Improving Textless Spoken Language Understanding with Discrete Units as Intermediate Target

- **Interspeech 2023**
- **NTU, Meta AL**
- **Paper**

Overview

- SLU task에서 intermediate target을 넣어서 성능을 높여보자
  - Intermediate target: discrete token from SSL model
    - HuBERT에서 clustering한 token들, 즉, phoneme tokens

- SLU를 학습할 때, discrete token을 Auxiliary guidance 삼아 학습
  - 대상 task: Spoken NER, Joint IC and Slot Filling, Spoken Semantic Parsing
  - Sequence generation task 형태로 학습
  - 여기서 discrete token은 HuBERT large를 K-means Clustering한 것

Motivation

- **End-to-end SLU**
  - **Cf) ASR – NLU pipeline**

- **이전의 guidance 방법**
  - **Pre-trained ASR model로 모델을 초기화한 후 학습**
  - **ASR/NLU와 SLU를 동시에 학습**
  - **즉, paired transcript 가 있어야 가능**

- **제안하는 guidance 방법**
  - **Paired transcript에 의존하지 않고, SSL feature를 이용해서 guidance를 주자**
  - ⇒ **Unsupervised target for capturing content information in SLU**

Model Structure

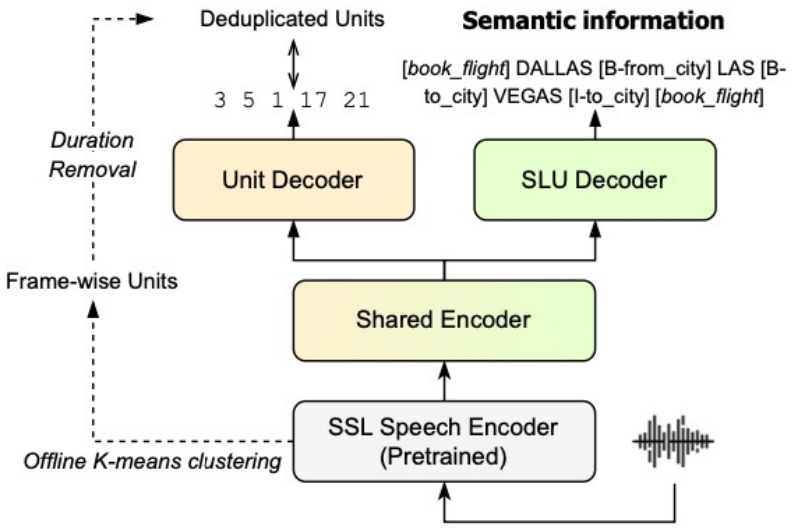

Figure 1: *The proposed textless SLU framework with discrete units as the intermediate target. The transformer encoder is shared between the unit decoder and SLU decoder. Noting that we train the model to predict only semantic information without using full transcripts of the utterances.*

- **SSL Speech Encoder(HuBERT-base)의 last hidden state가 shared encoder로 들어가고,**
- **Unit-decoder 는 HuBERT에서 뽑은 token을 생성**
- **SLU decoder는 각 task에 해당하는 정답을 생성**

$$\mathcal{L} = (1 - \lambda) \times \mathcal{L}_{slu} + \lambda \times \mathcal{L}_{aux}$$

- **Shared encoder, SSL encoder 모두 slu loss와 aux loss에 대해서 업데이트 됨**

- **Deduplication**
  - **Frame-wise unit은 20ms 단위**
  - **17 17 5 8 8 -> 17 5 8**
  - **반복되는 것은 하나로**

Experiment

- 사용 데이터셋
  - **Spoken NER : SLUE-SNER**
  - **Joint IC and Slot Filling : ATIS, SLURP, SNIPS**
  - **Spoken Semantic Parsing : STOP(SLU semantic parsing dataset)**

- **Previous: 기존 논문 성능**
- **Baseline: aux 없이 SLU만 학습**
- **Unit: 제안한 모델**
- **Text: transcript를 target으로 학습 (aux guidance target)**

Result

Table 2: *The performance of baseline, unit-guiding, and text-guiding approaches measured on test sets of ATIS, SLURP, SNIPS, STOP, and the development set of SLUE-SNER. The "Previous" results are sourced from [3] for ATIS, [18] for SLUE-SNER, [5] for SLURP, [10] for SNIPS, and [21] for STOP. The notation "*" with gray color indicates that the training uses ASR transcripts, which are not directly comparable to our setup. "N/A" means that we cannot find performance reported by prior works.*

