The IWSLT 2019 KIT Speech Translation System

[†]*Ngoc-Quan Pham,*[†]*Thai-Son Nguyen,* [†]*Thanh-Le Ha,* [†]*Juan Hussain,* [†]*Felix Schneider,*
[‡]*Jan Niehues,* [†]*Sebastian Stüker,* [*†]*Alexander Waibel*

[†]Institute for Anthropomatics and Robotics
KIT - Karlsruhe Institute of Technology, Germany
firstname.lastname@kit.edu
[‡]Department of Data Science
University of Maastricht, Netherlands
firstname.lastname@maastrichtuniversity.nl
[*] Language Technologies Institute
Carnegie Mellon University
firstname.lastname@cmu.edu

Abstract

This paper describes KIT's submission to the IWSLT 2019 Speech Translation task on two sub-tasks corresponding to two different datasets. We investigate different end-to-end architectures for the speech recognition module, including our new transformer-based architectures. Overall, our modules in the pipe-line are based on the transformer architecture which has recently achieved great results in various fields. In our systems, using transformer is also advantageous compared to traditional hybrid systems in term of simplicity while still having competent results.

1. Introduction

The Karlsruhe Institute of Technology (KIT) participated in the IWSLT 2019 Evaluation Campaign in two tasks: Speech Translation task (SLT) and Text Translation. This paper reports KIT systems for SLT task.

1.1. SLT Task

Different from previous years, this year's IWSLT SLT task focuses on the end-to-end performance of the speech translation systems on two datasets, or as we called them two subtasks: How2[1] and TED. This makes ways for building end-to-end systems as well as multi-modal systems since the participants can use different modalities such as speech, text and even video in training the translation systems. For the TED dataset, speech, alignment information and text data are provided within the MuST-C corpus[2] in English-German. For the How2 dataset, which is built based on the instructional videos and their transcripts in English and Portuguese, the visual-grounded information is also provided.

We have still, however, conducted our SLT systems for both datasets in a piper-lined manner: The input would go through our Speech Recognition module, and then go through segmentation and normalization prior to translation. Excepts the machine translation module which employs multilingual models, other modules use single input and output modalities corresponding to the task that they need to perform.

In this evaluation campaign, we built only sequence-to-sequence speech recognition (ASR) models with two different architectures. Similar to our previous year' systems, the segmentation and normalization are basically a monolingual system which translates from the disfluent, broken, uncased text (i.e. ASR outputs) to a more fluent, written-style with punctuations in order to match the data conditions of the translation system. Finally, our translation system have been implemented as a multilingual system using Transformer architecture on all the data we have. Thus, it could cover all the translation directions for both sub-tasks with a little bit further fine-tuning for each target languages.

This paper is structured as follows: In Section 2, we describe different speech recognition architectures we employed in the campaign. Afterwards, we give a detailed description of the segmentation approach in Section 3. Our multilingual machine translation system is described in Section 4. In Section 5 we report our results and give some insights on them. Finally, we draw our conclusions in the Section 6.

2. Speech Recognition

Data Preparation We used different training datasets and feature extraction approaches for the two SLT tasks. For TED translation, we collect the audio from the TED-LIUM and How2 corpora and then extract 40-dimensional log scale mel filterbank to generate input features for ASR training models. For How2 translation, we used only the data and the audio extracted features provided by the organization which contains 40 filterbanks coefficients plus the addition of 3 pitch features. To generate labels for sequence-to-sequence ASR models, we used the SentencePiece toolkit to train and generate 4000 different byte-pair-encoding (BPE) for all models.

Modeling In this year's evaluation, we have used only sequence-to-sequence encoder-decoder ASR models. We have investigated two different network architectures: long short-term memory (LSTM) and the Transformer. We follow the network architecture in [3] to construct LSTM-based models which consist of 6 bidirectional layers of 1024 units for the encoder and 2 unidirectional layers for the decoder. For the transformer-based models, we adopted the implementation presented in [4]. Basically, these transformer-based models take the audio features as the inputs, concatenate 4 consecutive features before combining them with the position information and putting them to the self-attention blocks. The architecture of our ASR transformer-based models is described in Figure 1, with totally 32 blocks for the encoder and 12 blocks for the decoder. To effectively train this deep architecture, beside other standard regularization techniques, we employed Stochastic Layers in our models. For more details, please refer to [3, 4].

