

On Knowledge Distillation for Translating Erroneous Speech Transcriptions

Ryo Fukuda¹, Katsuhito Sudoh^{1,2}, and Satoshi Nakamura^{1,2}

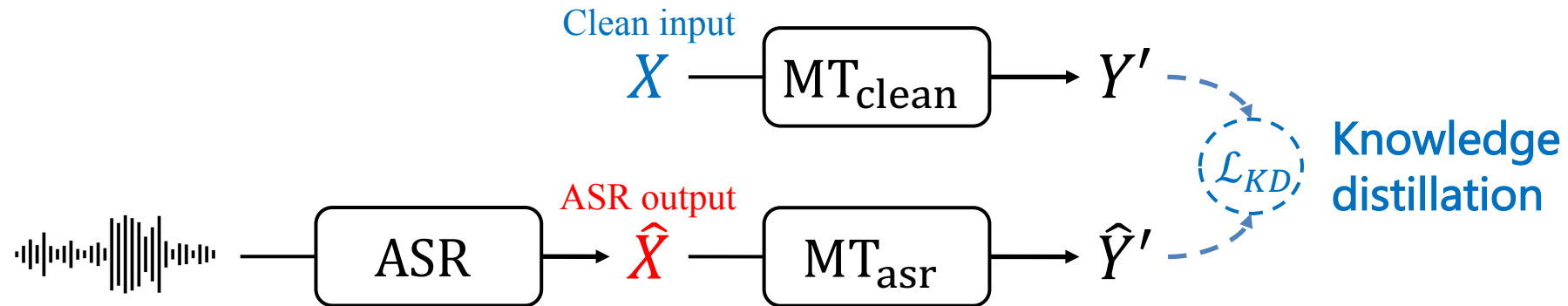
¹Nara Institute of Science and Technology, Japan

²AIP, RIKEN, Japan



Brief Overview

In this work, we investigate the effect of **knowledge distillation (KD)** with a speech translation using ASR and MT models.



- Experimental results demonstrated that KD brings 0.4-1.0 BLEU improvement.
- The combination of KD and fine-tuning (FT) consistently improved two language pairs (Must-C En-It and Fisher Es-En) up to 1.5 BLEU.

Background

Speech Translation (ST)

ST converts utterances in a source language into text in another language.

- **Cascade ST** consists of two components: ASR and MT.
 - the error propagation from ASR to MT
- **End-to-end ST** uses a single model to directly translate speech into text.
 - naive end-to-end ST without additional training remains inferior to a cascade ST
 - requires parallel data of the source language speech and the target language text, which cannot be obtained easily

We focus on the cascade approach due to performance advantage against end-to-end STs and tackle the problem of ASR error propagation.

ASR-based input for MT training

We can use (1) clean human transcripts or (2) erroneous transcriptions by ASR for the MT training on cascade ST, but

- (1) discrepancy of the inputs of training and inference leads to error propagation.
 - (2) realistic assumption bridges the gap between training and inference, but the training of noisy-to-clean text translation is difficult.
- We investigated how to use both types of input for training.
 - Our solution: knowledge distillation and fine-tuning.

