Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation

Ye Jia, Melvin Johnson, Wolfgang Macherey, **Ron Weiss**, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, Yonghui Wu

Google Research

@ICASSP 2019

End-to-End Speech-to-Text Translation (ST)

- Task: English speech to Spanish text translation
- End-to-end models have outperformed cascaded systems on small tasks
- Goal: Scale it up and see if this still holds
- Use "weakly supervised" ASR and MT data (spanning part of the task) via:
    1. pretraining network components
    2. multitask training
    3. synthetic target translation (~distillation) and
       synthetic source speech (~back translation)

Experiment data

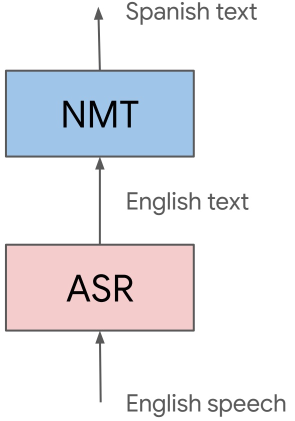
- Fully (1x) and weakly (100x) supervised training corpora
  - **ST-1**: 1M read English speech → Spanish text
    - conversational speech translation
  - **MT-70**: 70M English text → Spanish text
    - web text, superset of ST-1
  - **ASR-29**: 29M transcribed English utterances
    - anonymized voice search logs
- Evaluation
  - **In-domain:** read speech, held out portion of ST-1
  - **Out-of-domain**: spontaneous speech

Baseline: Cascade ST model

- Train ASR model on ASR-29 and ST-1, NMT model on MT-70
  - both sequence-to-sequence networks with attention
- Pro: easy to build from existing models
- Con: compounding errors, long latency
- Metrics (case-sensitive, including punctuation)

	In-domain	Out-of-domain
ASR (WER) *	13.7%	30.7%
NMT (BLEU)	78.8	35.6
ST Cascade (BLEU)	**56.9**	**21.1**

* ASR WER -- if case-insensitive w/o punctuation, 6.9% for in-domain and 14.1% for out-of-domain.

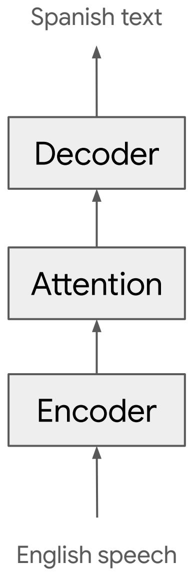


Cascaded model for En-Es speech translation

Fused end-to-end speech translation model

- Fuse recognition and translation into a single sequence-to-sequence model
- Smaller model, lower latency
- Challenge: training data expensive to collect
- Train on ST-1

	In-domain	Out-of-domain
Cascaded	**56.9**	**21.1**
Fused	49.1	12.1



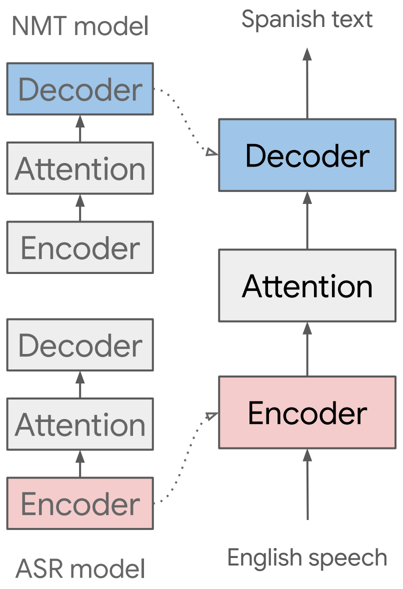
Fused model for En-Es speech translation

Berard, et al., A proof of concept for end-to-end speech-to-text translation. NeurIPS workshop 2016.
Weiss, et al., Sequence-to-sequence models can directly translate foreign speech. Interspeech 2017.

Strategy 1: Pretraining

- Pretrain encoder on ASR-29 task, decoder on MT-70 task
- *Fine-tune* on ST-1 data
- Model sees the same training data as cascade
- Simplest way to incorporate weakly supervised data

	In-domain	Out-of-domain
Cascaded	**56.9**	**21.1**
Fused	49.1	12.1
Fused + pretraining	54.6	18.2



