Data Augmentation Methods and Influence of Speech Recognition Performance for TED Talk's English to Japanese Speech Translation

Kento Masuda* , Kazumasa Yamamoto† and Seiichi Nakagawa ‡
Department of Computer Science, Chubu University, Japan
E-mail: {*tp22021-6334@sti, †kazumasayamamoto@isc, ‡nakagawa@isc}.chubu.ac.jp

***Abstract*—In this paper, we compare uni-directional and bi-directional translation models for English to Japanese and various data augmentation methods to improve the performance of translation models using the Transformer. In the augmentation methods, we report on the performance improvements in text translation using pseudo parallel corpus based on the Japanese translation of English monolingual corpus, pseudo parallel corpus based on the English translation of Japanese monolingual corpus, pseudo parallel corpus based on translations of one side of the English-Japanese parallel corpus, addition of auto-encoder part to the translation model, and use of sentences generated by the auto-encoder as pseudo corpus, and training multilingual translation models based on English-Korean parallel corpus. Next, a speech translation model was constructed by cascade connecting machine translation model and speech recognition model, and evaluated with various models of Kaldi and Whisper speech recognizers to analyze the relationship between speech recognition rate and speech translation performance.**

I. Introduction

For TED talks, English speech recognition has been frequently presented at the IWSLT (International Workshop on Spoken Language Translation), but there are fewer on English to Japanese translations. In 2021, simultaneous text translation for English to Japanese [2] and in 2022, simultaneous speech translation for English to Japanese was included as a task [3].

For TED speech recognition, Ueno et al. achieved a word error rate (WER) of 8.56% using a 12 layers Conformer encoder and a uni-directional LSTM decoder with Attention function, trained on TEDLIUM2 (211 hours of training data)[4]. Currently, it seems effective to incorporate self-supervised learning (SSL) feature vectors into Conformer. In this paper, we use Kaldi, a typical recognition tool based on the DNN-HMM method, as a speech recognizer. Additionally, we utilize Whisper, which has been widely used recently. Whisper, though not using SSL technology, is a large-scale model trained on over 600,000 hours of data, achieving extremely high recognition rates. Whisper offers models like tiny, small, medium, and large; in this paper, we use medium and large models to analyze the relationship between various speech recognition rates and translation performance.

For speech translation, there are two types of methods: a hybrid (cascade) method that connects speech recognition and text translation and an end-to-end method. Recently, the latter has made remarkable progress, closing the gap between the two [3,5]. For TED English to Japanese speech translation, Fukuda et al. reported BLEU scores of 11.6 for offline method and 10.6 for online simultaneous translation method [6]. The best result at IWSLT 2022 was the USTC (University of Science and Technology of China) system using large external speech and language resources, with a WER of about 5%, a text input BLEU score of 22, and an offline speech input BLEU score of about 19 using an emsemble method [11]. At IWSLT 2023, offline speech translation, simultaneous speech translation, and new TED test data were evaluated, and the ACL presentation speech was also included in the test [7]. Once again, the cascade method of speech recognition and machine translation outperformed the end-to-end method. Three teams participated in the offline English to Japanese speech translation task for TED talks, with BLEU scores of 10.6, 16.5, and 18.7, respectively. The best method from HW-TSC (Huawei Translation Service Center) involved fine-tuning Whisper (conformer) for speech recognition and using a one-to-many (1-to-3) Deep-Transformer model for translation [7].

To compensate for the lack of parallel data, there is a data augmentation method using monolingual corpus [8]. This method creates a pseudo parallel corpus by machine translating monolingual corpus and mixes them with the base parallel corpus for training. Yamagishi et al. used transcription data of academic conference lectures from the Corpus of Spontaneous Japanese (CSJ) for monolingual corpus in this method [9]. Another method converts the base parallel corpus into pseudo corpora based on certain rules, increasing the training data several times (parallel phrase model) [10]. Bao et al. generated multiple translations by adding keywords from source sentences to obtain high-quality pseudo corpus [11]. There are also various methods using auto-encoders. Cheng et al. used monolingual data and English to Chinese & Chinese to English translation models to generate target English sentences from source English sentences, that was, English sentences into Chinese sentences and then the Chinese sentences into English sentences, learning the translation models to make the source and target sentences identical, achieving higher effectiveness than standard back translation approach [12].