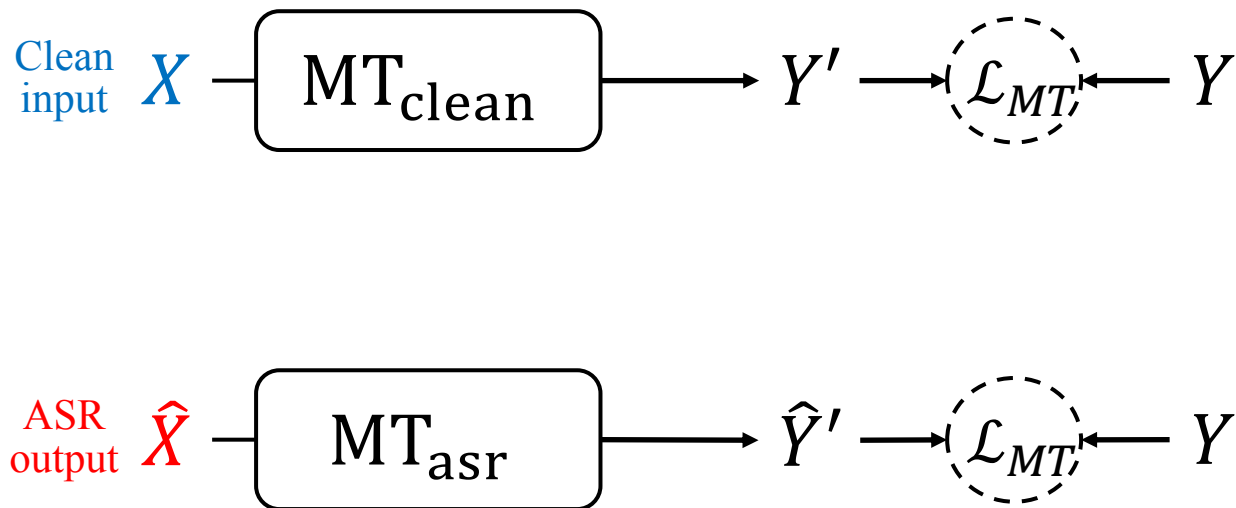
Related work

- Di Gangi et al. (2019):
showed that a model **fine-tuned** with ASR-based input becomes robust to erroneous ASR input for cascade ST.
 - Following this finding, we employ FT and investigate the joint use of KD and FT.
- Dakwale and Monz (2019):
proposed **knowledge distillation** as a remedy for the effective use of noisy parallel data for machine translation.
 - Unlike this study, we have loosely equivalent source sentences (clean or erroneous transcription).

Method

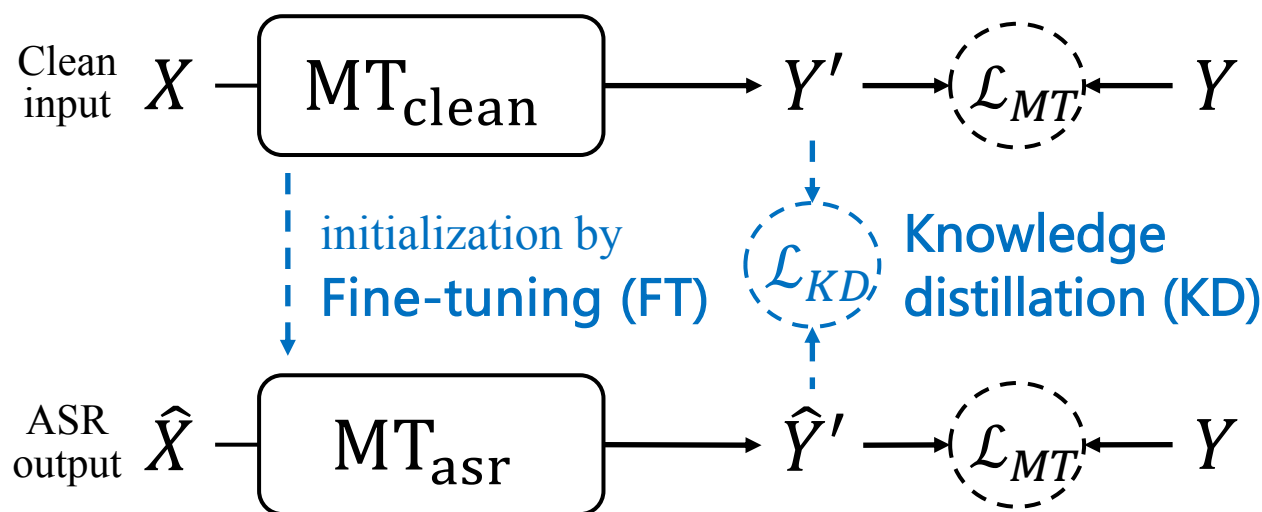
Training of MT for cascade ST

- We have clean transcripts of source language X and ASR output \hat{X} as input of MT, and translation Y as output of MT.
 - Normally, MT model is trained to generate Y by minimizing loss function \mathcal{L}_{MT} .



