The USTC-NEL Speech Translation system at IWSLT 2018

Dan Liu[1,2], *Junhua Liu*[1,2], *Wu Guo*[1]
Shifu Xiong[2], *Zhiqiang Ma*[2], *Rui Song*[2], *Chongliang Wu*[2], *Quan Liu*[2]

University of Science and Technology of China [1]
IFLYTEK Co. LTD. [2]

{danliu,jhliu}@mail.ustc.edu.cn wuguo@ustc.edu.cn
{danliu, jhliu, sfxiong, zqma2, ruisong, clwu4, quanliu}@iflytek.com

Abstract

This paper describes the USTC-NEL (short for "National Engineering Laboratory for Speech and Language Information Processing University of science and technology of china") system to the speech translation task of the IWSLT Evaluation 2018. The system is a conventional pipeline system which contains 3 modules: speech recognition, post-processing and machine translation. We train a group of hybrid-HMM models for our speech recognition, and for machine translation we train transformer based neural machine translation models with speech recognition output style text as input. Experiments conducted on the IWSLT 2018 task indicate that, compared to baseline system from KIT, our system achieved 14.9 BLEU improvement.

1. Introduction

Conventional speech translation systems consist of three components: source-language automatic speech recognition (ASR), post-processing over ASR outputs, and source-to-target text translation. This pipeline system suffers from error accumulation, which means speech recognition and translation models trained separately may perform well individually, but do not work well together because their error surface do not compose well [1].

In the most recent years, end-to-end speech translation based on encoder-decoder with attention mechanisms has been very promising for reducing accumulated errors [2, 1, 3]. However, parallel speech data is much smaller than those available to train text-based machine translation (MT) systems, particularly neural systems that needs to learn a relatively large parameters. As a result, an end-to-end speech translation system can often outperform pipeline systems with same training data, but is hard to beat pipeline system with dozens of training data [1].

In addition, to translate very long speech (e.g. translate a full talk), an end-to-end system must rely on voice activity detection (VAD) method to split raw audio into sentence-like fragments, in which mis-segmented sentence fragments are very likely to cause serious translation errors. Therefore, for pipeline systems, sentence re-segmentation based on ASR results may be done in post-processing step, which can improve performance significantly [4].

To reduce the error accumulation of pipeline systems, we introduce a data augmentation based solution to train translation model with ASR results as source directly, instead of normalize ASR results (e.g. insert punctuations, normalization for case, numerals, etc.) in post-processing. Text normalization cannot bring any new information, it just produces texts that translation system likes, and this may lead to additional errors. In our experiments, the data augmentation based solution performs significantly better than pipeline system with text normalization and end-to-end speech translation system.

This paper is organized as follows. We first describe the processing for speech and text training data in Section 2, following is our full system and training details. Our experiments are presented in Section 4.

2. Data Processing

We conduct experiments on IWSLT speech translation task [5] from English to German. All experiments were performed under requirements of IWSLT 2018 evaluation campaign speech translation task. The training data for speech recognition and translation after filtering are listed in Table 1 and Table 2.

Table 1: *speech training data.*

Corpus	# of seg.	Speech hours
TED LIUM2 [6]	92976	207h
Speech Translation	171121	272h

2.1. speech recognition training data

The speech data contains TED LIUM2 [6] and speech translation data by IWSLT evaluation campaign. In TED LIUM2, only raw wave files and manual transcriptions (without punctuation) were offered. And in speech translation data, raw wave files, English transcriptions and the corresponding German translations were offered, but some transcriptions is not

Table 2: *text training data.*

Corpus	raw	filtered
commoncrawl	2.39M	1.80M
rapid	1.32M	1.00M
europal	1.92M	1.81M
commentary	0.284M	0.233M
paracrawl	36.35M	12.35M
opensubtitles	22.51M	14.24M
WIT3(in domain)	0.209M	0.207M

match to there corresponding audio. Besides this, about 166 hours of audio in speech translation data were not labeled, we regard them as unsupervised data.

To utilize those data, we firstly train initial acoustic model based on TED LIUM2. Using this model, we do force alignment on IWSLT speech translation data, and discard utterances with significantly abnormal scores. After this process, the supervised data size of IWSLT has been reduced to 246 hours from 272 hours. Meanwhile, the unsupervised data is recognized by our initial model and filtered based on ASR confidence to expand the training set.

To further increase the amount of data in the training set, we perform data augmentation by noise and speed perturbations. For each speech signal, a noise version is created initially. Speed perturbation is then performed on the raw signals with speed factors 0.8 and 1.2. Eventually, up to (207+246+166)*4 hours of data may be used.

