Harshita **Diddee**

harshitadd@gmail.com/t-hdiddee@microsoft.com github.com/harshitadd Personal Website

Education

2017 2021 — B.Tech, Computer Science and Engineering, Guru Gobind Singh Indraprastha University
Major GPA - 9.5/10.0, CGPA - 9.46/10.0
Graduated as the Best Outgoing Student for the Class of 2021

Experience

Present / July 2021 — SCAI Centre Fellow, Microsoft Research India, Bangalore
- Developed Quantized and Distilled Machine Translation models (approximate size being <400MB) for extremelely low-resource languages (data <25K sentences).
- Evaluating the efficacy of using such lightweight models to provide dynamic assistance to data providers via assistive-translation interfaces.
- **Mentored by Kalika Bali,Principal Researcher**

Neural Machine Translation | Data Quality Estimation

August 2022 / June 2022 — Visting Pre-Doc for JSALT'22, Johns Hopkins University, Baltimore, USA
- Evaluated the generalizability of speech and text cross-lingual models to extremely low-resource languages.
- Exploring the development of mutlilingual variant of speechT5.
- **Participated in the Speech Translation for Under-Resourced Languages Track**

Neural Speech Translation

October 2020 / March 2021 — Intern, AI4Bharat, Remote
- Implemented a tesseract-based OCR pipeline;
- Assisted the mining of parallel sentences from a dense (12M+), monolingual embedding space (between a target and source language) using FAISS
- **Mentored by Dr.Mitesh M. Khapra,Assistant Professor (CSE),IIT Madras**

Neural Machine Translation | Optical Character Recognition

March 2020 / November 2020 — Research Intern, IIIT Delhi, Delhi
- Validating a Cross-Silo Federated Learning (FL) hypothesis for heterogeneous clients having distributed and exclusive feature space.
- **Mentored by Dr.Koteswar Rao Jerripothula, Assistant Professor (CSE), IIIT Delhi**

Cross-Silo Federated Learning | Dropout Regularization

May 2019 / October 2019 — Research Intern, Celestini Program India (IIT Delhi), Delhi
- Developed a federated learning enabled custom deep learning model that powers VisionAir, an android application that predicts the Real-Time Air Quality Index of an image.
- Work published by TensorFlow
- **Mentored by Dr.Aakanksha Chowdhery, Google Brain**

Cross-Device Federated Learning | Computer Vision

Honors and Grants

September 2020 — [**ACM**] Grant to present my work at The 35th IEEE/ACM International Conference on ASE
August 2020 — [**Google AI**] Selected to attend the Google AI Summer School 2020
March 2020 — [**Google LLC**] Travel Grant to attend the TensorFlow Dev Summit at Sunnyvale,CA
October 2019 — [**The Marconi Society, Google LLC**] Runner's Up at the Celestini Prize India 2019
October 2019 — [**Jointly by the Government of Singapore and India**] Awarded 8000 SGD as 1st Runner's Up at the Singapore India Hackathon
March 2019 — [**Government of India**] Winner at Nationals for the e-Yantra Robotics Competition

Relevant Publications

October 2022 **Too Brittle To Touch : Comparing the Stability of Quantization and Distillation Towards Developing Lightweight Low-Resource MT Models, To Appear : Conference in Machine Translation, 2022, Authored by : Diddee et. al.**
Evaluated mutliple priors on which the distillation of extremely low-resource language translation models depend. Provided a comparitive analysis with post-training quantization for the same set of 8 languages.
Preprint

June 2022 **The Six Conundrums of Building and Deploying Language Technologies for Social Good, ACM COMPASS, 2022, Authored by : Diddee et. al.**
I explored literature at the intersection of NLP and HCI to substantiate the importance of defining concrete tools that can assist Language Technologists in resolving non-trivial decision dilemmas during low resource, community-driven technologies.
ACM Digital Library

February 2022 **Samanantar : The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages, Transactions of ACL, Authored by : Ramesh et. al.**
I implemented a tesseract-based OCR pipeline; Additionally, I assisted the mining of parallel sentences from a dense (12M+), monolingual embedding space (between a target and source language) using FAISS.
Transactions of ACL

December 2020 **PsuedoProp at SemEval-2020 Task 11:Propaganda Span Detection using BERT-CRF and Ensemble Sentence Level Classifier, SemEval, 2020, Authored by Chauhan et. al.**
I implemented a sequential BERT-based model's inference pipeline. I also developed a post-processing pipeline that would bridge the structural difference between the outputs of the Sentence Level classifier and the fine-grained analysis BERT-CRF Model. We ranked 14th globally on Shared Task 11 of SemEval'2020 collocated with COLING.
ACL Anthology

Projects

INMT-Lite
July 2021 - Present

Codebase

Since there will always be languages which have little to no data : we are working with the Gondi Community, a severely under-resourced language spoken by 2.4M across India, to identify (a) if intermediate models (trained with < 25000) samples can be utilized to generate recommendations that can accelerate the data provision yield of a crowdsourced system targetting such a language's collection and (b) what is the best interface (dropdowns, Bag of Words, gisting, etc) to present such low-quality assistance so as to not make it disruptive for the user's ability to provide data.

Low-Resource Machine Translation | Quantization | Distillation

Model Selection for Low-Resource ASR
October 2022 - Present

Codebase

Evaluation metrics are often not accurate proxies of the performance of systems on low-resource languages so I'm exploring (a) under what conditions can these metrics be used optimal model selection and (b) are there more representative metrics that can be monitored during training to select the best model.

Automatic Speech Recognition