Fused model for En-Es speech translation

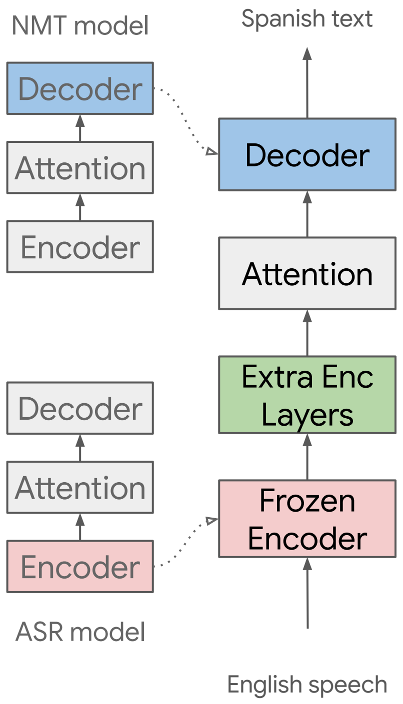
Google

Berard, et al., End-to-end automatic speech translation of audiobooks. ICASSP 2018.
Bansal, et al., Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. NAACL 2019.

Strategy 1: Pretraining - Freezing encoder layers

- Generalizes better to out-of-domain speech
  - will avoid overfitting to synthetic speech
- Append additional trainable layers,
  allowing adaptation to deep encoded representation

	In-domain	Out-of-domain
Cascaded	**56.9**	**21.1**
Fused	49.1	12.1
Fused + pretraining	54.6	18.2
Fused + pretraining w/frozen enc	54.5	19.5
Fused + pretraining w/frozen enc + 3 layers	55.9	19.5

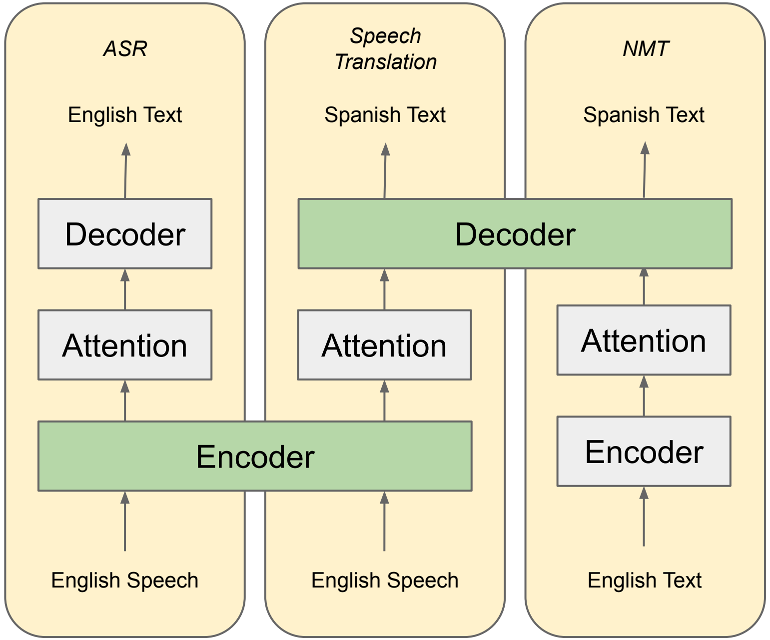


Google

Strategy 2. Multitask learning

- Train ST / ASR / NMT jointly, with shared components
  - sample task independently at each step
- Utilize all available datasets

	In-domain	Out-of-domain
Cascaded	56.9	21.1
Fused	49.1	12.1
Fused + pretraining	54.6	18.2
Fused + pretraining + extra enc	55.9	19.5
Fused + pretraining + multitask	**57.1**	**21.3**

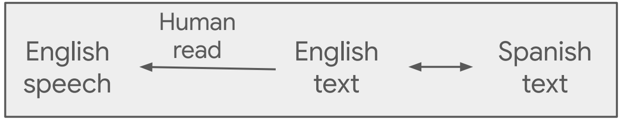


Weiss, et al., Sequence-to-sequence models can directly translate foreign speech. Interspeech 2017.
Anastasopoulos, et al., Tied multitask learning for neural speech translation. NAACL-HLT 2018.

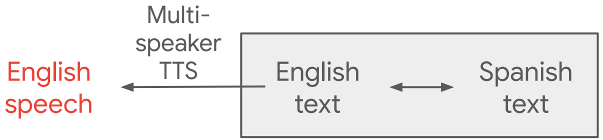
Strategy 3: Synthetic training data

- Utilize all available datasets
  - convert weakly supervised data to fully supervised
- From MT-70 dataset
  - synthesize source English speech with TTS using multispeaker Tacotron model
  - similar to back-translation
- From ASR-29 dataset
  - synthesize target Spanish translation with MT
  - similar to knowledge distillation

Real collected ST training set (ST-1):



Synthesized from MT training set (MT-70):



Synthesized from ASR training set (ASR-29):

Jia, et al., Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. NeurIPS 2018.
Sennrich, et al., Improving neural machine translation models with monolingual data, ACL 2016
Hinton, et al., Distilling the knowledge in a neural network, NeurIPS 2015