The method we attempted is to learn so that the source side sentence and target side sentence are identical, which is

a similar method as Currey et al [13]. However, Currey et al. did not evaluate the method with the Transformer or the bi-directional model. Nor did they compare their method with a pseudo parallel corpus using sentences generated by an auto-encoder.

TABLE I: Data sets

(a) Training dataset for machine translation model

Data set	Training # sentence pairs	Validation # sentence pairs
IWSLT2016 English-Japanese parallel corpus	233,108 pairs	871 pairs
	1,863 lectures	8 lectures

(b) Data set for data augmentation

Data set	# sentences (#lectures)
IWSLT2018 (new-English monolingual)	80,222 sent. (862 lec)
CSJ (new-Japanese monolingural)	219,229 sent. (1,565 lec)
IWSLT2016 (English-Korean pairs)	230,240 pairs (1,920 lec)
ASPEC (written English-Japanese pairs)	1,000,000 pairs

(c) Test data

Data set	Test Sentences (English)
IWSLT2016 English-Japanese	963 sentences
	10 lectures

II. Speech Recognition Models

A. Kaldi

Kaldi provides recipes for various corpora, and in this paper, we used the TEDLIUMv3 corresponding recipe for speech recognition. Kaldi employs a DNN-HMM architecture. In this paper, the DNN has 13 hidden layers, each with 1024 units. The number of units in the output layer corresponds to the shared tied state number of triphone HMMs. The input features to the network are 40-dimensional MFCCs (Mel-frequency cepstral coefficients). For feature preprocessing, we applied fMLLR, LDA, and SAT to extract speaker-independent features. We trained the DNN-HMM on 450 hours of TED talk data. The language model was created from the text of 450 hours of TED talks with a vocabulary size of about 150,000 words.

B. Whisper

Whisper, provided by OpenAI, is an open-source software for speech recognition. It is designed as an end-to-end model, meaning it directly converts input audio into text without needing intermediate representations. This architecture simplifies the processing pipeline and can potentially lead to higher accuracy and efficiency. It uses about 680,000 hours of multilingual audio data scraped from the web for training, enabling high-accuracy speech recognition. Using large and diverse training data is expected to improve robustness against background noise and specialized terminology. The basic processing involves converting input audio into acoustic features and segmenting the data. Specifically, input audio is transformed into a log-Mel spectrogram. The hyperparameters for this feature extraction are hardcoded in sorce code, resulting in 80-dimensional features, with a 10ms stride, resampled to 16kHz mono audio. Whisper processes audio in 30-second segments, padding with zeros if necessary, to ensure consistent segment length. Whisper offers models of various sizes, with transcription accuracy improving as the number of parameters increases. In this experiment, we used the medium model with 769M parameters and the large-v3 model with 1550M parameters. We used these pre-trained models without fine-tuning on TED data.

III. Translation Models

A. Transformer

The Transformer [14] model consists of encoder and decoder layers. The encoder is composed of stacked identical encoders, each consisting of a self-attention mechanism and a feed-forward neural network (FFNN). The decoder is similarly composed of stacked identical decoders. In this experiment, we used fairseq ver.0.12.0 [15].

B. Number of Encoder-Decoder Layers

A standard Transformer model has 6 encoder layers and 6 decoder layers, but the optimal configuration varies depending on the amount of training data. Preliminary experiments in previous research [1] showed a BLEU score of 0.80 with 6 encoder and 6 decoder layers and a score of 13.22 with 3 encoder and 3 decoder layers. In this experiment, we compared various data augmentation methods with different encoder-decoder layer numbers.

