Self-Training for End-to-End Speech Recognition

Jacob Kahn, Ann Lee, and Awni Hannun
Facebook AI Research
ICASSP 2020

Outline

1

Self-Training in End-to-End ASR

Motivating/defining the pipeline and related work.

2

Baseline Acoustic and Language Model, Filtering, and Ensembles

Key components for sequence-to-sequence models.

3

Results

WER on LibriSpeech datasets with pseudo-labeling, improving on prior results.

4

Future Work

Extending self-training-style techniques in speech.

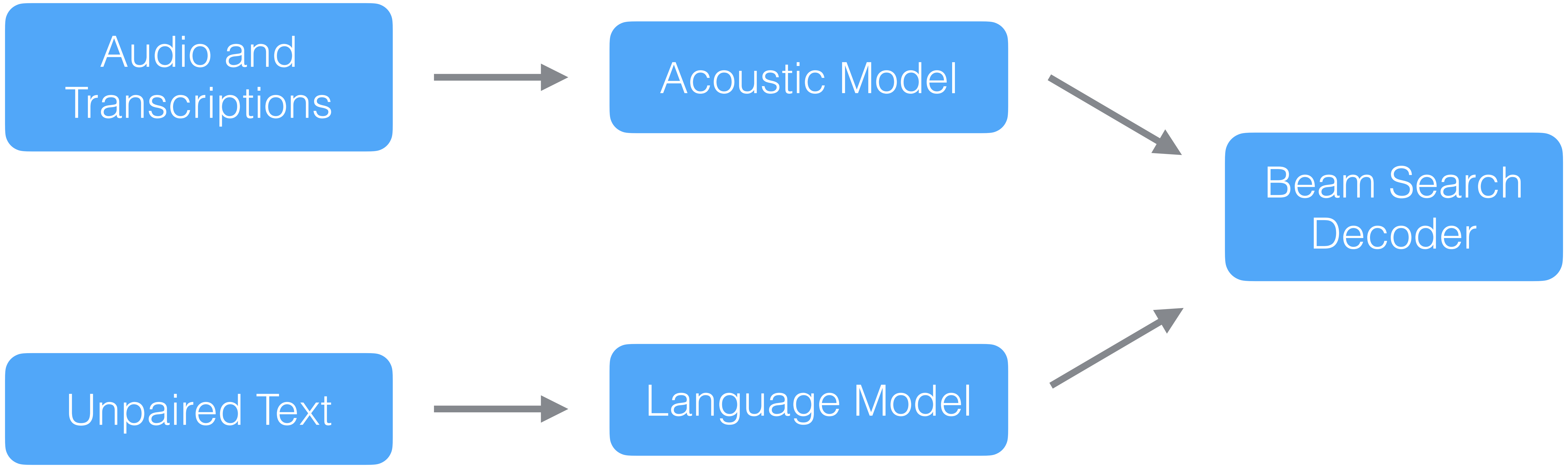
Motivation, Related Work, and Pipeline

An overview of self-training and prior work in speech.

— 1

Our goal is to leverage unlabeled audio and text.

End-to-end systems' performance degrades with less training data.



Our goal is to leverage unlabeled audio and text.

End-to-end systems' performance degrades with less training data.

Unlabeled

Audio and Transcriptions

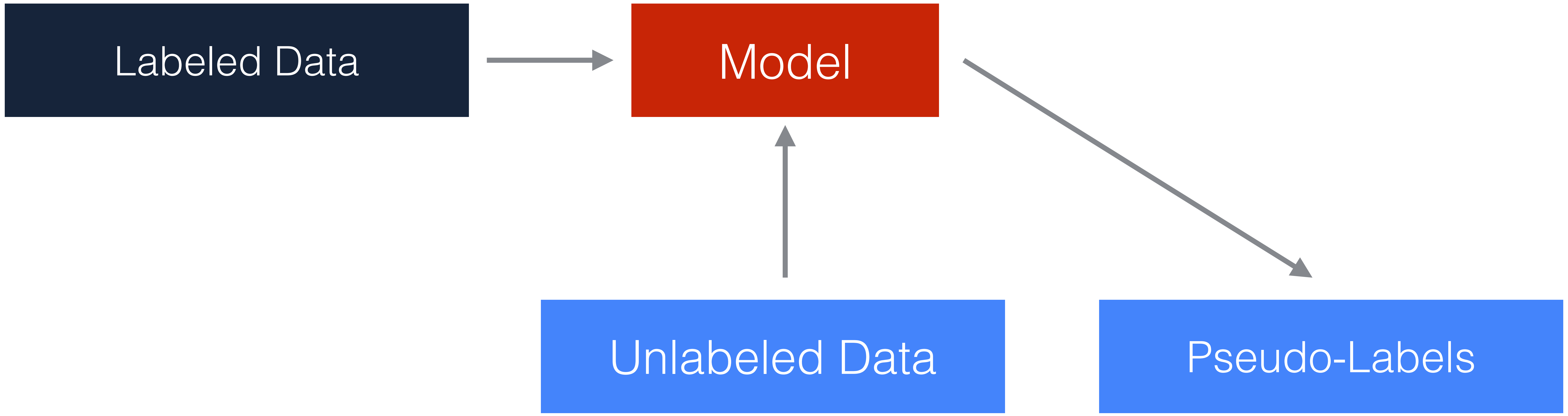
Unpaired Text

Self-Training

Use the same model to generate labels for unlabeled data; train on the resulting labels.

