Source and Target Bidirectional Knowledge Distillation for End-to-end Speech Translation

Hirofumi Inaguma[♣] **Tatsuya Kawahara**[♣] **Shinji Watanabe**[♠]
[♣] Kyoto University, Japan [♠] Johns Hopkins University, USA
inaguma@sap.ist.i.kyoto-u.ac.jp

Abstract

A conventional approach to improving the performance of end-to-end speech translation (E2E-ST) models is to leverage the source transcription via pre-training and joint training with automatic speech recognition (ASR) and neural machine translation (NMT) tasks. However, since the input modalities are different, it is difficult to leverage source language text successfully. In this work, we focus on sequence-level knowledge distillation (SeqKD) from external text-based NMT models. To leverage the full potential of the source language information, we propose *backward SeqKD*, SeqKD from a target-to-source backward NMT model. To this end, we train a bilingual E2E-ST model to predict paraphrased transcriptions as an auxiliary task with a single decoder. The paraphrases are generated from the translations in bitext via back-translation. We further propose *bidirectional SeqKD* in which SeqKD from both forward and backward NMT models is combined. Experimental evaluations on both autoregressive and non-autoregressive models show that SeqKD in each direction consistently improves the translation performance, and the effectiveness is complementary regardless of the model capacity.

1 Introduction

End-to-end speech translation (E2E-ST) (Bérard et al., 2016), which aims to convert source speech to text in another language directly, is an active research area. Because direct ST is a more difficult task than automatic speech recognition (ASR) and machine translation (MT), various techniques have been proposed to ease the training process by using source transcription. Examples include pre-training (Bérard et al., 2018; Wang et al., 2020c; Bansal et al., 2019; Wang et al., 2020d), multi-task learning (Weiss et al., 2017; Bérard et al., 2018; Bahar et al., 2019), knowledge distillation (Liu et al., 2019), meta-learning (Indurthi et al., 2020), two-pass decoding (Anastasopoulos and Chiang, 2018; Sperber et al., 2019), and interactive decoding (Liu et al., 2020; Le et al., 2020). However, as input modalities between ST and MT tasks are different, an auxiliary MT task is not always helpful, especially when additional bitext is not available (Bahar et al., 2019). Moreover, because monotonic speech-to-transcription alignments encourage the ASR task to see surface-level local information, an auxiliary ASR task helps the E2E-ST model to extract acoustic representations, not semantic ones, from speech.

Sequence-level knowledge distillation (SeqKD) (Kim and Rush, 2016) is another approach to transferring knowledge from one model to another. Recent studies have shown that SeqKD has the effect of reducing the complexity of training data and thus eases the training of student models, e.g., non-autoregressive (NAR) models (Gu et al., 2018; Zhou et al., 2019a; Ren et al., 2020).

Paraphrasing, which represents text in a different form but with the same meaning, can also be regarded as SeqKD when using neural paraphrasing via back-translation (Mallinson et al., 2017; Wieting et al., 2017; Federmann et al., 2019). It has been studied to improve the reference diversity for MT system evaluations (Thompson and Post, 2020; Bawden et al., 2020a,b) and the performance of low-resource neural MT (NMT) models (Zhou et al., 2019b; Khayrallah et al., 2020).

In this work, due to its simplicity and effectiveness, we focus on SeqKD from text-based NMT models to improve the performance of a bilingual E2E-ST model. In order to fully leverage source language information, we propose *backward SeqKD*, which targets paraphrased source transcriptions generated from a target-to-source backward NMT model as an auxiliary task. Then, a single ST decoder is trained to predict both source and target language text as in a multilingual setting (Inaguma et al., 2019). This way, the decoder is biased to capture semantic representations from speech, un-

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1872–1881
June 6–11, 2021.

like joint training with an auxiliary ASR task. We also propose *bidirectional SeqKD*, which combines SeqKD from two NMT models in both language directions. Therefore, the E2E-ST models can fully exploit the knowledge embedded in both forward and backward NMT models.

Experimental evaluations demonstrate that SeqKD from each direction consistently improves the translation performance of both autoregressive and non-autoregressive E2E-ST models. We also confirm that bidirectional SeqKD outperforms unidirectional SeqKD and that the effectiveness is maintained in large models.

2 Method

In this section, we propose bidirectional SeqKD from both forward and backward NMT models that leverages machine-generated source paraphrases as another target in addition to the distilled translation to enhance the training of a bilingual E2E-ST model. Let X denote input speech features in a source language and Y^{s} and Y^{t} denote the corresponding gold transcription and translation, respectively. Let $\mathcal{D}_{\mathrm{st}} = \{(X_i, Y_i^{\mathrm{s}}, Y_i^{\mathrm{t}})\}_{i=1}^{I}$ be an ST dataset including I samples, and $\mathcal{D}_{\mathrm{asr}} = \{(X_i, Y_i^{\mathrm{s}})\}_{i=1}^{I}$ and $\mathcal{D}_{\mathrm{mt}} = \{(Y_i^{\mathrm{s}}, Y_i^{\mathrm{t}})\}_{i=1}^{I}$ denote the corresponding ASR and MT datasets, respectively.[1] We drop the subscript i when it is obvious.