Dataset	ATIS			SLUE-SNER			SLURP		SNIPS			STOP
Metric	F1↑	ST-F1↑	INT-Acc↑	F1↑	ST-F1↑	SV-CER↓	SLU-F1↑	INT-Acc↑	ST-F1↑	SV-CER↓	INT-Acc↑	EM-Tree↑
Previous	76.6	N/A	93.2	70.3*	N/A	N/A	71.9*	77.0	89.8*	21.8*	N/A	82.9*
Baseline	79.1	84.3	96.5	64.8	74.1	35.2	63.2	78.7	77.6	42.9	96.7	80.0
Unit	**82.4**	**86.0**	**96.8**	**68.6**	**78.1**	**29.4**	**67.9**	**80.9**	**82.7**	**31.9**	**97.0**	**84.4**
Text	84.5	87.7	97.4	69.2	78.2	29.0	69.9	82.5	83.2	30.7	97.0	84.5

- 모든 Task에서 aux loss를 추가하였을 때 더 좋은 성능
- 'Text' 성능과는 작은 차이로 떨어짐
  - Paired transcript(supervised target)와의 비교(unit은 unsupervised target)
  - 참고: wav2vec2.0으로 해도 비슷한 양상을 보였음

Result

Table 3: *The performance of few-shot training with 10% of the training set on SLURP and SNIPS datasets. Values denoted in the δ column indicate the performance drop from 100% training data to 10% training data.*

Dataset	SLURP						SNIPS					
Metric	SLU-F1↑			INT-Acc↑			ST-F1↑			SV-CER↓		
Portion	100%	10%	δ	100%	10%	δ	100%	10%	δ	100%	10%	δ
Baseline	63.2	45.2	18.0	78.7	54.7	24.0	77.6	64.2	13.4	42.9	62.2	19.3
Unit	**67.9**	**53.7**	**14.2**	**80.9**	**69.5**	**11.4**	**82.7**	**78.0**	**4.7**	**31.9**	**40.3**	**8.4**

- **Few-shot training**
  - 훈련 set의 10%만 훈련하고 Test
  - 델타: 모든 훈련셋으로 훈련했을 때와의 차이

  - SNIPS의 경우, 100% 사용했을 때 보다 좋은 성능을 보임

Result

Table 4: *Results of the performance drop on the test set of SLURP dataset under different noisy configurations. G is Gaussian noise with the followed amplitude value. M is MUSAN background noise, with the dB value representing signal-to-noise ratio. Reverb represents the reverberation effect.*

Metric	SLU-F1↑						INT-Acc↑					
Noise	w/o	G-0.01	G-0.02	M-20dB	M-10dB	Reverb	w/o	G-0.01	G-0.02	M-20dB	M-10dB	Reverb
Baseline	63.2	-3.9	-7.7	-1.7	-5.4	-3.9	78.7	-4.8	-9.8	-2.2	-6.4	-4.2
Unit	**67.9**	**-3.2**	**-6.8**	**-1.4**	**-4.8**	**-2.6**	**80.9**	**-3.7**	**-5.3**	**-1.5**	**-5.2**	**-2.7**

Table 5: *Results of the performance drop on the test set of SNIPS dataset under different noisy configurations. The notations for noises (G, M, Reverb) are the same as Table 4.*

Metric	ST-F1↑						SV-CER↓					
Noise	w/o	G-0.005	G-0.01	M-20dB	M-10dB	Reverb	w/o	G-0.005	G-0.01	M-20dB	M-10dB	Reverb
Baseline	77.6	-6.9	-15.4	-3.1	-13.4	-2.3	42.9	+9.1	+19.0	+4.1	+16.2	-3.3
Unit	**82.7**	**-2.9**	**- 9.4**	**-1.9**	**-11.2**	**-2.1**	**31.9**	**+4.9**	**+14.4**	**+2.8**	**+15.5**	**-3.1**

- **Noisy 환경에서의 성능**
  - **Baseline에 비해 성능 저하 정도가 작음**
  - **Discrete unit이 모델이 useful content에 집중할 수 있도록 (noise말고) 돕는 regularization 역할을 한다고 볼 수 있음**

Conclusion

- **Textless SLU 의 performance를 paired transcript 없이도 높일 수 있는 방법**
- **Discrete unit as unsupervised target**

- **논문에서 말하는 future work**
  - **Input augmentation with clean unit guidance to enhance noise robustness**
  - **Intermediate target using unsupervised ASR**