The HW-TSC's Offline Speech Translation System for IWSLT 2022 Evaluation

Yinglu Li[1], **Minghan Wang**[1], **Jiaxin Guo**[1], **Xiaosong Qiao**[1], **Yuxia Wang**[2], **Daimeng Wei**[1], **Chang Su**[1], **Yimeng Chen**[1], **Min Zhang**[1], **Shimin Tao**[1], **Hao Yang**[1], **Ying Qin**[1]
[1]Huawei Translation Services Center, Beijing, China
[2]The University of Melbourne, Melbourne, Australia
{liyinglu,wangminghan,guojiaxin1,qiaoxiaosong,weidaimeng,suchang8,chenyimeng,zhangmin186,taoshimin,yanghao30,qinying}@huawei.com
yuxiaw@student.unimelb.edu.au

Abstract

This paper describes the HW-TSC's designation of the Offline Speech Translation System submitted for IWSLT 2022 Evaluation. We explored both cascade and end-to-end system on three language tracks (en-de, en-zh and en-ja), and we chose the cascade one as our primary submission. For the automatic speech recognition (ASR) model of cascade system, there are three ASR models including Conformer, S2T-Transformer and U2 trained on the mixture of five datasets. During inference, transcripts are generated with the help of domain controlled generation strategy. Context-aware reranking and ensemble based robustness enhancement strategy are proposed to produce better ASR outputs. For machine translation part, we pretrained three translation models on WMT21 dataset and fine-tuned them on in-domain corpora. Our cascade system shows more competitive performance than the known offline systems in the industry and academia.

1 Introduction

In recent years, end-to-end system and cascade system are fundamental pipelines for speech translation tasks. Traditional cascade system is comprised of continuing parts, automatic speech recognition (ASR) is responsible for generating transcripts from audios and machine translation model aims at translating ASR outputs from source language into target language. Obviously, the ASR part and MT part of this system are independent to some extent. Therefore, this paradigm enables people to utilise state-of-the-art ASR models and MT models and conduct experiments by different permutations and combinations. And those experiments can help us find the best combination of choice of ASR and MT model. ASR model like Conformer (Gulati et al., 2020) and S2T-Transformer (Synnaeve et al., 2019) are commonly used. MT models like Transformer (Vaswani et al., 2017) can be considered as a standard configuration.

On the contrary, there is also a disadvantage when applying cascade systems. The main aspect is that some important information such as the intonation and emphasis of speakers could not be explicitly expressed in the transcripts. This "missing information" might be the key to distinguish the gender of speaker, or the sarcasm and symbolism behind the texts. It means, there is a risk of losing important information under the condition of cascade system.

Correspondingly, end-to-end system preserves the competitive edge to learn the "missing information", because it is directly trained on the speech-to-text dataset without any transit process. Due to this property, end-to-end system has been paid attention in research and there is encouraging progress. For instance, Conformer (Gulati et al., 2020) can also be used in this task. However, there are some disadvantages for the end-to-end system. Firstly, due to the lack of large scale high quality bilingual speech translation datasets, training a productive end-to-end ST model can be non-trivial. Next, the mapping from speech space to the target language space is far more difficult than the mapping to the source language space, leading to greater demand on the scale of the training set.

This paper presents our work in IWSLT 2022 (Anastasopoulos et al., 2022) offline speech translation track. The main contribution of this paper can be summarized as follows:

1) We tested various combinations of ASR models, and finally found ensemble of Conformer and S2T-Transformer and filter by U2 can improve the ASR fluency and sentence expression.

2) Context-aware LM reranking can effectively improve the possibility to choose the best candidate in beam search.

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 239 - 246

Dataset	Number of Utterance	Duration(hrs)
LibriSpeech	281,241	960.85
MuST-C	340,421	590.67
IWSLT	170,229	254.41
CoVoST	1362,422	1802.52
TEDLIUM3	268,214	453.42

Table 1: Data statistics of our ASR corpora

Language	WMT Bilingual	In-domain Text
En-De	79M	459K
En-Zh	96M	590K
En-Ja	42M	552K