2.1 Sequence-level knowledge distillation

We first train a text-based source-to-target forward NMT model $\mathcal{M}_{\mathrm{fwd}}$ with $\mathcal{D}_{\mathrm{mt}}$.[2] Then, we perform beam search decoding with $\mathcal{M}_{\mathrm{fwd}}$ on $\mathcal{D}_{\mathrm{st}}$ to create a new dataset $\mathcal{D}_{\mathrm{st}}^{\mathrm{fwd}} = \{(X_i, Y_i^{\mathrm{s}}, \hat{Y}_i^{\mathrm{t}})\}_{i=1}^{I}$, where $\hat{Y}_i^{\mathrm{t}}$ is a distilled translation. $\mathcal{D}_{\mathrm{st}}^{\mathrm{fwd}}$ is used to train the E2E-ST models, referred to as *forward SeqKD* (or fwd SeqKD).

2.2 Paraphrase generation

To exploit semantic information in the source language, we leverage machine-generated paraphrases of source transcriptions. We train a text-based target-to-source backward NMT model $\mathcal{M}_{\mathrm{bwd}}$ with $\mathcal{D}_{\mathrm{mt}}$ and then generate a new dataset $\mathcal{D}_{\mathrm{st}}^{\mathrm{bwd}} = \{(X_i, \hat{Y}_i^{\mathrm{s}}, Y_i^{\mathrm{t}})\}_{i=1}^{I}$, where $\hat{Y}_i^{\mathrm{s}}$ is a paraphrase of Y_i^{s}. We use $\mathcal{D}_{\mathrm{st}}^{\mathrm{bwd}}$ for training the E2E-ST models. As neural paraphrasing can be regarded as SeqKD from $\mathcal{M}_{\mathrm{bwd}}$, we referred to it as *backward SeqKD* (or bwd SeqKD). In this work, we do not use large paraphrase datasets (Wieting and Gimpel, 2018; Hu et al., 2019) because their availability depends on languages and domains. Moreover, neural paraphrasing is applicable to any source languages that lack a sufficient amount of paired paraphrase data.

We also propose combining forward SeqKD with backward SeqKD, referred to as *bidirectional SeqKD* (or bidir SeqKD), and construct a new dataset $\mathcal{D}_{\mathrm{st}}^{\mathrm{bidir}} = \{(X_i, \hat{Y}_i^{\mathrm{s}}, \hat{Y}_i^{\mathrm{t}})\}_{i=1}^{I}$. When using two references per utterance (*2ref* training) (Gordon and Duh, 2019), we concatenate $\mathcal{D}_{\mathrm{st}}^{\mathrm{fwd}}$ and $\mathcal{D}_{\mathrm{st}}^{\mathrm{bwd}}$, and the suitable combination is analyzed in Section 4.3. This way, we can distill the knowledge of both $\mathcal{M}_{\mathrm{fwd}}$ and $\mathcal{M}_{\mathrm{bwd}}$ to a single E2E-ST model.

2.3 Training

We train an E2E-ST model with a direct ST objective $\mathcal{L}_{\mathrm{st}}(Y^{\mathrm{t}}$ or $\hat{Y}^{\mathrm{t}}|X)$ and an auxiliary speech-to-source text objective $\mathcal{L}_{\mathrm{src}}(Y^{\mathrm{s}}$ or $\hat{Y}^{\mathrm{s}}|X)$. We refer to joint training with $\mathcal{L}_{\mathrm{src}}(Y^{\mathrm{s}}|X)$ as *joint ASR* and with $\mathcal{L}_{\mathrm{src}}(\hat{Y}^{\mathrm{s}}|X)$ as *backward SeqKD*. Both losses are calculated from the same ST decoder. To bias the model to generate the desired target language, we add language embedding to token embedding at *every* token position in the decoder (Conneau and Lample, 2019).[3] We then apply bidirectional SeqKD to both autoregressive (AR) and non-autoregressive (NAR) E2E-ST models.

Autoregressive E2E-ST model

We use the speech Transformer architecture in (Karita et al., 2019) with an additional language embedding. The total training objective is formulated with a hyperparameter $\lambda_{\mathrm{src}}(\geq 0)$ as

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{st}} + \lambda_{\mathrm{src}}\mathcal{L}_{\mathrm{src}}, \tag{1}$$

where both $\mathcal{L}_{\mathrm{st}}$ and $\mathcal{L}_{\mathrm{src}}$ are defined as cross-entropy losses. The entire encoder-decoder parameters are shared in both tasks.

Non-autoregressive E2E-ST model

We adopt Orthors (Inaguma et al., 2021), in which a decoder based on a conditional masked language model (CMLM) (Ghazvininejad et al., 2019) is jointly trained with an additional AR decoder

[1] We focus on a complete triplet of $(X, Y^{\mathrm{s}}, Y^{\mathrm{t}})$ only. However, the proposed method can easily be extended to a semi-supervised setting featuring additional ASR and MT pair data.