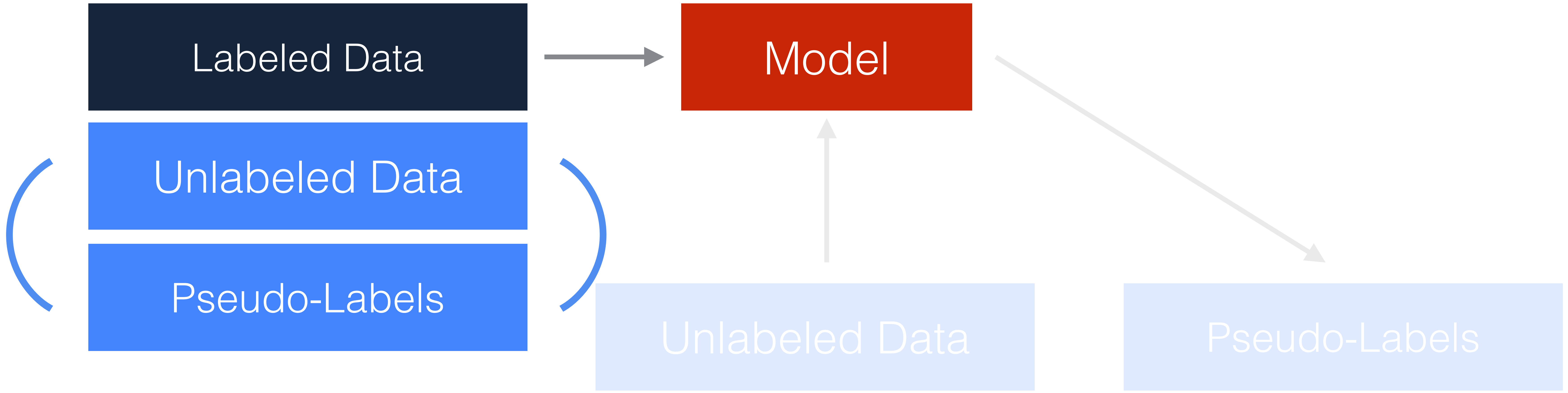
Self-Training

Use the same model to generate labels for unlabeled data; train on the resulting labels.



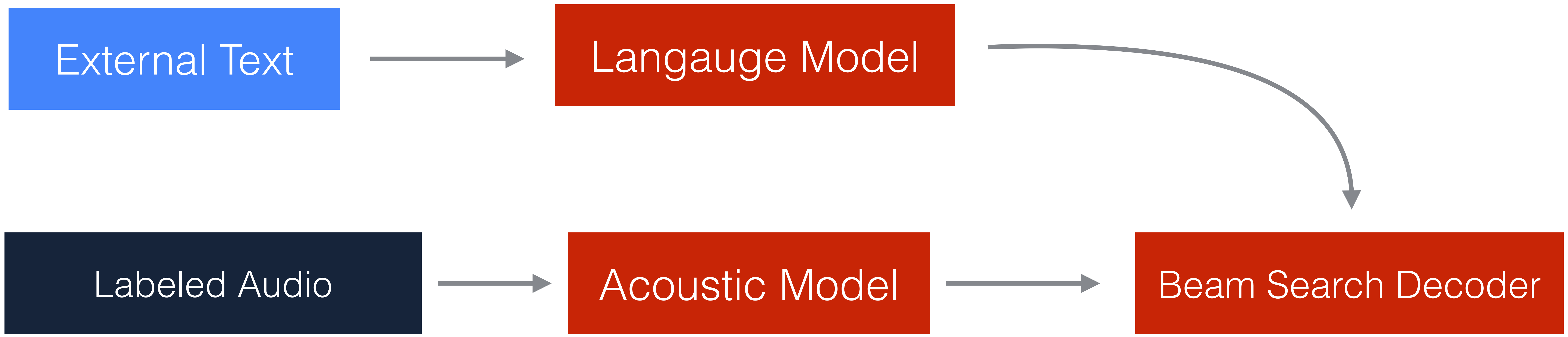
Self-Training

Use the same model to generate labels for unlabeled data; train on the resulting labels.



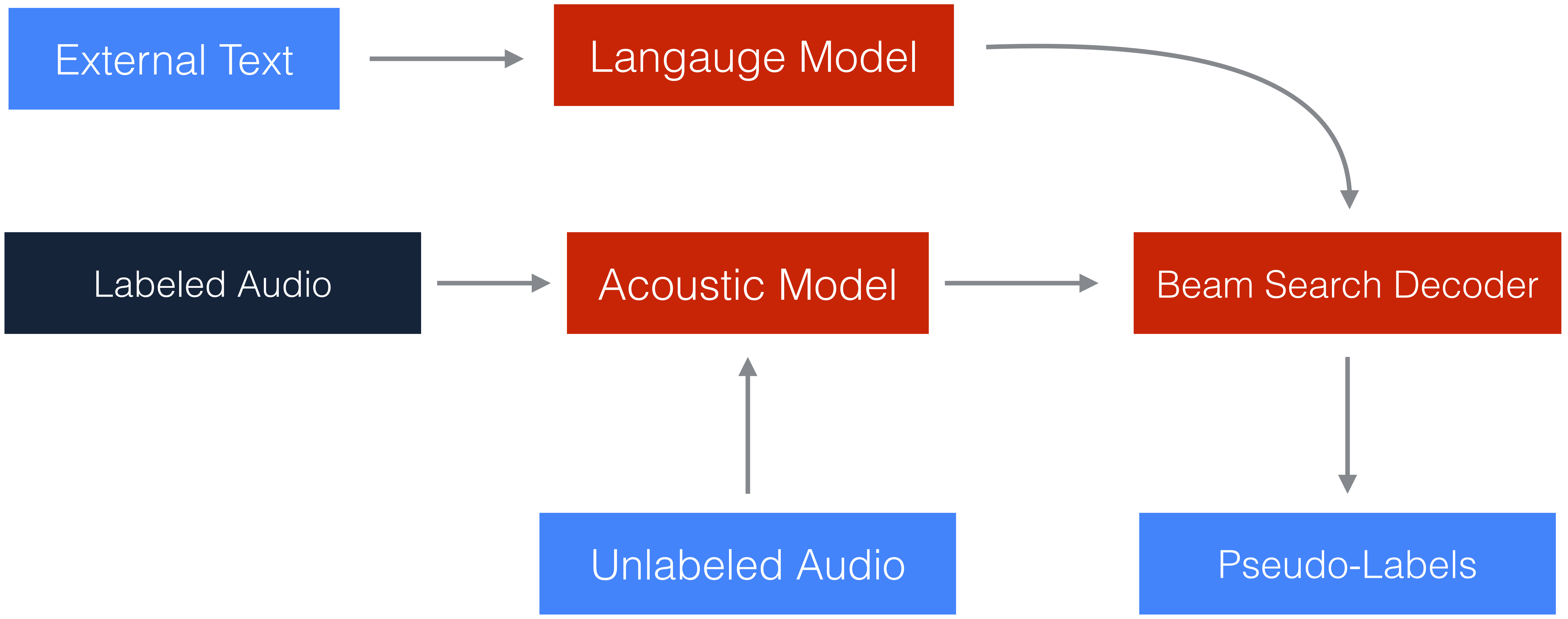
Self-Training in End-to-End ASR

Leverage unpaired audio and unpaired data.



