The NYA's Offline Speech Translation System for IWSLT 2024

Yingxin Zhang, Guodong Ma, Binbin Du
NetEase YiDun AI Lab, Hangzhou, China
{zhangyingxin03,maguodong,dubinbin}@corp.netease.com

Abstract

This paper reports the NYA's submissions to IWSLT 2024 Offline Speech Translation (ST) task on the sub-tasks including English to Chinese, Japanese, and German. In detail, we participate in the unconstrained training track using the cascaded ST structure. For the automatic speech recognition (ASR) model, we use the Whisper large-v3 model. For the neural machine translation (NMT) model, the wider and deeper Transformer is adapted as the backbone model. Furthermore, we use data augmentation technologies to augment training data and data filtering strategies to improve the quality of training data. In addition, we explore many MT technologies such as Back Translation, Forward Translation, R-Drop, and Domain Adaptation. Moreover, our model is a one-to-many ST system that utilizes flags for different tasks. Experimental results on the tst2022 test set demonstrate that our model achieves 36.37, 20.92, and 24.28 BLEU in En2Zh, En2Ja, and En2De, respectively.

1 Introduction

The Offline Speech Translation (ST) Task translates the source audio into target text. Currently, there are two leading solutions for ST. The first is the traditional cascade system (Matusov et al., 2005a), which decouples the ST task into an automatic speech recognition (ASR) and a neural machine translation (NMT) task. In the traditional cascade system, when translating, the source speech is recognized into source text, and then the NMT model is used to translate the source text into target text. However, it often leads to higher architectural complexity and error propagation (Duong et al., 2016), affecting subsequent NMT tasks. In order to alleviate this problem, the end-to-end (E2E) ST architecture (Bérard et al., 2016) is proposed. The E2E ST combines ASR and NMT modeling to establish the map between the source audio and the target text.

For the E2E ST architecture, one disadvantage is the lack of parallel training data. For the traditional cascade ST system, sufficient training can obtain high-accuracy ASR and MT systems due to the large ASR and MT datasets. Therefore, the traditional cascade ST system generally achieves better performance than the E2E ST. At the same time, in the recent offline track of IWSLT evaluation (Anastasopoulos et al., 2021, 2022; Agarwal et al., 2023), we can see that the cascade ST system is better than the E2E ST system. Thus, in this work, we use the traditional cascaded ST scheme.

Specifically, in the ASR task, we directly adopt the Whisper (Radford et al., 2023) large-v3 model, which can achieve a strong comprehensive ASR performance. We also explore sharding strategies, such as Supervised Hybrid Audio Segmentation (SHAS) (Tsiamas et al., 2022), to segment the source audio for better ST results. In the MT task, we use the Transformer architecture (Vaswani et al., 2017) as the backbone model. To ensure the MT model is fully trained, we meticulously collect a large amount of parallel data and monolingual data from various data sources. Furthermore, we delve into many MT technologies such as Back Translation (Sennrich et al., 2016), Forward Translation, R-Drop (Wu et al., 2021), Domain Adaptation, and Ensemble (Ganaie et al., 2022). Moreover, we compare the two solutions: one-to-one and one-to-many ST, and we find that one-to-many is better.

Through the above explorations, our model finally achieves good ST performance. In detail, experimental results on the tst2022 test set demonstrate that our model achieves 36.37, 20.92, and 24.28 BLEU in En2Zh, En2Ja, and En2De, respectively.

The rest of this paper is organized as follows. Section 2 describes the datasets and data preprocessing. Section 3 describes our speech translation system, which includes ASR and MT models.

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 87–93
August 15-16, 2024

Corpus	En2Zh	En2Ja	En2De
CoVoST (Wang et al., 2020)	171K	191K	220K
MuST-C v3 (Cattoni et al., 2021)	296K	251K	238K
NewsCommentary (Tiedemann, 2012)	400K	-	345K
OpenSubtitles (Lison and Tiedemann, 2016)	4.9M	832K	12M
Tatoeba (Tiedemann, 2012)	-	193K	302K
GigaST (Ye et al., 2023)	6.2M	-	6.3M
JParaCrawl (Morishita et al., 2020)	-	6.4M	-
Total	12M	8.2M	19.5M

Table 1: Data statistics on MT datasets.

