Spoken MASSIVE: A Multilingual Spoken Language Understanding Dataset

Chutong Meng   Milton Lin

May 5, 2024

Introduction

Spoken language understanding (SLU) is the foundation of voice-based virtual assistants like Alexa, Siri, and Google Assistant.

Two SLU Tasks

- **Input**: an utterance "does dominoes do takeaway"
- **Outputs**:
  - **Intent Classification**: "takeaway_query" – *we focus on this*
  - **Slot Filling**: "does [food_type : dominoes] do [order_type : takeaway] ".

SLU Challenges

Cascaded models face issues from the disconnection between ASR and NLU communities.

- Traditional cascaded approach: ASR + NLU.
- Potential issues:
  - ASR errors, especially in low-resource languages, affect NLU accuracy.
  - Important speech details (tempo, pitch) are lost after ASR.

Solution: End-to-End SLU Models

E2E models process speech directly to determine intent and slots. However,

multilingual E2E SLU is largely unexplored with no multilingual datasets available.

Contributions of This Work

1. Introduce first multilingual SLU dataset.
2. Provide baseline methods and results.

Two Proposed Methods:

- **Mine** speech data from existing multilingual ASR datasets.
- **Synthesize** speech for multilingual NLU datasets.

Related work

Various datasets covering single languages but no multilingual SLU datasets.

- S2IDataset: **Indian** accented English from the banking domain.
- Timers and Such: **English** voice commands on numerical data.
- VoxPopuli and Common Voice: multilingual ASR datasets, covering multiple European and other languages.

Mine Real Speech from Multilingual ASR Datasets

- Multilingual ASR dataset: CommonVoice [1].
- Search for transcriptions that are semantically close to sentences in MASSIVE.
- Apply LASER2 sentence embeddings [3] to embed CommonVoice and MASSIVE sentences.

$$\mathrm{score}(x, y) = \frac{\cos(x, y)}{\sum_{z \in NN_k(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in NN_k(y)} \frac{\cos(y,z)}{2k}} \tag{1}$$

where $NN_k(x)$ denotes the k nearest neighbors of x in the other side. We use $k = 16$.

margin score	MASSIVE sentence	CommonVoice transcript
1.5656	hey i missed you	I missed you.
1.4087	good evening	Good evening.
1.3398	i can't hear you	I couldn't hear you.
1.2819	please mute the sound	Please mute the sound.
1.2110	what is the definition for this object	What is the dimension of this object?

Generate Synthesized Speech

- Single speaker TTS model: Fairseq's MMS [5]; supports 1107 languages.
  - Easily overfit. Not useful.
- Multilingual TTS with *voice cloning*: XTTS[1].
  - 13 languages
  - Source speakers come from VCTK[2], an English multi-speaker dataset.
  - 110 speakers = 63 F + 47 M

[1] https://coqui.ai/blog/tts/open_xtts

[2] https://datashare.ed.ac.uk/handle/10283/3443

Dataset Statistics of Synthesized Speech

Subset	Speakers per sample	Duration (hours)	#Samples
Train	random 4 out of (10M+10F)	31.25	46,056
Dev	2M+2F	5.41	8,132
Test	2M+2F	8.09	1,1896

Table: Statistics for English.

- **Split of speakers:** same set of speakers across languages; different set of speakers across subsets.
- **Languages Covered:** 4/51: Chinese, English, German and Spanish.
- **Source:** Derived from the MASSIVE dataset, originally localized from SLURP. 60 intents and 55 slot types.

Evaluation Strategies

- **Human Evaluation:** Rate synthesized speech against real speech in a blind manner, although prone to errors [2].
- **Train on Synthetic, Test on Real:** Approach limited to English datasets which matches with SLURP. – **adopted**
- **Evaluation with Pre-trained Models:** Assess speech quality using available ASR models.

Audio!

- Voice cloning is good.
- 4 languages.

Experiments: Intent Classification

- Model: XLSR + classifier
  - XLSR: wav2vec2.0 model pretrained on speech of 53 languages
- Data: Multi-speaker TTS data
  - + mined speech data with different margin-score thresholds

Model Performance Across Languages

Language	Accuracy	F-1 Score
English	0.8198	0.8223
Spanish	0.8004	0.8005
German	-	-
Chinese	0.7540	0.7561

Table: Performance of models on the synthetic MASSIVE test set across different languages.

- Accuracy and F-1 are quite high
- En, Es perform better than Zh. XLSR model was pretrained on more En and Es data than Zh.

Model Performance on English SLURP Test Set

Dataset Type	Accuracy	F-1 Score
Real SLURP	0.6757	0.6880
Synthetic SLURP	0.8198	0.8223

Table: Comparison of model performance on real vs. synthetic SLURP test sets for English.

Could be improved if we use

- More speakers
- Data augmentation: e.g. noises, speed perturbation, etc.

Effect of Using Mined Data

Training Set	Accuracy	F-1 Score
Synthetic MASSIVE	0.6757	0.6880
+ margin-score $\geq$ 1.09	0.6695	0.6716
+ margin-score $\geq$ 1.00	0.6453	0.6494

Table: Model performance on SLURP test set when adding different amount of mined speech.

Adding mined data hurts the performance.

Exploring New Frontiers in SLU Model Development

- **Architectural Variations:** Experiment with both cascaded and end-to-end (E2E) SLU models to compare performance across monolingual and multilingual setups [6].
- **Data Ratio Studies:** Conduct ablation studies to analyze the impact of varying synthetic to real data ratios during training, [4].
- **Robustness and Generalization:** Evaluate model robustness under real-world conditions such as noisy environments and varying accents, focusing on models trained with synthetic data.

References I

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber.
Common voice: A massively-multilingual speech corpus.
In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

Chris Donahue, Julian McAuley, and Miller Puckette.
Adversarial audio synthesis.
In *International Conference on Learning Representations*, 2018.

References II

Kevin Heffernan, Onur Çelebi, and Holger Schwenk.
Bitext mining using distilled sentence representations for low-resource languages.
In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Ting-Yao Hu, Mohammadreza Armandpour, Ashish Shrivastava, Jen-Hao Rick Chang, Hema Koppula, and Oncel Tuzel.
Synt++: Utilizing imperfect synthetic data to improve speech recognition, 2021.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli.
Scaling speech technology to 1,000+ languages.
arXiv, 2023.

Yao Qian, Ximo Bian, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng.
Speech-language Pre-training for End-to-end Spoken Language Understanding, 2021.