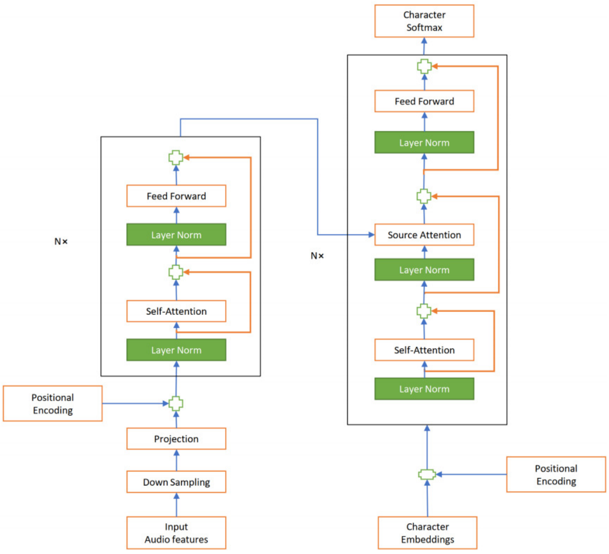

Figure 1: *Our ASR Transformer-based Architecture.*

3. Segmentation and Normalization

The output of Automatic Speech Recognition (ASR) systems is traditionally generated without the information about punctuation which segments the streams into sentences or phrases, and casing (for example proper names or words that are supposed to be upper cased). More importantly, the basic sequential units in ASR - utterances are not on the same scale with sentences. On the other hand, machine translation models are conscious about these phenomenons which leads to the mismatch in the interface between ASR and MT. Monolingual translation models are developed as a solution to smooth out the transition between ASR and MT.

As the name suggests, these models restore the casing and punctuation information from the ASR transcriptions. The training data, on the other hand, can be created from existing monolingual text corpus of the recognized language. The corpus is first randomly shuffled into randomized segments, each of which contains from 20 to 30 words. For the sequence-to-sequence model which learns to restore the casing information, the model inputs are the segments with the punctuations being removed and the words being lowercased. The target side is simplified into a set of output units including casing to be restored and punctuation to be added. Starting from the input text, we replace each word with its casing (**U** for upper-cased or **L** for lower-cased). The punctuation marks that follow the word in the text (pre-tokenized) are directly attached to **U** or **L**.

The training data is the combination of the EPPS corpus, the News-Commentary corpus and the paraCrawl corpus (which we filtered using model pre-training as for Machine Translation). The model configuration is a Transformer model with 12 layers with the base model size of 512 and the inner-size of the feed-forward layers is 1024.

It is crucial that the decoding process requires the sliding window technique as described in [5]. The data is "cloned" with segments of length 10 starting with every word in the data. The monolingual translation model then generates the target features (U or L with punctuations) for these segments. The final step is to aggregate the information from the cloned segments. The punctuation mark is added to the final sequence if there is at least one punctuation generated after the particular word in any of the cloned sequences. If different punctuation marks are predicted, we take the most frequent one. Finally, if the punctuation mark is an end of sentence punctuation mark ”.”,”!”,”?”, we also start a new segment. The segmented test data with case and punctuation information is passed on to the machine translation system. The main difference between TED Talks and How2 test sets is that, the latter is pre-segmented by the organizer. Therefore, re-segmentation is only required for TED Talks.

4. Machine Translation

Data Preprocessing. The training data includes the indomain corpora for TED translation (TED Talks) and How2 video translation tasks. We utilize the ability of the large neural models to translate between multiple languages [6, 7] by mixing the English-German and English-Portuguese data into one single training corpus. The English-German data is comprised of the Europarl, News Commentary, Rapid, Common crawl and OpenSubtitles. The data is enhanced by the massive amount of Paracrawl which was filtered by pre-training a translation model to identify the low quality sentence pairs [8]. Moreover, we pre-trained a German-

English model to back-translate 39M sentences from the German news monolingual data provided by WMT 2018. The back-translation data is generated using the sampling approach as examined in [9].
For the English-Portuguese part, we also used the Europarl, OpenSubtitles corpora and the data from the How2 video translation dataset for training. The validation data interestingly has both English-Portuguese and English German directions which makes the latter out-of-domain.
All data is preprocessed using tokenization, true-casing and BPE-splitting. One single BPE model is trained for all three languages with 40K codes.

Modeling. The main models are built based on the Transformer [10] with self-attention encoder and decoder. Following the work of deep Transformer models [8, 4], the Transformer model was implemented with the 'Big' configuration (1024 model size, 4096 inner size for the feed-forward networks and 16 attention heads) and its depth is enhanced to 12 layers. The learning rate is gradually increased in 8000 steps and then linearly decreased. The model is trained for 400000 steps, each of which takes 25000 tokens into one single update. Regularization is applied to the model with Dropout probability of 10% and label smoothing with ϵ of 0.1. In order to control the target language (German or Portuguese), we simply use a simple token for each language to start the sequence accordingly. During the decoding, the beam size is 8 and we found that just normalizing the probabilities without length penalty gave us the best performance.

Domain and language adaptation. Domain is an important factor to consider improving the performance of the translation model which is initially trained on a large amount of data. We applied fine-tuning on the TED Talks for English-German and How2 dataset for English-Portuguese. We observed that fine-tuning basically made the model forget the other language when it focuses on the main language. Notably, this process ends up with two different models specialized for German and Portuguese specifically.