C. Uni-directional and Bi-directional Translation Models

In an uni-directional translation model, the translation model is created by training on pairs of source (English) and target (Japanese) languages. In a bi-directional translation model, the model is created by training on datasets consisting of pairs of source and target languages and pairs where the source and target languages are swapped. This model performs both English to Japanese and Japanese to English translations for improving the robustness.

IV. Data Augmentation Methods

A. Data Augmentation Using Forward/Backward Translation of Monolingual Corpus

(a) Utilization from other than base corpus

Due to the limited amount of English and Japanese parallel corpora in the IWSLT base corpus of TED lectures used in this experiment, we created pseudo-parallel corpora by translating English or Japanese monolingual corpora forward (English to Japanese) / backward (Japanese to English) using the base model, thereby performing data augmentation [1]. As the monolingual corpora, we used the monolingual corpora used the English side of the IWSLT2018 English-Spanish parallel corpus and the simulated lectures from the CSJ Japanese corpus.

(b) Utilization from base corpus

Translating the English or Japanese side of the base corpus forward/backward using the model with the highest translation accuracy from previous research [1], the translated sentence with the other original sentence of the base parallel corpus is combined to create pseudo parallel corpora.

B. Data Augmentation Using Auto-Encoders

English to English and Japanese to Japanese parallel corpora for training (autoencoder [14]) are also included. During this process, language pair tags were added to the source language side to discriminate language pairs (such as EJ, JE, EE, JJ) as shown in Figure 1. Figure 1 shows the schematic illustration of the traing data setup using original parallel data and auto-encoder method.

(c) Parallel corpus with identical source and target sentences
For the base corpus, we added an autoencoder component with identical source and target sentences for training. In the case of an English to Japanese uni-directional translation model, we simultaneously trained the model to translate English into Japanese and added parts where English sentences were translated into the same English sentences or Japanese sentences into the same Japanese sentences. For the English to Japanese/Japanese to English bi-directional translation model, in addition to bi-directional translation between English and Japanese, we trained the model with the autoencoder functions of English to English and Japanese to Japanese.

(d) Using pseudo source sentences generated by auto-encoder
A pseudo-parallel corpus was created using an autoencoder trained with the same source and target sentences of the base parallel corpus and used for training as a data augmentation, that is, generated English-Japanese or English-generated Japanese in addition to English-Japanese.

(e) Addition of auto-encoder part from monolingual corpus other than the base corpus
As similar to the method (c), but using a monolingual corpus other than the base corpus (b), an auto-encoder component was added to the translation model.

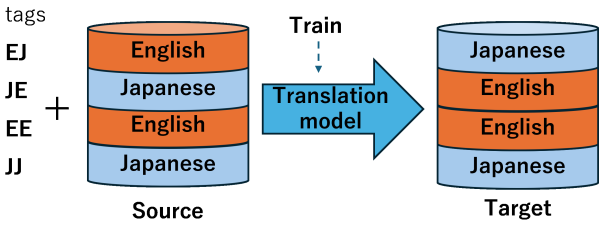


Fig. 1: Schematic illustration of the training data setup using the auto-encoder method

C. Data Augmentation Using English-Korean Parallel Data

(f) Considering that multilingual models are effective for low-resource learning data [14], we applied this approach to English to Japanese and Japanese to English translation tasks by training with a parallel corpus of English and Korean, which has a similar grammar to Japanese, in addition to the base parallel corpus.

D. Transfer Learning from ASPEC Corpus Model

(g) Although the ASPEC (Asian Scientific Paper Excerpt Corpus) corpus consists of written language rather than spoken language like TED, it contains a large-scale parallel corpus of 1 million sentence pairs. We trained uni-directional and bi-directional models using the ASPEC corpus, and used these parameters as initial values to train the translation model with the IWSLT base corpus and the data-augmented dataset. Note that the vocabulary of TED's spoken language was set as the vocabulary during the translation model training using the ASPEC corpus [1].