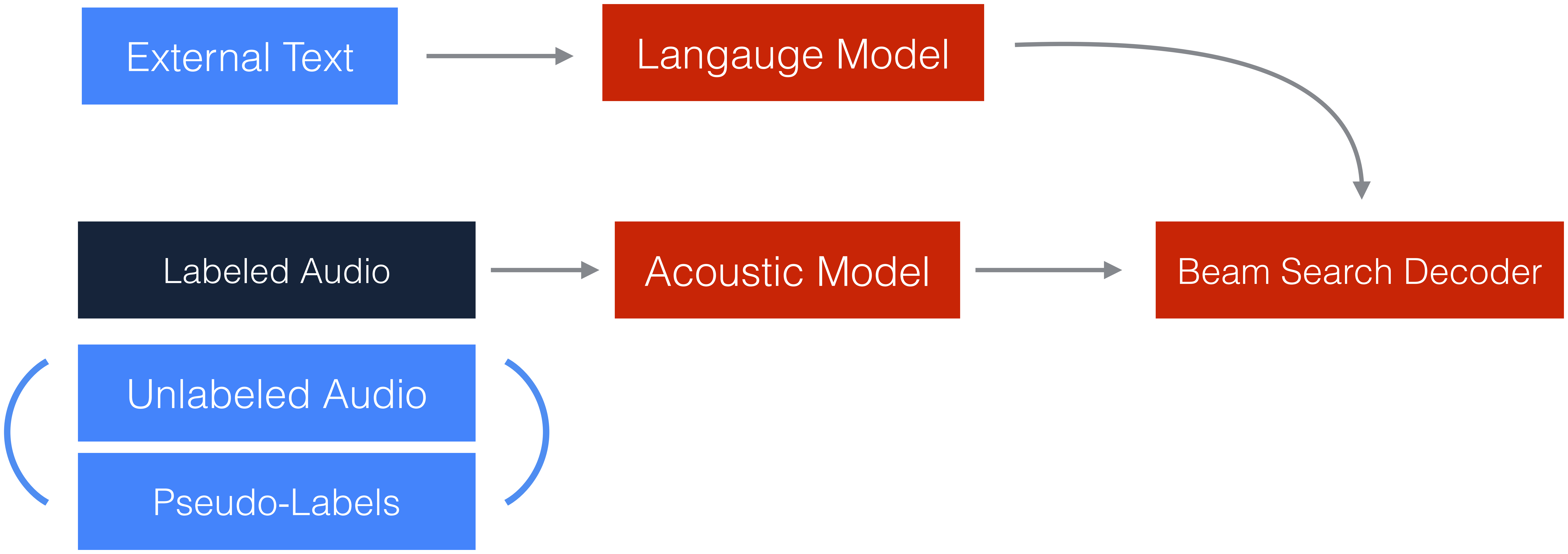
Self-Training in End-to-End ASR

Leverage unpaired audio and unpaired data.



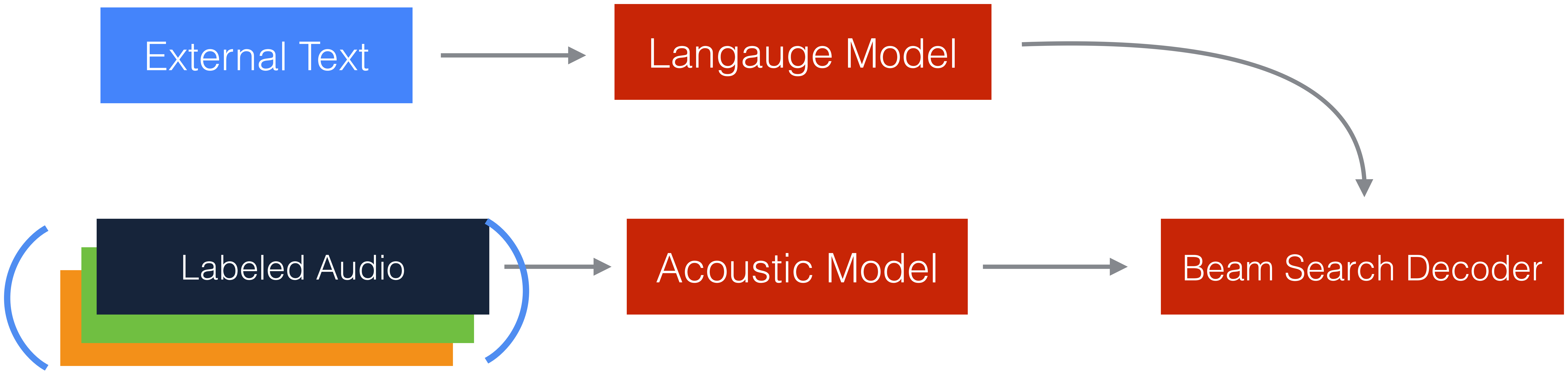
Self-Training in End-to-End ASR

Leverage unpaired audio and unpaired data.



Self-Training in End-to-End ASR

Leverage unpaired audio and unpaired data.



- Unlabeled audio is equally-weighted.
- Model trained on pseudo-labeled audio is trained from scratch, rather than fine-tuning.

Self-Training in End-to-End ASR

Formulation

Given a paired dataset $\mathscr{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and pseudo-labeled data $\bar{\mathscr{D}} = \{(X_i, \bar{Y}_i) \mid X_i \in \mathscr{X}\}$, where $\mathscr{X}$ is a set of paired audio data.

We maximize the following equally-weighted objective:

$$\sum_{(X,Y)\in\mathscr{D}} \log\left(P(Y \mid X)\right) + \sum_{(X,\bar{Y})\in\bar{\mathscr{D}}} \log\left(P(\bar{Y} \mid X)\right)$$

Related Work

- **Self-training in hybrid speech recognition**
  - Focus on data filtering to improve PL quality (Charlet, 2001; Wessel, 2004; Vesely, 2013, Vesely 2017)
  - Confidence-based filtering (Charlet, 2001; Wessel, 2004) or agreement-based selection (Vesely, 2013)
- **Student-teacher models and cycle-consistency loss with TTS/speech generation**
  - Cycle consistency (Hori, 2019)
  - Teacher output labels/posteriors used to train a student model (Hari, 2019)
- **Backtranslation or continuous embedding-style techniques**
  - Data augmentation with backtranslation (Hayashi, 2018)
  - TTS-based techniques (Baskar, 2019)
  - Audio and text in the same embedding space (Karita, 2018)

Delphine Charlet, *Confidence-measure-driven unsupervised incremental adaptation for hmm-based speech recognition*, ICASSP 2001
Wessel et al. *Unsupervised training of acoustic models for large vocabulary continuous speech recognition*, IEEE Trans Speech Audio Process, 2004
Vesely et al. *Semi-supervised training of deep neural networks*, ASRU 2013
Vesely et al. *Semisupervised DNN training with word selection for ASR*, Interspeech 2017
Hayashi et al. *Back-translation-style data augmentation for end-to-end ASR*, SLT 2018
Hari et al. *Lessons from building acoustic models with a million hours of speech*, ICASSP 2019
Karita et al. *Semi-supervised end-to-end speech recognition*, Interspeech 2018

Highlights

Our contributions are in three areas:

- **A strong baseline model.**
  - A well-performing end-to-end, sequence-to-sequence model trained on 100 hours of clean speech from LibriSpeech.
- **Filtering techniques.**
  - Heuristic filtering in addition to confidence-based filtering for mitigating common pitfalls with sequence-to-sequence models.
- **A novel ensemble approach.**
  - Increasing pseudo-label diversity improves results.

