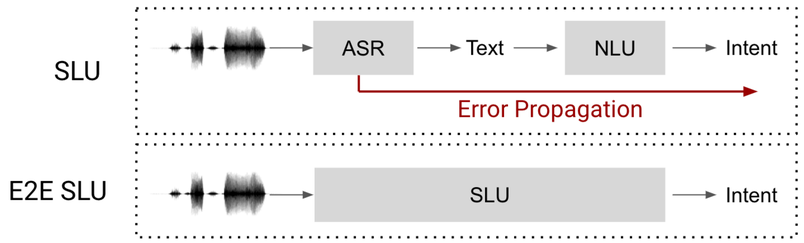
ST-BERT: Cross-Modal Language Model Pre-Training for End-to-end Spoken Language Understanding

Minjeong Kim[3*], Gyuwan Kim[1,2], Sang-Woo Lee[1,2], Jung-Woo Ha[1,2]
[1]NAVER Clova, [2]NAVER AI Lab, [3]Upstage AI Research

#3009

End-to-end Spoken Language Understanding

Goal: Understanding the speaker's intent

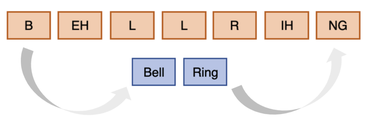


E2E SLU can understand the speaker's intent
without translating the speech to a text transcript

Motivation

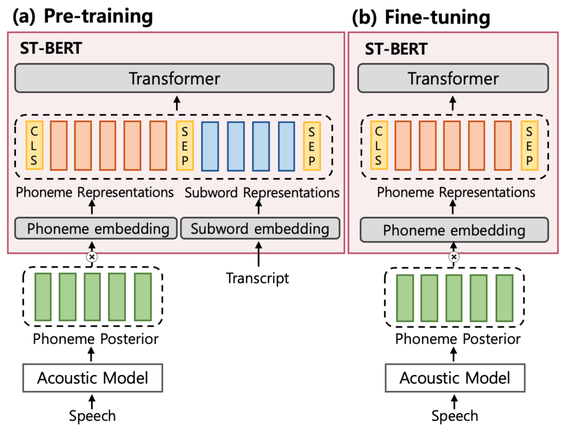
1. Pre-trained Language Models (PLMs)
- Learn rich textual information
- Show the outstanding performance for most NLP task!

2. Speech & Transcript
- Share the same meaning
- Can be translated to each other!



Leverage PLM's textual knowledge
by cross-modal language model pre-training

Model Architecture



1. Acoustic model (AM)
- Estimates the phoneme posterior from raw waveform
- Trained with ASR task before ST-BERT pre-training
- During the ST-BERT training, AM remains frozen

2. Phoneme and subword embedding
- Subword embedding is obtained from subword embedding matrix via one-hot encoding method
- Phoneme embedding is a weighted sum of phoneme embedding vectors based on the phoneme posterior
- When pre-training, we concatenate both embeddings for the input of transformer layers

3. Transformer layers
- We followed BERT structure

Experiments

Model	Full	10%	1%
Lugosch et al. [2]	98.80%	97.96%	82.78%
Wang et al. [3] (BERT)	98.95%	-	-
Wang et al. [3] (ERNIE)	99.02%	-	-
Cho et al. [5]	98.98%	98.12%	83.12%
Price [31]	99.3%	-	-
ST-BERT (Ours)	**99.50%**	**99.13%**	95.64%
- CM-CLM	99.42%	99.09%	**96.51%**
- text data (pre-training)	99.39%	99.04%	89.81%
+ DAPT	**99.59%**	**99.25%**	95.83%

Table 1. Fluent Speech Command (FSC) Dataset

Model	SmartLights Far	SmartLights Close	Snips -
Huang and Chen [28]	47.38%	67.98%	89.55%
ST-BERT (Ours)	**60.98%**	**84.65%**	**96.21%**
- CM-CLM	57.06%	81.97%	95.07%
- text data (pre-training)	54.93%	81.22%	95.64%
+ DAPT	**69.40%**	**86.91%**	96.07%

