Large Margin Training Improves Language Models For ASR

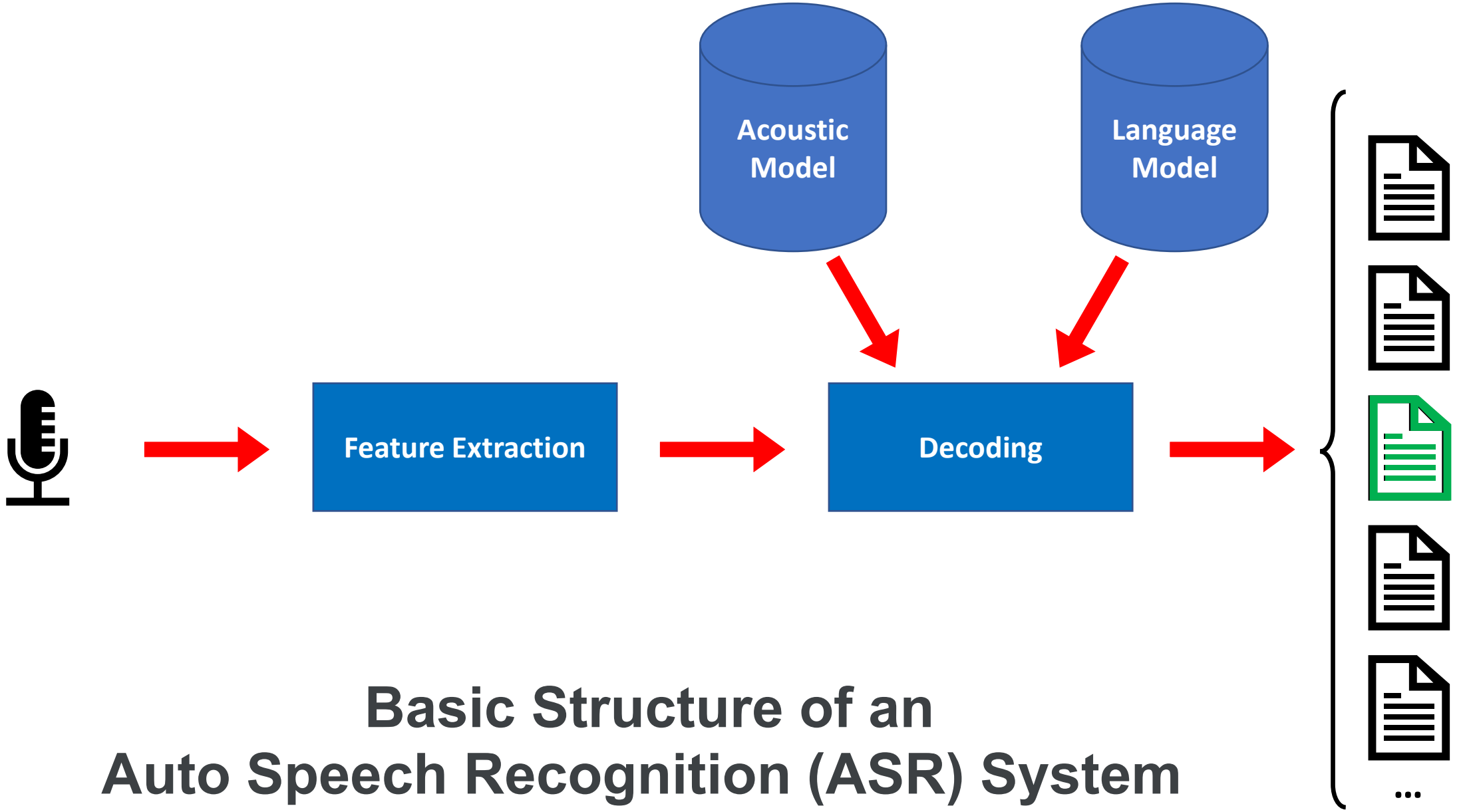
Jilin Wang

Boston University
Boston, MA, USA

Jiaji Huang
Kenneth Ward Church

Baidu Research,
Sunnyvale, CA, USA

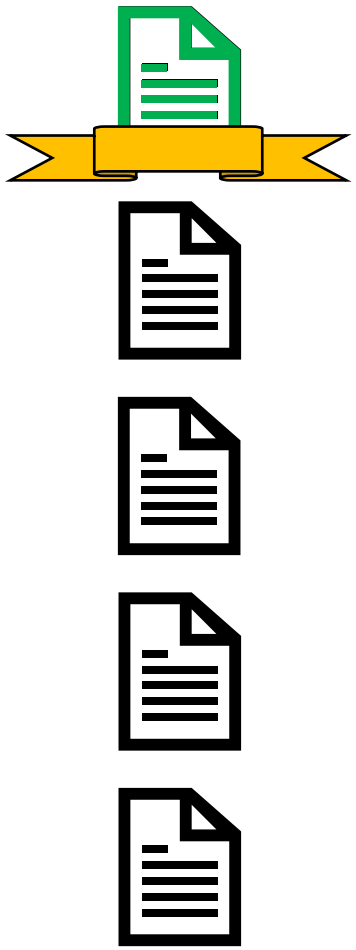
2021 IEEE International Conference on Acoustics, Speech and Signal Processing



Basic Structure of an Auto Speech Recognition (ASR) System

N-Best Rescoring

Language Model



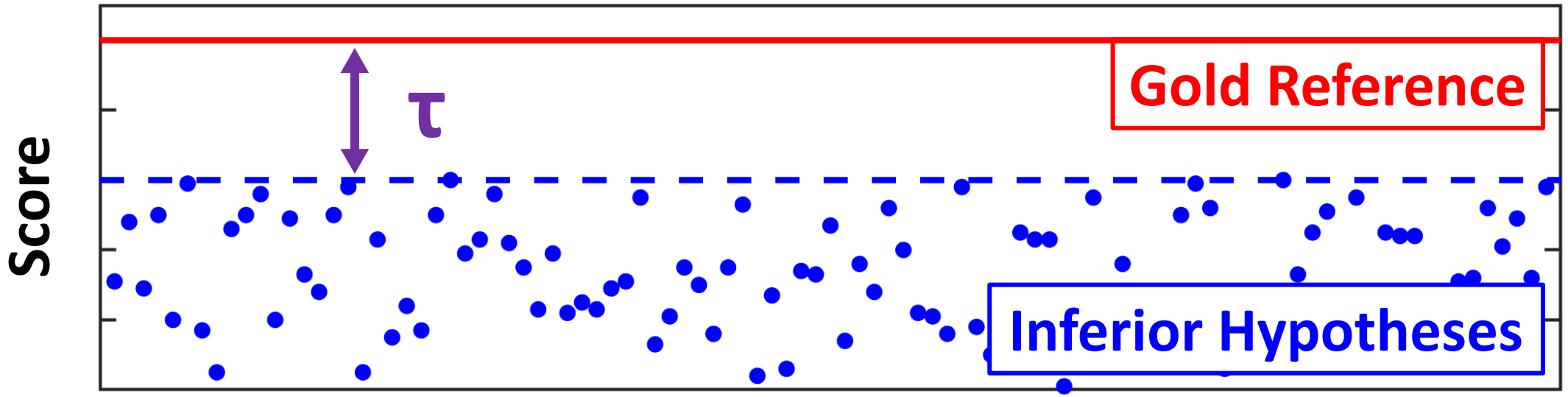
Perplexity

$$PPL = \exp\left\{-\frac{1}{|X|}Score(X)\right\}$$

