# Improve Speech Translation Through Text Rewrite

**Jing Wu** [1], **Shushu Wang** [2], **Kai Fan** [1], **Wei Luo** [1], **Minpeng Liao** [1], **Zhongqiang Huang** [1]
[1]Tongyi Lab, [2] Zhejiang University
{lz.wujing, k.fan, muzhuo.lw, minpeng.lmp, z.huang}@alibaba-inc.com
{wangshushu0213}@zju.edu.cn

## Abstract

Despite recent progress in Speech Translation (ST) research, the challenges posed by inherent speech phenomena that distinguish transcribed speech from written text are not well addressed. The informal and erroneous nature of spontaneous speech is inadequately represented in the typical parallel text available for building translation models. We propose to address these issues through a text rewrite approach that aims to transform transcribed speech into a cleaner style more in line with the expectations of translation models built from written text. Moreover, the advantages of the rewrite model can be effectively distilled into a standalone translation model. Experiments on several benchmarks, using both publicly available and in-house translation models, demonstrate that adding a rewrite model to a traditional ST pipeline is a cost-effect way to address a variety of speech irregularities and improve the speech translation quality for multiple language directions and domains.

## 1 Introduction

Much progress has been made in recent years in speech translation, from cascade systems (Sperber et al., 2017; Matusov et al., 2018; Zhao et al., 2020; Xu et al., 2021; Papi et al., 2021) to end-to-end systems (Bérard et al., 2016), and large language model systems (Chen et al., 2024). However, the unique characteristics of spontaneous speech, including accents and presentation quality (disfluencies, grammar errors, etc.), and challenges arising from errors in automatic speech recognition (ASR) and punctuation prediction with cascade systems, continue to pose significant challenges.

In an effort to bridge the gap between spoken and written languages, prior studies have explored various methods, including disfluency detection (Johnson and Charniak, 2004) and removal (Wang et al., 2010), recognition error correction (Guo et al., 2019), and grammar error correction (GEC) (Rothe et al., 2021). Each of these approaches aims to address specific aspects of spoken language as independent tasks.

It is a common practice among expert human interpreters to skip redundant or incomprehensible parts (Liu, 2008) and summarize speech fragments into unambiguous segments (Al-Khanji et al., 2000; He et al., 2016), so that they can focus on the meaning of the source messages and generate accurate translations (Camayd-Freixas, 2011). However, due to limited working memory, real-time interpreting tend to over-compress information (Sridhar et al., 2013). Meanwhile, high-quality offline interpreting annotation is expensive, especially for multilingual translation directions.

Our pilot study shows that monolingual human annotators possess the ability to apply the aforementioned interpreting strategies to an erroneous speech transcript generated by an automatic system, and produce a high-quality rewritten transcript that effectively preserves the original meaning. Building upon this observation, we propose a novel approach to model the human rewrite process as a generation task, where a supervised model is trained using annotated rewrite data and learns to directly generate the rewritten transcript, eliminating the need for annotators to label each individual operation separately. As illustrated in Figure 1, this rewrite model can be integrated as a component within a cascade speech translation system or can be distilled into a standalone translation model.

The significant advantage of our approach lies in its efficiency. Rather than relying on costly bilingual data or deploying separate models for different types of irregularities, our proposed approach requires only monolingual annotation to serve multiple target language speech translations, and handles various irregularities all at once. Aiming at more effective application, we propose a



*Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 331–342
January 19–24, 2025. 

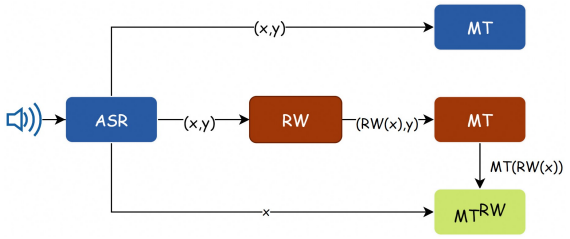


Figure 1: The blue pipeline represents basic cascade speech translation. The red ST pipeline integrates the rewrite model as a component. The green pipeline shows that the student translation model learns the rewrite process through knowledge distillation training without needing a rewrite component.

rewrite distillation method that seamlessly integrates into the speech translation pipeline without introducing additional components or incurring extra latency and inference cost. Additionally, we demonstrate through testing on multiple datasets that our method significantly improves speech translation performance while maintaining efficiency. To facilitate future research in related areas, we open-source our annotated rewrite data of the BSTC at https://huggingface.co/datasets/have-to-name/TextRewrite.

## 2 Related Works

In cascade speech translation systems, detecting and removing classical types of disfluencies (Chen et al., 2020) is widely used to bridge the differences between spoken and written languages. Most previous disfluency detection works with a sequence labeling model trained and evaluated on the English Switchboard corpus (Godfrey and Holliman, 1993). For directly removing disfluency, Dong et al. (2019) substitutes the multiple disfluency labels in Switchboard with one tag to generate end-to-end disfluency data. Synthesized disfluent-to-fluent data were created by inserting typical disfluency types of repeat, filler and restart into unlabeled corpus (Wang et al., 2020b; Passali et al., 2022), or unsupervised style-transfer approach through back-translation (Saini et al., 2021). Human annotated data has also been used by Cho et al. (2016) and Salesky et al. (2019). Word error correction models have also been applied in ST pipelines (Rothe et al., 2021; Guo et al., 2019; Hrinchuk et al., 2019; Cui et al., 2021). However, researchers find that speech recognition errors are sparse and that word error correction models tend to introduce new errors by modifying too many originally correct words while correcting errors (Leng et al., 2021).

