The HW-TSC Video Speech Translation system at IWSLT 2020

Minghan Wang, Hao Yang, Yao Deng, Ying Qin, Lizhi Lei,
Daimeng Wei, Hengchao Shang, Jiaxin Guo, Ning Xie, Xiaochun Li
Huawei Translation Service Center, Beijing, China
{wangminghan,yanghao30,dengyao3,qinying,leilizhi
weidaimeng,shanghengchao,guojiaxin1,
nicolas.xie,carol.lixiaochun}@huawei.com

Abstract

In this paper, we present details of our system in the IWSLT Video Speech Translation evaluation. The system works in a cascade form, which contains three modules: 1) A proprietary ASR system. 2) A disfluency correction system aims to remove interregnums or other disfluent expressions with a fine-tuned BERT and a series of rule based algorithms. 3) An NMT System based on the Transformer and trained with massive publicly available corpus is used for translation.

1 Introduction

There has been great advances in Neural Machine Translations (NMT) in recent years which also promotes Multimodal translation including image (Specia et al., 2016), speech and video translation (Wang et al., 2020; Imankulova et al., 2020; Wu et al., 2019). For speech and video translation, there are basically two types of systems (i.e. cascade and end-to-end). Cascade systems are often composed of several independent modules, where each one can be trained with intermediate inputs and labels. End-to-end systems can be fully differentiable and trained with original multimodality data.

In the IWSLT 2020 Video Speech Translation task (Ansari et al., 2020), participants are required to develop systems to translate speeches in the video from source language into target language. However, we consider that visual contexts are not necessarily important for a translation task, therefore we only use audio as contexts, and treat it as a offline speech translation task, in addition, our system is built mainly for Chinese to English translation.

We choose to build our system in a cascade manner because training an end-to-end system requires massive aligned audio and text data, which is hard to find. On the other hand, a cascade system allows us to train each part separately, which is more feasible.

Our system is composed with three modules. 1) A proprietary ASR system. 2) A disfluency correction system (Wang et al., 2019). 3) An NMT model based on the Transformer (Vaswani et al., 2017).

2 System Architecture

2.1 ASR

For the task, we simply extract the sound tracks from videos, then feed them to our proprietary ASR system and proceed transcripts to downstream modules.

2.2 Disfluency Correction

A major flaw of the cascade system is the error propagation from the ASR to the NMT system. For example, interregnums like “uh”, “you know” should not be translated, or repeated words like “I wanna wanna ...” due to disfluency (Shriberg, 1994; Wang et al., 2018). To deal with this problem, we developed the disfluency correction system to de-noise the ASR outputs so that the NMT model could generate more fluent translations. The disfluency correction system is based on fine-tuning BERT for sequence tagging and incorrect N-gram mining. BERT is fine-tuned to predict whether a token generated by the ASR system should be kept or deleted. The N-gram model is able to mine high-frequency mistakes, e.g. “a i”(AI), “r and d”(R&D).

2.2.1 Data

The dataset used to train the disfluency detection model is collected from internal meeting recordings, which are transcribed by human. We create the golden set based on these transcripts. It con-

Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT), pages 187–190

https://doi.org/10.18653/v1/P17

Strategy	Train	Dev	Test
Baseline	35.2	36.3	33.1
+ Domain rule	31.8 (-3.4)	31.9 (-4.4)	30.1 (-3.0)
+ BERT model	28.3 (-3.5)	28.3 (-3.6)	26.7 (-3.4)
+ N-gram mining	26.4 (-1.9)	25.9 (-2.4)	24.9 (-1.8)

Table 1: The performance of each strategy evaluated on our own dataset

tains approximately 200,000 tokens and are split into train/dev/test set with the proportion of 7:1:2.

We use sclite from SCTK to score the ASR outputs, the sclite is able to output four types of actions for tokens and gaps in the ASR outputs. Then, those actions will be used to automatically label the ASR outputs with following rules:

- Tokens are scored as C (correct) by the sclite means that the ASR system outputs a correct token, compared to the ground truth sequence. These tokens will be labeled as **OK** and should be preserved in post-processing.
- Tokens are scored as S (substitute) by the sclite means the ASR system outputs an incorrect (substituted) token at current position, compared to the ground truth sequence. These tokens are labeled as **BAD** and will be deleted. Note that we don't consider predicting the correct token because of the complexity.
- Tokens are scored as I (insertion) means the ASR system inserts an unnecessary token at current position. These tokens will be labeled as **BAD** and should be deleted.
- Gaps are scored as D (deletion) means the ASR fails to output a necessary token at the gap. This situation also involves predicting the missing token therefore is not considered.

Based on these rules, the ASR outputs are labeled with OK and BAD for each token, which will be used to fine-tune the BERT model.