Table 2: Data statistics of our MT corpora

2 Method

2.1 Data Preparation and Preprocessing

There are five different datasets used in the training of our ASR models and ST models, such as MuST-C V2 (Cattoni et al., 2021), LibriSpeech(Panayotov et al., 2015), TED-LIUM 3 (Hernandez et al., 2018), CoVoST (Wang et al., 2020), IWSLT, as described in the left sub-plot of Figure 1. For the training dataset we extracted 80-dimensional filter bank features from the raw waveform firstly. Then, the dataset was cleaned in a fine-grained process. The training set was filtered on the criteria of absolute frame size (within 50 to 3000), number of tokens (within 1 to 150) and speed of the speech (within $\mu(\tau) \pm 4 \times \sigma(\tau)$), where $\tau = \frac{\#\text{ frames}}{\#\text{ tokens}}$. The detailed attributes such as the number of utterance and the duration of training datasets are shown in table 1. For test set, each TED talk was segmented into several utterances (no more than 20 seconds) with the officially provided segmentation tool (LIUM_SpkDiarization.jar).

We use the exactly same corpus to train our MT models following the configuration of (Wei et al., 2021), with the scale of the dataset showing in Tabel 2.

2.2 Automatic Speech Recognition

There are three types of basic ASR models Conformer (Gulati et al., 2020), S2T-Transformer (Synnaeve et al., 2019) and U2 (Zhang et al., 2020) used to recognize the speech and get transcripts. The first two models are standard autoregressive ASR models built upon the Transformer architecture (Vaswani et al., 2017). The last one is a unified model that can perform both streaming and non-streaming ASR, supported by the dynamic chunking training strategy (Zhang et al., 2020). During the training and decoding process, there are three important strategies we used to generate ASR results of these models as follows.

Domain controlled training and decoding By observing the corpus in the training set, we find that the style of text and the domain of the speech can be different between each dataset. Although the model is able to learn such difference implicitly, there are still some confusing patterns like case sensitivity and existence of punctuation that can not be easily learned. Therefore, we add the domain tag as the prefix token, acting as a known condition to guide the model to generate texts in required domain and style. It means, the model can learn the pattern given more prior knowledge. For example, the tag "<MC>" provides an instruction to the model to generate texts in the MuST-C style, or we can also use <LS> to make the model to generate LibriSpeech alike transcripts. The strategy also had a positive effect in our offline task submission of IWSLT 2021 (Wang et al., 2021). For Conformer and S2T-Transformer, since they are autoregressive generative models, we simply use the domain tag as the prefix token. However, this is not feasible for U2 with the CTC decoder. Therefore, we propose to first encode the domain tag with the input-embedding of the attention-based decoder of U2, then, adding the encoded tag to the down-sampled features element-wise, being together fed into attention layers of the encoder.

Context-aware LM reranking In order to take benefits from both Conformer and S2T-Transformer which has different model architecture, we ensemble them by averaging the predicted probabilities while generation. However, the ensemble doesn't solve a key problem comes from the independence assumption on each utterance. In other words, we translate each utterance in a TED talk speech independently without considering context information, which often cause inconsistent prediction on named entities such as person names. To this end, we adopt a language model (LM) to rerank beam candidates conditioned on a fixed length window of generated contexts.

Specifically, a Transformer-LM was trained on

Algorithm 1 Context-aware LM reranking

Require: ASR, LM, context length, beam size, utterance list: ϕ, Q, N, k, U
  Initialize: Context Buffer $C \leftarrow \{\}$
  Initialize: utterance index $i \leftarrow 0$
  while $i \neq |U| - 1$ **do**
    $\hat{Y}, P_\phi \leftarrow \phi(u_i, k)$: propose candidates
    if $i < N$ **then**
      $P_Q \leftarrow Q(\hat{Y}, C)$
    else
      $P_Q \leftarrow Q(\hat{Y}, C_{[-N:]})$
    end if
    $\hat{y}^* \leftarrow \arg\max_{\hat{y}} \sum_{m \in \{Q,\phi\}} w_m P_m$
    $C \leftarrow C \cup \{\hat{y}^*\}$
    $i \leftarrow i + 1$
  end while
  return C

the WMT21 monolingual English dataset, providing the perplexity score of each ASR beam candidate from the ensemble models by taking N previous generated sentence into account, ($N = 3$ obtains the best result). This method is commonly used to optimize document-level translation (Yu et al., 2020). A detailed explanation is presented in Algo 1 and the right sub-plot of Figure 1, which actually works like performing context-aware greedy search in the sentence-level. Besides the PPL (converted to the log probability) estimated by the LM, we also take the log probability of each beam candidate output by ASR models into account, combining them with a weighted sum (best combination searched in the experiment: $w_{LM} = 0.6, w_{ASR} = 0.4$).