Previous works also apply aligned interpreting corpora to improve the quality of speech translation. Zhang et al. (2021) builds a speech translation corpus where speech irregularities are kept in transcription while omitted in translation. Zhao et al. (2021) uses the interpretation corpus EPTIC to fine-tune the MT model and achieves an improvement on the interpreting test set collected from the European Parliament. However, a significant decline in performance on the corresponding translation test set is observed due to too much missing information in the training interpretation.

Knowledge Distillation(KD) approaches aim to transfer knowledge from a teacher model to a student model (Hinton et al., 2015). Kim and Rush (2016) first applied sequence-level knowledge distillation to NMT models.

## 3 Our Approach

### 3.1 Text Rewrite Annotation

We propose an annotation task to emulate human strategies during rewriting ASR transcript to high quality human speech manuscripts.

Annotators are presented with the texts generated from automatic speech recognition with machine-induced punctuation. Ultimately, they are asked to rewrite the texts into a fluent and grammatically correct form that maintains the original meaning and can be used as a prepared speech. They are instructed to perform segmentation at first. During the annotation, the annotator can combine more relevant context by re-segmenting the ASR transcription, or choose not to use the sample to do annotation. We provide two segments in Table 1 and tag the operations that have been observed during human rewrite. Note that We display more information of rewrite annotation guideline in Appendix A.1.

Red tags in Table 1 show removal of disfluencies and non-translatable content that cased by factors such as unprepared speakers, automatic transcription errors, and improper punctuation segmentation.

Green tags show word error correction. Intuitively, any errors in the text caused by the speaker, the audio recording environment, or the recognition system should be corrected. However, we found that audio-based word correction that is detached from the context can easily lead to halluci-

| Systems | Examples |
|---|---|
| ASR | In this life, they are they they are, um, {they 7 ? much, look, yeah, the new hats, so much}. There are three colors. {One is we call turn green, it's the color, like turquoise.} |
| - RW | In this live stream, {There are so many new hats.} There are three colors. { One color is turquoise green because the color looks like turquoise.} |
| - MT$_a$ | In diesem Leben sind sie, sie sind, ähm,. blau sie 7 ? viel, schau, ja, die neuen Hüte, so viel. Es gibt drei Farben. Einer ist, dass wir anrufen Grün werden, das ist die Farbe, wie Türkis. |
| - RW - MT$_a$ | Grün werden, das ist die Farbe, wie Türkis. In diesem Live-Stream gibt es so viele neue Hüte. Es gibt drei Farben. Eine Farbe ist Türkisgrün, weil die Farbe wie Türkis aussieht. |
| ASR | 就是旁边。对。他是一个多功能{多时期}的手表，可以展显示多个时区。{它们都是白色的一个，主要的一个带的搭配。} |
| - RW | 它是一个多功能{多时区}的手表，可以显示多个时区。{它们都搭配白色的表带。} |
| - MT$_a$ | is next to it. right. He is a multi-functional multi-period watch that can display multiple time zones. They're both the white one, with the main one strapped to match. |
| - RW - MT$_a$ | It is a multi-functional multi-time zone watch that can display multiple time zones. They all come with a white strap. |

Table 1: Examples of rewrite annotation and translation examples, with operations highlighted in respective colors and explained in 3.1.

nation issue, resulting in wrongly corrections or introduce new errors. To alleviate this issue, annotators are instructed to make corrections only based on the context within the given segments of automatic transcript. Please refer to Appendix A.2 for details about our annotations to minimize hallucination problems.

Blue tags represent the comprehensive operations that simulate the interpreting process, and annotators must use a combination of operations to complete the rewrite annotation task. We make it clear that excessive compression or loss of important information, commonly seen in human simultaneous interpretation, should be avoided.

### 3.2 Text Rewrite Through Knowledge Distillation

The crux of our approach is to enable the model to learn from human-generated rewrite annotations, and then incorporate the automatic rewrite results into the speech translation pipeline without incurring additional components.

Firstly, we train a text rewrite model to learn from human rewrite. In our practice, the rewrite model uses a typical encoder-decoder transformer architecture (Vaswani et al., 2017). Let RW stand for text ReWrite model, s denote a source speech input, $\mathrm{x} = (x_1, \cdots, x_m)$, and let $\mathrm{y} = (y_1, \cdots, y_l)$ and $\mathrm{z} = (z_1, \cdots, z_n)$, denote the corresponding ASR transcript, translation reference, and the rewritten text of the transcript, respectively. It is first initialized from a pre-trained language model and then fine-tuned on labeled rewrite data. Given the annotated rewrite training data $\mathcal{D}_{\mathrm{rw}} = \{(\mathrm{x}, \mathrm{z})\}$, the training objective of the rewrite model is defined as follows:

$$\hat{\theta}_{\mathrm{rw}} = \arg\max_{\theta} \sum_{(\mathrm{x},\mathrm{z})\in\mathcal{D}_{\mathrm{rw}}} \log P(\mathrm{z}|\mathrm{x};\theta) \quad (1)$$

If we add the RW component in the cascade ST pipeline, it will incur extra latency. We address this problem by performing a sequence-level knowledge distillation (Hinton et al., 2015; Kim and Rush, 2016; Pino et al., 2020). As shown in Figure 1, we employ the cascade sub-system composed of the RW and MT components to generate pseudo parallel sentence $(\mathrm{x}, \mathrm{y})$ by pairing the ASR output x with the MT output $\mathrm{y} = \mathrm{MT}(\mathrm{RW}(\mathrm{x}))$.

