corpus brought a greater quality reduction compared to using corpus including the ASR and golden transcriptions. Comparing with models fine-tuned by golden transcription and model fine-tuned by filtered golden transcription, we can find that although the number of sentences in sub-sentences corpus has decreased after filtering, it has obtained a relatively high score, which reflects the effectiveness of the filtering operation.

The main reason for the unsatisfactory fine-tuning effect may because the sub-sentence corpus contains too much noise. It may be difficult to obtain high-quality segmentation results by the word alignment results. Although we have filtered many inappropriate sentences, there is still a lot of noise in the sub-sentence corpus. And because the sub-sentences are shorter, the translation errors of the sentence pair in fine-tuning corpus will have a greater negative impacts on translation model.

Here is an example to explain the difficulty of sentence division. In the sentence showed in table 4, we list the source sentence and target sentence, and also direct translation for each phrase just for understanding the meaning of Chinese words. From the perspective of word alignment, it can be easily divided from the comma position to obtain two sub-sentences. For the first sub-sentence pair, the Chinese and English sub-sentences contain same information, and good English translation results can be easily obtained according to Chinese. But for the second sub-sentence pair, it's hard to obtain golden translation relay only on Chinese sub-sentence. If you directly translate the Chinese, you may get a translation result similar to "amazing by hearing. ". This is because the result of golden translation is obtained with full sentence, and in order to make the translated English expression more fluent, free translation is carried out. If the translation model only reads the second sub-sentence, it is difficult to obtain a suitable translation result relative to the golden result.

5 Related work

This article uses segmentation strategy to achieve low-latency simultaneous translation. There are also some similar works use segmentation strategy to divide long sentences into segments for translation, (Xiong et al., 2019) focus on improving the coherence of the sub-sentences translation results, (Zhang et al., 2020) focus on solving the problem of long-distance reordering in simultaneous translation.

In addition, there are two different strategies for achieving simultaneous translation: one is a more flexible translation strategy based on sentence prefixes. The process of simultaneous translation is defined as a read-write action sequence from the perspective of behavior. It is necessary to define a suitable strategy to find out the action sequence, and adjust the translator to make the model more suitable for the translation of sentence prefixes (Ma et al., 2018)(Arivazhagan et al., 2019). Another type is translation based on dynamic refresh without the need to adjust the machine translation model. Whenever the input increases, translate all input and overwrite the translation result that has been generated last time (Niehues et al., 2016)(Arivazhagan et al., 2020b)(Arivazhagan et al., 2020a).

6 Conclusion

In this paper we describe a simultaneous translation method that reduces translation

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

Table 3: Performance of each model on the development set. AL is latency metric and BLEU is text quality metric.

Source sentence	这些东西	都是	大自然奇特的物产	,	听听都很奇特。
Literal translation	These things	are all	nature's amazing creations	,	amazing by hearing.
Target sentence	These	are all	amazing creations of the nature	,	you can tell just from their names .

Table 4: A example hard to segment. The sentence can be segmented by comma. The literal translation of second sub-sentence is quite different from the target.

delay by cutting the full sentence into sub-sentences.We fine-tune a pre-trained translation model in terms of domain and sentence length. The sub-sentence corpus is constructed by word alignment, we found that directly using all the sub-sentences we obtained has a negative impact on translation performance, but it can be improved after filtering. In the end, we obtained translation results that exceeded the baseline model.

Acknowledgements

Supported by the National Key Research and Development Program of China (No. 2016YFB0801200)

References

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. *arXiv preprint arXiv:1906.05218*.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020a. Re-translation versus streaming for simultaneous translation. *arXiv preprint arXiv:2004.03643*.

Naveen Arivazhagan, Colin Cherry, Isabelle Te, Wolfgang Macherey, Pallavi Baljekar, and George Foster. 2020b. Re-translation strategies for long form, simultaneous, spoken language translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7919–7923. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, et al. 2018. Stacl: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. *arXiv preprint arXiv:1810.08398*.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic transcription for low-latency speech translation. In *Interspeech*, pages 2513–2517.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 230–238, Atlanta, Georgia. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang.

2019. Dutongchuan: Context-aware translation model for simultaneous interpreting. *arXiv preprint arXiv:1907.12984*.

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*.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020. Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2280–2289.

A Development results

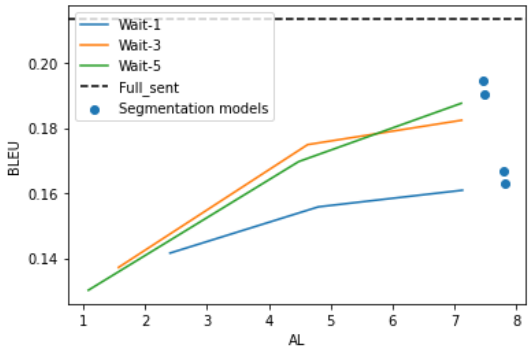

Figure 3: Development results

The results of each model on the development set are shown in the figure 3, where each curve of wait-1, wait-3, wait-5 and full-sent is the wait-k series model and full-sentence model provided by the organizer. Each model is a transformer neural machine translation model. Each scattered point represents a segmentation model in this article. According to the results, it can be seen that the domain fine-tuning model and a better-performed sub-sentence fine-tuning model are better than the wait-k series model.