V. Experimental Results

A. Data used

The IWSLT2016 English-Japanese parallel corpus was used as the base corpus for this study. The English side of the IWSLT2018 English-Spanish parallel corpus and the Japanese spoken corpus CSJ, a monolingual corpus of Japanese, were used as a monolingual corpus for data augmentation, and the IWSLT2016 English-Korean parallel corpus was used for the multilingual model. Table I shows the number of training, development, and test sentences.

B. Speech Recognition Results

The TED lectures vary in difficulty depending on their content, affecting translation performance. Additionally, the speakers differ for each lecture, resulting in significant variations in speech recognition performance by speaker. Whisper performed speech recognition for each lecture as a whole, rather than for each utterance. Whisper sometimes recognizes meaningless word sequences during silent intervals and may output repeated word sequences. Table IVshows the word error rate (WER=100 - error rates of insertion, deletion, and substitution) for each lecture and speaker using Kaldi and Whisper. In Kaldi, the WER ranged from about 10% to 33%, with an average of 15.37%. With Whisper's Medium model, the WER ranged from 3.4% to 13.34%, with an average of 8.85%. The Large model had a WER ranging from 3.14% to 11.76%, with an average of 7.21%.

C. Text Translation Results

The number of encoder and decoder layers was set to 3 or 6 layers, and both un-idirectional and bi-directional models were trained and evaluated. Optimization was performed using Adam with a learning rate of 0.0005, and training was conducted for up to 20 epochs. BLEU (BiLingual Evaluation Understudy) was used to calculate the similarity between

TABLE II: Results of the 10 TED talks English-to-Japanese translation evaluation experiment (BLEU)
The parenthesis () denotes result by fine-tuning model
"not well done" means thath the translation model was not well learned

model		Uni-direction		Bi-direction	
method \ #layers		3 layers	6 layers	3 layers	6 layers
Base line		**13.78**	10.67	13.53 (**14.40**)	not well done
b-1	+ Pseudo English	15.72	16.04	15.88 (16.11)	
b-2	+ Pseudo Japanese	**16.52**	16.19	15.63 (**16.23**)	
b-3	+ Pseudo English & Japanese	16.01	16.36	15.30 (15.79)	
c-1	+ English-English	13.73	15.47	13.70 (13.89)	15.81 (16.21)
c-2	+ Japanese-Japanese	13.48	15.19	13.80 (12.78)	16.14 (16.37)
c-3	+ English-English & Japanese-Japanese	13.66	15.22	13.26 (13.63)	15.67 (16.17)
d-1	+ Generated English	13.69	12.13	14.39 (14.70)	not well done
d-2	+ Generated Japanese	12.89	11.97	14.12 (14.39)	
d-3	+ Generated English + Generated Japanese	13.54	12.12	14.56 (14.79)	
e-1	+ new English-new English	13.50	14.81	13.65 (14.09)	14.76 (15.13)
e-2	+ new Japanese-new Japanese	12.87	14.65	13.44 (13.71)	14.14 (14.56)
e-3	+ New English-New English & New Japanese-New Japanese	13.98	15.03	13.88 (14.36)	15.59 (16.18)
f	+ English-Korean			14.66 (15.15)	not well done
h-1	+New English (a) + New Japanese (a) + ASPEC (g)			14.76 (15.48)	16.16 (**16.50**)
h-2	(h-1)+(b-3)+(c-3)			16.13 (**16.77**)	not well done