When we focus on the utility of the manually annotated rewrite training data $\mathcal{D}_{rw}$, for each sample $(\mathrm{x}, \mathrm{z}) \in \mathcal{D}_{rw}$, we translate z to y using the trained MT model and obtain a pseudo parallel corpus $\mathcal{D}_{rw|mt}$:

$$\{(\mathrm{x}, \mathrm{y}) : (\mathrm{x}, \mathrm{z}) \in \mathcal{D}_{rw}, \mathrm{y} = \mathrm{MT}(\mathrm{z})\} \quad (2)$$

which is used to train a new MT model that can perform text rewrite implicitly during translation.

## 4 Experiments Settings

### 4.1 Data Sets

We construct the human rewrite annotation set based on two widely used datasets in speech translation studies: BSTC (Zhang et al., 2021) and MuST-C (Di Gangi et al., 2019), and an in-house monolingual dataset. The statistics of the rewrite

training data are summarized in Table 2. We conducted an analysis of the statistical changes before and after rewriting in Appendix A.3.

We evaluate the effect of text rewrite on a variety of speech translation test sets in multiple translation directions. In addition to the aforementioned BSTC, ECLS and MuST-C corpora that are used in the rewrite annotation, we also evaluate on MSLT (Federmann and Lewis, 2017) and CoVoST 2 (Wang et al., 2020a) test sets.

- **ECLS**: this dataset is constructed internally from the e-commerce live streaming audio recordings in Mandarin and English. The hosts in audio recordings of Mandarin speech are native speakers, some of whom have strong dialectal accent. The English-speaking hosts have varying levels of fluency, ranging from native speakers with regional accents to non-native speakers with strong accents and often choppy utterances. We present a Chinese to English test set of ECLS that consists of six audio files, each of which contains an ASR transcript, a human transcript, and three human translations. Translators are instructed to produce high quality translations directly from the audio files and deal with the irregularities in speeches with interpreting strategies. The source of the test set is the ASR transcripts, which had an average CER (character error rate) of 23.78.

- **BSTC**: this dataset is constructed for Chinese-English speech translation. The training set is based on 68 hours of Mandarin speech from videos of talks. Since the BSTC test set is not available for public use, we evaluate on the released development set for Chinese-English translation. Speech irregularities are removed from human translation on this development set. We utilize the ASR transcription with a CER of 14.8 for this development set.

- **MuST-C**: the dataset comprises several hundred hours of audio recordings from English TED Talks, we use ten hours of which for English rewrite annotation. The WER of its ASR transcription in test set is 10.7.

- **MSLT**: this test set is constructed from spontaneous conversations on Skype. It provides raw, verbatim human transcripts with eleven kinds of irregularities tagged. Human translators are instructed to produce high-quality translations without translating the irregularities. The ASR transcripts are used as the source for the test set, with the measured WER being 17.47 for the English-Chinese, Japanese test sets and 28.25 for the English-French, German test sets.

- **CoVoST 2**: this English-German speech translation test set contains fewer irregularities than the other datasets. The WER of each ASR transcription in CoVoST 2 is 22.7.

| Dateset | Lang. | Seg. | Average tokens per Seg. | | |
|---|---|---|---|---|---|
| | | | w/o RW | w/ RW | Diff. |
| BSTC | Zh | 22,000 | 60.8 | 56.5 | -4.3 |
| MuST-C | En | 4,084 | 89.5 | 87.4 | -2.1 |
| ECLS | Zh | 33,395 | 68.1 | 57.8 | -10.3 |
| | En | 17,117 | 92.9 | 68.4 | -24.5 |

Table 2: Summary of rewrite training sets.

| Dataset | Language | Segments | Tokens |
|---|---|---|---|
| BSTC | Zh-En | 956 | 26,059 |
| ECLS | Zh-En | 1,000 | 28,716 |
| MSLT | En-Zh/Ja | 2,217 | 38,990 |
| | En-De/Fr | 3,133 | 52,280 |
| MuST-C | En-De | 2,534 | 51,592 |
| CoVoST 2 | En-De | 15,530 | 166,337 |

Table 3: Summary of ST test sets.

### 4.2 Metrics

To report the character error rate (CER) on Chinese (which does not employ word segmentation) and the word error rate (WER) on English, we use jiwer[1]. For a comprehensive and fair evaluation for the ST translation, we adopt three distinct metrics. First, we employ the detokenized, case-insensitive sacreBLEU[2] (Post, 2018) with default options. Secondly, we use BLEURT (Pu et al., 2021) with the BLEURT-20 model. Lastly, we apply COMET (Rei et al., 2020) with the wmt22-comet-da model as released in (Rei et al., 2022).