2.2. speech translation training data

The speech translation training data is the same as the speech recognition training data. The target references for LIUM2 and unsupervised data are generated by our best text machine translation system.

2.3. text translation training data

The text translation training data contains parallel data and monolingual training data. As for parallel data, we use all of the allowed training data for Speech Translation Task which includes TED corpus, data provided by WMT 2018 and OpenSubtitles2018 [7]. The data is pre-processed before training and translation. Sentences longer than 100 words and duplicated sentence pairs are removed. Also, numbers are normalized in order to match the ASR outputs. NMT systems are more vulnerable to noisy training data, rare occurrences in the data, and the training data quality in general. So we measure the cross-lingual similarities between source and target sentences, and then reject sentences with similarity below a specified threshold. After filtering, we can get relevant and high quality data. The training data after filtering are listed in Table 2.

As for monolingual training data provided by WMT 2018, we clean the noisy data for English and German, and then we use the supervised convolutional neural network method [8] to select monolingual training data that are close to the TED domain. After this processes, we select 91M monolingual English data and 43M mono-lingual German data for language model training.

3. System Description

3.1. speech recognition

The primary system of our speech recognition is a hybrid-HMM system. The acoustic model contains multiple deep neural networks based on CNN and LSTM structure. State level posterior fusion technique is used for the final ASR results. The details of model structure and training criterion are as following:

1. DenseNet [9]: DenseNet with 13 dense connection blocks and 3 max-pooling steps with stride 2 on both time and frequency domain, trained with cross-entropy (CE) and sequence-discriminative training (SDT) criterion [10].

2. BiLSTM [11]: 3 layers BiLSTM network trained with CE and SDT criterion.

3. CLDNN [12]: CNN-BiLSTM-DNN structure trained with CE and SDT criterion.

The language models are trained on English monolingual data described in Section 2.3. The first-pass decoding is performed with the HMM and 3-gram LM. A 4-gram LM is used for second-pass decoding and followed by a LSTM-based LM.

In this task we should do speech recognition on full talk, so we have to split the raw audio into sentence-like pieces for speech recognition. We do speech segmentation with LSTM based VAD model, which is trained on TED LIUM2 dataset with speech/nonspeech labels extracted by force alignment with our hybrid-HMM model.

3.2. post-processing vs data augmentation

It has been shown that post-processing is crucial for achieving good speech translation performance [4], this comes from two aspects. First, segmentation boundaries for ASR are based on VAD, which inevitably leads to fragments with incomplete semantics, and sentence re-segmentation based on ASR results is needed. Second, translation models are trained with written text as input, which means text normalization of ASR results is essential for conventional systems.

We know punctuations may contain rich semantic information, but in post-processing for speech translation, punctuations are only generated from ASR output word sequences. In this case, these punctuations can not bring more information than words. The main goal of post-processing is just to produce text suitable for machine translation. However, it should be noted that errors in punctuation prediction may be propagated in machine translation process.

Here we introduce a new solution with respect to mismatch between ASR results and machine translation inputs. Instead of transform ASR results to written text on decoding step, we transform the source text for machine translation training data into the style of ASR results on training step. The difficulty of normalizing ASR results to written text seems equal to the difficulty of normalizing written text to ASR results. However, data augmentation with fake ASR results for machine translation is more robust for errors compared to text normalization on decoding step.

We train a neural machine translation (NMT) model to translate written text into ASR results. To build the training data, we process the English written data by rule (remove punctuations, lower case and translate Arabic numerals into English words), the generated text is similar to ASR results except for recognition errors. We also build real data with the ASR results and source written texts provided in speech translation dataset. The NMT model from written text to ASR results are trained on these two dataset and fine-tuned on only real data. This model may generate ASR output style text with common ASR errors. And we augment the text machine translation dataset by translating the source written texts into ASR output style texts. As a comparison, we also trained an inverted NMT for text normalization.

The data augmented based solution can translate directly from ASR result, which reduces errors caused by text normalization. Besides this, our model has the ability to tolerate common recognition errors. E.g., our ASR system may mistake "two" to "to" in some special contexts, and our NMT system may translate "top to percent" to "top zwei Prozent".

Sentence re-segmentation are still important to speech translation system, because training data for machine translation are all semantically complete sentences. Data augmentation with semantic incomplete sentence fragments may suffer from reordering between source and target language. So we train a LSTM based model to re-segmented sentences based only on text infomation. This model is trained on TED and OpenSubtitle dataset, with one whole paragraph as input, and the punctuation ".!?" as sentence boundaries.