TABLE III: Multi-reference translation evaluation results for 2 TED talks (BLEU). ***() denotes fine-tuing model***

model		Uni-direction		Bi-direction	
method	#references \ #layers	3 layers	6 layers	3 layers	6 layers
Base line	single	**11.16**	8.75	11.69 (**12.18**)	not well done
	3 references	**24.47**	18.35	**25.94** (24.94)	
b-3	single	**12.41**	11.19	10.54 (12.90)	
	3 references	**27.34**	26.97	25.55 (27.32)	
c-3	single	9.96	10.94	11.37 (11.08)	12.19 (12.30)
	3 references	23.24	26.45	24.88 (24.59)	28.08 (26.62)
d-3	single	10.38	10.02	11.47 (11.71)	not well done
	3 references	23.20	22.85	24.55 (24.99)	
e-3	single	9.93	10.31	10.91 (11.16)	11.61 (11.86)
	3 references	22.74	25.98	24.29 (23.30)	27.71 (27.44)
f	single			12.06 (13.39)	not well done
	3 references			26.88 (27.04)	
h-2	single			12.44 (**13.09**)	
	3 references			28.15 (**30.94**)	

machine translation and reference, with values up to 4-gram used in this study. The English-Japanese text translation results for 10 lectures are shown in Table Ⅱ. Additionally, human translations by two translation agencies were obtained for two lectures, resulting in three references including the original reference, as shown in Table Ⅲ.

Fine-tuning of the bi-directional translation model involves additional re-training of the model with the base parallel corpus after model training using data augmentation. The baseline refers to the model trained with the IWSLT2016 English-Japanese base parallel corpus using both uni-directional and bi-directional translation models. "+Pseudo-English" refers to the pairs of translated English sentences and original Japanese sentences of the base parallel corpus using the method of Section Ⅳ(b) as a pseudo-parallel corpus for training. The same applies to "+Pseudo-Japanese." The English part of IWSLT2018 used as a monolingual corpus is referred to as "new-English," and the CSJ corpus as "new-Japanese."

The baseline results were BLEU values of 13.78 for the uni-directional model and 14.40 for the bi-directional model. The method of data augmentation of pseudo-corpora by machine translation of the base corpus (b) exceeded the baseline BLEU values for both pseudo-English and pseudo-Japanese. The uni-directional translation model improved the BLEU value by +2.74 (to 16.52). The bi-directional translation model also showed further improvement in translation accuracy with fine-tuning, with a BLEU value improvement of +1.83(to 16.23) compared to the bi-directional baseline. The augmentation method of adding an autoencoder part with identical source and target sentences to the base parallel corpus showed high translation accuracy for all models fine-tuned with 6 layers, achieving BLEU values of 16.00 or higher in text translation. This suggests that the translation model obtained robust internal representations through the added autoencoder part.

The method of using sentences generated by the autoencoder as pseudo-parallel corpora (d) did not show significant differences from the baseline. This is because the performance of the autoencoder was too good, resulting in high similarity between the generated pseudo-parallel corpus and the base parallel corpus (generated English: BLEU = 99.47, generated Japanese: BLEU = 98.84), and thus, generated sentences including errors was not mostly achieved.

The method of adding monolingual corpora other than the base corpus as the autoencoder part (e) showed higher improvement for the model (e-3) that added both the English part of IWSLT2018 and the CSJ corpus. However, the degree of translation accuracy improvement was lower compared to the model (c) that added the autoencoder part to the base

TABLE IV: Results of WER and BLEU by speech translation/text translation using Kaldi and Whisper Medium/Large models.
The parenthesis () denotes BLEU by baseline-model ; BLEU by best model (BLEU by baseline model)

Lecture	Kaldi WER	Medium WER	Large-v3 WER	Kaldi BLEU	Medium BLEU		Large-v3 BLEU		Text BLEU
1	9.53	10.87	9.58	13.37	13.57	(11.13)	13.89	(12.41)	16.82 (12.64)
2	11.00	11.19	5.10	11.58	9.87	(8.38)	10.09	(9.67)	12.55 (12.01)
3	11.30	9.78	8.68	10.46	11.54	(9.87)	11.98	(11.16)	12.57 (11.59)
4	11.49	3.40	3.14	14.14	15.02	(14.12)	15.44	(15.33)	15.54 (14.63)
5	11.71	4.27	4.18	16.77	18.54	(17.19)	18.65	(18.21)	21.37 (18.84)
6	12.93	10.43	6.70	8.69	8.75	(9.02)	9.22	(10.27)	11.87 (11.61)
7	14.79	8.41	7.61	11.24	10.90	(7.92)	11.20	(9.54)	11.76 (11.10)
8	16.38	13.34	11.76	17.79	15.03	(13.52)	15.87	(15.03)	21.21 (17.54)
9	21.42	9.79	9.67	11.12	14.01	(12.28)	14.07	(13.71)	16.49 (15.37)
10	33.17	6.99	5.67	11.27	15.30	(11.25)	15.75	(12.44)	15.71 (12.43)
Ave.	15.37	8.85	7.21	12.64	13.25	(11.47)	13.61	(12.77)	15.58 (13.81)

