Multilingual Spoken Language Understanding dataset

Milton Lin

February 14, 2024

Problems between ASR and SLU community:

- ASR (Automatic Speech Recognition, speech-text) and NLU (Natural Language Understanding, text-intent) communities don't talk, [FH21].
- Only a handful of works on End-to-End (E2E) Spoken Language Understanding (SLU) (speech-intent) models. This is theoretically more desirable:
  - ASR transcription often contains errors, which cascades to NLU module,
  - Even if the transcription is perfect, the rich information of speech (e.g., tempo, pitch, and intonation) is lost after ASR.
- Why people don't use E2E SLU: lack of training data. Best performing is still ASR+ NLU.

Goal

Fill the gap of the multilingual dataset for an end-to-end SLU by creating a multilingual speech to the intent dataset.

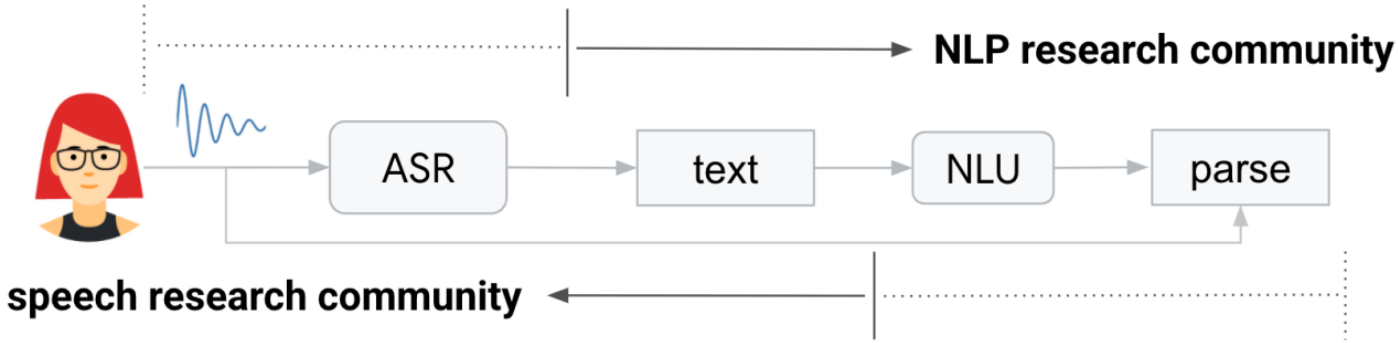

Figure: [FH21] Bridging ASR and NLU community

How do current datasets do?

- Currently *no* multilingual E2E SLU dataset. There is a Chinese data set from CAT SLU.
- monolingual speech to intent dataset, Fluent Speech Command (FSC) [Lug+19].
- In the field of ASR, is nicely depicted in [Con+22]

Dataset	#Languages	Total Duration	Domains	Speech Type	Transcripts
BABEL [13]	17	1k hours	Conversational	Spontaneous	Yes
CommonVoice [12]	93	15k hours	Open domain	Read	Yes
CMU Wilderness [15]	700	14k hours	Religion	Read	Yes
MLS [8]	8	50.5k hours	Audiobook	Read	Yes
CoVoST-2 [11]	22	2.9k hours	Open domain	Read	Yes
Voxlingua-107 [14]	107	6.6k hours	YouTube	Spontaneous	No
Europarl-ST [16]	6	500 hours	Parliament	Spontaneous	Yes
MuST-C [17]	9	385 hours	TED talks	Spontaneous	Yes
mTEDx [18]	9	1k hours	TED talks	Spontaneous	Yes
VoxPopuli [9]	24	400k hours	Parliament	Spontaneous	Partial
CVSS [19]	22	1.1k hours	Open domain	Read/Synthetic	Yes
FLEURS (this work)	102	1.4k hours	Wikipedia	Read	Yes

What can you do with these multilingual SLU?

