LatticeBART: Lattice-to-Lattice pre-training for Speech Recognition

SJTU Cross Media Language Intelligence Lab
上海交通大学跨媒体语言智能实验室

Lingfeng Dai, Lu Chen†, Zhikai Zhou, Kai Yu†
X-LANCE Lab, Department of Computer Science and Engineering
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
Shanghai Jiao Tong University, Shanghai, China

Poster number: 8432

Motivation

- WFST-based ASR systems are still widely-used.
- The direct output of WFST-based ASR systems is hard to read.
  - Missing punctuation
  - Without capitalization
  - Containing recognition errors
- The post-processing system is needed.
  - Rule-based
  - Sequence-to-sequence (seq2seq) model

Highlights

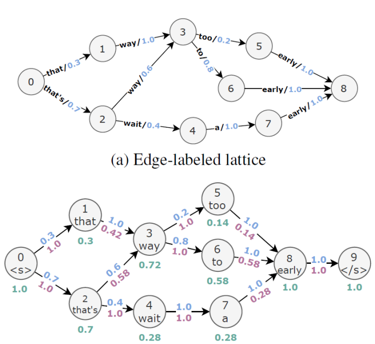
LatticeBART (proposed) can:

- use the knowledgeable pre-trained language models like BART.
- be pre-trained in a lattice-to-lattice (L2L) unsupervised method.

Lattice Process Methods

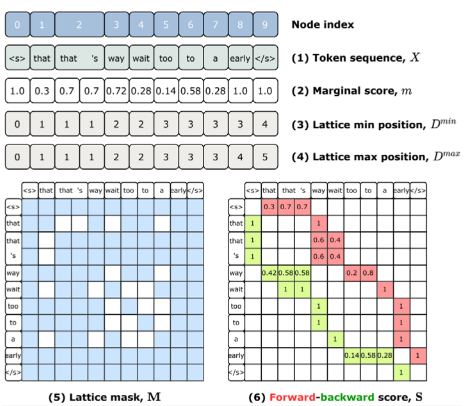
- Using line-graph method to convert edge-labeled lattice to node-labeled lattice.
- Convert the node-labeled to the data format that can be input into the LatticeBART.
  - Using topological sorting to obtain a sequence of nodes
  - Adding forward-backward, marginal scores
  - Adding lattice mask and lattice positional embedding

$$\mathbf{H_i} = Softmax\left(\frac{\mathbf{Q_i K_i^T}}{\sqrt{d_k}} + \mathbf{S} + \mathbf{M}\right)(\mathbf{V_i} + \mathbf{mZ})$$



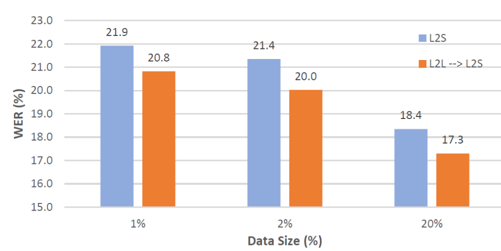
(a) Edge-labeled lattice

(b) Node-labeled lattice.



Training Method

- Lattice-to-Sequence training
- Lattice-to-Lattice pre-training
  - Add noise to the inputs of encoder:
    - Homophone substitution
    - Token masking
    - Path masking
    - Depth offset
    - Weights perturbation
  - Using causal lattice mask to the decoder



Performance comparison of models with small-scale data.

Experiments Setup

- ASR system
  - Phone-based CTC ASR system
    - 5-layer BiLSTM acoustic model with hidden size of 320 is trained on SWBD 300 hours speech
    - 3-gram language model is trained on SWBD-Fisher 2000 hours transcripts
- Test on: eval2000, rt03
- LatticeBART
  - Use BART-base parameters (6-layer encoder and 6-layer decoder)
    - Learning rate: 1e-5
  - Control group with random initialized parameters
    - Learning rate: 1e-4
  - AdamW optimizer with cosine learning rate decay

Table 1. WER (%) results on eval2000 and rt03.

	Model	eval2000		rt03		Avg.
		Callhome	SWBD	Fisher	SWBD	
1	lattice best path	22.5	12.4	17.1	26.0	19.5
2	10-best rescore	23.0	12.6	17.3	26.1	19.8
3	20-best rescore	22.9	12.6	17.3	26.1	19.7
4	$L2S^*_{20\%}$	22.9	13.3	17.6	26.4	20.1
5	$L2S^*_{100\%}$	21.1	12.2	15.9	24.4	18.4
6	$L2S_{20\%}$	21.4	11.5	15.9	24.6	18.4
7	$L2S_{100\%}$	19.1	10.0	13.7	22.4	16.3
8	$L2L_{20\%} \rightarrow L2S_{20\%}$	20.1	10.5	14.8	23.8	17.3
9	$L2L_{80\%} \rightarrow L2S_{20\%}$	20.1	10.6	14.7	23.6	17.3

Table 2. Effect of different beam widths on WER (%).

Beam width	eval2000		rt03		Avg.
	Callhome	SWBD	Fisher	SWBD	
1	21.0	10.9	15.5	24.1	17.9
2	20.3	10.5	14.7	23.2	17.2
3	19.3	10.2	14.2	22.7	16.6
4	19.1	10.0	13.7	22.4	16.3
5	18.9	10.0	13.7	22.2	16.2
6	18.7	9.9	13.6	22.1	16.1