[2] All NMT models are autoregressive in this paper.

[3] We found this was more effective than replacing the start-of-sentence symbol with a language ID (Inaguma et al., 2019; Wang et al., 2020b; Le et al., 2020) as done in previous multilingual E2E-ST studies.

Language direction	**BLEU** (↑)	**TER** (↓)
De → En	43.49	38.60
Fr → En	48.55	34.30

Table 1: Quality of paraphrases in the training set

on the shared speech encoder. The training of the NAR decoder is further enhanced with semi-autoregressive training (SMART) (Ghazvininejad et al., 2020). $\mathcal{L}_{\text{st}}$ in Eq. (1) is modified as

$$\mathcal{L}_{\text{st}} = \mathcal{L}_{\text{cmlm}} + \lambda_{\text{ar}}\mathcal{L}_{\text{ar}} + \lambda_{\text{lp}}\mathcal{L}_{\text{lp}}, \quad (2)$$

where $\mathcal{L}_{\text{cmlm}}$, $\mathcal{L}_{\text{ar}}$, and $\mathcal{L}_{\text{lp}}$ are losses in NAR E2E-ST, AR E2E-ST, and length prediction tasks, respectively. λ_* is the corresponding tunable loss weight. During inference, the mask-predict algorithm is used for T iterations with a length beam width of l (Ghazvininejad et al., 2019). The best candidate at the last iteration is selected from the NAR decoder based on scores from the AR decoder (Inaguma et al., 2021). Note that we apply $\mathcal{L}_{\text{src}}$ to the NAR decoder only.

3 Experimental setting

Data We used Must-C En-De (408 hours) and En-Fr (492 hours) datasets (Di Gangi et al., 2019). Both language pairs consist of a triplet of (X, Y^{s}, Y^{t}). We performed the same data preprocessing as (Inaguma et al., 2020) (see details in Appendix A.1). We report case-sensitive detokenized BLEU scores (Papineni et al., 2002) on the `tst-COMMON` set with the `multi-bleu-detok.perl` script in Moses (Koehn et al., 2007).

Model configuration We used the Transformer (Vaswani et al., 2017) architecture having 12 encoder layers following two CNN blocks and six decoder layers for the ASR and E2E-ST tasks. For the MT models, we used six encoder layers. We built our models with the ESPnet-ST toolkit (Inaguma et al., 2020). See details in Appendix A.2.

Training We always initialized the encoder parameters of the E2E-ST model by those of the corresponding pre-trained ASR model (Bérard et al., 2018). We follow the same optimization strategies as in (Inaguma et al., 2021, 2020). When using joint ASR or backward SeqKD, we set λ_{src} to 0.3. More details are described in Appendix A.3 and A.4.

ID	Model	**BLEU** (Δ) (↑)			
		En-De		En-Fr	
–	ESPnet-ST†	22.91		32.69	
	Fairseq-S2T‡	22.7		32.9	
	+ Multilingual◇	24.5		34.9	
A1	Baseline	22.77		33.51	
A2	+ MT pre-training	**23.12**	(+0.35)	**33.84**	(+0.33)
A3	+ Joint ASR	22.97	(+0.20)	33.37	(–0.14)
A4	+ Bwd SeqKD	**23.11**	(+0.34)	**33.78**	(+0.23)
B1	A1 + Fwd SeqKD	24.42	(+1.65)	34.66	(+1.15)
B2	+ MT pre-training	24.68	(+1.91)	34.57	(+1.06)
B3	+ Joint ASR	24.67	(+1.90)	34.68	(+1.17)
B4	+ Original (2ref)	**24.83**	**(+2.06)**	**34.92**	**(+1.41)**
C1	A1 + Bidir SeqKD	**24.83**	**(+2.06)**	**34.78**	**(+1.27)**
C2	+ Original (2ref)	**25.28**	**(+2.51)**	**35.29**	**(+1.78)**

Table 2: BLEU scores of AR models on Must-C `tst-COMMON` set. † (Inaguma et al., 2020), ‡ (Wang et al., 2020a). ◇Large model trained with eight language pairs (Wang et al., 2020a).

Inference For the AR models, we used a beam width of 4. For the NAR models, we set $T = \{4, 10\}$ and $l = 9$ as in (Inaguma et al., 2021).

4 Results

4.1 Main results

We first report the paraphrasing quality, which is shown in Table 1. As confirmed by the BLEU and translation edit rate (TER) scores (Snover et al., 2006), the paraphrased source text was not just a simple copy of the transcription (see examples in Appendix A.5).

Autoregressive models The results are shown in Table 2. Pre-training the ST decoder with the forward MT decoder (A2) improved the baseline performance (A1). Joint ASR showed a marginal improvement on En-De but a degraded performance on En-Fr (A3). We attribute this to the fact that the ASR task was more trivial than the ST task and biased the shared decoder to capture surface-level textual information. In contrast, backward SeqKD showed small but consistent improvements in both language directions (A4), and it was as effective as MT pre-training. As the encoder was already pre-trained with the ASR model, paraphrases had an additional positive effect on the BLEU improvement.