### 4.3 System Settings

We generate Chinese and English automatic speech transcripts for all datasets using an open-

[1] https://github.com/jitsi/jiwer
[2] https://github.com/mjpost/sacrebleu

source ASR API [3] developed by Gao et al. (2020). The API offers an optional punctuation model developed by Chen et al. (2020). The respective recognition errors on the test sets are shown in Table 4 and Table 5.

For the implementation of the rewrite model, we leverage the pre-trained language model of PALM[4] (Bi et al., 2020) to initialize the parameters of our rewrite model. PALM is built upon the encoder-decoder Transformer architecture and specifically designed for context-conditioned generation. After initialization, the rewrite model is subsequently fine-tuned with the rewrite training data. The base-size PALM model incorporates a 12-layer encoder, a 12-layer decoder, 768 embedding/hidden size, 3072 feed-forward filter size, and 12 attention heads. We use 8 NVIDIA P100 GPUs and a beam search with a size of 4 during inference. We also experimented with mBART (Liu et al., 2020) as a pre-training model for the rewrite fine-tuning and found its performance to be on par with PALM. Considering that PALM requires fewer training parameters, we have elected to present our experiments using the base-size of PALM only.

As our rewrite approach does not depend on any specific MT models, we trained a neural machine translation model, denoted as $\text{MT}_w$, on the WMT22 Chinese-English translation dataset[5] for text rewrite distillation fine-tuning. $\text{MT}_w$ adopts the base transformer model (Vaswani et al., 2017) with a BPE (Sennrich et al., 2015) vocabulary of 32,000 tokens. The base transformer architecture is a 6-layer encoder-decoder model with an embedding size of 512. We set the learning rate to 0.0001, the dropout to 0.1, the warming-up steps to 4,000, the batch size to 2,000 tokens, and the label smoothing to 0.1 for the cross-entropy loss. The training of the $\text{MT}_w$ model is conducted using eight NVIDIA P100 GPUs, and a beam size of 4 is employed during inference.

[3] https://www.modelscope.cn/models/damo/speech_UniASR_asr_2pass-en-16k-common-vocab1080-tensorflow1-offline/

[4] The English and Chinese PALM models can be accessed respectively at https://github.com/overwindows/PALM and https://modelscope.cn/models/damo/nlp_palm2.0_pretrained_chinese-base/summary.

[5] https://www.statmt.org/wmt22/

| Systems | BSTC | | | ECLS | | |
|---|---|---|---|---|---|---|
| ASR | 14.8 | | | 23.8 | | |
| -$\text{MT}_w$ | 16.2 | 57.6 | 71.7 | 11.3 | 56.0 | 64.5 |
| **-RW-$\text{MT}_w$** | **17.7** | **59.6** | **72.9** | **12.8** | **57.7** | **66.8** |
| **-$\text{MT}_w^{\text{RW}}$** | **17.9** | **59.8** | **74.1** | **12.6** | **57.8** | **67.2** |

Table 4: Main results on **Zh-En** ST test sets. The metric is CER for Chinese ASR, and BLEU (↑), BLEURT (↑) and COMET (↑) in the order from left to right for the ST systems.

## 5 Results and Analysis

### 5.1 Main Results

We assess the effectiveness of our method and present the results in Table 4. We focus on the conversion of text rewrite annotations $\mathcal{D}_{rw}$, collected from BSTC and ECLS for this experiment, into pseudo parallel sentences. We compare two pipeline systems to the baseline pipeline ASR-$\text{MT}_w$: (1) pipeline system ASR-RW-$\text{MT}_w$, where RW is the text rewrite model trained on the rewrite annotations $\mathcal{D}_{rw}$, and (2) system ASR-$\text{MT}_w^{\text{RW}}$ with a standalone translation model $\text{MT}_w^{\text{RW}}$ that is fine-tuned on pseudo parallel data $\{(\text{x}, \text{MT}(\text{z}))\}$, where each sample is created by using the base translation model $\text{MT}_w$ to translate a rewritten text $z$ in $\mathcal{D}_{rw}$. The results in Table 4 indicate that the pipelined system ASR-RW-$\text{MT}_w$ and distilled system ASR-$\text{MT}_w^{\text{RW}}$ approach achieve comparable performance and significantly outperform the baseline system ASR-$\text{MT}_w$, demonstrating the effectiveness of the distillation approach.

We also conduct a comparison by directly fine-tuning the MT model with interpreting corpora of BSTC in Appendix B. In Appendix C, we illustrate an evaluation that shows how the knowledge distillation method leverages unlabeled text. We also evaluate the domain robustness of the text rewrite models in Appendix D.

