ICASSP 2023 SPOKEN LANGUAGE UNDERSTANDING GRAND CHALLENGE

Akshat Shrivastava[1], *Suyoun Kim*[1], *Paden Tomasello*[1], *Ali Elkahky*[1], *Daniel Lazar*[1], *Trang Le*[1], *Shan Jiang*[1], *Duc Le*[2], *Aleksandr Livshits*[1], *Ahmed Aly*[1]

[1] Meta, USA, [2]TikTok, USA
akshats@meta.com

ABSTRACT

Spoken language understanding (SLU) is a important field between the Speech and NLP community focused on converting a users' speech utterance into an executable semantic parse. In order to facilitate open research in this space, we introduce the 1st Spoken Language Understanding challenge hosted at ICASSP 2023. We leverage the newly released SLU dataset STOP [1]. In this challenge, participants are asked to compete in 3 tracks of SLU relevant to the field (1) Quality: build the highest performance model (2) On-device: build the highest quality model under 15M parameters and (3) Low-resource: Achieve the highest quality in a low-resource setting. While participants have made significant strides in the challenge, there is still a long way to go in building data and compute efficient SLU models.

Index Terms— Spoken Language Understanding, NLU

1. INTRODUCTION

Task-oriented conversational assistants have gained popularity in recent years for enabling speech-based interactions to accomplish tasks such as sending messages, retrieving weather information, and controlling devices. These systems achieve this by converting a user's speech into a semantic structure through Spoken Language Understanding (SLU).

SLU systems typically consist of first Automatic Speech Recognition (ASR) to convert speech to text and then Natural Language Understanding (NLU) to convert text to a semantic parse. Today many advances in this system consist of independent improvements to ASR and NLU components. However recently there has become an increased interest in End-to-End SLU systems with the promise to improve the performance by leveraging acoustic information lost in the intermediate textual representation and preventing cascading errors from ASR. Further, having one unified model has efficiency advantages when deploying assistant systems on-device for low power / mobile devices. In order to facilitate further progression in the SLU community we hosted the 1st Spoken Language Understanding Challenge at ICASSP 2023, leveraging the Spoken Task Oriented Parsing (STOP) dataset [1], the largest and most complex open source SLU dataset.

In this challenge participants are tasked with exploring the SLU space on 3 tracks. (1) How do we build the highest quality models §3.1 (2) How do we build high quality size efficient on-device models §3.2 and (3) How do we build data efficient models for new domains §3.3.

2. TASK DEFINITION

Semantic parsing is the task of converting a user's request into a structured format for executing tasks. Examples are given in the table below from the STOP dataset. Each parse consists of 2 key components: Intent and Slot. As an example consider "what is the weather in seattle"

Intent: Intention of a user based on predefined intent labels (intent = get weather)

Slot: Slots correspond to relevant arguments for each intent (slot = location: weather)

As queries become more complex we require **Compositionality [7]**: Slots can contain nested intent structures capturing multi-intent queries.

For this task, we leverage the **STOP (spoken task oriented parsing) dataset** [1], the largest most semantically complex end-to-end spoken language dataset. It contains over 200,000 audio files from over 800 different speakers. The text utterances and semantic parses are taken from TOPv2 [8] which contains 125K unique utterance parse pairs, across 8 domains, alarm, event, messaging, music, navigation, reminder, timer and weather. Crowd workers were requested to record themselves speacking each utterance.

The objective of the task is to maximize the **Exact Match (EM) Metric** Computed by comparing the exact string match (equivalence) of the predictoin and target intent/slot labels and leaf slot text.

Duc Le performed the work while at Meta.

