Improving Textless Spoken Language Understanding with Discrete Units as Intermediate Target

Paper

Guan-Wei Wu[1*], Guan-Ting Lin[1*], Shang-Wen Li[2], Hung-yi Lee[1]

[1]National Taiwan University [2]Meta AI (*Equal contribution)

INTERSPEECH 2023

{b08901019, f10942104}@ntu.edu.tw

Meta AI

Introduction

❑ **Background**

- End-to-end SLU aims to predict semantic labels directly from speech features.
- Previous studies rely on using a *pre-trained ASR model as initialization* or *jointly training ASR/NLU and SLU* with paired transcripts guidance.

❑ **Goal**

To alleviate the reliance on paired transcripts, the goal of **Textless SLU** is to extract the semantic information without paired transcripts.

❑ **Main Contribution**

Leverage self-supervised discovered speech units as the intermediate target to improve the performance of end-to-end textless SLU.

Tasks

❑ **Data**

- Speech Name Entity Recognition
  - SLUE-SNER
- Intent Classification & Slot Filling
  - ATIS, SLURP, SNIPS
- Speech Semantic Parsing
  - STOP

*If the task requires to predict entity names, such as slot filling, the entity names are still utilized for training, while all other parts of transcripts are discarded.

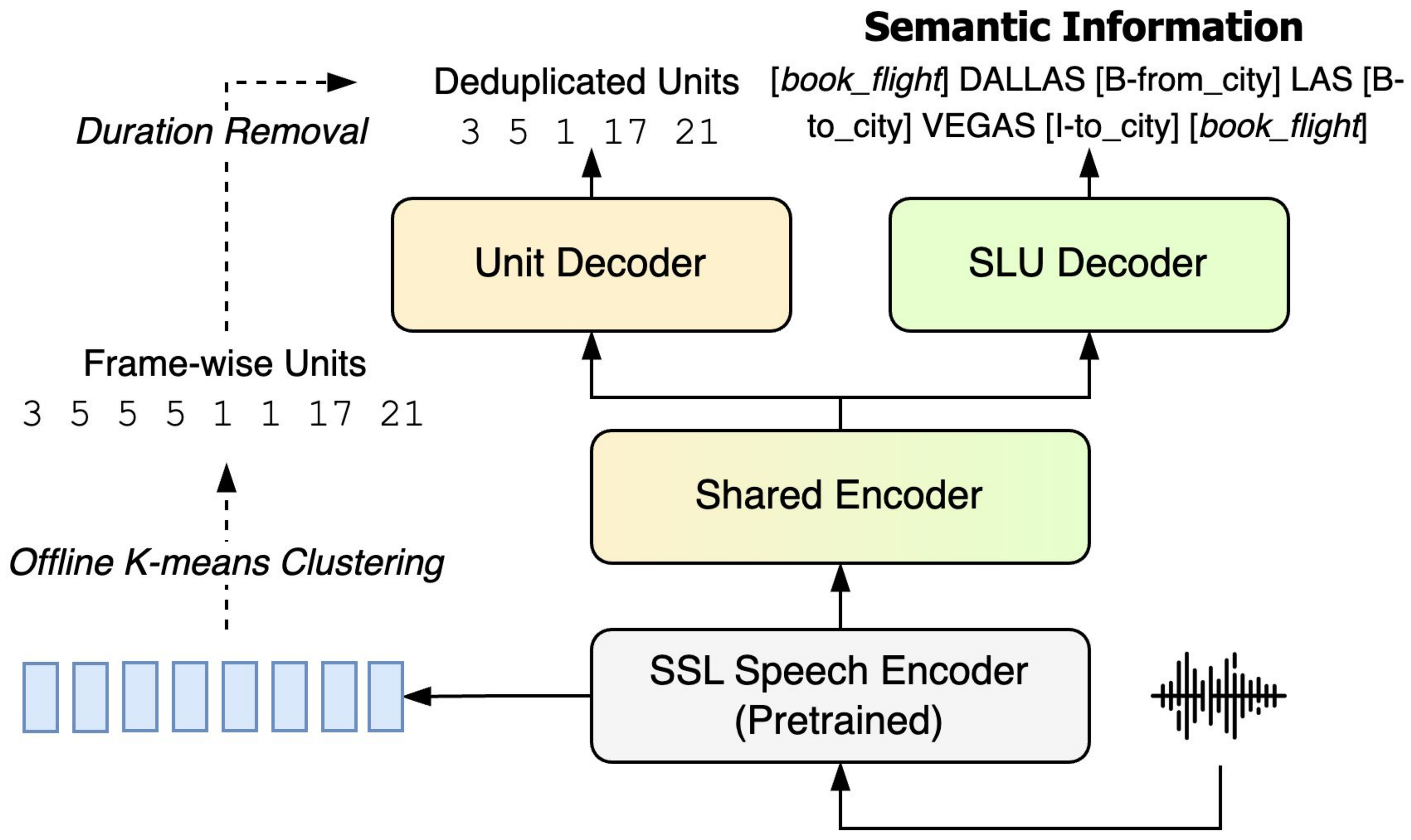
- **Transcript** : we cannot influence tariff reductions that would benefit the people of the united kingdom
- **SNER** : united **B-PLACE** kingdom **I-PLACE**

- **Transcript** : i want to fly from baltimore to dallas round trip
- **ICSF** : **atis_flight** baltimore **B-fromloc.city_name** dallas **B-toloc.city_name** round **B-round_trip** trip **I-round_trip atis_flight**

- **Transcript** : stop the work timer
- **Parse Tree** : **[IN:PAUSE_TIMER** stop the work **[SL:METHOD_TIMER** timer **]]**
- **SSP** : **[IN:PAUSE_TIMER [SL:METHOD_TIMER]]**

Proposed Method

Motivation: Discrete units mainly contain **content-related** information -> serve as the regularization target for SLU.



❑ **Training Objective**

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

$\mathcal{L}_{slu}$: cross-entropy loss for sequence generation

$\mathcal{L}_{aux}$: cross-entropy loss for unit sequence prediction

❑ **Different methods**

- **Baseline:** only the main task (i.e., λ = 0).
- **Unit (Prosposed):** use discrete units as the target of the auxiliary task with λ = 0.5.
- **Text (Upper bound):** use transcripts as the target instead of units.

Results

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

On all five SLU corpora, leveraging "**unit as intermediate target**" significantly outperforms the baseline method, even reaching similar performance as using ground truth text transcripts.

Discussions

❑ **Few-shot Capability**

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**

% means the portion of original training data
δ is performance drop from 100% to 10% training data.

The proposed method "Unit" has less performance drop compared to the Baseline method, achieving even better performance than Baseline (full data) with just 10% training data on SNIPS.

❑ **Noise Robustness (Performance Drop compared to clean input)**

G: Gaussian noise with the followed amplitude value. M: MUSAN background noise. dB: signal-to-noise ratio. Reverb: reverberation effect.

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**

Unit prediction improves the noise robustness for all types of noises compared to Baseline method, especially on Gaussian noise.