2.2.2 BERT based Disfluency Detection

BERT is widely used in many NLP tasks thanks to its flexible architecture and pre-trained weights contributed by the community. As described previously, we formulate the task as a sequential labelling problem so that the pre-trained BERT can be fine-tuned easily. In the post-processing, we simply remove tokens which are predicted as BAD. The model used for fine-tuning is provided by transformers' (Wolf et al., 2019) BERT-base.

2.2.3 N-Gram Disfluency Mining

Except from training a detection model, we also create a mistake table with n-grams (where n=1,2,3,4), aiming to correct high-frequency mistakes. The table statistics the frequency of incorrect transcripts, and top 10 mistakes are used as a rule based mapping. We mainly use it to solve some situations where the pronunciation and text are not the same, this situation mainly appears in terminologies.

2.2.4 Disfluency Correction Experiments

We preform several experiments on the created dataset with the combination of methods mentioned above to evaluate the effectiveness. We use WER as the evaluation metric. Detailed results are presented in Table 1.

2.3 NMT

The NMT model is based on a Transformer (Vaswani et al., 2017) model with some modifications that will be introduced later. The model is trained with approximately 26M publicly available parallel corpora.

2.3.1 Data

There is no officially published in-domain text data to train and evaluate the model, therefore, we use the WMT 2019 Chinese to English news translation corpora which is composed with 6.5M CWMT and 20M UN sentence pairs. The test set of CCMT 2018 news translation is used as our test set for the experiment.

First of all, we clean up the dataset as follows:

- Remove duplicated sentences.
- Sentences that are too short (e.g. less than two tokens) are removed.
- Sentences that are too long (e.g. greater than 300 tokens) are removed.
- Parallel sentence pairs with abnormal length ratio (e.g. greater than 3 times standard deviation) are removed.

Strategy	BLEU
Transformer-big	36.27
+ Domain Classification	37.63
+ Ensemble	39.25

Table 2: The performance of the nmt model evaluated on CCMT dataset.

- Sentences with abnormal characters are considered as HTML entities, e.g. , are removed.

Subword-nmt (Sennrich et al., 2016) is used to tokenize English sentences, while character based tokenization is used for Chinese sentences.

2.3.2 Model

Transformer Big (Vaswani et al., 2017) is used in our NMT system, which has same number layers compared to Transformer base but with wider embedding and FFN layers, additional normalization and dropout layers are also added.

As we mentioned in Section 2.3.1, we mix the CWMT and UN corpus in which sentences may draw from different distributions and thus may have significant differences in the quality, domain as well as the style. These differences may degrade the performance of a NMT model. To deal with such problem, inspired by (Britz et al., 2017), we add a domain discrimination tag (token) at the start of the sentence for target sentences, here, we use [CWMT] and [UN] to represent two domains (data sources). The initial tag will be used to calculate a discrimination loss thus makes the model trained in a multi-task setting. In the inference phase, the generated tag will be removed in the post-processing.

2.3.3 NMT experiments

We perform experiments with three strategies: Transformer-big with and without domain classification as well as an ensemble model. We use BLEU (Papineni et al., 2002) as the evaluation metric. Details are shown in Table 2.

3 Conclusion

In this paper, we introduce our system for IWSLT 2020 Video Speech Translation evaluation. Our cascade system is developed and evaluated separately, we select the best strategy for each module to integrate a pipeline system which finally makes predictions for our submission.

References

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondrej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, and Changhan Wang. 2020. Findings of the IWSLT 2020 Evaluation Campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT 2020)*, Seattle, USA.

Denny Britz, Quoc V. Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 118–126.

Aizhan Imankulova, Masahiro Kaneko, Tosho Hirasawa, and Mamoru Komachi. 2020. Towards multimodal simultaneous neural machine translation. *CoRR*, abs/2004.03180.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. 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*.

Elizabeth Shriberg. 1994. Preliminaries to a theory of speech disfluencies. phd, University of California, Berkeley.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 543–553.

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, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020. Curriculum pre-training for end-to-end speech translation. *CoRR*, abs/2004.10093.

Feng Wang, Wei Chen, Zhen Yang, Qianqian Dong, Shuang Xu, and Bo Xu. 2018. Semi-supervised disfluency detection. In *Proceedings of the 27th International Conference on Computational Linguistics,*

COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 3529–3538.

Shaolei Wang, Wanxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. 2019. Multi-task self-supervised learning for disfluency detection. *CoRR*, abs/1908.05378.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zixiu Wu, Ozan Caglayan, Julia Ive, Josiah Wang, and Lucia Specia. 2019. Transformer-based cascaded multimodal speech translation. *CoRR*, abs/1910.13215.