Outline

1

Self-Training in End-to-End ASR

Motivating/defining the pipeline and related work.

2

Baseline Acoustic and Language Model, Filtering, and Ensembles

Key components for sequence-to-sequence models.

3

Results

WER on LibriSpeech datasets with pseudo-labeling, improving on prior results.

4

Future Work

Extending self-training-style techniques in speech.

Baseline AM and LM and Ensemble & Filtering Techniques

2

Key components for self-training with sequence-to-sequence models.

Audio Data

LibriSpeech audio books and language model corpuses.

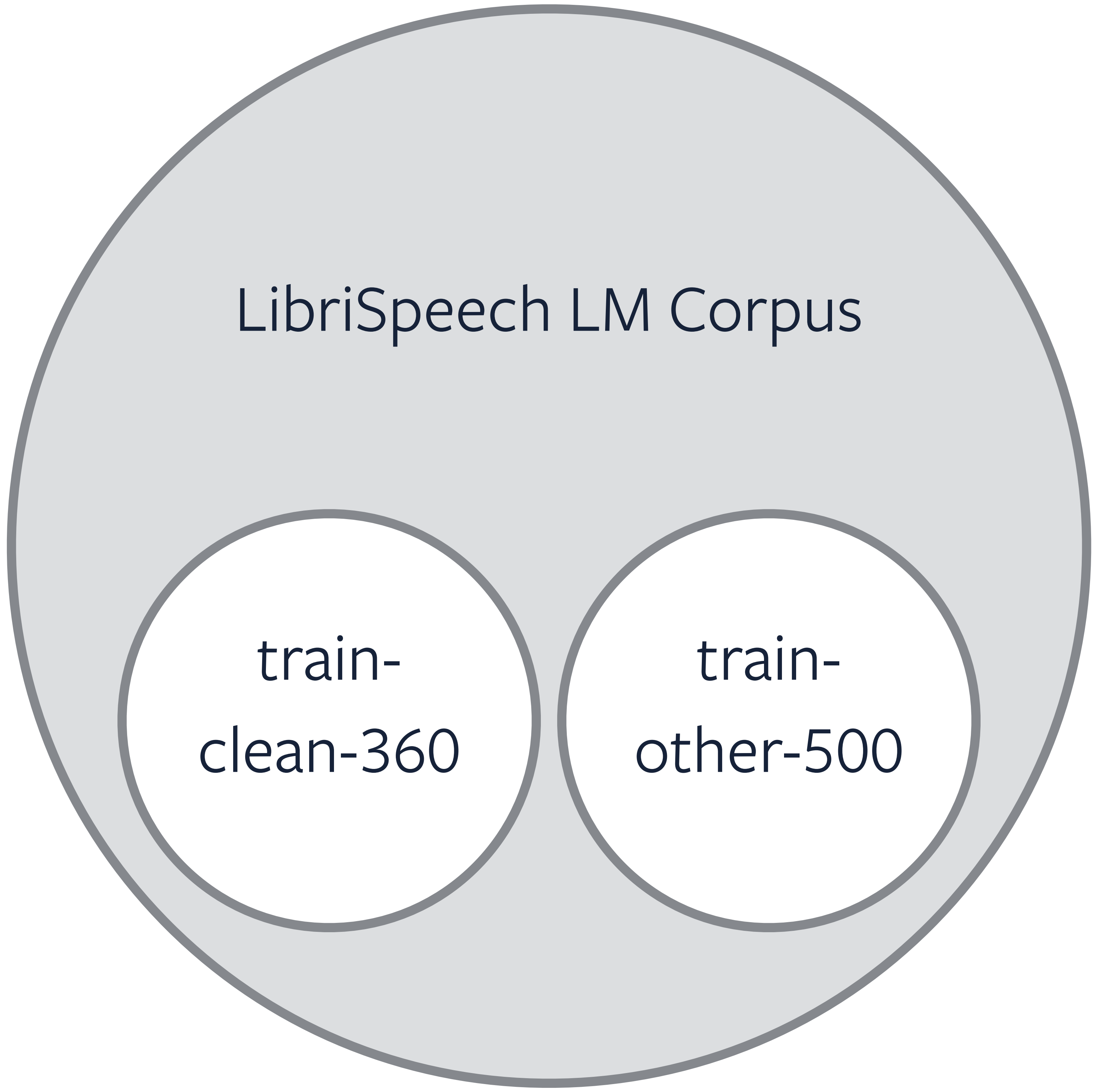
- **LibriSpeech** (Panayotov et al. 2015)
  - Labeled (“paired”) audio
    - LibriSpeech **train-clean-100** (100 hours)
  - Unlabeled (“unpaired”) audio
    - LibriSpeech **train-clean-360** — clean speech (360 hours)
    - LibriSpeech **train-other-500** — noisy speech (500 hours)
- **LibriSpeech LM Corpus**
  - Text from 14k books.

Panayotov et al. Librispeech: an ASR corpus based on public domain audio books, ICASSP 2015

Language Model Training Corpus

Carefully remove text for which there is pseudo-unpaired audio — i.e. some train sets.

Remove text which corresponds to audio in our unpaired audio sets.



Training a Strong Baseline Model

- **Time depthwise-separable convolutions with sequence-to-sequence loss (Hannun et al. 2019)**
  - Sequence-to-sequence decoder with attention
  - Stable beam search decoding with a language model
- **GCNN language model (Dauphin et al. 2017)**
  - Follows the recipe from Zeghidour et al. 2018.

	Dev WER		Test WER	
	clean	other	clean	other
Liu et al. [24]	21.6	-	21.7	-
Hayashi et al. [22]	24.9	-	25.2	-
Lüscher et al. [1]	14.7	38.5	14.7	40.8
Our model	14.0	37.0	14.9	40.0

WER for end-to-end models trained on LibriSpeech train-clean-100, with no external LM.