Forward SeqKD significantly improved the performance, as previously reported in (Inaguma et al., 2021). However, the gains by MT pre-training and joint ASR were diminished. Forward SeqKD was more effective than backward SeqKD solely (A4 vs. B1). However, backward SeqKD was still

Model	T	BLEU (↑)	
		En-De	En-Fr
Fwd SeqKD	4	21.93	30.46
+ Joint ASR		22.13	30.80
Bidir SeqKD		**22.22**	**31.21**
(Inaguma et al., 2021)	10	22.88	32.20
Fwd SeqKD (ours)		22.96	32.42
+ Joint ASR		23.31	32.41
Bidir SeqKD		**23.41**	**32.64**

Table 3: BLEU scores of NAR models on Must-C `tst-COMMON` set. All methods used forward SeqKD.

beneficial on top of forward SeqKD (`C1`, i.e., bidirectional SeqKD) while joint ASR was less so (`B3`). We also augmented the target translations by concatenating $\mathcal{D}_{\text{st}}$ and $\mathcal{D}_{\text{st}}^{\text{fwd}}$ (*2ref* training), which further improved forward SeqKD (`B4`). Nevertheless, a combination of *2ref* training and backward SeqKD (i.e., bidirectional SeqKD with $\mathcal{D}_{\text{st}}^{\text{fwd}} \cup \mathcal{D}_{\text{st}}^{\text{bwd}}$) had a complementary effect and showed the best result (`C2`). It even outperformed larger multilingual models (Wang et al., 2020a) without using additional data in other language pairs.

Non-autoregressive models The results are presented in Table 3. Following the standard practice in NAR models (Gu et al., 2018), we always used forward SeqKD. We did not use *2ref* training for the NAR models because it increases the multimodality. Joint ASR improved the performance on all NAR models, except for En-Fr with the number of iterations $T = 10$. However, bidirectional SeqKD with $\mathcal{D}_{\text{st}}^{\text{bidir}}$ further improved the performance consistently regardless of T. Since NAR models assume conditional independence for every token, they prefer monotonic input-output alignments with lower alignment complexity in theory. However, paraphrasing collapses the monotonicity of the ASR task and increases the alignment complexity, making the auxiliary speech-to-source text task non-trivial. Nevertheless, BLEU scores were improved by adding backward SeqKD. This was probably because the complexity of transcriptions in the training data was reduced at the cost of the alignment complexity, which was more effective for the NAR models.

4.2 Analysis

We analyze the performance of bidirectional SeqKD through a lens of complexity in the training data following (Zhou et al., 2019a). We aligned words in every source and target sentence pair with

Condition	Entropy (↑ more complex)	
	En-De	En-Fr
$\mathcal{C}(\overrightarrow{\mathcal{D}_{\text{st}}})$ (Real)	0.70	0.65
$\mathcal{C}(\overrightarrow{\mathcal{D}_{\text{st}}^{\text{fwd}}})$ (Fwd SeqKD)	**0.52**	**0.47**
$\mathcal{C}(\overrightarrow{\mathcal{D}_{\text{st}}^{\text{bwd}}})$ (Bwd SeqKD)	0.54	0.47
$\mathcal{C}(\overrightarrow{\mathcal{D}_{\text{st}}^{\text{bidir}}})$ (Bidir SeqKD)	0.63	0.61
$\mathcal{C}(\overleftarrow{\mathcal{D}_{\text{st}}})$ (Real)	0.40	0.54
$\mathcal{C}(\overleftarrow{\mathcal{D}_{\text{st}}^{\text{fwd}}})$ (Fwd SeqKD)	0.28	0.36
$\mathcal{C}(\overleftarrow{\mathcal{D}_{\text{st}}^{\text{bwd}}})$ (Bwd SeqKD)	**0.25**	**0.31**
$\mathcal{C}(\overleftarrow{\mathcal{D}_{\text{st}}^{\text{bidir}}})$ (Bidir SeqKD)	0.37	0.49

Table 4: Corpus-level conditional entropy

Condition	Faithfulness (↓ more faithful)	
	En-De	En-Fr
$\mathcal{F}(\overrightarrow{\mathcal{D}_{\text{st}}^{\text{fwd}}})$ (Fwd SeqKD)	12.61	11.65
$\mathcal{F}(\overrightarrow{\mathcal{D}_{\text{st}}^{\text{bwd}}})$ (Bwd SeqKD)	**9.31**	**8.67**
$\mathcal{F}(\overrightarrow{\mathcal{D}_{\text{st}}^{\text{bidir}}})$ (Bidir SeqKD)	11.42	10.72
$\mathcal{F}(\overleftarrow{\mathcal{D}_{\text{st}}^{\text{fwd}}})$ (Fwd SeqKD)	**9.58**	**8.48**
$\mathcal{F}(\overleftarrow{\mathcal{D}_{\text{st}}^{\text{bwd}}})$ (Bwd SeqKD)	12.97	10.70
$\mathcal{F}(\overleftarrow{\mathcal{D}_{\text{st}}^{\text{bidir}}})$ (Bidir SeqKD)	11.23	9.98