### 5.2 Evaluation on Various ST Directions

Conventional speech translation systems directly integrate disfluency model as an additional component. While this increases computational cost and latency, it allows easy application to multi-language translation directions. To verify the effectiveness of our approach across multi-language speech translation, and to compare text rewrite to disfluency handling, we also directly integrated the rewrite model into the ST pipeline. We uti-

| Systems | MSLT | | | | | | | | | | | | CoVoST 2 | | | MuST-C | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | En-Zh | | | En-Ja | | | En-Fr | | | En-De | | | En-De | | | En-De | | |
| ASR | 17.5 | | | 17.5 | | | 28.3 | | | 28.3 | | | 22.7 | | | 10.7 | | |
| -MT$_a$ | 38.6 | 64.5 | 81.2 | 22.3 | 53.9 | 79.0 | 32.9 | 51.2 | 73.5 | 27.3 | 56.2 | 73.6 | 27.1 | 59.5 | 73.4 | 24.6 | 59.6 | 73.5 |
| -DR-MT$_a$ | 40.1 | 64.7 | 81.6 | 22.9 | 54.3 | 80.1 | 35.0 | 51.6 | 74.0 | 29.1 | 56.4 | 74.2 | – | | | – | | |
| **-RW-MT$_a$** | **41.0** | **65.8** | **82.1** | **23.7** | **55.0** | **81.1** | **35.9** | **53.1** | **75.0** | **29.7** | **57.9** | **75.2** | **27.8** | **60.5** | **74.2** | **24.9** | **60.5** | **74.4** |
| -MT$_g$ | 37.9 | 65.0 | 80.8 | 22.2 | 57.6 | 80.6 | 33.7 | 52.9 | 74.4 | 28.2 | 58.5 | 75.5 | 31.8 | 65.5 | 78.0 | 27.7 | 65.2 | 77.5 |
| -DR-MT$_g$ | 39.9 | 65.1 | 81.0 | 23.6 | 58.1 | 80.6 | 35.8 | 53.2 | 74.6 | 30.0 | 58.7 | 75.9 | – | | | – | | |
| **-RW-MT$_g$** | **41.2** | **65.7** | **82.0** | **24.2** | **58.9** | **81.5** | **36.6** | **54.9** | **75.9** | **30.3** | **60.1** | **77.0** | **32.4** | **66.8** | **78.7** | **28.2** | **66.9** | **78.6** |

Table 5: Main results of text rewrite and its comparison with disfluency detection on **En-X** ST test sets. The metric is WER for English ASR, and BLEU (↑), BLEURT (↑) and COMET (↑) in the order from left to right for the ST systems.

.

| Rewrite vs Original | Transcription | Translation |
|---|---|---|
| Better | 74.5% | 28.0 % |
| Equal | 22.5% | 68.5% |
| Worse | 3.0% | 3.5% |

Table 6: Human evaluation of model rewrite on a partial ECLS test set.

lized two popular commercial MT engines, referred to as MT$_a$[6] and MT$_g$[7] for multi-language translation ST pipeline ASR-RW-MT$_a$ and ASR-RW-MT$_g$. Due to the lack of a public disfluency system, we opted to manually annotate the MSLT test set by asking the annotators to label and remove the classic types of disfluency listed in the annotation sets of MSLT such as repeats, fillers, restarts, and non-speech noise, as well as repairing improper punctuation. However, the more flexible combination of operations newly defined in this paper, as shown in Table 1, are not allowed. We denoted the ST pipeline with disfluency as ASR-DR-MT$_a$ and ASR-DR-MT$_g$.

Table 5 shows the results of various translation directions of En-X across multiple test sets. For En-{Zh, Ja, Fr, De} language pairs, significant improvements are obtained by the RW model on the most task related test sets MSLT, with the average BLEU score improved by 2.3 and 2.6 with MT$_a$ and MT$_g$ respectively. Consistent results were also observed with the BLEURT and COMET metrics. We also observed modest improvements on the CoVoST2 and MuST-C En-De test sets, which contain fewer speech irregularities.

As shown in Tables 5, the model-based rewrite approach outperformed manual disfluency removal by an average of 0.79 and 0.76 BLEU when combined with MT$_a$ and MT$_g$, respectively. Feedback from annotators indicated that the text-rewrite annotation allowed for more flexibility in transforming the original noisy transcript into a cleaner representation.

### 5.3 Human Evaluation and Case Study

We randomly selected 200 segments from the Chinese ECLS test set[8] and asked two experts to: 1) compare the quality of speech transcriptions before and after rewriting, and 2) rate the translation quality (produced by MT$_w$) on a scale of 1 to 5 before or after model rewrite, without knowing which system produced the results.

As illustrated in Table 6, 74.5% rewrite of the rewrite outputs were judged to be better than the original transcription, with 58.5% receiving two votes for higher quality, and merely 3% of the rewrites were rated worse, due to missing words or hallucinations. In terms of translation evaluation, 28% of the rewrite outputs resulted in better translation quality, compared to 3.5% that result in worse quality. The human evaluations indicate that model rewrite significantly improves translation quality.

A few examples of text rewrite and corresponding translations generated from our rewrite method are presented in Appendix E.

## 6 Conclusion

Our rewrite annotation mimics human interpretation to handle various irregularities and can be

[6] https://translate.alibaba.com/

[7] https://translate.google.com/

[8] Table 4 has automatic evaluation results on the full set.

integrated into a translation model using knowledge distillation. Consequently, our proposed rewrite approach offers a cost-efficient way to significantly enhance the speech translation quality across multiple language directions. In our practice, we have utilized the encoder-decoder architecture for both rewrite and translation models. However, our approach can also be easily applied with decoder-only models for automatic text rewrite and translation fine-tuning.

## Acknowledgments

This work was supported by Alibaba Innovative Research Program.

## References

Raja Al-Khanji, Said El-Shiyab, and Riyadh Hussein. 2000. On the use of compensatory strategies in simultaneous interpretation. *Journal des Traducteurs 45(3):548–577*.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*.