*Current state of the art on train-clean-100 is from Irie et al. with 12.7, 33.9, 12.9, and 35.5 on dev-clean, dev-other, test-clean, and test-other, respectively.

Hannun et al. *Sequence-to-Sequence Speech Recognition with Time-Depth Separable Convolutions*, Interspeech 2019
Dauphin et al. *Language modeling with gated convolutional networks*, ICML 2017
Zeghidour et al. *Fully convolutional speech recognition*, 2018
[24] Liu et al. *Adversarial training of end-to-end speech recognition using a criticizing language model*, ICASSP 2019
[22] Hayashi et al. *Back-translation-style data augmentation for end-to-end ASR*, SLT 2018
[1] Lüscher et al. *RWTH ASR systems for LibriSpeech: Hybrid vs attention*, Interspeech 2019
Irie et al. *On the Choice of Modeling Unit for Sequence-to-Sequence Speech Recognition*, Interspeech 2019

Filtering Pseudo-Labels

Sequence-to-sequence models with attention can fail catastrophically.

Ground truth: `I went to the store then I went to my house <EOS>`

- Looping:

`I went to the store then I went to the store then I went to the store ...`

- Early stopping:

`I went to the store <EOS>`

Filtering Pseudo-Labels

Two approaches to filtering:

- **Heuristic**
  - Filter based on looping and early stopping
    - Remove examples with *n*-grams that repeat more than *k* times
    - Remove examples with early stopping
- **Generic**
  - Use a threshold based on scores output by the model

Let $X = [X_1, ..., X_T]$ be frames of speech with predicted transcriptions $Y = [Y_1, ..., Y_T]$:

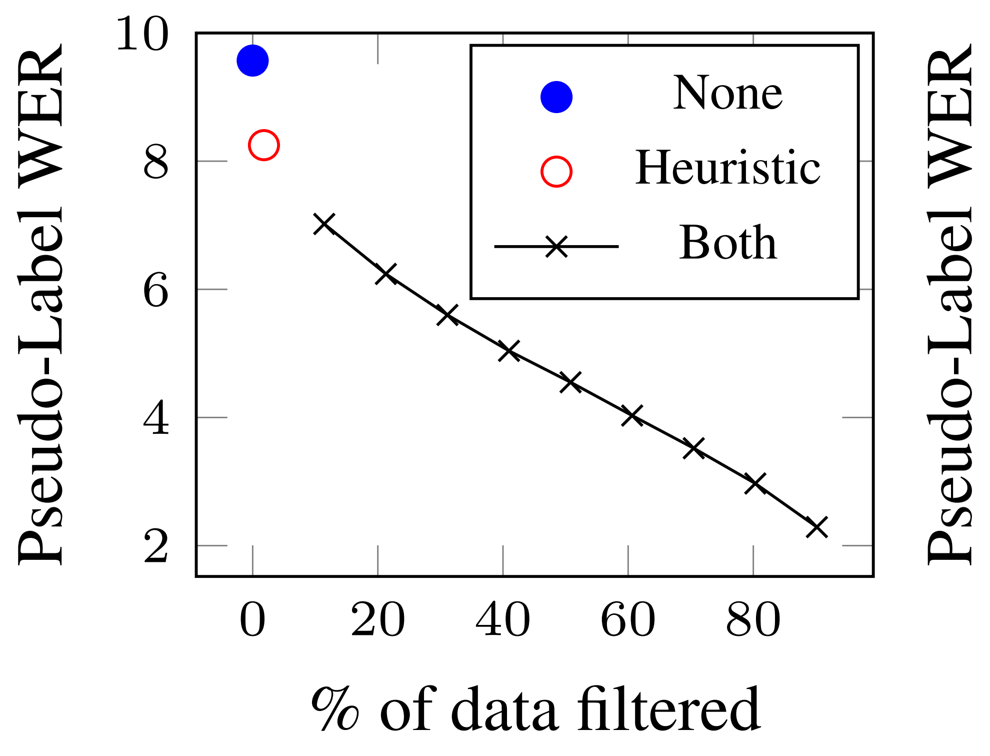
$$\text{ConfidenceScore}(\bar{Y}_i) = \frac{\log\left(P_{AM}\,(Y_i \mid X_i)\right)}{|\,\bar{Y}_i\,|}$$

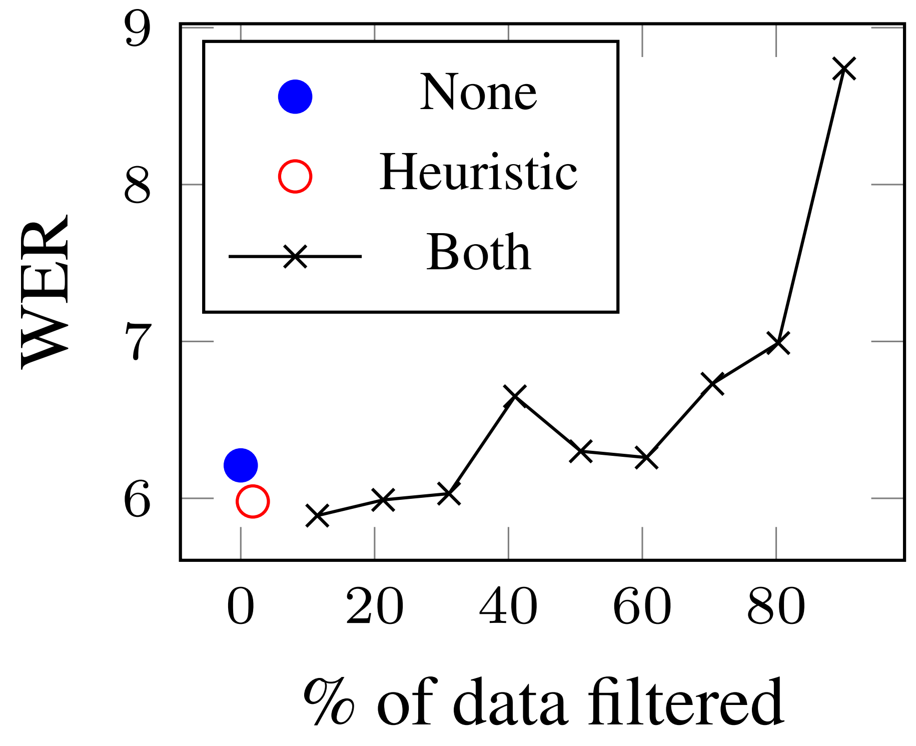
where $|\,\bar{Y}_i\,|$ is the number of tokens in the utterance.