parallel corpus.

The use of the English-Korean parallel corpus (f) improved English to Japanese translation accuracy with a multilingual model incorporating the parallel corpus of English-Korean, a language with similar grammar to Japanese.

Finally, as various combinations, Table Ⅳ(h-2) shows the translation evaluation results of models that performed data augmentation using the method of adding pseudo-parallel corpora from monolingual corpora (a) and combining initial parameters from the ASPEC corpus model (g), with the method of using pseudo-corpora from the base corpus (b) and adding the autoencoder part to the base corpus (c). By combining the proposed method, which showed a high effect of translation accuracy improvement, with the best model from previous research [1], a high evaluation result of BLEU 16.77 was obtained in text translation.

The original reference is close to free translation of spoken language, limiting the BLEU improvement by machine translation. Therefore, when the reference is made multi (three) references, the BLEU value improves significantly (Table Ⅲ).

D. Speech Translation Results

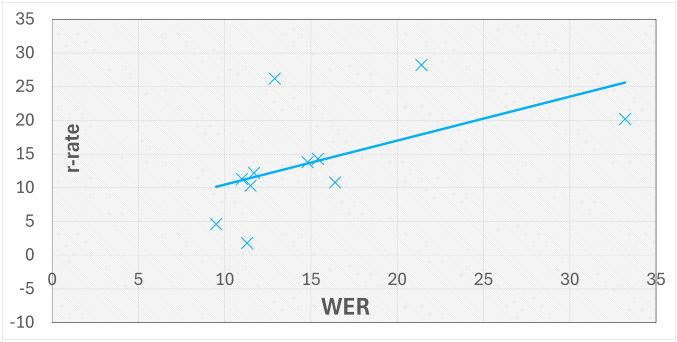
The speech translation results are shown in Table Ⅳ. The translation models are the baseline model and the best model (h-2). The speech recognizers are Kaldi and Whisper's medium model and large model. We should notice that the BLEU scores for Table Ⅳare slightly different from those of Table Ⅱ, because the BLEU in Table Ⅲis calculated for all lectures as a whole, but the BLEU in Table Ⅳis calculated for each lecture and then averaged. Compared to the BLEU values of text translation, the BLEU values of speech translation with speech input decreased by 1 to 5 points. Figure 2 shows the relationship between speech recognition performance (word error rate; WER) and the reduction rate in translation performance. The r-rate indicates the reduction rate of the BLEU value of speech translation results from the BLEU value of text translation results and is calculated by $(\frac{\text{Text BLEU}-\text{ASR BLEU}}{\text{Text BLEU}}) \times 100$. For Kaldi's speech recognition results, only the best translation model was evaluated. The correlation between WER and the reduction rate of BLEU in the best model was 0.545 for Kaldi model, 0.47 for Whisper's Large-v3 model, 0.75 for the Medium model, and 0.63 for Whisper overall, respectively. In the baseline model, the correlation between WER and the reduction rate of BLEU was 0.4 for the Large-v3 model, 0.73 for the Medium model, and 0.63 for Whisper overall, respectively. As seen in Figure 2 (b) and (c), both the baseline and best translation models show that the overall linear approximation almost coincides with recognition results of Whisper. This means that regardless of the translation model, the relationship between speech recognition accuracy (if the recognizer is the same) and the reduction in translation performance shows the same trend. For example, if the WER is about 7%, the reduction in translation performance due to speech input compared to text input is about 10% in BLEU value. However, for Kaldi's speech recognition results, if the WER is 10%, the reduction rate in translation performance is about 10%. This different trend may be caused by different ASR architectures which yield proper misrecgonition results.