Erik Camayd-Freixas. 2011. Cognitive theory of simultaneous interpreting and training. *In Proceedings of the 52nd Conference of the American Translators Association*.

Qian Chen, Mengzhe Chen, Bo Li, and Wen Wang. 2020. Controllable time-delay transformer for real-time punctuation prediction and disfluency detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8069–8073. IEEE.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE.

Eunah Cho, Jan Niehues, Thanh-Le Ha, and Alex Waibel. 2016. Multilingual disfluency removal using nmt. In *Proceedings of the 13th International Conference on Spoken Language Translation*.

Tong Cui, Jinghui Xiao, Liangyou Li, Xin Jiang, and Qun Liu. 2021. An approach to improve robustness of nlp systems against asr errors. *arXiv preprint arXiv:2103.13610*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics.

Qianqian Dong, Feng Wang, Zhen Yang, Wei Chen, Shuang Xu, and Bo Xu. 2019. Adapting translation models for transcript disfluency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6351–6358.

Christian Federmann and William D. Lewis. 2017. The Microsoft Speech Language Translation (MSLT) Corpus for chinese and japanese: Conversational test data for machine translation and speech recognition. In *Proceedings of the 16th Machine Translation Summit (MT Summit XVI)*, Nagoya, Japan.

Zhifu Gao, Shiliang Zhang, Ming Lei, and Ian McLoughlin. 2020. Universal asr: Unifying streaming and non-streaming asr using a single encoder-decoder model. *arXiv preprint arXiv:2010.14099*.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. *Linguistic Data Consortium Philadelphia, USA*.

Jinxi Guo, Tara N Sainath, and Ron J Weiss. 2019. A spelling correction model for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5651–5655. IEEE.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–976.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Oleksii Hrinchuk, Mariya Popova, and Boris Ginsburg. 2019. Correction of automatic speech recognition with transformer sequence-to-sequence model. *arXiv preprint arXiv:1910.10697*.

Mark Johnson and Eugene Charniak. 2004. A tag-based noisy-channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 33–39.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiangyang Li, Edward Lin, and Tie-Yan Liu. 2021. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. *Advances in Neural Information Processing Systems*, 34:21708–21719.

Minhua Liu. 2008. How do experts interpret. *Implications from research in interpreting studies and cognitive*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Evgeny Matusov, Patrick Wilken, Parnia Bahar, Julian Schamper, Pavel Golik, Albert Zeyer, Joan Albert Silvestre-Cerda, Adria Martinez-Villaronga, Hendrik Pesch, and Jan-Thorsten Peter. 2018. Neural speech translation at apptek. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 104–111.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2021. Speechformer: Reducing information loss in direct speech translation. *arXiv preprint arXiv:2109.04574*.

Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos, and Stefanos Vrochidis. 2022. Lard: Large-scale artificial disfluency generation. *arXiv preprint arXiv:2201.05041*.

Juan Miguel Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. In *INTERSPEECH*.

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, Brussels, Belgium. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Annual Meeting of the Association for Computational Linguistics*.

Nikhil Saini, Drumil Trivedi, Shreya Khare, Tejas Dhamecha, Preethi Jyothi, Samarth Bharadwaj, and Pushpak Bhattacharyya. 2021. Disfluency correction using unsupervised and semi-supervised learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3421–3427.

Elizabeth Salesky, Matthias Sperber, and Alex Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech translation. *arXiv preprint arXiv:1906.00556*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 90–96.

Vivek Kumar Rangarajan Sridhar, John Chen, and Srinivas Bangalore. 2013. Corpus analysis of simultaneous interpretation data for improving real time speech translation. In *INTERSPEECH*, pages 3468–3472.

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. 2020a. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Shaolei Wang, Wanxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. 2020b. Multi-task self-supervised learning for disfluency detection. In *AAAI*.

Wen Wang, Gokhan Tur, Jing Zheng, and Necip Ayan. 2010. Automatic disfluency removal for improving spoken language translation. In *Proc. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5214 – 5217.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Qi Ju, Tong Xiao, Jingbo Zhu, et al. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. *arXiv preprint arXiv:2105.05752*.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. Bstc: A large-scale chinese-english speech translation dataset. *arXiv preprint arXiv:2104.03575*.

Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2020. Neurst: Neural speech translation toolkit. *arXiv preprint arXiv:2012.10018*.

Jinming Zhao, Philip Arthur, Gholamreza Haffari, Trevor Cohn, and Ehsan Shareghi. 2021. It is not as good as you think! evaluating simultaneous machine translation on interpretation data. *arXiv preprint arXiv:2110.05213*.

# A Annotation Details

## A.1 Annotation Guideline

The detailed text rewrite annotation guidelines were developed by incorporating practical strategies employed by expert interpreters (Al-Khanji et al., 2000; Liu, 2008; He et al., 2016), as well as referencing previous research on disfluency and word error correction. Annotators are provided with examples as shown in Table 1, and asked to perform the rewrite task defined as follows:

> *Based on your understanding of the original text, within its context as an error-prone automatic transcript of human speech, rewrite it into a fluent and grammatically correct form that maintains the original meaning and can be used as a prepared speech in the corresponding application scenario.*

The cost of rewrite annotation is $0.011/word for English and $0.0025/character for Mandarin. In contrast, the cost of translation annotation is $0.055 /word and /character respectively. To ensure quality, each ASR transcript is processed by one annotator and cross-checked by another who can make further edits or discard a sample of annotation altogether. The overall annotation process took around six months due to multiple iterative cycles to improve the quality of annotation.