3.3. machine translation

3.3.1. text machine translation

Transformer [13] is adopted as our baseline, all experiments use the following hyper-paramter settings based on Tensor2Tensor transformer_relative_big settings [1]. This corresponds to a 6-layer transformer with a model size of 1024, a feed forward network size of 8192, and 16 heads relative attention. Model is trained on the full dataset described in Section 2.3 and fine-tuned on speech translation dataset. We trained both conventional NMT model and NMT model with augmented data described in Section 3.2.

3.3.2. end-to-end speech translation

For our end-to-end speech translation model, DenseNet described in Section 3.1 followed by one BiLSTM layer is employed as encoder, and the decoder is same as transformer model in Section 3.3.1. It is difficult to train speech translation model from random initialization parameters, for reordering between source and target language are difficult to align with frame based speech representations. Pre-training with speech recognition task significantly improves the performance. And this encoder-decoder based ASR model is used for rescoring our final ASR results.

End-to-end speech translation system has no chance to re-segment sentences. We found splicing audio segments acquired by VAD may improve the translation performance, but still has a significant gap to performance based on sentence re-segmentation.

4. Experimental Results

In this section, we present a summary of our experiments for the IWSLT 2018 speech translation evaluation task. We test WER (word error rate) for our speech recognition system on dev2010, which is the only dataset with CTM format transcriptions. And we test our speech translation systems on IWSLT dev2010, tst2010, tst2013, tst2014 and tst2015. Case sensitive BLEU based on realigning system outputs to reference by minimizing WER [14] is used for our speech translation evaluation metric.

4.1. Results of Speech Recognition

In this section, we demonstrate the results of our ASR system. The acoustic model of our primary system is the deep CNN model, and we decode with 3-gram for first-pass decoding and 4-gram for second-pass. We test our performance in dev2010. First, we compare the impact of training data in Table 3. Here "spv." represents supervised data, "usv." represents unsupervised data and "spd." represents speed disturbed data. As show in Table 4, by training with noisy data, the WER is relatively reduced by 7.32%.

Table 3: *WER for speech recognition with different training data on dev2010*

Training Data	WER
spv.	9.7
noisy spv.	8.99
noisy spv. usv.	8.92
noisy spd. spv. usv.	8.86

Based on the above results, we train three acoustic models with different structures. Further promotion is achieved by fusing multiple acoustic models, rescoring with RNN-LM. We also test the encoder-decoder based speech recognition model described in Section 3.3.2, which performs significantly worse than our hybrid-HMM systems. But rescoring

[1] https://github.com/tensorflow/tensor2tensor/tree/v1.6.3

with encoder-decoder system brings a small improvement. Details are showed in Table 4.

Table 4: *Results of fusion of different models for speech recognition on dev2010 .*

DenseNet	8.86
BiLSTM	8.72
CLDNN	8.40
Encoder-Decoder	14.64
DenseNet +BiLSTM + CLDNN	8.22
+ RNN	7.61
+Encoder-Decoder	7.3

4.2. Results of End-to-end Speech Translation

In this section, we describe our experiments on end-to-end speech translation. The average BLEU score of our baseline end-to-end speech translation system is 20.50, which is significantly worse than our pipeline system (Tabel 6). The degradation comes from two aspects. Firstly, our encoder-decoder speech recognition performs worse than baseline speech recognition system (WER 7.61% to14.64%). Secondly, the end-to-end system has no chance to re-segment sentences based on source recognition results.

To reduce the influence of incomplete sentence fragments caused by VAD, we splice the VAD fragments to at least 10 seconds, which brings the improvements of about 1 BLEU. For comparison, we present the performance of a system that re-segment audio based on speech recognition results, which brings another 1.3 BLEU gain, but this is not a ”end-to-end” system. At last, the ensemble of 4 different models improves about 1 BLEU compared to corresponding single model. The details are showed in Table 5.

4.3. Results of Pipeline Speech Translation

In this section, experiments are all based on the best ASR results described in Section 4.1. At test time, we use a beam size of 80 and a length penalty of 0.6. All data used for training are described in Section 2. All reported scores are computed using IWSLT speech translation evaluation metric.

4.3.1. post processing

The post processing procedure.includes two parts: sentence re-segmentation and text normalization. And we introduced one data augmentation based solution to remove text normalization. We compare the performance for different solutions in Table 6.

We see sentence re-segmentation has a huge impact on performance. Since sentence-like pieces obtained by VAD do not carry any semantic information, it is very unfavorable for machine translation. Other than this, our data augmentation based solution achieves a average BLEU score of 28.76, 1.3 BLEU higher over system with post processing. And we found the models with text regularization and data augmentation can be combined to get better results.