Table 5: Faithfulness to training data distribution

fast_align[4] (Dyer et al., 2013). Then, we calculated corpus-level conditional entropy $\mathcal{C}(\mathcal{D})$ and faithfulness $\mathcal{F}(\mathcal{D})$ for both forward ($\overrightarrow{\mathcal{D}}$) and backward ($\overleftarrow{\mathcal{D}}$) language directions to evaluate the multimodality. In short, conditional entropy measures uncertainty of translation, and faithfulness is defined as Kullback–Leibler divergence and measures how close the distilled data distribution is to the real data distribution. See the mathematical definition in Appendix A.6.

The results of entropy and faithfulness are shown in Tables 4 and 5, respectively. Consistent with (Zhou et al., 2019a), the entropy of target translations was reduced by forward SeqKD, indicating target translations were converted into a more deterministic and simplified form. Interestingly, the entropy of the original translations was also reduced by backward SeqKD. In other words, backward SeqKD modified transcriptions so that the target translations can be predicted easier. This would help E2E-ST models learn relationships between source and target languages *from speech* because E2E-ST models are not conditioned on text in another language explicitly. Therefore, we presume that the encoder representations were enhanced by back-

[4] https://github.com/clab/fast_align

Training data	Target1	Target2	BLEU (↑)	
			En-De	En-Fr
$\mathcal{D}_{\text{st}} \cup \mathcal{D}_{\text{st}}^{\text{fwd}}$ (B4 + Joint ASR)	$(\boldsymbol{Y}^{\text{s}}, Y^{\text{t}})$	$(\boldsymbol{Y}^{\text{s}}, \hat{Y}^{\text{t}})$	25.00	35.05
$\mathcal{D}_{\text{st}} \cup \mathcal{D}_{\text{st}}^{\text{bidir}}$	$(\boldsymbol{Y}^{\text{s}}, Y^{\text{t}})$	$(\hat{\boldsymbol{Y}}^{\text{s}}, \hat{Y}^{\text{t}})$	25.21	35.17
$\mathcal{D}_{\text{st}}^{\text{bwd}} \cup \mathcal{D}_{\text{st}}^{\text{bidir}}$	$(\hat{\boldsymbol{Y}}^{\text{s}}, Y^{\text{t}})$	$(\hat{\boldsymbol{Y}}^{\text{s}}, \hat{Y}^{\text{t}})$	25.01	35.22
$\mathcal{D}_{\text{st}}^{\text{fwd}} \cup \mathcal{D}_{\text{st}}^{\text{bwd}}$ (C2)	$(\hat{\boldsymbol{Y}}^{\text{s}}, Y^{\text{t}})$	$(\boldsymbol{Y}^{\text{s}}, \hat{Y}^{\text{t}})$	**25.28**	**35.29**

Table 6: Ablation study of dataset concatenation on Must-C tst-COMMON set. *2ref* training was used.

Model	BLEU (↑)	
	En-De	En-Fr
Transformer Large + Fwd SeqKD	25.19	35.47
+ Bidir SeqKD	**25.62**	**35.74**
Conformer + Fwd SeqKD	26.81	37.23
+ Bidir SeqKD	**27.01**	**37.33**
Text-based NMT (WER: 0%)†	27.56	39.09

Table 7: BLEU scores of large AR models on Must-C tst-COMMON set. *2ref* training was used. † Punctuation and case information is removed on the source side.

ward SeqKD. Using machine-generated sequences in both languages increased the entropy, probably due to error accumulation. However, E2E-ST models do not suffer from it because they are conditioned on the source speech. We also confirmed similar trends in the reverse language direction.

Regarding faithfulness, distilled target sequences degraded faithfulness as expected. However, an interesting finding was that the faithfulness of bidirectional SeqKD was better than that of forward SeqKD, meaning that the former reflected the true word alignment distribution more faithfully than the latter. Although lexical choice might be degraded by targeting distilled text in both languages (Ding et al., 2021), mixing the original and distilled text by *2ref* training would recover it.

4.3 Ablation study

We conduct an ablation study to verify the analysis in the previous section. In Table 4, we observed that it was better to have the original reference in the target sequence of either the source or target language. For example, to reduce the entropy of German text in the training set, it was best to condition the distilled German translation on the original English transcription, and vice versa. Therefore, we hypothesize that the best way to reduce the entropy in both source and target languages during *2ref* training is to combine $(\hat{Y}^{\text{s}}, Y^{\text{t}})$ and $(Y^{\text{s}}, \hat{Y}^{\text{t}})$ for each sample. We compared four ways to leverage source text: gold transcription Y^{s} only, distilled paraphrase $\hat{Y}^{\text{s}}$ only, and both.[5] The results are shown in Table 6. We confirmed that the model trained with the original reference in either language for every target achieved the best BLEU score, which verifies our hypothesis.

