Shikhar Bharadwaj

Pre-Doctoral Researcher, **Google Research**

🌐 shikhar-s.github.io ✉ shikhar.coder@gmail.com 🎓 Google Scholar ⭘ github.com/Shikhar-S

EDUCATION

Indian Institute of Science	Bengaluru, India
M.Tech (Research) in Intelligent Systems; CGPA: 8.8/10.0	*Aug. 2019 – May 2022*
Birla Institute of Technology and Science	Hyderabad, India
B.E. (Honors) in Computer Science Engineering; CGPA 9.74/10.0	*Aug. 2014 – July. 2018*

RESEARCH EXPERIENCE

Google Research	Bengaluru, India
Pre-Doctoral Researcher	*May 2022 - Present*
Microsoft Research	Bengaluru, India
Research Intern	*Summer 2018*

PUBLICATIONS   S=IN SUBMISSION, C=CONFERENCE, W=WORKSHOP, * DENOTES EQUAL CONTRIBUTION

[C.3] **Label Aware Speech Representation Learning For Language Identification** 📄
Shikhar Vashishth, Shikhar Bharadwaj, Sriram Ganapathy, Ankur Bapna, Min Ma, Wei Han, Vera Axelrod, Partha Talukdar
INTERSPEECH 2023

[C.2] **Efficient Constituency Tree based Encoding for Natural Language to Bash Translation** 📄
Shikhar Bharadwaj and Shirish Shevade
Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022)

[C.1] **Explainable Natural Language to Bash Translation using Abstract Syntax Tree** 📄
Shikhar Bharadwaj and Shirish Shevade
Conference on Computational Natural Language Learning (CoNLL 2021)

[W.3] **MASR: Multi-Label Aware Speech Representation** 📄
Anjali Raj, Shikhar Bharadwaj, Sriram Ganapathy, Min Ma, Shikhar Vashishth
Workshop on Automatic Speech Recognition and Understanding (ASRU 2023)

[W.2] **GitHub Issue Classification Using BERT-Style Models** 📄
Shikhar Bharadwaj* and Tushar Kadam*
Second prize - Competition track at NLBSE workshop, ICSE 2022

[W.1] **An extraction based approach to keyword generation and precedence retrieval** 📄
G. V. Sandeep and Shikhar Bharadwaj
Forum for Information Retrieval Evaluation workshop (FIRE 2017)

[S.2] **Multimodal Modeling For Spoken Language Identification** 📄
Shikhar Bharadwaj*, Min Ma*, Shikhar Vashishth*, Ankur Bapna, Sriram Ganapathy, Vera Axelrod, Siddharth Dalmia, Wei Han, Yu Zhang, Daan van Esch, Sandy Ritchie, Partha Talukdar, Jason Riesa

[S.1] **CodeQueries: A Dataset of Semantic Queries over Code** 📄
Surya Prakash Sahu, Madhurima Mandal, Shikhar Bharadwaj, Aditya Kanade, Petros Maniatis, Shirish Shevade

SELECTED RESEARCH PROJECTS

- **Speech to Text Transfer Learning in Multimodal models**   *June 2023 - Present*
  ***Advisors**: Dr. Partha Talukdar, Dr. Sriram Ganapathy*
  - The multimodal transfer hypothesis: knowledge transfer from speech to text modality for machine translation task in languages with limited-text.
  - Designed experimental settings and set up evaluation pipeline.

- Currently testing large scale PaLM2 baselines (upto 8B parameters).

- **Project Vaani: Data collection for Indic Languages.** *January 2023 - Present*
  Advisors: Dr. Partha Talukdar, Dr. Sriram Ganapathy
  - Project Vaani aims to cover the language landscape of India by region anchored speech data collection. ▶
  - Benchmarked and analysed results from internal ASR and Language Identification models on Vaani data.
  - Results show that Vaani is a challenging dataset for ASR models because of dialectical variations.

- **Spoken Language Identification** *July 2022 - September 2023*
  Advisors: Dr. Partha Talukdar, Dr. Sriram Ganapathy, Ankur Bapna
  - Built **Museli [S.2]** - a multi-modal framework for language identification of YouTube videos. Our model beats the speech-only baselines by 6% (absolute F1 score) and achieves SOTA performance on public language identification datasets. This method scales even better on internal YouTube datasets spanning over 500 languages.
  - Worked towards learning language aware speech representations for better spoken language identification. The LASR model [**C.3**] uses contrastive loss in addition to MLM based losses for learning language information.
  - MASR [**W.3**] is an extension to LASR that includes external knowledge in the form of lang2vec vectors. Worked on the holistic evaluation of the MASR method to include ASR and non-semantic speech evaluations.

- **Semantic Queries over Code** *March 2022 - May 2022*
  Advisor: Dr. Aditya Kanade, Dr. Shirish Shevade
  - Helped in creating a benchmark for question answering over code, and evaluated models like GraphCodeBERT and CodeBERT on this dataset. [**S.1**]

- **Natural Language to Bash Translation** 📄 *January 2021 - March 2022*
  Advisor: Dr. Shirish Shevade
  - Developed an algorithm for translating Natural Language to Bash commands by utilizing command Abstract Syntax Tree and Bash manual page data, resulting in explainable predictions beating baselines like T5 and Seq2Seq with attention. [**C.1**]
  - Developed a novel method for Natural Language to Bash command translation using constituency tree structure of the input invocation. Results include a 1.8x improvement in inference time, 5x reduction in model parameters (compared to the Transformer) and SOTA performance. [**C.2**]

Awards and Recognition

- **Institute Merit Scholarship**: Top 2% of the batch for 7 semesters and top 1% for 1 semester at BITS
- **NTSE Scholarship**: Awarded by Govt of India for qualifying National Talent Search Examination
- **INSPIRE Scholarship**: Awarded by Govt of India to students with top 1% percentile marks in ISC
- **National Standard Examination in Physics**: Qualified NSEP (top 2.5% in India)
- **Media Coverage**: Audio collection effort in collaboration with the *Indian Institute of Science* has been covered by The White House and The Economic Times 📰 ▶
- **Google Internal Awards**: Three bonuses for contributions to the annual Google4India event and Google-USM

Professional Experience

Myntra Designs Pvt Ltd. Remote, India
Machine Learning Intern *May 2020 - July 2020*

Media Net Mumbai, India
Platform Engineer *Aug. 2018 - July 2019*

Other Skills

- **Teaching Assistantships**: Computer Programming (CS F111), Object Oriented Programming (CS F213)
- **Programming Languages**: Python, C++, Java, SQL, Bash
- **Relevant Coursework**: *Graduate Level:* Deep Learning for NLP, Data Analytics, Linear Algebra and Probability, Computational Methods of Optimization, Design and Analysis of Algorithms; *Undergraduate Level:* Machine Learning, Data Mining, Data Structures and Algorithms, Object Oriented Programming
- **Tools and Frameworks**: Pytorch, Pytorch-Lightning, OpenNMT, Pandas, Matplotlib, Mlpack, Stanford Parser