Section 4 reports the experimental results. Finally, we conclude in Section 5.

2 Dataset

2.1 Text Data

The dataset used for machine translation is shown in Table 1, which contains both speech-to-text-parallel and text-parallel data types of all language pairs allowed by IWSLT 2024. Additionally, we employ the GigaST dataset to expand our text training data. sBERT (Reimers and Gurevych, 2019, 2020) is used for calculating sentence representations. We compute sentence embeddings for all parallel text data and remove sentences pairs that lower than 0.7 cosine similarity. The data statistics in table represent the number of sentences remaining in each dataset after sBERT filtering.

2.2 Data pre-processing

We perform the following preprocessing steps to filter all text-parallel data:

- Remove empty sentences and duplicate sentences.
- Remove sentences containing invalid characters and HTML tags.
- Remove sentences longer than 200 tokens or shorter than 3 tokens.
- Remove sentences with unbalanced source-target token ratio.
- Remove sentences with too much punctuation.
- Remove sentences where the source or target language constitutes a low percentage.
- Remove sentences with mismatched punctuation marks, such as quotation marks.

Then we apply mosesdecoder toolkits[1] (Koehn et al., 2007) for punctuation, space and case normalization. The sentences are then tokenized using joint SentencePiece model (SPM) (Kudo and Richardson, 2018). The vocabulary size of joint SPM is about 130,000, with 40k in English, 40k in Chinese, 30k in German, and 20k in Japanese, both source and target side share the same dictionary.

3 Speech translation system

3.1 ASR model

Whisper[2] (Radford et al., 2023) is an excellent multilingual ASR system trained on 680,000 hours of multilingual and multitask supervision data. It still shows strong robustness in various audio scenes, such as accent speech and background noise, and achieves good recognition results. It adopts the Encoder-Decoder architecture (Dong et al., 2018), and the training data has an extraordinarily structured design. In addition, it uses a method similar to prompt during the training process. The open-source Whisper models have five sizes of models: tiny, base, small, medium, and large. It is worth noting that the OpenAI has recently updated the Whisper large model to form a more effective large-v3 version model. In this work, we adopt the Whisper large-v3 version as the ASR part of our ST system.

3.2 MT model

3.2.1 Model structure

We adopt Transformer model (Vaswani et al., 2017) to build our machine translation system and implemente them on Fairseq toolkits (Ott et al., 2019). More specifically, we adopt a wider and deeper Transformer model which contains 18-layer encoder, 6-layer decoder, 16 self-attention heads and

[1] https://github.com/moses-smt/mosesdecoder
[2] https://github.com/openai/whisper

Language	Raw data	Filter data
Chinese	22M	9M
Japanese	30M	15M
English	8M	4.1M

Table 2: Data statistics on monolingual corpus.

FFN with 4096 dimensions. We utilize all provided parallel data from three language directions (En2Zh, En2De, En2Ja) for model training, and derived a one-to-many MT model.

3.2.2 R-Drop

The Dropout method (Srivastava et al., 2014; Gao et al., 2022) is an influential strategy for the regularization of deep neural networks. While it enhances the efficacy of the training process, the stochastic nature of dropouts might result in discrepancies between the training and inference phases. R-Drop, as introduced by Wu et al. (2021), ensures consistency among the output distributions of the sub-models generated by dropout. To enhance the consistency within our model, we implement the R-Drop algorithm and set weight factor α to 5. Consequently, the R-Drop training strategy significantly improves the performance of our baseline model.

Furthermore, when using the R-drop mechanism to train models, the model computation increases exponentially, which will consume more training time and GPU resources. Given the limitation of time and resources, we adopt it solely for our foundational model, and integrate the R-Drop-augmented model into ST system by using model ensemble approach during the evaluation stage.

3.2.3 Data Augmentation

Previous works (Edunov et al., 2018) has demonstrated that the incorporation of synthetic data can significantly enhance the efficacy of machine translation systems. We implement following data augmentation methodologies to further refine our translation models.

Forward translation (FT) is a process of transforming source language into target language using MT model. On the contrary, backward translation (BT) (Sennrich et al., 2016) is the translation of target language back into source language, forcing the model to learn a more robust representation of the source language. Both methods use additional monolingual resources to create bilingual data.