4.4 Increasing model capacity

Finally, we investigate the effectiveness of bidirectional Seq-KD with *2ref* training when increasing the model capacity in Table 7. The purpose of this experiment is to verify our expectation that large models can model complex target distributions in multi-referenced training better. In addition to simply increasing the model dimensions, we also investigate Conformer (Gulati et al., 2020), a Transformer encoder augmented by a convolution module. We confirmed that bidirectional SeqKD always outperformed forward SeqKD in both language directions regardless of model configurations. We also found that the Conformer encoder significantly boosted the translation performance of forward SeqKD, but the gains of bidirectional SeqKD were transferred.

[5]Both gold translation Y^{t} and distilled translation $\hat{Y}^{\text{t}}$ were always used as target sequences.

5 Conclusion

To fully leverage knowledge in both source and target language directions for bilingual E2E-ST models, we have proposed bidirectional SeqKD, in which both forward SeqKD from a source-to-target NMT model and backward SeqKD from a target-to-source NMT model are combined. Backward SeqKD is performed by targeting source paraphrases generated via back-translation from the original translations in bitext. Then, the E2E-ST model is enhanced by training to generate both source and target language text with a single decoder. We experimentally confirmed that SeqKD from each direction boosted the translation performance of both autoregressive and non-autoregressive E2E-ST models, and the effectiveness was additive. Multi-referenced training with the original and distilled text gave further gains. We also showed that bidirectional SeqKD was effective regardless of model sizes.

Acknowledgement

The authors thank the anonymous reviewers for useful suggestions and Siddharth Dalmia, Brian Yan, and Pengcheng Guo for helpful discussions.

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019. A comparative study on end-to-end speech to text translation. In *Proceedings of ASRU*, pages 792–799.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Rachel Bawden, Biao Zhang, Andre Tättar, and Matt Post. 2020a. ParBLEU: Augmenting metrics with automatic paraphrases for the wmt'20 metrics shared task. In *5th Conference on Machine Translation*.

Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. 2020b. A study in improving BLEU reference coverage with diverse automatic paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 918–932, Online. Association for Computational Linguistics.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *Proceedings of ICASSP*, pages 6224–6228.

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. In *Proceedings of NeurIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, pages 7059–7069.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Understanding and improving lexical choice in non-autoregressive translation. In *Proceedings of ICLR*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China. Association for Computational Linguistics.

Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, and Omer Levy. 2020. Semi-autoregressive training improves mask-predict decoding. *arXiv preprint arXiv:2001.08785*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.

Mitchell A Gordon and Kevin Duh. 2019. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. *arXiv preprint arXiv:1912.03334*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *Proceedings of ICLR*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for speech recognition. In *Proceedings of Interspeech*, pages 5036–5040.

J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. ParaBank: Monolingual bitext generation and sentential paraphrasing via

lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *Proceedings of ASRU*, pages 570–577.

Hirofumi Inaguma, Yosuke Higuchi, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2021. Orthros: Non-autoregressive end-to-end speech translation with dual-decoder. In *Proceedings of ICASSP*.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. End-end speech-to-text translation with modality agnostic meta-learning. In *Proceedings of ICASSP*, pages 7904–7908.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on Transformer vs RNN in speech applications. In *Proceedings of ASRU*, pages 499–456.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics.

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

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *Proceedings of ICLR*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proceedings of Interspeech*, pages 3586–3589.

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. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. In *Proceedings of Interspeech*, pages 1128–1132.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *Proceedings of AAAI*, pages 8417–8424.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech*, pages 2613–2617.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *Proceedings of ASRU*.

Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, Sheng Zhao, and Tie-Yan Liu. 2020. A study of non-autoregressive model for sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 149–159, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7:313–325.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proc. of CVPR*, pages 2818–2826.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

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

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020c. Bridging the gap between pretraining and fine-tuning for end-to-end speech translation. In *Proceedings of AAAI*, pages 9161–9168.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020d. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3728–3738, Online. Association for Computational Linguistics.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of Interspeech*, pages 2625–2629.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark. Association for Computational Linguistics.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2019a. Understanding knowledge distillation in non-autoregressive machine translation. In *Proceedings of ICLR*.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2019b. Paraphrases as foreign languages in multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 113–122, Florence, Italy. Association for Computational Linguistics.