Ensemble based robustness enhancement strategy Compared with ASR results generated from different ASR models, an interesting pattern can be found that U2 prefers to predict blank lines when facing with some hard samples. Hard samples, such as laughter and applause always confused S2T-Transformer and Conformer and they are more likely to output incorrectly. For instance, S2T-Transformer always outputs "*thank you very much indeed*" and Conformer generates "*There's many a slip, twixt cup and the lip.*" when the input is the audio which contained only the applause of audiences. This phenomenon can be explained by the reason that U2 is more robust to interference than S2T-Transformer and Conformer. Consequently, the strategy that U2 could be utilised to filter the noise of ASR results from Conformer and S2T-Transformer. In other words, we extracted the blank lines of prediction of U2 as the standard to correct the results of other two models. The process provides our system with more robustness to non-speech or background noise.

2.3 Machine Translation

In an cascade system, the input of machine translation (MT) model is the ASR results. In order to obtain the translated results, we use the WMT21 news corpora to train three individual MT models for each language (En-De, En-Zh, En-Ja). Then these MT models are fine-tuned on the combination of MuST-C and IWSLT dataset. After applying the MT models on the ensembling ASR results above, the final results, also called hypothesis were obtained in our experiment.

2.4 Multilingual E2E-ST

In the ene-to-end system, the ASR model and machine translation model trained on bilingual corpora are not the continents of the system. The E2E model can be directly trained on the bilingual/multilingual speech corpora. However, only MuST-C and COVOST provides the translation of some language pairs, which might not be enough. Therefor, we propose to use the MT model to generate translations in specific language for all ASR training corpora, and then combined them together including the ASR (English) text, tagged with domain and language abbreviations like "<MC_en>", "<LS_zh>", etc. This is commonly considered as sequence level knowledge distillation (KD) (Kim and Rush, 2016). Next, a multilingual speech translation (ST) model is trained on the corpora, which can be used in both ASR and translation in an end-to-end paradigm by giving required language and domain tag.

3 Experiments

3.1 Settings

Model Configurations Sentencepiece (Kudo and Richardson, 2018) is utilised for tokenization on ASR texts with a learned vocabulary restricted to 20000 sub-tokens. ASR models are configured as: $n_{\text{encoder\_layers}}$ = 16, $n_{\text{decoder\_layers}}$ = 6, n_{heads} = 16, d_{hidden} = 1024, d_{FFN} = 4096 for Conformer, $n_{\text{encoder\_layers}}$ = 12, $n_{\text{decoder\_layers}}$ = 6, n_{heads} = 16, d_{hidden} = 1024, d_{FFN} = 4096 for S2T-Transformer and $n_{\text{encoder\_layers}}$ = 12, $n_{\text{decoder\_layers}}$ = 6, n_{heads} =