## A.2 Word Error Correction Annotation

Consider the example of an audio segment with a reference transcript "She looked for her watch for an hour", which was transcribed by an ASR system as "She looked for her wallet for an hour". In this case, it is not reasonable to expect any text-only rewrite model to correct "wallet" to "watch", as it lacks informative context. If annotators make such annotation data based on the audio files, it could lead to hallucination problems, as the text rewrite model may learn to change "wallet" to "watch" in similar contexts, even though the correct transcript is actually "wallet.".

To alleviate this issue, annotators are instructed to make corrections only based on the context within the given segments of automatic transcript. In the first sentence in Table 1, based on the context, the annotators correct 'turn green' to 'turquoise green.' With such annotations, the transformer model can learn to make word modifica-

tions based on the context, thereby reducing hallucination issues.

### A.3 Annotation Analysis

The statistics of the rewrite training data are summarized in Table 2. As expected, the text after human rewrite is shorter than the original text, and the difference is highly dependent on the domain and the quality of the original text. Among the three datasets, the largest difference in length is observed in the ECLS data set. This is because the majority of hosts in e-commerce live streams are merchants themselves who are not trained professionally in live streaming. This results in more disfluent transcripts, especially in the English transcripts as many hosts are not native speakers.

## B Comparison with Translation Fine-tuning

Table 7 provides a performance comparison between text rewrite and direct MT fine-tuning using speech translation annotations. The data used in translation fine-tuning is the Zh-En BSTC training set, and the evaluation is carried out on the BSTC development set.

In the volume-equivalent setting, we fine-tune the translation model $\text{MT}_w$ using 28K parallel sentences from the BSTC parallel training set, matching the amount of text rewrite annotations collected on BSTC. The resulting translation model is denoted as $\text{MT}_w^{\text{FT}_{\text{vol}}}$.

Since our rewrite annotation is faster and cheaper than interpretation annotation, to account for the cost difference, we introduce a cost-equivalent setting. We randomly sample 3K parallel sentences from the BSTC parallel training set for direct fine-tuning, aiming to match the budget allocated for rewrite annotations on 28K sentences, as estimated through quotes from our vendors. The resulting translation model is labeled $\text{MT}_w^{\text{FT}_{\text{cost}}}$.

Table 7 illustrates that while our proposed system ASR-$\text{MT}_w^{\text{RW}}$ yields a slightly lower BLEU score compared to system ASR-$\text{MT}_w^{\text{FT}_{\text{vol}}}$ that is directly fine-tuning on full manual parallel training data, it outperforms the system ASR-$\text{MT}_w^{\text{FT}_{\text{cost}}}$ in translation quality in the cost-equivalent setting. Furthermore, unlike the direct fine-tuning approach that requires manual annotations for each translation direction, our approach requires only monolingual data annotation and can benefit translations to multiple target languages.

| Methods | Systems | BSTC | | |
|---|---|---|---|---|
| | ASR-$\text{MT}_w$ | 16.2 | 57.6 | 71.7 |
| RW | -RW-$\text{MT}_w$ | 17.7 | 59.6 | 72.9 |
| | -$\text{MT}_w^{\text{RW}}$ | 17.9 | 59.8 | 74.1 |
| FT | -$\text{MT}_w^{\text{FT}_{\text{cost}}}$ | 17.6 | 59.7 | 73.1 |
| | -$\text{MT}_w^{\text{FT}_{\text{vol}}}$ | 18.1 | 60.6 | 74.3 |

Table 7: Comparison of the RW method and FT method on the BSTC development set. The metrics from left to right are BLEU (↑), BLEURT (↑) and COMET (↑).

## C Knowledge Distillation with Unlabeled Data

We focus on distilling an existing rewrite model $\text{RW}_{\text{ECLS}}$, which is trained on rewrite annotations collected from ECLS for this experiment, on the BSTC dataset. We again compare two systems: (1) pipelined system ASR-$\text{RW}_{\text{ECLS}}$-$\text{MT}_w$, and (2) system ASR-$\text{MT}_w^{\text{RW}_{\text{ECLS}}}$ with a standalone translation model $\text{MT}_w(\text{RW}_{\text{ECLS}})$ that is fine-tuned on pseudo parallel data, in which each sample $(\text{x}, \text{MT}(\text{RW}'(\text{x})))$ is created by first rewriting a source sentence x from $\mathcal{D}_{rw}$ using the rewrite model RW' and then translating it using the base translation model. Once again, the results in Table 7 demonstrate that the ASR-$\text{MT}_w^{\text{RW}_{\text{ECLS}}}$ approach achieves comparable or even better performance than the pipeline approach across all evaluation metrics.

| Method | BSTC | | |
|---|---|---|---|
| ASR-$\text{MT}_w$ | 16.2 | 57.6 | 71.7 |
| -RW-$\text{MT}_w$ | 17.7 | 59.6 | 72.9 |
| -$\text{MT}_w^{\text{RW}}$ | 17.9 | 59.8 | 74.1 |
| -$\text{RW}_{\text{ECLS}}$-$\text{MT}_w$ | 17.2 | 58.9 | 72.0 |
| -$\text{MT}_w^{\text{RW}_{\text{ECLS}}}$ | 17.3 | 59.0 | 72.2 |

Table 8: ST performance comparisons of knowledge distillation RW method with unlabeled data on the BSTC development set. The metrics are BLEU (↑), BLEURT (↑) and COMET (↑).