Reference1	She took our order, and then went to the couple in the booth next to us, and she lowered her voice so much, I had to really strain to hear what she was saying.
Paraphrase1 (Backward NMT)	She picked up our order, and then went to the pair in the niche next to us and lowered her voice so much that I had to really try to understand them.
Reference2	And she said "Yes, that's former Vice President Al Gore and his wife, Tipper." And the man said, "He's come down a long way, hasn't he?" (Laughter)
Paraphrase2 (Backward NMT)	She said, "Yes, that's ex-vice President Al Gore and his wife Tipper." And the man said, "It's a nice gap, what?" (Laughter)

Table 8: Examples of source paraphrases on the Must-C En-De training set

A Appendix

A.1 Data preprocessing

All sentences were tokenized with the `tokenizer.perl` script in Moses (Koehn et al., 2007). Non-verbal speech labels such as "*(Applause)*" and "*(Laughter)*" were removed during evaluation (Di Gangi et al., 2019; Inaguma et al., 2021; Le et al., 2020). We built output vocabularies based on the byte pair encoding (BPE) algorithm (Sennrich et al., 2016) with the Sentencepiece toolkit (Kudo, 2018)[6]. The joint source and target vocabularies were constructed in the ST and MT tasks, while the vocabularies in the ASR task were constructed with transcriptions only. For autoregressive models, we used 5k for ASR models and 8k for E2E-ST and MT models. We used 16k vocabularies for non-autoregressive E2E-ST models (Inaguma et al., 2021).

For input speech features, we extracted 80-channel log-mel filterbank coefficients computed with a 25-ms window size and shifted every 10ms with 3-dimensional pitch features using Kaldi (Povey et al., 2011). This results in 83-dimensional features for every frame. The features were normalized by the mean and the standard deviation for each training set. To avoid overfitting, training data was augmented by a factor of 3 with speed perturbation (Ko et al., 2015) and SpecAugment (Park et al., 2019). We used $(m_T, m_F, T, F) = (2, 2, 40, 30)$ for the hyperparameters in SpecAugment.

A.2 Model configuration

We used the Transformer (Vaswani et al., 2017) architecture implemented with the ESPnet-ST toolkit (Inaguma et al., 2020) for all tasks. ASR and E2E-ST models consisted of 12 speech encoder blocks and six decoder blocks. The speech encoders had two CNN blocks with a kernel size of 3 and a channel size of 256 before the first Transformer encoder layer, which resulted in 4-fold downsampling in the time and frequency axes. The text encoder in the MT models consisted of six Transformer blocks. The dimensions of the self-attention layer d_{model} and feed-forward network d_{ff} were set to 256 and 2048, respectively, and the number of attention heads H was set to 4. For a large Transformer model configuration, we increased d_{ff} from 256 to 512 and H from 4 to 8. For a Conformer model configuration, we set $d_{\text{model}} = 256$, $d_{\text{ff}} = 2048$, and $H = 4$. The kernel size of depthwise separable convolution was set to 15. None of the other training or decoding hyperparameters were modified.

A.3 Initialization

In addition to initializing the encoder parameters of the E2E-ST model by those of the pre-trained ASR model, the auxiliary AR decoder parameters of the NAR models were initialized by those of the corresponding pre-trained AR MT model (Inaguma et al., 2021). The other decoder parameters of both the AR and NAR models were initialized as in BERT (Devlin et al., 2019; Ghazvininejad et al., 2019; Inaguma et al., 2021), where weight parameters were sampled from $\mathcal{N}(0, 0.02)$, biases were set to zero, and layer normalization parameters were set to $\beta = 0$, $\gamma = 1$. Note that we did not use additional data for pre-training.

A.4 Training

The Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ was used for training with a Noam learning rate schedule (Vaswani et al., 2017). We used dropout and label smoothing (Szegedy et al., 2016) with a probability of 0.1 and 0.1, respectively. The other training configurations for all tasks are summarized in Table 9. We removed utterances having more than 3000 input speech frames or more than 400 characters due to the GPU memory capacity. The last five best checkpoints based on the validation performance were used for model averaging.

[6] https://github.com/google/sentencepiece

Configuration	ASR	E2E-ST		MT
		AR	NAR	
Warmup step	25k	25k	25k	8k
Learning rate factor	5.0 (2.0)	2.5	5.0	1.0
Batch size × accum	128	128	256	96
Epoch	45 (30)	30	50	100
Validation metric	Accuracy	BLEU	BLEU	BLEU

Table 9: Summary of training configuration. Numbers inside parentheses correspond to Conformer.

For the training of ASR models used for E2E-ST encoder pre-training, we removed case and punctuation information from transcriptions and then applied a joint CTC/Attention objective (Watanabe et al., 2017). However, we retained this information in the transcriptions and paraphrases used for training the E2E-ST and MT models.

A.5 Case study

We present examples of generated paraphrases on the Must-C En-De training set in Table 8. We observed that most paraphrases kept the original meaning while some words were simplified to alternatives having a similar meaning. We also found that the first conjunction in an utterance was more likely to be omitted via paraphrasing.

A.6 Mathematical formulation of complexity and faithfulness

In this section, we mathematically formulate the corpus-level complexity and faithfulness given $\mathcal{D} \in \{\mathcal{D}_{\text{st}}, \mathcal{D}_{\text{st}}^{\text{fwd}}, \mathcal{D}_{\text{st}}^{\text{bwd}}, \mathcal{D}_{\text{st}}^{\text{bidir}}\}$. Our formulation follows (Zhou et al., 2019a), but we also consider the reverse language direction.