- Scores of beams candidate from decoder are given by their likelihood
- Fine-tune an LM by minimizing the **Perplexity (PPL)** on the “gold” references could fit it to the ground-truth transcriptions

- **No information from ASR beam candidates utilized**
- **Sometimes propose ``bad'' hypotheses -> give a higher score on inferior hypotheses than the “gold” reference**

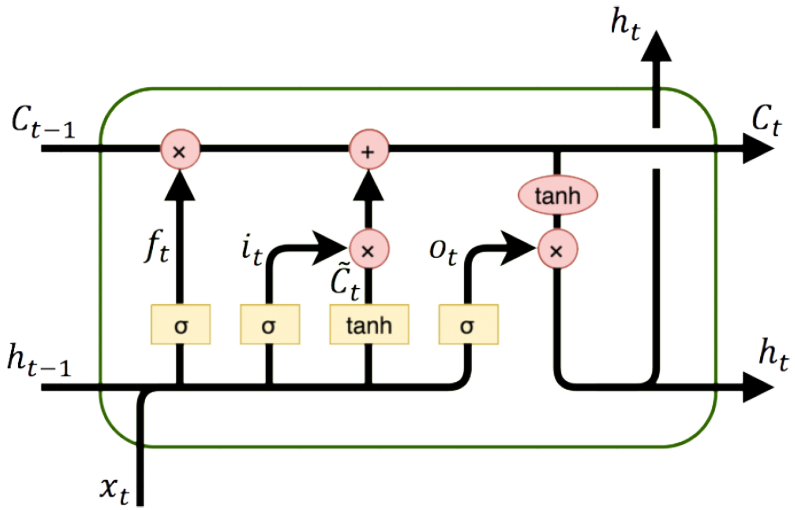
Large Margin Language Model (LMLM)



$$LMLM = \sum_{i=1}^{K} \sum_{j=1}^{N} \max\{0, \tau - (Score(X_i) - Score(X_{i,j})\}$$