## D Domain Robustness

To verify that the effectiveness of our method is not confined to a specific training set, as well as to assess whether the rewrite model can learn general linguistic phenomena across different domains,

we compared the performance of two rewrite models. One is trained on annotated segments from BSTC, denoted as $\mathcal{D}_{\text{rw}} = \{\text{x}, \text{z}\}$, and the other was trained on an equivalent amount of annotated segments from ECLS, with $\mathcal{D}_{\text{rw}} = \{\text{x}^{\text{e}}, \text{z}^{\text{e}}\}$. We cross-tested these models on each other. As shown in Table 9, although both models perform better on the domain they were trained on, both significantly improve the translation quality on both domains against the baseline.

| Systems | BSTC | | | ECLS | | |
|---|---|---|---|---|---|---|
| ASR-MT$_w$ | 16.2 | 57.6 | 71.7 | 11.3 | 56.0 | 64.5 |
| -RW$_{\text{BSTC}}$-MT$_w$ | 17.3 | 59.1 | 72.2 | 12.4 | 57.2 | 65.5 |
| -RW$_{\text{ECLS}}$-MT$_w$ | 17.2 | 58.9 | 72.0 | 13.0 | 58.0 | 67.0 |

Table 9: Cross-domain evaluation between BSTC and ECLS. The metrics are BLEU (↑), BLEURT (↑) and COMET (↑) in the order from left to right.

## E Examples from Text Rewrite Model

Table 10 presents the text rewrites and their corresponding translations generated by our rewrite method using MT$_g$ and MT$_a$. Table 11 provides examples generated from the distilled translation system MT$_w^{\text{RW}}$.

| Systems | Rewrite Examples from En-Zh on ECLS training set |
|---|---|
| ASR | Tell me which kind of backpack you looking for, we sell women . Backpack man backpacks , handbags. |
| -MT$_a$ | 告诉我您要找哪种背包，我们卖女款。背包男士背包，手袋。 |
| -RW | Tell me which kind of backpack you looking for. We sell women backpacks, men backpacks, handbags. |
| -RW-MT$_a$ | 告诉我你要找哪种背包。我们出售女士背包、男士背包、手提包。 |
| ASR | And uh. In this life, they are they they are they 7 ? You uh , we don't have a giveaway in this live stream. |
| -MT$_g$ | 嗯。这辈子，他们是他们他们是他们 7？你呃，我们这次直播没有赠品。 |
| -RW | We don't have a giveaway in this live stream. |
| -RW-MT$_g$ | 我们在这个直播中没有赠品。 |
| Systems | Rewrite Examples from En-De on MSLT test set |
| ASR | Exactly. And, you know, actually that bring you bring up a good point. |
| -MT$_a$ | Genau. Und wissen Sie, das bringt Sie tatsächlich auf einen guten Punkt. |
| -RW | Exactly, you bring up a good point. |
| -RW-MT$_a$ | Genau, Sie sprechen einen guten Punkt an. |
| ASR | It's a like the hotel is a very it's a very old Italian hotel and it only has a few rooms. |
| -MT$_g$ | Es ist, als wäre das Hotel ein sehr altes italienisches Hotel und es hat nur ein paar Zimmer. |
| -RW | It's a very old Italian hotel and it only has a few rooms. |
| -RW-MT$_g$ | Es ist ein sehr altes italienisches Hotel und es hat nur wenige Zimmer. |

Table 10: Rewrite examples generated from the rewrite model and corresponding translation models of MT$_a$ and MT$_g$ on the sets of MSLT and ECLS.

| Systems | Rewrite Examples from Zh-En on BSTC development set |
|---|---|
| ASR | 走到了那么就要引导一下用户是 okay，我们不能支持你的意思。 |
| -MT$_w$ | When you get there, you need to guide the user to be okay. We can't support what you mean. |
| -MT$_w^{\mathrm{RW}}$ | Then we need to guide the users that we can't support your meaning. |
| ASR | 你也开发者在初次接触这两个指标的时候，说这两个指标到底应该怎么计算。 |
| -MT$_w$ | When you first came into contact with these two indicators, you also said how to calculate these two indicators. |
| -MT$_w^{\mathrm{RW}}$ | When developers first come into contact with these two indicators, how should they be calculated? |
| Systems | Rewrite Examples from Zh-En on ECLS test set |
| ASR | 因为它是一个新品，所以我们新品推广西的时候，它是非常优惠的。 |
| -MT$_w$ | Because it is a new product, it is very favorable when our new product is promoted to the west. |
| -MT$_w^{\mathrm{RW}}$ | Because it is a new product, it is very favorable when we promote the new product. |

Table 11: Translation examples generated from the distilled translation model MT$_w^{\mathrm{RW}}$ on the sets of BSTC and ECLS.