As shown in Table 2, we select 22M sentences of Chinese, 8M sentences of English and 30M sentences of Japanese of monolingual data from public datasets, such as Common Crawl and News Crawl corpus. Moreover, to make our MT model have better results in ACL scenarios, we adopt the scientific English monolingual corpus from Rohatgi et al. (2023). After data pre-processing pipeline mentioned above, approximately 40%-50% of the sentences from the original data are retained for each language. BT model is trained separately for each language pair, and then the monolingual data is used for backward translation. We employ an iterative forward-backward translation approach to progressively enhance the translation quality of both the FT model and BT model. As shown in figure 1, the FT model and BT model generated pseudo-labels *target'* and *source'* respectively. We mix them with labelled text pairs *(source, target)* to update our BT model and FT model. As the BLEU scores of BT model increased, the positive impact of the back-translated data on the FT model also becomes more pronounced.

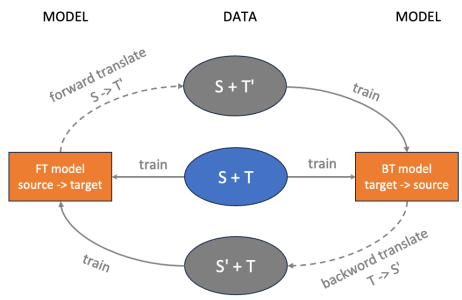

Figure 1: The iterative updating process for FT and BT model.

When using data generated by BT model, we refer to the tagged BT method (Caswell et al., 2019), adding a special token <BT> at the beginning of source sentence.

We also convert numerical expressions in English sentences into forms that more closely match the ASR transcription results, e.g., converting '21' to 'twenty-one', '2018' to 'two thousand and eighteen'. Additionally, we randomly discard punctuation marks within sentences to enable the model to generalize well across varying punctuation styles. These transformed sentences are merged with the original sentences to obtain an augmented dataset.

3.2.4 Domain adaptation

Considering the quality of machine translation models is easily influenced by specific domain, we also select in-domain data and fine-tune the model

	System	En2Zh	En2Ja	En2De
1	Baseline model	35.04	18.75	23.14
2	+ R-drop	35.67	19.36	23.71
3	+ GigaST	35.42	19.21	23.70
4	+ Backward translation	35.71	19.77	23.94
5	+ Domain adaptation	35.44	19.90	23.97
	Ensemble(2,4)	36.33	20.90	24.26
	Ensemble(2,4,5)	**36.37**	**20.92**	**24.28**

Table 3: Main results with BLEU scores on IWSLT tst2022 datasets

System	En2Zh	En2Ja
one-to-one	32.77	18.38
one-to-many	**35.04**	**18.75**

Table 4: BLEU scores on IWSLT tst2022 datasets (one-to-one vs. one-to-many ST)

System	En2Zh	En2Ja	En2De
Baseline	35.42	19.21	23.70
+ BT-Ja	35.37	19.71	**24.00**
+ BT-Zh	**35.71**	**19.77**	23.94

Table 5: BLEU scores on IWSLT tst2022 datasets with different BT data

to enhance in-domain performance. We use MUST-C data (Cattoni et al., 2021) as domain-specific dataset to train monolingual language models separately, and then use them to score all language pairs. We set specific thresholds to filter parallel data closer to the domain, with higher scores implying better quality, and train incrementally to get domain-specific model. The filtered in-domain data is about 5-10% of the total data.

3.2.5 ASR output adaptation

For ST dataset, we use ASR models to transcribe the audio data and replace their source side label with ASR recognition results, and finally obtain an augmented dataset containing ASR noise. ASR model may produce incorrect transcriptions for words with similar pronunciations, which, despite reducing the quality of MT training dataset, also bolster the robustness of the ST system. For this part of data, we also add a special tag <ASR> at the beginning of source sentence.