VI. Conclusions

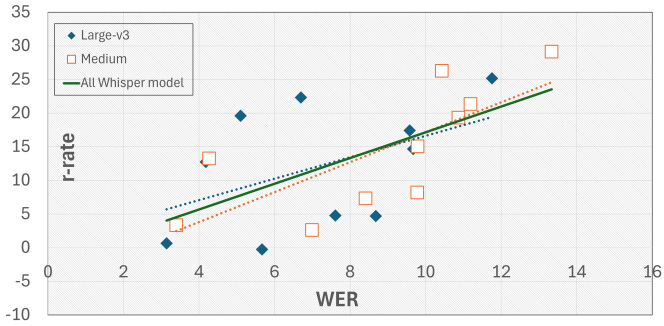
For the English-Japanese TED talk, using Transformer based machine translation models trained by the corpus consisting of 220,000 sentence pairs, translation performance was improved for both unidirectional and bidirectional models by incorporating auto-encoder components and augmenting the base corpus with machine-translated pseudo parallel corpora. For the 10 lectures, the best data augmentation method improved the BLEU score for text translation from a baseline of 14.40 to 16.77, and for speech translation from a baseline of 12.03 to 14.32.

For speech translation, we used Kaldi and Whisper as speech recognizers in a cascade setup with the text translation model. Whisper's word error rate ranged from 3% to 14%, and the translation performance degradation for speech input compared to text input was 1 to 3 BLEU points, or a 0% to 20% decrease. We demonstrated that with a speech recognition accuracy of around 95% (word error rate of 5%), the BLEU score degradation compared to text translation performance is approximately 10%. In Appendix, examples of ASR results and translation results are shown. In future work, we will develop simultaneous speech translation system.

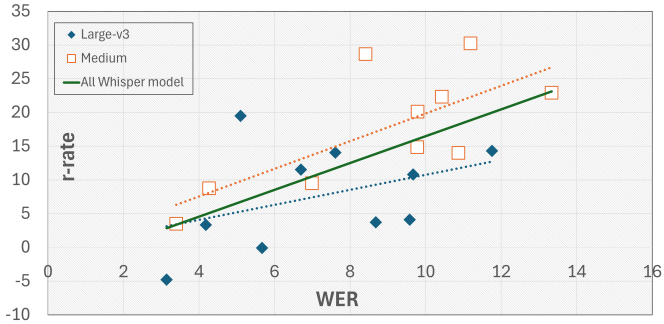
t



(a) Relation of WER of Kaldi ASR and reduction rate of BLEU



(b) Relation of WER of Whisper ASR and best translation model



(c) Relation of WER of Whisper ASR and baseline translation model

Fig. 2: Relation of WER of ASR and reduction rate of BLEU by speech translation to text translation

ACKNOWLEGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP18H01062, JP1919K12027, JP22K12084.