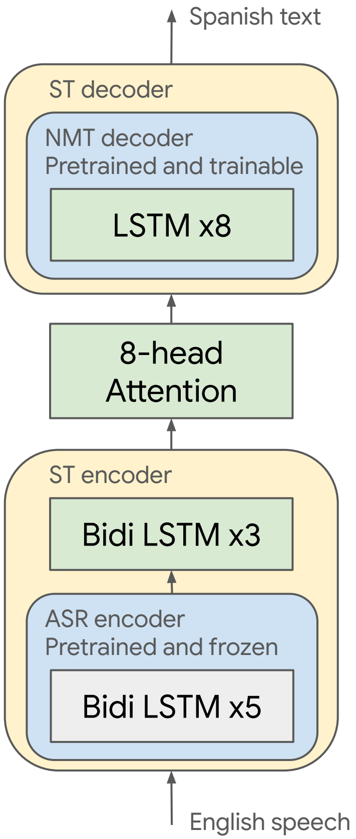
Strategy 3: Synthetic training data - Results

- Sample dataset independently at each step
- Significantly outperforms baseline
- Synthetic text gives bigger improvement on out-of-domain set
    - better match to spontaneous speech

	Fine-tuning set	In-domain	Out-of-domain
Cascaded		56.9	21.1
Fused		49.1	12.1
Fused + pretraining	ST-1 + synthetic speech	**59.5**	22.7
Fused + pretraining	ST-1 + synthetic text	57.9	26.2
Fused + pretraining	ST-1 + synthetic speech / text	**59.5**	**26.7**

Final model architecture

- Sequence-to-sequence with attention
  - 8-layer encoder
  - 8-layer decoder
  - 8-head additive attention
- Pretraining
  - Lower encoder layers pretrained on ASR-29
    - frozen during ST training
  - Decoder pretrained on MT-70
    - fine tuned during ST training
- No multitask learning



Fine-tuning with only synthetic data

- Can train a speech translation model without any fully supervised data!
  - distillation from pre-existing ASR and MT models

	Fine-tuning set	In-domain	Out-of-domain
Cascaded		56.9	21.1
Fused		49.1	12.1
Fused + pretraining	ST-1 + synthetic speech / text	**59.5**	26.7
Fused + pretraining	synthetic speech / text	55.6	**27.0**

Training with unsupervised data

- From unlabeled speech:
    - Synthesize target translation with cascaded ST system
- From unlabeled text:
    - Synthesize source speech with TTS, synthesize target translation with NMT
- Significantly improves over fused model trained only on ST-1

	Training set	In-domain	Out-of-domain
Fused	ST-1	49.1	12.1
Fused	ST-1 + synthetic from unlabeled speech	52.4	15.3
Fused	ST-1 + synthetic from unlabeled text	**55.9**	**19.4**
Fused	ST-1 + synthetic from unlabeled speech / text	55.8	16.9

Synthetic data: Encoder ablation

- Fully trainable encoder overfits if fine-tuned on synthetic speech alone

Fine-tuning set	Encoder	In-domain	Out-of-domain
ST-1 + synthetic speech	freeze first 5 layers	**59.5**	**22.7**
ST-1 + synthetic speech	fully trainable	58.7	21.4
synthetic speech	freeze first 5 layers	53.9	20.8
synthetic speech	fully trainable	35.1	9.8

Synthetic data: TTS ablation

- Model overfits if fine-tuned using a single-speaker Tacotron 2 TTS model
  - worse on out-of-domain speech, prosody closer to read speech?

Fine-tuning set	TTS model	In-domain	Out-of-domain
ST-1 + synthetic speech	multispeaker	**59.5**	**22.7**
ST-1 + synthetic speech	single speaker	**59.5**	19.5
synthetic speech	multispeaker	53.9	20.8
synthetic speech	single speaker	38.5	13.8

Jia, et al., Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. NeurIPS 2018.
Shen, et al., Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, ICASSP 2018.

Summary

- Train an end-to-end speech translation (ST) model on 1M parallel examples
    - Underperforms cascade of ASR and NMT models
- Recipe for building ST model with minimal (or no) parallel training data:
    - Can outperform cascade by pretraining ST components, and fine-tuning on
        - ***back-translated*** TTS speech from MT-70 training set
        - ***distilled*** translations for ASR-29 training set
    - Fine-tuning without <u>any</u> real parallel examples still perform wells

	Fine-tuning set	In-domain	Out-of-domain
Cascaded		56.9	21.1
Fused		49.1	12.1
Fused + pretraining	ST-1 + synthetic speech/text	**59.5**	26.7
Fused + pretraining	synthetic speech/text	55.6	**27.0**