4 Experiments and results

All models are implemented on Fairseq toolkits (Ott et al., 2019) and trained on four NVIDIA A100 GPUs. The IWSLT test sets of tst2022 are used to evaluate the translation performance at sentence level. The mwerSegmenter toolkit[3] (Matusov et al., 2005b) is used to resegment and align translation results and then SacreBLEU[4] (Post, 2018) is used to compute BLEU scores. For the Japanese text, tokenization is performed using the Mecab, while for the Chinese text, tokenization is executed at character level. We apply SHAS[5] (Tsiamas et al., 2022) for audio segmentation and try a variety of combinations for min and max segment length, the optimal parameters is 5-30 secs for TED domain.

The table 4 presents a comparative analysis between the one-to-one and the one-to-many systems, specifically their performance on En2Zh and En2Ja. In the one-to-one system, each source language corresponds to only one target language, with BLEUs of 32.77 in En2Zh and 18.38 in En2Ja. In the one-to-many system, a source language text can correspond to multiple target language texts. The system trains data from English to three target languages (En2Zh , En2Ja , En2De) simultaneously and distinguishes the target language type by adding <zh>/<ja>/<de> tags. The performance of the one-to-many system improves to 35.04 in En2Zh and 18.75 in En2Ja. These scores indicate that one-to-many system outperforms the one-to-one system.

For the one-to-many system in Table 3, we first train a baseline model with all constrained data. We find that introducing R-drop mechanism positively affects model performance. Then, we add GigaST dataset for incremental training, which enriches the data diversity but also leads to a dramatic increase in the training data. We observe that as the amount of training data increases, R-drop no longer benefits model performance while consuming more training time, so we remove the R-drop mechanism

[3] https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz
[4] https://github.com/mjpost/sacrebleu
[5] https://github.com/mt-upc/SHAS

in subsequent stages.

In the forth stage, we collect monolingual data in Chinese and Japanese and perform back translation. As shown in table 5, the model performance is incrementally enhanced by incorporating back translation data into training dataset. Specifically, after adding BT-Ja data, the BLEU score for En2Ja improves significantly from 19.21 to 19.71, while En2Zh slightly decreases to 35.37. The addition of BT-Zh data enhances En2Zh to 35.71 and En2Ja to 19.77. Notably, although no BT data is added for En2De, its BLEU score still improves by 0.24, demonstrating a positive impact of back translation data on the overall model performance. Finally, domain adaptation brings some improvements in En2Ja and En2De.

Finally, we integrate the baseline model, which is enhanced by the R-drop mechanism, with fine-tuned models that leverage additional data, backward translation, and adaptation techniques. The ensemble of model (2, 4) achieves notable improvements, with BLEU scores of 36.33 for En2Zh, 20.90 for En2Ja, and 24.26 for En2De. Furthermore, the ensemble of model (2, 4, 5) slightly surpasses the ensemble of model (2, 4), reaching scores of 36.37 for En2Zh, 20.92 for En2Ja, and 24.28 for En2De. This indicates the effectiveness of model ensemble in boosting translation quality.

5 Conclusion

This paper describes our submission to the IWSLT24 offline speech translation task. We collect a large amount of parallel and monolingual data from the public data sources and adopt the traditional cascade ST architecture for the unconstrained training track. For the ASR model, we use the excellent Whisper large-v3 model, which is trained on 680,000 hours of multilingual and multitask supervision data. It shows strong robustness in various audio scenes. For the MT model, we explore a wider and deeper Transformer model using Fairseq tookit. To make the model fully trained, we carefully experiment many MT technologies, such as Back Translation, Forward Translation, Domain Adaptation, and R-Drop. Experimental results on the tst2022 test set show that our model achieves 36.37, 20.92, and 24.28 BLEU in En2Zh, En2Ja, and En2De, respectively.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander 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)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

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. *Computer speech & language*, 66:101155.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. 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*, pages 949–959, San Diego, California. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Bi-simcut: A simple strategy for boosting neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3938–3948.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. 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: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

E. Matusov, S. Kanthak, and Hermann Ney. 2005a. On the integration of speech recognition and statistical machine translation. In *Proc. Interspeech 2005*, pages 3177–3180.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005b. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

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 (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The acl ocl corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In *Proc. Interspeech 2022*, pages 106–110.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2023. GigaST: A 10,000-hour Pseudo Speech Translation Corpus. In *Proc. INTERSPEECH 2023*, pages 2168–2172.