- Relationship between robustness and multilinguality.
  - due to "cascading" errors - would we expect some size model of E2E SLU perform better in robustness tests?
  - does the property of robustness transfer across different language?
  - does training on more multilingual audio (or video) help with robustness performance? There is recent dataset of Common Phone, [Klu+22]. See also, [Gon+23].
- What features drive *audio* cross-lingual transfer ?
  - In text, [Cha+23] disuccses the geographic proximity of languages , shared writing systems or morphological systems etc.
- Run the same experiments that one does for XNLI.[1]
  - Do we also have the *curse of multilinguality*? [Con+18].
  - Comparison of the multilingual capabilities of cascading vs E2E SLU.

[1] We have different modality now

How to collect the data? Crowdsourcing

Crowd sourcing, use same method as [Con+22] and [Lug+19], we expand upon [Sch+19]. A data set of 57k annotated utterances in English (43k), Spanish (8.6k), and Thai (5k) across the domains weather, alarm, and reminder. The dataset is publicly available.

1. Crowdsourcing platform. We may follow CrowdSpeech, [PSU21], using Tolaka.
2. Each speaker was recorded saying each wording for each intent twice.
3. A separate validation phase will involve crowdsourced workers who will review the audio recordings.
4. Document and release anonymized demographic information of the speakers without compromising privacy.

How to collect the data? Synethetic generation

Synthetic generation, [Li+18]. The dataset was created by converting the passage part of (Stanford Question and Answering) SQuAD dataset into speech by using Google Text-to-Speech.

- How much better does a model trained perform when trained on human speech vs synthetic speech ?

References I

[Cha+23] Chang, Tyler A. et al. *When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages*. 2023. arXiv: 2311.09205 [cs.CL] (cit. on p. 5).

[Con+18] Conneau, Alexis et al. "XNLI: Evaluating Cross-lingual Sentence Representations". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018 (cit. on p. 5).

[Con+22] Conneau, Alexis et al. "FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech". In: *2022 IEEE Spoken Language Technology Workshop (SLT)* (2022), pp. 798–805. URL: https://api.semanticscholar.org/CorpusID:249062909 (cit. on pp. 4, 6).

References II

[FH21] Faruqui, Manaal and Hakkani-Tür, Dilek Z. "Revisiting the Boundary between ASR and NLU in the Age of Conversational Dialog Systems". In: *Computational Linguistics* 48 (2021), pp. 221–232. URL: https://api.semanticscholar.org/CorpusID:245124569 (cit. on pp. 2, 3).

[Gon+23] Gong, Yuan et al. "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers". In: *INTERSPEECH 2023*. interspeech2023. ISCA, Aug. 2023. URL: http://dx.doi.org/10.21437/Interspeech.2023-2193 (cit. on p. 5).

References III

[Klu+22] Klumpp, Philipp et al. "Common Phone: A Multilingual Dataset for Robust Acoustic Modelling". In: *International Conference on Language Resources and Evaluation*. 2022. URL: https://api.semanticscholar.org/CorpusID:246015467 (cit. on p. 5).

[Li+18] Li, Chia-Hsuan et al. *Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension*. 2018. arXiv: 1804.00320 [cs.CL] (cit. on p. 7).

[Lug+19] Lugosch, Loren et al. "Speech Model Pre-training for End-to-End Spoken Language Understanding". In: *ArXiv* abs/1904.03670 (2019). URL: https://api.semanticscholar.org/CorpusID:102352396 (cit. on pp. 4, 6).

References IV

[PSU21] Pavlichenko, Nikita, Stelmakh, Ivan, and Ustalov, Dmitry. "CrowdSpeech and Vox DIY: Benchmark Dataset for Crowdsourced Audio Transcription". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. 2021. arXiv: 2107.01091 [cs.SD]. URL: https://openreview.net/forum?id=3_hgF1NAXU7. Forthcoming (cit. on p. 6).

[Sch+19] Schuster, Sebastian et al. *Cross-Lingual Transfer Learning for Multilingual Task Oriented Dialog*. 2019. arXiv: 1810.13327 [cs.CL] (cit. on p. 6).