Fine-tuning and Knowledge Distillation

- We have clean transcripts of source language X and ASR output \hat{X} as input of MT, and translation Y as output of MT.
 - Normally, MT model is trained to generate Y by minimizing loss function \mathcal{L}_{MT} .
 - Additionally, we introduce two training techniques: Fine-tuning (FT) and Knowledge Distillation (KD).



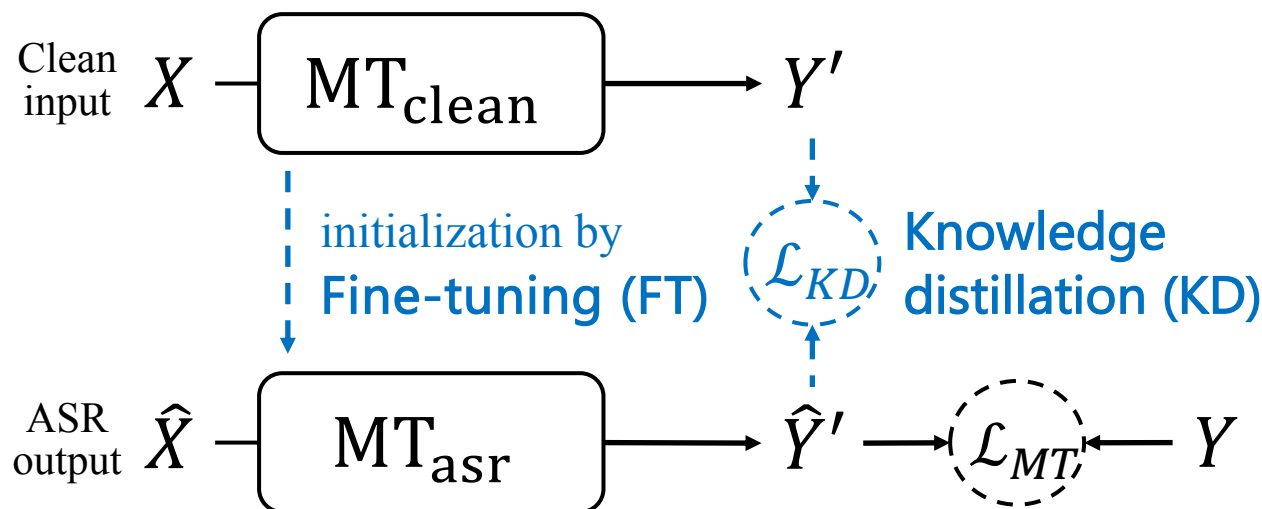
Joint use of KD and FT

- We examined possible combinations of KD and FT:
 - (1) FT + KD. Apply these techniques at the same time.
 - (2) KD \rightarrow FT. Perform additional training with \mathcal{L}_{MT} to model trained by KD.
 - (3) FT \rightarrow KD. Perform additional training with \mathcal{L}_{KD} to model trained by FT.

Joint use of KD and FT (1) FT + KD

(1) FT + KD. Apply these techniques at the same time.