REFERENCES

[1] H. Banno, H. Sakurai, J. Adachi, K. Yamamoto, S. Nakagawa, Consideration on Speech Recognition, Speech Translation and Speech Summarization for TED English Lectures, Proc. NLP, pp.1417–1422, 2023.3 (in Japanese)
[2] A. Anastasopoulos, O. Bojar, et al. : Findings of the IWSLT 2021 Evaluation Campaign, Proc. IWSLT-2021, pp.1–29, 2021.
[3] A. Anastasopoulos, L. Barrault, et al. : Findings of the IWSLT 2022 Evaluation Campaign, Proc. IWSLT-2022, pp.98–158, 2022.
[4] S. Ueno, A. Lee, T. Kawahara, Enhancing Synthesized Speech using Speaker Information and Phone Masking for Data Augmentation of Speech Recognition, Proc. Autumn Meeting of ASJ, pp.1149–1150, 2022.8 (in Japanese)
[5] W. Zhang, Z. Ye, et al. : The USTC-NELSLIP Offline Speech Translation Systems for IWSLT 2022, Proc. IWSLT, pp.198–207, 2022.
[6] R. Fukuda, Y. Ko, Y. Kano, et al. : NAIST Simultaneous Speech-to-Text Translation for IWSLT 2022, Proc. IWSLT 2022, pp.286–292 , 2022.
[7] M.Agarwal, et al. : Findings of the IWSLT 2023 Evaluation Campaign, Proc. IWSLT-2023, pp.1–61, 2023.
[8] Rico Sennrich, Barry Haddow, Alexandra Birch: Improving Neural Machine Translation Models with Monolingual Data, Proc. ACL, pp.86–96, 2016.
[9] Y. Yamagishi, T. Akiba, H. Tsukada, English–Japanese Machine Translation for Lecture Subtitles using Back-translation and Transfer Learning, Proc. 9-th GCCE, 2022
[10] J.E. Hu, et al. : Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting, Proc. Conf. NAACL, pp.839–850, 2019.
[11] Guangsheng Bao, Zhiyang Teng, Zhang, Yue: Target-SideAugmentation for Document-Level Machine Translation, Proc. ACL, pp.10725–10742, 2023.
[12] Y. Cheng, et al. : Semi-Supervised Learning for Neural Machine Translation, Proc. ACL, pp.1965–1974, 2016.
[13] A. Currey, et al. : Copied Monolingual Data Improves Low-Resource Neural Machine Translation, Proc. ACL, pp.148–156, 2017.
[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob, Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin: Attention is All You Need, Conference on Neural Information Processing Systems (NIPS2017), 30, 2017.
[15] M. Ott, et al. : fairseq: A Fast, Extensible Toolkit for Sequence Modeling, Proc. NAACL 2019, pp.48–53, 2019.

APPENDIX

Appendix: Example of ASR Result and Translation Result

Input English	I am learning that it's a genius idea to use a pair of barbecue tongs to pick up things that you dropped. I'm learning that nifty trick where you can charge your mobile phone battery from your chair battery.
Japanese Reference	私 が 学ん で いる の は 天才 的 な アイデア です 落とし た 物 を 拾う の に トング を 使う と か ね 他 に は 車 椅子 の バッテリー で 携帯 の 充電 を する 粋 な アイデア
Text Translation (BLEU=38.78)	私 は この アイデア を 学び まし た バーベキュー ・ トング を 使っ て 自分 が 落とし た もの を 手 に れる こと が できる と いう アイデア です 私 は 素晴らしい 技 学び まし た 皆 さん の 椅子 から 携帯 電話 を 充電 できる と いう の です
Whisper-Medium ASR (WER=5.13)	I am learning that it's a genius idea to use a pair of barbecue tongs to pick up things that you drop. I'm learning that nifty trick where you can charge your mobile phone battery from your chair battery.
Whisper-Medium Translation (baseline model) (BLEU=8.27)	私 は トング と いう 天才 的 な アイデア を 学び まし た バーベキュー ・ トング の 2 つ の バーベキュー セット を 使っ て あなた が 椅子 から 携帯 電話 を 充電 できる よう に する ため です
Whisper-Large-v3 ASR (WER=17.95)	I am learning that it's a genius idea to use a pair of paws, and to use your own barbecue tongs to pick up things that you drop. I'm learning that nifty trick where you can charge your mobile phone battery from your chair battery.
Whisper-Large-v3 Translation (best model) (BLEU=10.52)	私 が 学ん だ の は 義足 を 使う の は 天才 だ と いう こと そして 自分 の バーベキュー トング を 使っ て 落とし た もの を 手 に 入れる こと です 私 が 学ん だ の は スマート に 製 の トリック で 椅子 の バッテリー で 充電 できる と いう こと です