Effect of Filtering

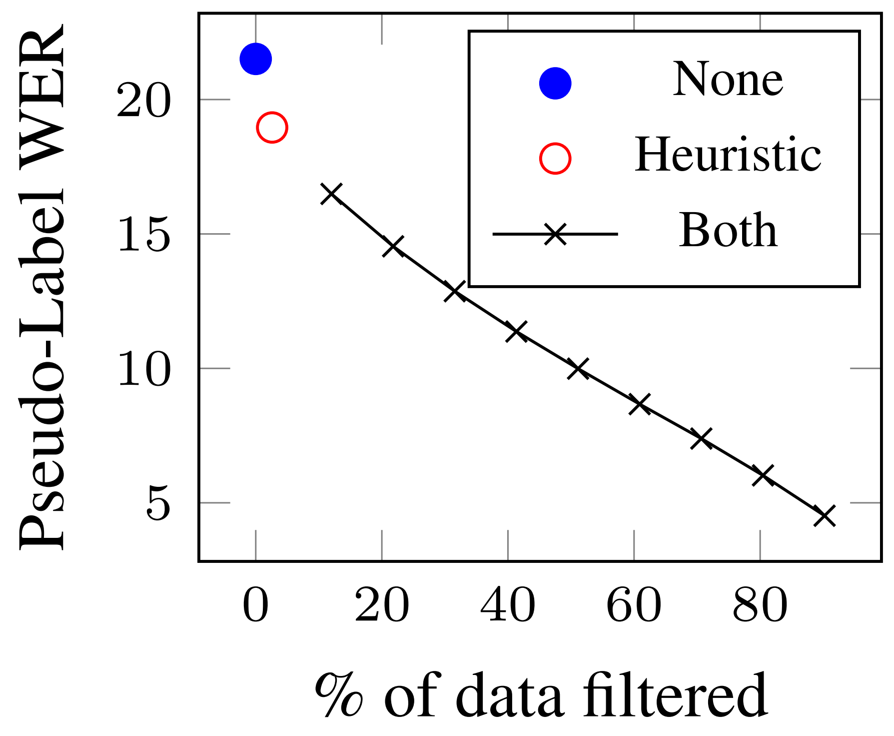
Confidence score-based filtering helps significantly with noisy data and marginally with clean.

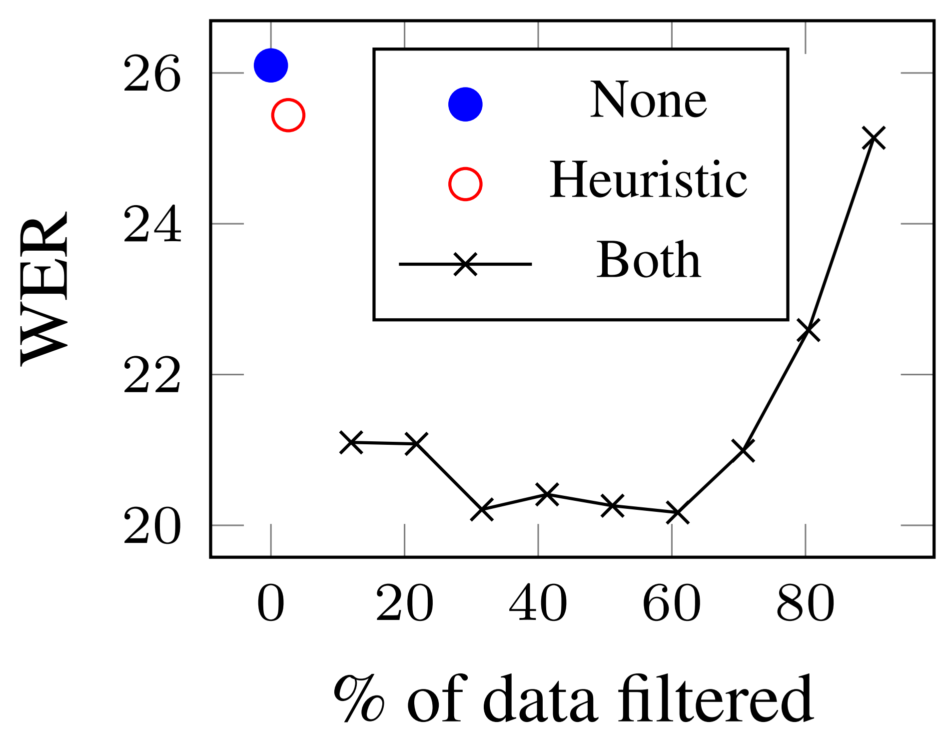
Clean pseudo-labels evaluated on clean audio





Noisy pseudo-labels evaluated on noisy audio



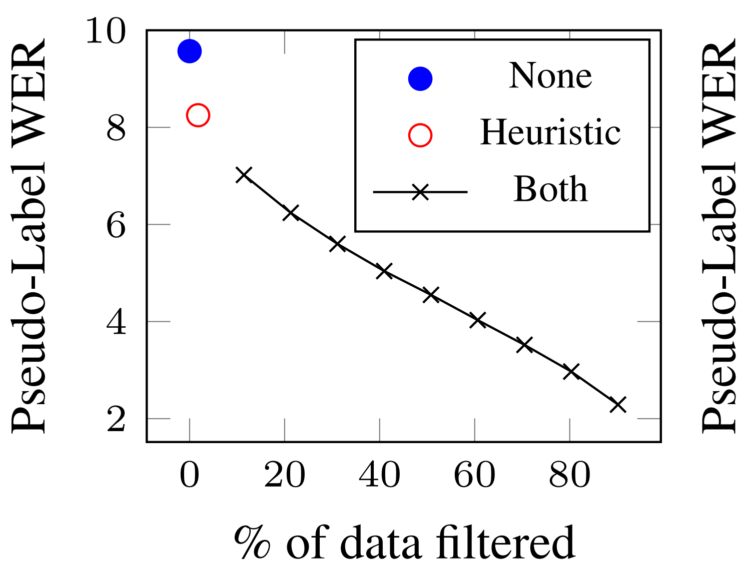


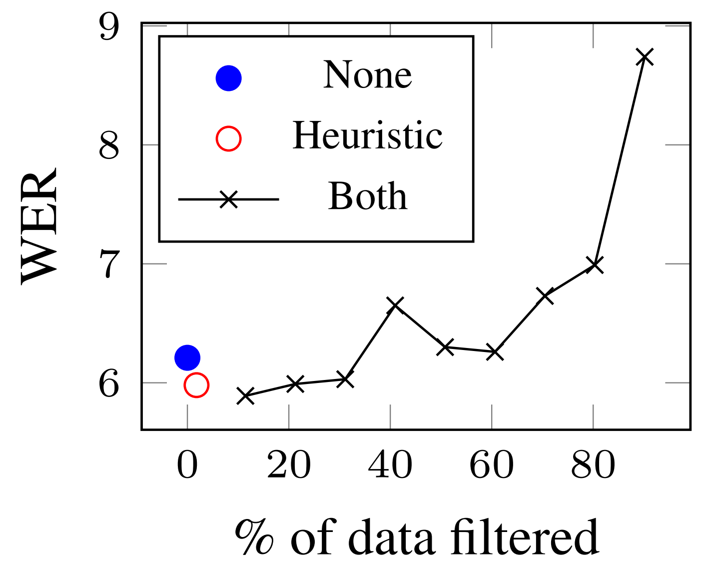
Heuristic filtering: repeated n-gram and early-stopping filters. **"Both"** adds confidence-based filtering on top of heuristic filters.