Table 2. SmartLights and Snips Datasets

1. Implementation

(1) Pre-training
- Pre-training on Librispeech
- Curriculum pre-training
- Initialized from pre-trained BERT-base

(2) Domain-Adaptive Pre-training (DAPT)
- Continues pre-training with domain-specific speech-text pair data, when those are available

(3) Fine-tuning
- Fine-tuning on FSC, SmartLights and Snips datasets
- Use synthesized audio for Snips dataset
- Measure performance twice

2. Results

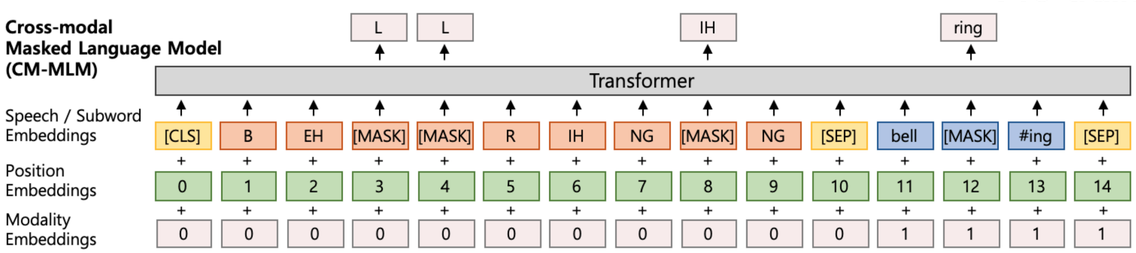
(1) Main results
- For all datasets, our model shows remarkable results

(2) Ablation studies
- [- CM-CLM] : Pre-train with CM-MLM only
- [- text data] : Pre-train with MLM on speech only
- Both pre-training methods are effective
- The usage of the synthesized voice affects to the result of Snips dataset

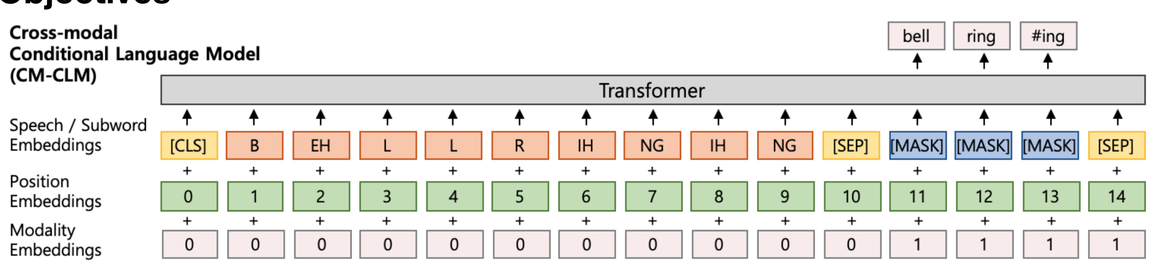
(3) Domain-Adaptive Pre-training (DAPT)
- Leads to further improvement
- Remarkably effective on a more noisy SLU data (far field in SmartLisght dataset was recorded from 2 meters away)

(4) Data shortage scenario
- ST-BERT shows comparatively marginal performance degradation
- Uni-modal pre-training on speech suffers from a relatively large performance drop
- Leveraging textual information during pre-training is critical when limited SLU data are available

Pre-training Objectives



1. Cross-Modal Masked Language Modeling (CM-MLM)
- Randomly samples 15% of input elements and replaces them with [MASK] token
- Model needs to predict a masked phoneme or subword
- To prevent model from easily predicting the masked phoneme from its neighbor, masks out the entire phoneme embedding span that shares the same target phoneme



2. Cross-Modal Conditional Language Modeling (CM-CLM)
- Masks out the entire sequence of target modality
- Model needs to predict the masked representation solely conditioned on the source modality
- Two types of CM-CLM exist: speech-to-text and text-to-speech CM-CLM
- More challenging task than CM-MLM

*This work was mainly performed in NAVER Clova