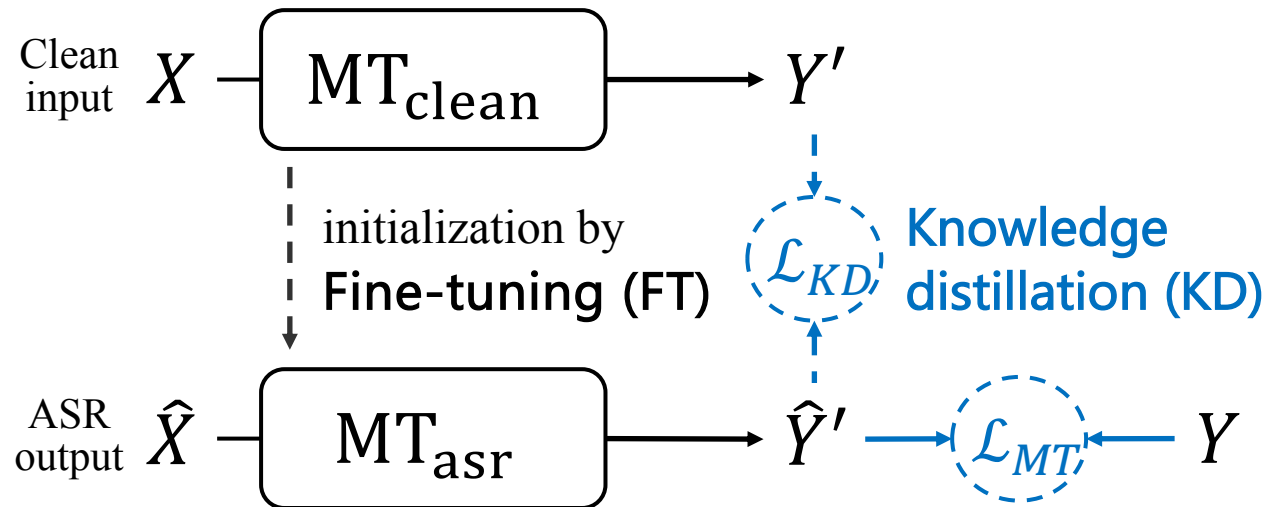
1. the teacher model is trained with clean input X and loss \mathcal{L}_{MT} .
2. the student model is trained with ASR-based input \hat{X} and loss \mathcal{L}_{KD} , inheriting the parameters of the teacher model.



Joint use of KD and FT (2) KD \rightarrow FT

(2) KD \rightarrow FT. Perform additional training with \mathcal{L}_{MT} to model trained by KD.

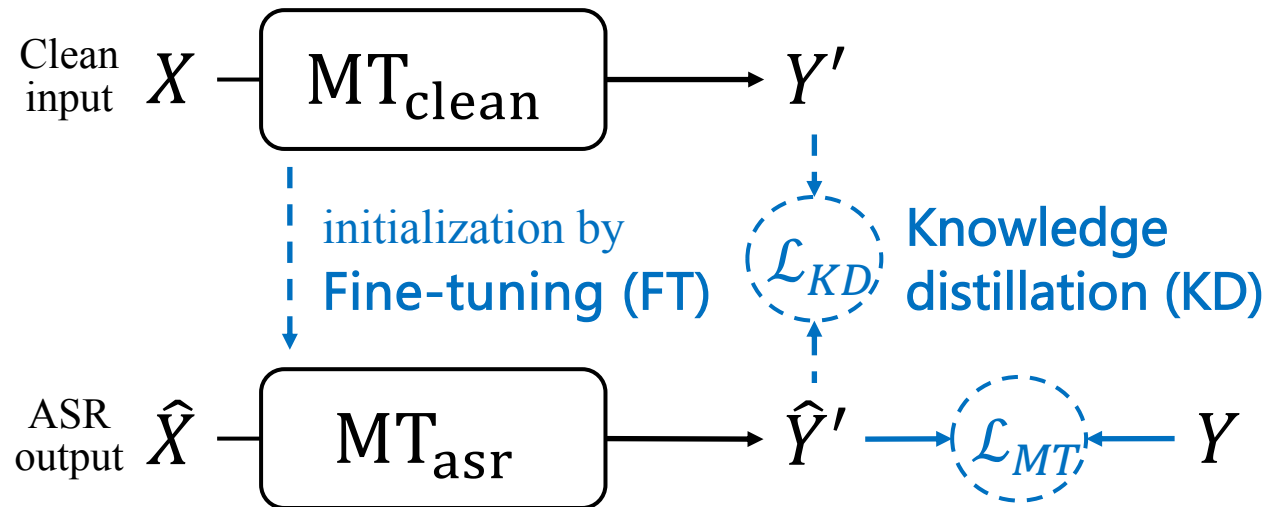
1. the student model is trained with \hat{X} and \mathcal{L}_{KD} .
2. fine-tune the model with \hat{X} and \mathcal{L}_{MT} .