Effect of Filtering

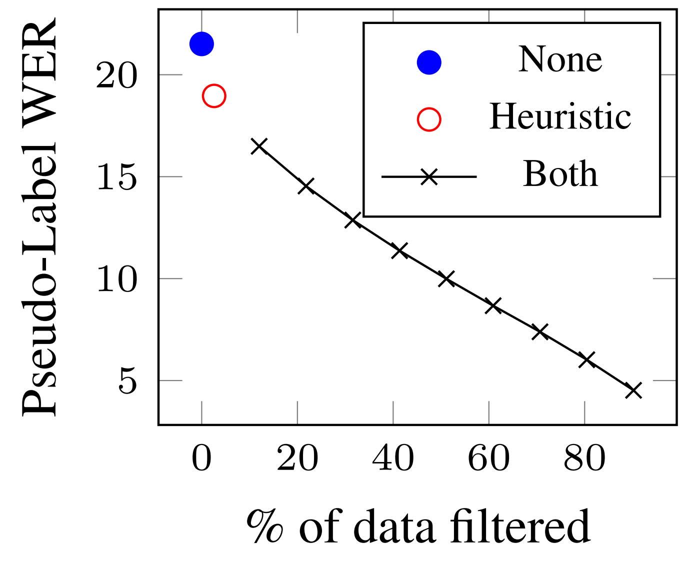
Confidence score-based filtering helps significantly with noisy data and marginally with clean.

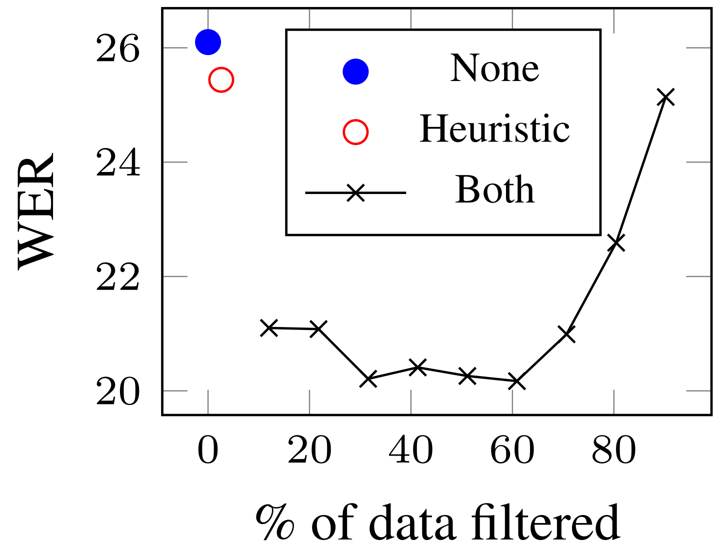
Clean pseudo-labels evaluated on clean audio





Noisy pseudo-labels evaluated on noisy audio





Heuristic filtering: repeated n-gram and early-stopping filters. **"Both"** adds confidence-based filtering on top of heuristic filters.

- **Clean setting** — data starvation quickly occurs with confidence-based filtering.
- **Noisy setting** — confidence-based filtering has a large impact and returns diminish after a larger percentage of audio is filtered. Data starvation still eventually occurs.

Pseudo-Label Ensembles

Increase label diversity by sampling from different pseudo-labels for the same audio.

- **Train K models.**

  - Using different seeds to change initialization can provide diversity.

- **Generate pseudo-labels $\bar{\mathscr{D}}_m$ for each of the M models.**

  - Apply filtering criterion on resulting pseudo-labels as needed.

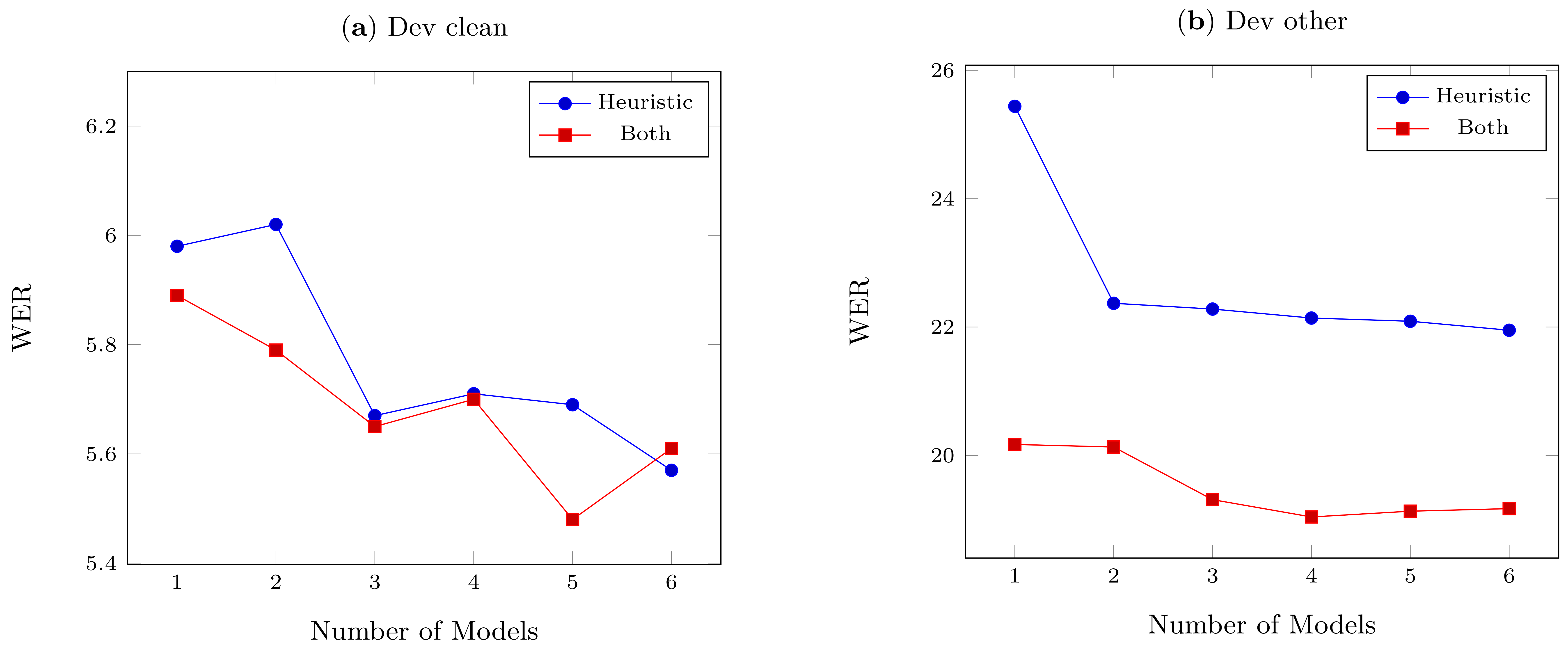
- **Train a new model on the resulting pseudo-labels — sample from the available k labels for each example.**