4.3.2. fusion of different models

We train 3 groups of different models, one for text regularization and two for data augmentation (L2R and R2L, which denotes the target order left to right and right to left). For each group we train 4 models with different initialized parameters, and decoded with the ensemble models to get 80-best hypotheses with beam size of 80. The 3 groups of hypotheses are merged and rescored by all translation models, target language model and end-to-end speech translation model. Performances are shown in Table 7.

4.4. Submission Results

We submitted 3 systems for speech translation task. The primary system is the best fusion system demonstrated at row 7 in Table 7, and the contrastive systems are all based on encoder-decoder model from audio features. Contrastive0 is based on sentence re-segmentation with source speech recognition results, which is not real ”end-to-end”, while contrastive2 is real end-to-end systems with only single model.

We compared our submitted systems to KIT speech translation system (noted as ”Baseline_KIT”)[2], which is the baseline system provided by KIT, performance is shown in Table 7. Our primary system achieves a average BLEU of 30.26, which is 14.9 BLEU higher than baseline from KIT.

5. Conclusion

In this paper we presented our speech translation systems for IWSLT 2018 evalution. Our results indicated that the end-to-end system still performs significantly worse than the conventional pipeline system, and NMT with data augmentation performs better than solutions with text regularization. Our best ensemble system achieved 14.9 BLEU improvement compared to baseline system from KIT.

6. References

[1] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, “Sequence-to-sequence models can directly translate foreign speech,” *arXiv preprint arXiv:1703.08581*, 2017.

[2] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, “An attentional model for speech translation without transcription,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 949–959.

[3] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, “End-to-end automatic speech transla-

[2] https://github.com/jniehues-kit/SLT.KIT

Table 5: *BLEU scores for end-to-end speech translation .*

system	dev2010	tst2010	tst2013	tst2014	tst2015	average
VAD	21.45	21.41	21.76	20.06	17.83	20.50
splice 10s	22.14	22.16	22.76	21.00	19.52	21.52
re-segment	23.79	24.18	24.18	22.22	20.07	22.89
ensemble(splice)	23.43	22.97	23.58	21.96	20.67	22.52
ensemble(re-segment)	24.78	24.92	25.41	23.23	21.01	23.87

Table 6: *BLEU scores for pipeline speech translation system*

re-segment	text regularization	data augmentation	dev2010	tst2010	tst2013	tst2014	tst2015	average
N	Y	N	26.47	27.71	28.04	25.65	24.00	26.37
N	N	Y	27.58	28.26	29.81	26.79	25.65	27.62
Y	Y	N	27.75	28.90	29.01	26.88	24.52	27.41
Y	N	Y	28.98	29.98	30.69	28.19	25.99	28.76

Table 7: *BLEU scores for fusion systems*

system	dev2010	tst2010	tst2013	tst2014	tst2015	average
text normalization	28.64	29.41	29.59	27.37	25.13	28.03
augment L2R	29.45	30.01	30.78	28.37	26.14	28.95
augment R2L	28.42	29.58	30.88	27.98	26.47	28.66
fusion	30.28	31.01	32.28	29.38	27.40	30.07
+target LM	30.30	31.00	32.37	29.44	28.14	30.25
+e2e model	30.50	31.06	32.31	29.35	28.06	30.26

Table 8: *performance of submitted systems*

system	end2end	single model	dev2010	test2010	test2013	test2014	test2015	average
Baseline_KIT	N	Y	17.07	12.37	16.59	15.42	15.15	15.32
PRIMARY	N	N	30.50	31.06	32.31	29.35	28.06	30.26
Contrastive0	N	N	24.78	24.92	25.41	23.23	21.01	23.87
Contrastive2	Y	Y	22.14	22.16	22.76	21.00	19.52	21.52

tion of audiobooks," *arXiv preprint arXiv:1802.04200*, 2018.

[4] E. Cho, J. Niehues, and A. Waibel, "Nmt-based segmentation and punctuation insertion for real-time spoken language translation," *Proc. Interspeech 2017*, pp. 2645–2649, 2017.

[5] C. Mauro, F. Marcello, B. Luisa, N. Jan, S. Sebastian, S. Katsuitho, Y. Koichiro, and F. Christian, "Overview of the iwslt 2017 evaluation campaign," in *International Workshop on Spoken Language Translation*, 2017, pp. 2–14.

[6] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks." in *LREC*, 2014, pp. 3935–3939.

[7] P. Lison, J. Tiedemann, and M. Kouylekov, "Opensubtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora," in *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan.(accepted)*, 2018.

[8] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[10] K. Vesel`y, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Interspeech*, 2013, pp. 2345–2349.

[11] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[12] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[14] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating machine translation output with automatic sentence segmentation," in *International Workshop on Spoken Language Translation (IWSLT) 2005*, 2005.