Joint use of KD and FT (3) FT \rightarrow KD

(3) FT \rightarrow KD. Perform additional training with \mathcal{L}_{KD} to model trained by FT.

1. the student model is trained with \hat{X} and \mathcal{L}_{MT} , inheriting the parameters of the teacher model.
2. fine-tune the model with \hat{X} and \mathcal{L}_{KD} .



Experiments

Evaluation

- Experiment 1: English to Italian NMT
 - We used MuST-C, which contains triplets of about 250K segments of English speeches, transcripts, and Italian translations.
 - ASR results contained erroneous transcriptions of WER 14.49 (lower WER condition).
- Experiment 2: Spanish to English NMT
 - We used LDC Fisher Spanish speech with fluent English translations, which has about 140K segments.
 - ASR-based inputs included in the dataset have many erroneous transcriptions of WER 36.5 (higher WER condition).

Experiment 1: MuST-C English to Italian

System	Test data	
	ASR-based input	Clean input
MT _{clean}	22.4	29.7
MT _{asr}	22.1	27.2
MT _{asr} + FT	23.2 ↑1.1	29.8
MT _{asr} + KD	22.5 ↑0.4	28.2
MT _{asr} + FT + KD	23.4	29.9
MT _{asr} + KD → FT	23.1	29.3
MT _{asr} + FT → KD	23.5 ↑1.4	30.2

- KD produced a slight improvement than FT for the ASR-based input of WER 14.49.
- Joint use of them provided more improvement for both the ASR-based and the clean input.

Experiment 2: Fisher Spanish to English

System	Test data	
	ASR-based input	Clean input
MT _{clean}	17.5	26.8
MT _{asr}	17.5	17.6
MT _{asr} + FT	18.3 ↑0.8	24.9
MT _{asr} + KD	18.5 ↑1.0	16.5
MT _{asr} + FT + KD	18.8	25.2
MT _{asr} + KD → FT	17.8	15.7
MT _{asr} + FT → KD	19.0 ↑1.5	25.2

- KD was superior to FT for the ASR-based input of WER 36.5.
- Joint use of them provided greater improvement for the ASR-based input.

The results of the two experiments suggest that the more colloquial the speech, the more beneficial the KD training may be.

Discussion: Effect of Knowledge Distillation

- KD forces the student model to **mimic literal teacher translations** that may include some errors instead of reproducing translations of colloquial spoken utterances.

Input: le ayuda si si, no es, no es interesante pero entonces, a ba- entonces ya despues cuando eso termino, tiene que escribir varios asi, ensayos, hacer un analisis

Translation: You have to write some essays like that, to make an analysis

KD teacher: It helps her yes, it's not interesting but then, when I finish, you have to write several, you have to make an analysis

Discussion: Effect of Fine-tuning

- FT for the erroneous ASR outputs may have provided robustness against common errors.

Correct input: Eh, para mi pues, eh, tengo como diez mil canciones en, en el, en la **Ipod**

ASR output: eh para mi pues eh tengo como diez mil canciones en en la **epod**

Correct translation: I have ten thousand songs in the **Ipod**.

MT w/o FT: To me, I have about ten thousand songs in the **ethics**

MT w/ FT: I have about ten thousand songs in the **Ipod**

Conclusion

We presented and discussed the benefits of using two types of inputs in cascade ST: **clean transcript and ASR output**.

- The experiments results demonstrated that the **KD is beneficial for a cascade ST**.
- The combination of KD and fine-tuning (FT) consistently improved two language pairs with high and low WER conditions.

In future work, we will incorporate our findings into an end-to-end ST to grow speech translation.