Language Model

LSTM



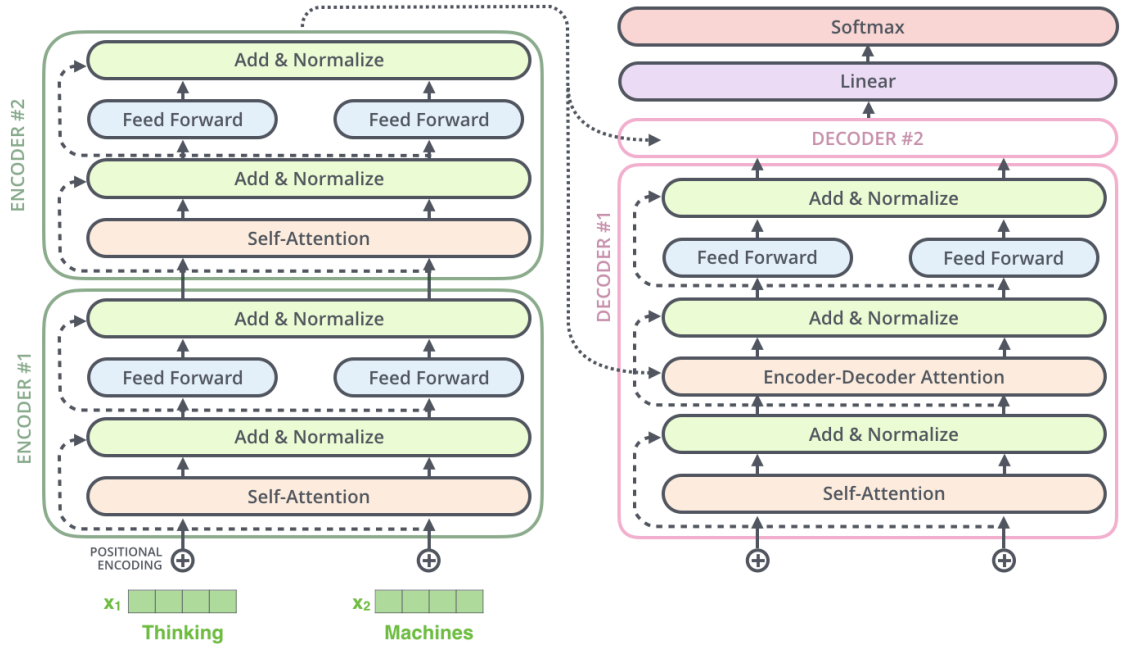
$$h_t = f(h_{t-1}, x_t)$$

Causal

Transformer

Encoder

Decoder



$$h_t = g(x_t, self\ attention(x_t, X_{context}))$$

Non-Causal

Causal

Score (Likelihood) of a Sentence

Causal

$$Score^{c}(X) = \sum_{t=1}^{|X|} \log P(x_t | X_{<t}; \theta)$$

$$X_{<t} = [X_1, \ldots, X_{t-1}]$$

Non-Causal

$$Score^{m}(X) = \sum_{t=1}^{|X|} \log P(x_t | X_{\setminus t}; \theta)$$

$$X_{\setminus t} = [X_1, \ldots, X_{t-1}, X_{t+1}, \ldots, X_{|X|}]$$

Experiment

- Experiment with **LibriSpeech** benchmark.
- Baseline Decoder:

  Acoustic model: chain system based on Factorized Time Delay Neural Network (TDNN-F)

  Language model: Trigram LM
- Language models for rescoring:

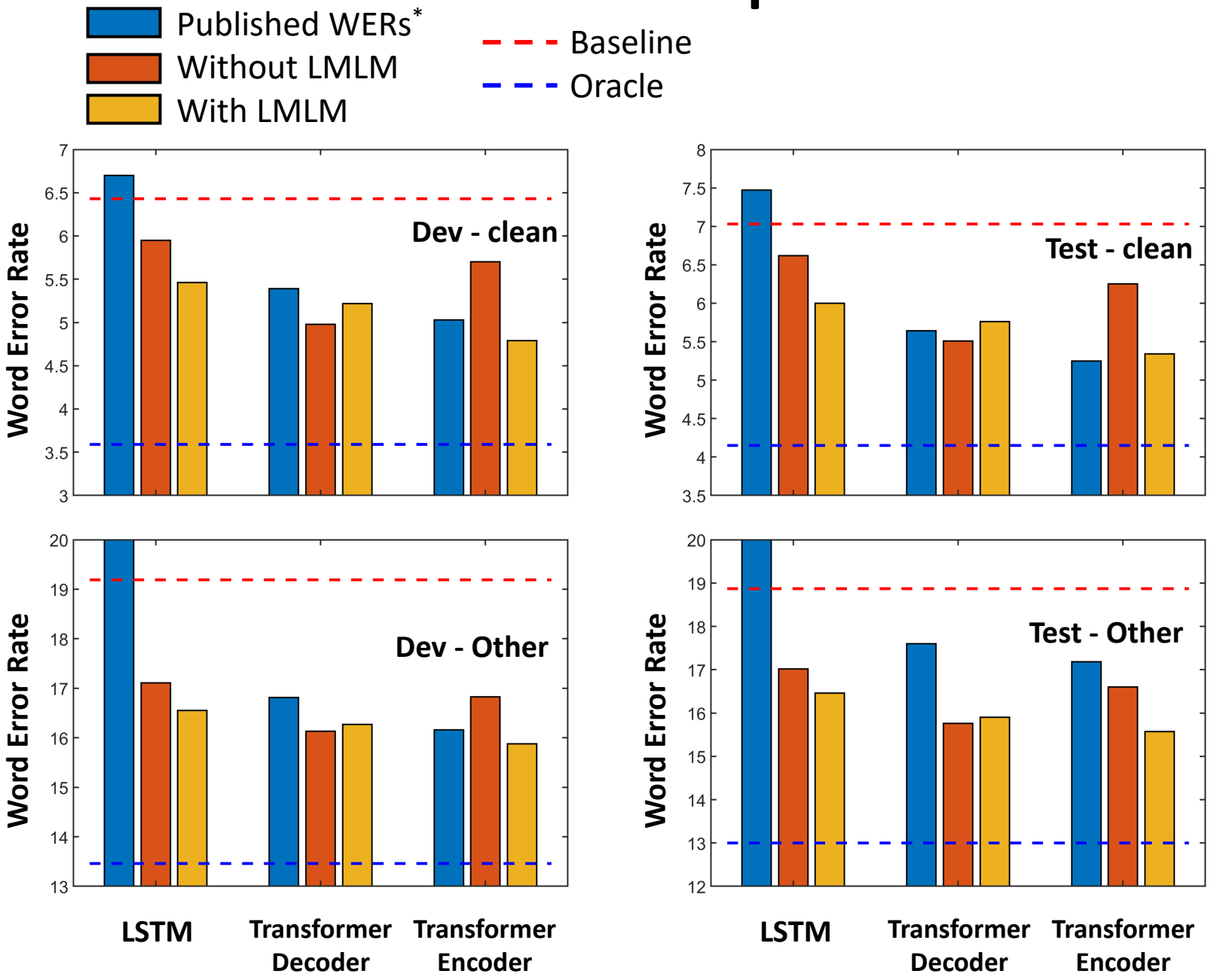
  LSTM: Causal, 4 layers, 512 hidden dimension

  Transformer Decoder: Causal, 12 layers, 768 hidden dimension, 12 self-attention heads

  Transformer Encoder: Non-causal, 12 layers, 768 hidden dimension, 12 self-attention heads

All neural LMs are pretrained on a joint of enWiki and bookCorpus.

Empirical Results



- Lowest WER is achieved with Transformer Encoder+LMLM training
- LMLM training significantly decreases WER for LSTM and Transformer Encoder
- Transformer Decoder without LMLM training is already very competitive.
- May be caused by the fundamental difference between causal LM score and non-causal LM score.

* Julian Salazar, et.al "Masked language model scoring," 58th ACL, 2019, pp. 2699–2712.
* Joonbo Shin, et.al "Effective sentence scoring method using bert for speech recognition," ACML, 2019, pp. 1081–1093
* Lu Huang, et.al. "An improved residual lstm architecture for acoustic modeling," 2nd ICCCS. IEEE, 2017, pp. 101–105.

Thank you!