When training, we effectively maximize the objective:

$$\sum_{(X,Y)\in\mathscr{D}} \log\left(P(Y \mid X)\right) + \frac{1}{M}\sum_{m=1}^{M} \sum_{(X,\bar{Y})\in\bar{\mathscr{D}}_m} \log\left(P(\bar{Y} \mid X)\right)$$

Pseudo-Label Ensembles

Increase label diversity by sampling from different pseudo-labels for the same audio.



- Increasing the number of models in the ensemble improves performance.
- Using an ensemble on top of both heuristic and confidence based filters is best in the noisy setting.

Outline

1

Self-Training in End-to-End ASR

Motivating/defining the pipeline and related work.

2

Baseline Acoustic and Language Model, Filtering, and Ensembles

Key components for sequence-to-sequence models.

3

Results

WER on LibriSpeech datasets with pseudo-labeling, improving on prior results.

4

Future Work

Extending self-training-style techniques in speech.

Results

Word error rate improvements with pseudo-labeled clean and noisy audio.

3

Measuring Improvements

- **Oracle word error rate ("Oracle WER")** is the word error rate for a model trained on ground truth labels for a given pseudo-label set and represents a palatable upper bound on model performance.

$$\textbf{Word error rate recovery rate — "WERR"} = \frac{\text{baseline WER} - \text{semi-supervised WER}}{\text{baseline WER} - \text{Oracle WER}}$$

- **Pseudo-label word error rate** is the word error rate of generated pseudo-labels with respect to ground truth labels, after filtering.
  - In our set up, since we know the ground truth labels for the audio on which we generate pseudo-labels, we can compute this.

Results — 360 hours of *clean* unlabeled audio

	Clean Test Set WER %	Noisy Test Set WER %
Baseline (100 hours, labeled)	8.06	30.44
Pseudo-label (100 hrs labeled + 360 hrs unlabeled)	6.46	22.90
Pseudo-label Ensemble* (100 hrs labeled + 360 hrs unlabeled)	5.79	21.63
Oracle (100 hrs labeled + 360 hrs labeled)	4.23	17.36

*Using an ensemble with 5 models

Results — 500 hours of *noisy* unlabeled audio

	Clean Test Set WER	Noisy Test Set WER
Baseline (100 hours, labeled)	8.06	30.44
Pseudo-label (100 hrs labeled + 500 hrs unlabeled)	6.56	22.09
Pseudo-label Ensemble* (100 hrs labeled + 500 hrs unlabeled)	6.20	20.11
Oracle (100 hrs labeled + 500 hrs labeled)	3.83	11.28

*Using an ensemble with 4 models

Results — Recovery Rates

WER	Clean Test Set WER	Noisy Test Set WER
Baseline (100 hours, labeled)	8.06	30.44
Pseudo-label Ensemble (100 hrs labeled + 360 hrs unlabeled)	5.79	21.63
Oracle (100 hrs labeled + 360 hrs unlabeled)	4.23	17.36
Pseudo-label Ensemble (100 hrs labeled + 500 hrs unlabeled)	6.20	20.11
Oracle (100 hrs labeled + 500 hrs unlabeled)	3.83	11.28
WERR	**Clean Test Set WERR**	**Noisy Test Set WERR**
Pseudo-label Ensemble (100 hrs labeled + 360 hrs unlabeled)	59.3%	67.4%
Pseudo-label Ensemble (100 hrs labeled + 500 hrs unlabeled)	44.0%	53.9%

Results — Recovery Rates

- **100 hours of paired audio, 360 hours of clean unpaired audio.**
  - WRR = word error rate recovery
    - Computed using an oracle that does not include LM decoding.

Method	Text (# words)	No LM Test clean WER (WRR)	With LM Test clean WER (WRR)
Cycle TTE [9]	4.8M	21.5 (27.6%)	19.5 (30.6%*)
ASR+TTS [10]	3.6M	17.5 (38.0%)	16.6 (-)
this work	842.5M	**9.62** (**76.2%**)	**5.79** (**59.3%**)

[9] Hori et al. *Cycle-consistency training for end-to-end speech recognition*, ICASSP 2019
[10] Karthick et al. *Semi-supervised sequence-to-sequence ASR using unpaired speech and text*, Interspeech 2019

Outline

1

Self-Training in End-to-End ASR

Motivating/defining the pipeline and related work.

2

Baseline Acoustic and Language Model, Filtering, and Ensembles

Key components for sequence-to-sequence models.

3

Results

WER on LibriSpeech datasets with pseudo-labeling, improving on prior results.

4

Future Work

Extending self-training-style techniques in speech.

Future Work

Extending self-training and semi-supervision with unlabeled audio and an external LM.

Iterative pseudo-labeling.

Can we generate higher-quality pseudo-labels using a model already bootstrapped on pseudo-labels?

More continuous relaxations/integration of language model information.

The language model helps — can information from it be integrated more continuously during training?

Increasing the amount of unlabeled audio.

Performance deteriorated when too many pseudo-labels was filtered — can using unlabeled audio mitigate this?

Increasing the quality of the baseline model.

Does a significantly better baseline model generate better pseudo-labels?

Thanks!

Reproduce: github.com/facebookresearch/wav2letter → recipes/models/self_training

facebook
Artificial Intelligence Research