Noise adaptation. One of the main difficulties of training SLT cascade models is that the model has to process the natural speech transcribed by the ASR model. In order to somewhat simulate this condition during training the MT models, the source sentences are corrupted following the SwitchOut algorithm [11] which randomly samples the number of corrupted positions, and also randomly samples a substitution from the vocabulary for each position. The SwitchOut coefficient was set to 0.95.

5. Experiments

5.1. Automatic Speech Recognition

In Table 1, we provide the word-error-rate (WER) performance of our ASR systems for both TED talk and How2 evaluation sets. The best WER system is the ensemble of LSTM-based and Transformer-based sequence-to-sequence models in which, we achieved 4.1% and 10.6% WERs respectively for two translation tasks.

Table 1: ASR performance on tst2015 and how2 tst2019 sets

Model	tst2015	How2
LSTM-based	4.5	11.5
Transformer-based	6.5	12.5
Ensemble	4.1	10.6

5.2. Machine Translation

For text-based translation performance, our Transformer model was able to reach 32 BLEU points on the newstest2017 (English-German) test set which showed 2 BLEU points improvement compared to the deep model in [8].

Table 2: SLT BLEU scores on tst2014 (En-De)

ASR	MT	BLEU
Hybrid	Deep Base Transformer	22.6
S2S	Deep Big Transformer + Adaptation	25.2
S2S	+ SwitchOut	25.7

Table 2 shows that the upgraded speech recognition system combined with the new big Transformer was able to improve the overall SLT performance by 2.6 BLEU points compared to last year. More importantly, we showed the gain of 0.5 BLEU points by using SwitchOut to make the model more robust to the disfluency of natural speech.

Table 3: SLT BLEU scores on How2 (En-Pt)

ASR	MT	BLEU
Oracle	Deep Big Transformer	58.0
Oracle	Deep Big Transformer (SwitchOut)	60.0
S2S	Deep Big Transformer (SwitchOut)	46.2

For the How2 dataset, there are both German and Portuguese translations for the development data. For the Portuguese side, since the model has access to the indomain training data, it was able to reach 58 BLEU points on text translation (the Oracle speech output as in table 3), which was further improved by SwitchOut to 60. However the performance for English-German on this dev set is only 13.0 BLEU points, which showed that the model is not robust enough to transfer the knowledge from Portuguese to German, which is inline with other multilingual translation research [7] in which translation from other languages to English is more favourable. We attempted to generate the German translation for the training data, which however does not help the model to generalize for this particular domain and language. In conclusion, we achieved adapted models which are very competitive for the specialized tasks (TED translation and How2 video translation from English to Portuguese), but further works are necessary to improve the model robustness, especially when adaptation makes the models forget the other languages very quickly. We also found out that the models with SwitchOut greatly outperform other variations (such as Word Dropout [12]). This is the reason why our final submission does not include model ensembling or rescoring.

6. Conclusions

We have built a pipe-lined system for IWSLT19 evaluation campaign's SLT task. We have conducted the speech recognition experiments with two different end-to-end architectures. The final model is the emsemble of those two architecture models, where it has achieved the best results on the development sets of SLT sub-tasks. For the machine translation part, we have employed a Transformer-based multilingual model, thus, we are able to produce the translations of all the sub-tasks with a single model.

For the future work, we would like to exploit the potential of a multi-modal end-to-end speech translation system using transformer architectures and compare it with our pipe-lined systems.

7. Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement n°825460 and the Federal Ministry of Education and Research (Germany)/DLR Projektträger Bereich Gesundheit under grant agreement n° 01EF1803B.

8. References

[1] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," *arXiv preprint arXiv:1811.00347*, 2018.

[2] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June 2019.

[3] T.-S. Nguyen, S. Stueker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," 2019.

[4] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Muller, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," *arXiv preprint arXiv:1904.13377*, 2019.

[5] E. Cho, J. Niehues, and A. Waibel, "Segmentation and punctuation prediction in speech language translation using a monolingual translation system," in *International Workshop on Spoken Language Translation (IWSLT) 2012*, 2012.

[6] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viegas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *CoRR*, vol. abs/1611.04558, 2016.

[7] T.-L. Ha, J. Niehues, and A. Waibel, "Toward multilingual neural machine translation with universal encoder and decoder," in *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, USA, 2016.

[8] N.-Q. Pham, J. Niehues, and A. Waibel, "The karlsruhe institute of technology systems for the news translation task in WMT 2018," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 467–472. [Online]. Available: https://www.aclweb.org/anthology/W18-6422

[9] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, "Facebook FAIR's WMT19 news translation task submission," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 314–319. [Online]. Available: https://www.aclweb.org/anthology/W19-5333

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[11] X. Wang, H. Pham, Z. Dai, and G. Neubig, "Switchout: an efficient data augmentation algorithm for neural machine translation," *arXiv preprint arXiv:1808.07512*, 2018.

[12] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Neural Information Processing Systems Conference (NIPS)*, Barcelona, Spain, 2016.