Conditional entropy (complexity) The corpus-level complexity of $\mathcal{D}$ in the forward language direction, $\mathcal{C}(\overrightarrow{\mathcal{D}})$, is defined as the conditional entropy $\mathcal{H}(Y^{\text{t}}|Y^{\text{s}})$ normalized over all samples. $\mathcal{H}(Y^{\text{t}}|Y^{\text{s}})$ is defined as

$$
\begin{aligned}
&\mathcal{H}(Y^{\text{t}}|Y^{\text{s}}) \\
&= -\sum_{i=1}^{I} p(Y_i^{\text{t}}|Y_i^{\text{s}}) \cdot \log p(Y_i^{\text{t}}|Y_i^{\text{s}}) \\
&\approx -\sum_{i=1}^{I} \Big(\prod_{k=1}^{|Y_i^{\text{t}}|} p(Y_{i,k}^{\text{t}}|Y_i^{\text{s}})\Big) \cdot \sum_{k=1}^{|Y_i^{\text{t}}|} \log p(Y_{i,k}^{\text{t}}|Y_i^{\text{s}}) \\
&\approx -\sum_{k=1}^{T^{\text{t}}} \sum_{y_k^{\text{t}} \in \mathcal{A}(Y_i^{\text{s}})} p(y_k^{\text{t}}|\text{Align}(y_k^{\text{t}})) \cdot \log p(y_k^{\text{t}}|\text{Align}(y_k^{\text{t}})) \\
&= \sum_{k=1}^{T^{\text{s}}} \mathcal{H}(y^{\text{t}}|Y_{i,k}^{\text{s}}),
\end{aligned}
$$

where $\mathcal{A}$ is an external alignment model, and T^{s} and T^{t} are the source and target sequence lengths, respectively. We make two assumptions: (1) conditional independence of target tokens given the source text sequence, and (2) the distribution of $p(y^{\text{t}}|Y^{\text{s}})$ follows the alignment model $\mathcal{A}$. Then, $\mathcal{C}(\overrightarrow{\mathcal{D}})$ is calculated as

$$
\mathcal{C}(\overrightarrow{\mathcal{D}}) = \frac{1}{|\mathcal{V}^{\text{s}}|} \sum_{y^{\text{s}} \in \mathcal{V}^{\text{s}}} \mathcal{H}(y^{\text{t}}|y^{\text{s}}),
$$

where $\mathcal{V}^{\text{s}}$ is a set of all words in the source language. Division by $|\mathcal{V}^{\text{s}}|$ is important to normalize frequent source words.

The corpus-level complexity of $\mathcal{D}$ in the backward language direction, $\mathcal{C}(\overleftarrow{\mathcal{D}})$, is defined similarly as

$$
\mathcal{C}(\overleftarrow{\mathcal{D}}) = \frac{1}{|\mathcal{V}^{\text{t}}|} \sum_{y^{\text{t}} \in \mathcal{V}^{\text{t}}} \mathcal{H}(y^{\text{s}}|y^{\text{t}}),
$$

where $\mathcal{V}^{\text{t}}$ is a set of all words in the target language.

Faithfulness Although the corpus-level conditional entropy can be used to evaluate the complexity of the training data, there are also trivial solutions to generate new data with smaller complexity when target translations are not adequate. Faithfulness is a good measure to assess how close the distilled data distribution is to the real (original) data distribution. The faithfulness of $\mathcal{D}$ in a forward language direction $\mathcal{F}(\overrightarrow{\mathcal{D}})$ and a backward language direction $\mathcal{F}(\overleftarrow{\mathcal{D}})$ is defined as the KL-divergence of the alignment distribution between the real dataset and a distilled dataset, as

$$
\mathcal{F}(\overrightarrow{\mathcal{D}}) = \frac{1}{|\mathcal{V}^{\text{s}}|} \sum_{y^{\text{s}} \in \mathcal{V}^{\text{s}}} \sum_{y^{\text{t}} \in \mathcal{V}^{\text{t}}} p_{\text{r}}(y^{\text{t}}|y^{\text{s}}) \log \frac{p_{\text{r}}(y^{\text{t}}|y^{\text{s}})}{p_{\text{d}}(y^{\text{t}}|y^{\text{s}})},
$$

$$
\mathcal{F}(\overleftarrow{\mathcal{D}}) = \frac{1}{|\mathcal{V}^{\text{t}}|} \sum_{y^{\text{t}} \in \mathcal{V}^{\text{t}}} \sum_{y^{\text{s}} \in \mathcal{V}^{\text{s}}} p_{\text{r}}(y^{\text{s}}|y^{\text{t}}) \log \frac{p_{\text{r}}(y^{\text{s}}|y^{\text{t}})}{p_{\text{d}}(y^{\text{s}}|y^{\text{t}})},
$$

where p_{r} and p_{d} are alignment distributions of the real and distilled data, respectively. Therefore, when $\mathcal{D} = \mathcal{D}_{\text{st}}$, $\mathcal{F}(\overrightarrow{\mathcal{D}}) = \mathcal{F}(\overleftarrow{\mathcal{D}}) = 0$, and it was omitted in Table 5.