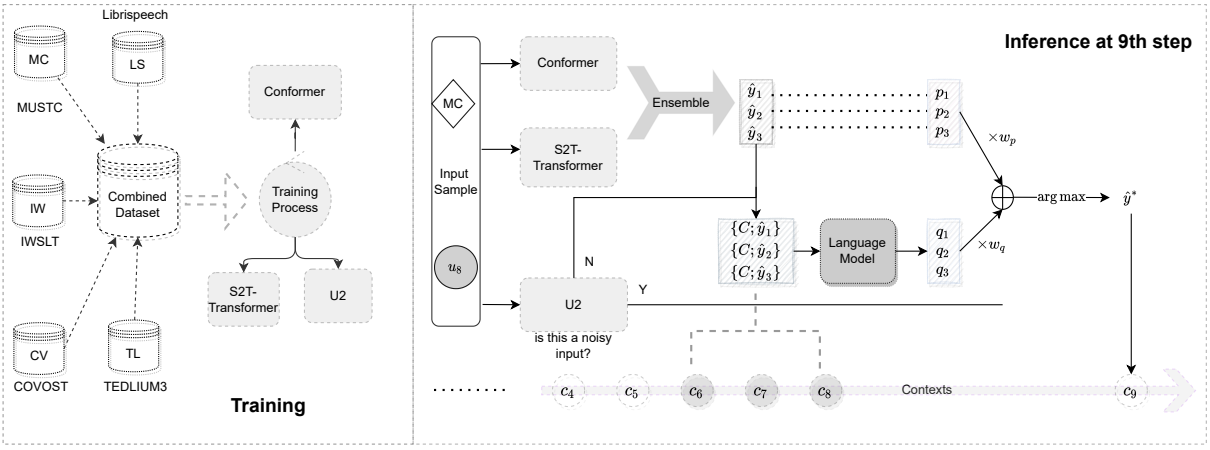

Figure 1: This figure presents the example of the training of our ASR models (left) as well as the inference of our cascade system (right). In the example of inference, input features and domain tags are feed into ASR models, being decoded by the ensemble of Conformer and S2T-Transformer and cleaned by U2. Then, beam candidates (k=3 here) are scored together with contexts (6 to 8) by the language model. Finally, the optimal candidate is selected according to modulated scores and becomes the new context.

ASR Model	CoVoST	MuST-C	TEDLIUM3	LibriSpeech
Conformer	11.27	6.31	5.33	4.39
S2T-Transformer	13.46	9.01	6.30	5.67
U2	14.68	9.71	11.93	5.79

Table 3: Comparison of wer scores of Conformer, S2T-Transformer and U2 trained on test sets of each individual dataset.

16, $d_{\text{hidden}} = 1024$, $d_{\text{FFN}} = 4096$ for U2. The NMT model has the standard Transformer-big configuration but with d_{FFN} set to 8192 (Ng et al., 2019). The language model is a standard Transformer language model with the configuration of: $n_{\text{layers}} = 12$, $n_{\text{heads}} = 16$, $d_{\text{hidden}} = 1024$, $d_{\text{FFN}} = 4096$. All models are implemented with `fairseq` (Ott et al., 2019).

During the training of ASR models, we set the batch size to the maximum of 20,000 frames per card. Inverse sqrt is used for lr scheduling with warm-up steps set to 10,000 and peak lr set as 5e-4. Adam is used as the optimizer. All ASR models are trained on 8 V100 GPUs for 50 epochs. Parameters for last 5 epochs are averaged. Audio features are normalized with utterance-level CMVN for Conformer and S2T-Transformer, and with global CMVN for U2. All audio inputs are augmented with spectral augmentation (Park et al., 2019).

We followed the work of Wei et al. (2021) on the pretraining of all NMT models. All of them are fine-tuned on in-domain corpus for 10 steps.

We use the toolkit from the SLT.KIT[1] for evaluation on all development set, which produces metrics including BLEU (Papineni et al., 2002), TER (Snover et al., 2006), BEER (Stanojevic and Sima'an, 2014) and CharacTER (Wang et al., 2016).

3.2 Results

Comparison of ASR models on each individual dataset We tested three ASR models (Conformer, U2 and S2T-Transformer) on four individual test sets, CoVoST, MuST-C, TEDLIUM and LibriSpeech. In Table 3, Conformer shows the best results in each column, which are 11.27, 6.31, 5.33 and 4.39 WERs in each dataset. It is obvious that Conformer has the significant advantage compared to other two models. However, after manually evaluating some samples, we find that Conformer is easier to over-fit the training corpora. Therefore, we decide to ensemble it with the S2T-Transformer during inference.

Comparison of our approach on past years' test sets In Table 4, we tested the performance of our cascade system on datasets of all past years, by providing 6 metrics evaluated by the SLT.KIT toolkit. By comparing these results with our last

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

SET	BLEU	BLEU (last year)	TER	BEER	CharacTER	BLEU(ci)	TER(ci)
dev2010	27.19 (+1.19)	26.00	60.61	53.10	48.27	28.73	58.21
tst2010	27.51 (+1.14)	26.37	60.66	52.57	48.90	29.13	58.14
tst2013	29.38 (-0.51)	29.89	60.94	53.70	47.07	30.7	58.83
tst2014	28 (-0.03)	28.03	61.19	52.90	47.95	28.93	59.51
tst2015	24.06 (+0.86)	23.20	77.89	50.20	50.86	24.94	76.77
tst2018	23.12 (+0.99)	22.13	73.65	51.33	51.50	23.92	71.23
tst2019	25.92	-	62.11	52.22	48.96	27.13	60.08
tst2021 (En-De)	27.5/21.2/39.9						
tst2022 (En-De)	24.2/20.8/33.5						
tst2022 (En-Zh)	34.6/33.4/42.1						
tst2022 (En-Ja)	23.3/14.3/31.0						