Table 1: Tables present our challenge results across all of our tracks

(a) Quality Track Results

	Test EM
HuBERT E2E [1]	69.23
Cascaded [1]	72.36
Quality 1 [2]	**80.80**
Quality 2 [3]	75.90

(b) On-device Track Results

	Params	EM
On-device 1 [4]	**13.38**	**71.97**
On-device 2 [5]	15.00	70.90

(c) Low-Resource Track Results

	Weather 25 SPIS EM	Reminder 25 SPIS EM
HuBERT E2E [1]	46.77	15.38
LR 1 [6]	**75.00**	**63.30**

3. CHALLENGE TRACKS AND RESULTS

3.1. Quality

Our first track is focused purely on improving model quality. The target is to build the system whether E2E or cascaded that achieves the highest quality models defined by the highest exact match of the semantic parse tree on the test set. **Results**: We received 3 total submissions for this track with 2 winners for 1st [2] and 2nd place [3].

3.2. On-device

An important use case for end-to-end spoken language understanding is on-device modeling, facilitating the building of compressed systems. The goal of this track is to build the highest quality (EM) model with a limit of 15 million parameters. **Results** We received 2 submissions [4, 5] for this track both out performing the baseline E2E systems provided in [1].

3.3. Low-resource

Due to the multi-domain nature of the STOP and the relevance of domain scaling in language understanding [8], our third track consists of low-resource domain transfer. Given 6 held-in domains (alarm, event, messaging, music, navigation, timer), generalize to held-out domains (weather, reminder) by building a domain specific model. The new domain has 25 samples per intent/slot. **Results** We received 1 submissions for this track [6], greatly surpassing our provided baselines.

4. CONCLUSION

With the success of our first challenge and all our winners surpassing the baselines set as part of the STOP dataset [1], we have seen exciting progress in the space of spoken language understanding. Throughout the 3 tracks we had noticied a couple of interesting trends (1) pipeline systems still appear to be the top of the leader board, there needs to be more work done for E2E systems to surpass their pipeline counter parts. (2) Leveraging powerful pre-trained speech models such as Whisper [9] seems to significanatly improve performance. In the next SLU challenge we will focus on providing a more rigorous benchmarking for on-device models, and a finer grained metric collection to measure more granular progress.

5. REFERENCES

[1] Paden Tomasello, Akshat Shrivastava, Daniel Lazar, Po chun Hsu, Duc Le, Adithya Sagar, Ali Mamdouh Elkahky, Jade Copet, Wei-Ning Hsu, Yossef Mordechay, Robin Algayres, Tu Nguyen, Emmanuel Dupoux, Luke Zettlemoyer, and Abdel rahman Mohamed, "Stop: A dataset for spoken task oriented semantic parsing," *ArXiv*, vol. abs/2207.10643, 2022.

[2] Siddhant Arora, Hayato Futami, Shih-Lun Wu, Jessica Huynh, Yifan Peng, Yosuke Kashiwagi, Emiru Tsunoo, Brian Yan, and Shinji Watanabe, "A study on the integration of pipeline and e2e slu systems for spoken semantic parsing toward stop quality challenge," 2023.

[3] Othman Istaiteh, Yasmeen Kussad, Yahya Daqour, Maria Habib, Mohammad Habash, and Dhananjaya Gowda, "Srjo at icassp 2023 grand challenge: Spoken task oriented parsing," 2023.

[4] Zhang Gaosheng, Shilei Miao, Tang Linghui, and Qian Peijia, "A two-stage system for spoken language understanding," 2023.

[5] Yosuke Kashiwagi, Siddhant Arora, Hayato Futami, Jessica Huynh, Shih-Lun Wu, Yifan Peng, Brian Yan, Emiru Tsunoo, and Shinji Watanabe, "E-branchformer-based e2e slu toward stop on-device challenge," 2023.

[6] Hayato Futami, Jessica Huynh Siddhant Arora, Shih-Lun Wu, Yosuke Kashiwagi, Yifan Peng, Brian Yan, Emiru Tsunoo, and Shinji Watanabe, "The pipeline system of asr and nlu with mlm-based data augmentation toward stop low-resource challenge," 2023.

[7] Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis, "Semantic parsing for task oriented dialog using hierarchical representations," *CoRR*, vol. abs/1810.07942, 2018.

[8] Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta, "Low-resource domain

adaptation for compositional task-oriented semantic parsing," *CoRR*, vol. abs/2010.03546, 2020.

[9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022, vol. abs/2212.04356.