Table 4: Overall results comparison on dev and test sets from 2010 to this year with the full use of our strategies (The results of 2010-2019 are all in En-De). For the column of BLEU, we also presents the improvements compared to our last year's BLEU score. The lower part of the table presents our submission results in this year, values from left to right are BLEU-ref1, BLEU-ref2 and BLEU both, respectively.

years' report (Wang et al., 2021), we find that our strategy used in this year provides significant improvements on most of datasets, demonstrating their efficiency.

In order to illustrate the difference between ASR results of Conformer, S2T-Transformer and U2, we choose some representative cases in Tab 5. Case 1 presents three sentences generated from three ASR models given an audio segment which only contains background music and applause. Obviously Conformer and S2T-Transformer both outputs wrong sentences, because nothing should be generated in the decoding process. Contrarily, U2 outputs the blank line which indicates the robustness of the model itself. Case 2 provides the transcripts that Conformer and S2T-Transformer outputs the correct results. However, U2 made some mistakes on uppercase and punctuation marks even though the contents are generally correct, which shows that U2 is not sensible with case or punctuation; This actually caused by the multi-modality problem (Gu et al., 2018), which is faced by all non-autoregressive generation models. Since the prediction of each token are independently modeled in U2 (conditional independence assumption used by the CTC decoder), the prediction of tokens with one-to-many mappings (usually referred to as capitalism or existence of punctuation) can be difficult to learn without visible contexts (compared to autoregressive models). Case 3 presents that the results of Conformer and S2T-Transformer contains different errors. The Conformer misunderstood the "an ex-boyfriend" for "a next boyfriend", and S2T Transformer made a mistake on "cuss words". By fixing the different mistakes, we successfully obtain the correct sentence in the ensemble results.

3.3 Ablation

Effectiveness of context-aware reranking We investigated and demonstrated whether the context-aware ASR reranking strategy works well and the results are indicated in Table 6. As we can see, we experimented the weight combination like $w_{\text{LM}} = \{0.0, 0.5, 0.6, 1.0\}$, $w_{\text{ASR}} = \{1.0, 0.5, 0.4, 0.0\}$, and several context length including $N = \{3, 4, 5\}$.

The higher the w_{LM} is, the more contribution does the LM provides to the scoring. The ablation study shows that context length at 3 is the best choice for reranking, since the results with context length at 4 or 5 both indicates lower BLEU scores. We suspect that longer contexts often misleads the scoring processing due to the unstable estimation of PPL on beam candidates of current utterance, resulting in non-convincing reranked results. Meanwhile, we find that the best combination of the weight on LM and ASR is 0.6 and 0.4, indicating that scoring only with LM cannot always produce promising estimation on the quality of the sentence.

Performance of Translation models We used the ASR results generated from Conformer on MuST-C tst-COMMON dataset to measure the performance of two text MT models and an end-to-end ST model, i.e. the MT model pretrained

	ASR model	Sentences
Case 1	**Conformer**	*There's many a slip, twixt cup and the lip.*
	S2T-Transformer	*Thank you very much indeed.*
	U2	-
	Ensemble	-
Case 2	**Conformer**	***And I** predict that in 10 years, we will lose our bees.*
	S2T-Transformer	***And I** predict that in 10 years, we will lose our bees.*
	U2	*and i predict that in ten years we will lose our bees*
	Ensemble	***And I** predict that in 10 years, we will lose our bees.*
Case 3	**Conformer**	*... the language that a next boyfriend taught you, where you learned all the **cuss** words ...*
	S2T-Transformer	*... the language that **an ex-boyfriend** taught you, where you learned all the cusp words ...*
	U2	*... the language that an ex-boy taught you or you learned all the cus words ...*
	Ensemble	*... the language that **an ex-boyfriend** taught you, where you learned all the **cuss** words ...*

Table 5: The table presents three cases to compare the difference when generating ASR results. Those words or sentences marked by underline represents the mistakes. Case 1 shows that U2 predict more robust result than Conformer and S2T-Transformer if the input audio is filled with applause; Case 2 shows the transcripts that Conformer and S2T outputs the correct results but U2 is not sensible with uppercase and punctuation marks; Case 3 presents that the results of Conformer and S2T-Transformer both contains error, but ensemble strategy successfully obtain the correct sentence.

Hyper-Parameters	**N=3**	**N=4**	**N=5**
$w_{\text{LM}} = 0.0, w_{\text{ASR}} = 1.0$		25.12	
$w_{\text{LM}} = 0.5, w_{\text{ASR}} = 0.5$	25.66	25.65	25.70
$w_{\text{LM}} = 0.6, w_{\text{ASR}} = 0.4$	**25.92**	25.76	25.73
$w_{\text{LM}} = 1.0, w_{\text{ASR}} = 0.0$	25.58	25.48	25.52

Table 6: This table shows the BLEU score evaluated on IWSLT tst2019 En-De dataset with different combination of LM reranking weight (w) and context length (N).

Model	**En-De**	**En-Zh**	**En-Ja**
Pretrained MT	33.1	24.1	14.8
In-domain FT MT	**33.3**	**24.6**	**15.1**
Multilingual E2E ST	30.8	22.3	13.0

Table 7: This table presents the BLEU score evaluated on MuST-C tst-COMMON dataset with our pretrained and in-domain fine-tuned MT model, note that the source texts comes from the same Conformer ASR model instead of the oracle text. The last row is performance of our end-to-end multilingual ST model evaluated with the speech input.

on WMT news corpora, the in-domain fine-tuned MT model and our multilingual ST model. The in-domain FT MT was trained on the combination of MuST-C and IWSLT text corpora, providing the best BLEU scores compared with other two models. The result demonstrates that the in-domain fine-tuning is effective to generate the reasonable translation hypothesis. On the other hand, End-to-End multilingual ST proves to be a competitive model since the results are relatively close to those of the baseline pretrained MT model. More importantly, the E2E ST was only trained once on the combination of all language pairs, without further fine-tuning on any of them.

4 Conclusion

This paper presents our offline speech translation systems in the IWSLT 2022 evaluation. We explored different strategies in the pipeline of building the cascade and end-to-end system. In the data preprocessing, we adopt efficient cleansing approaches to build the training set collected from different data sources. Domain controlled generation was used in the training and decoding of ASR models to fit the requirement of the evaluation test set. We also investigated the positive effect of context-aware LM reranking aiming at improving the quality and consistency of ASR outputs. Fi-

nally, we demonstrated that the cascade system consisted of reranking ASR system and MT model has the best performance than end-to-end system. In our future works, we would like to investigate more strategies on improving the consistency of ASR outputs beyond reranking, as well as better training and data augmentation strategies for end-to-end models.

References

Antonios Anastasopoulos, Luisa Bentivogli, Marcely Z. Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Marcello Federico, Christian Federmann, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M. Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Juan Pino, Elizabeth Salesky, Jiatong Shi, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alex Waibel, Changhan Wang, and Shinji Watanabe. 2022. FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland. Association for Computational Linguistics.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Comput. Speech Lang.*, 66:101155.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings*, volume 11096 of *Lecture Notes in Computer Science*, pages 198–208. Springer.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1317–1327. The Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 314–319. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

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. ACL.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study

of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Milos Stanojevic and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 202–206. ACL.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end ASR: from supervised to semi-supervised learning with modern architectures. *CoRR*, abs/1911.08460.

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

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus.

Minghan Wang, Yuxia Wang, Chang Su, Jiaxin Guo, Yingtao Zhang, Yujia Liu, Min Zhang, Shimin Tao, Xingshan Zeng, Liangyou Li, Hao Yang, and Ying Qin. 2021. The hw-tsc's offline speech translation systems for IWSLT 2021 evaluation. *CoRR*, abs/2108.03845.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 505–510. The Association for Computer Linguistics.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 2021. Hw-tsc's participation in the WMT 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 225–231. Association for Computational Linguistics.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes' rule. *Trans. Assoc. Comput. Linguistics*, 8:346–360.

Binbin Zhang, Di Wu, Zhuoyuan Yao, Xiong Wang, Fan Yu, Chao Yang, Liyong Guo, Yaguang Hu, Lei Xie, and Xin Lei. 2020. Unified streaming and non-streaming two-pass end-to-end model for speech recognition. *CoRR*, abs/2012.05481.