Simran Khanuja

PhD in Computer Science | Carnegie Mellon University

🌐 simran-khanuja.github.io @ khanuja.simran7@gmail.com github.com/simran-khanuja
🎓 Google Scholar twitter.com/simi_97k

Education

-
Aug 2022 | **Carnegie Mellon University** | **Pittsburgh, Pennsylvania**
PhD in Computer Science (NLP), *QPA: 4.08/4.00*
Advisor: Prof. Graham Neubig

Aug 2020
Aug 2015 | **Birla Institute of Technology and Science, Pilani** | **Goa, India**
B.E. (Honors) Computer Science; M.Sc.(Hons.) Economics, *CGPA: 8.81/10.00*

Experience

Aug 2022
Aug 2020 | **Google Research | Natural Language Understanding [🌐]** | **Bangalore, India**
Pre-Doctoral Researcher | Advisor: Dr. Partha Talukdar
Collaborators: Dr. Sebastian Ruder, Dr. Alexis Conneau
Projects: Multilingual Representations for Indian Languages (*MuRIL*), Merging Pre-trained Language Models using Distillation (*MergeDistill*), Multilingual Neural Semantic Parsing for Google Assistant (*mNSP*), Cross-Lingual Text Speech Embeddings (*XTeSE*)

Jul 2020
Jul 2019 | **Microsoft Research | Project Mélange [🌐]** | **Bangalore, India**
Research Intern (Bachelor Thesis) | Advisors: Dr. Sunayana Sitaram, Dr. Monojit Choudhury
Projects: General Language Understanding and Evaluation for Code-Switching (*GLUECoS*), Code-Mixed Natural Language Inference, Adapting TULR for Code-Mixing

Mar 2019
Aug 2018 | **Birla Institute of Technology and Science, Pilani [🌐]** | **Goa, India**
Undergraduate Research | Advisors: Dr. Sreejith V., Dr. Aswani Kumar Mishra
Projects: Multi-modal Emotion Aware Connected Healthcare (*EACH*), Quantitative Analysis of Equity Ownership and Firm Performance of the Indian Manufacturing Sector

Aug 2018
May 2018 | **International Institute of Information Technology | LTRC [🌐]** | **Hyderabad, India**
Summer Research Intern | Advisor: Dr. Dipti Misra Sharma
Project: Generating Synthetic Code-Mixed data using Syntactic Parse Trees

Publications

S=In Submission, C=Conference, W=Workshop

[S.2] **MuRIL: Multilingual Representations for Indian Languages** [PDF]
Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, Partha Talukdar
large | base [*Coverage*: Economic Times | Indian Express | Google AI Blog]
[Working Paper]

[S.1] **Evaluating Inclusivity, Equity, and Accessibility of NLP Technology: A Case Study for Indian Languages**
Simran Khanuja, Sebastian Ruder, Partha Talukdar
[Under Review] **[ARR November '21]**

[C.3] **MergeDistill: Merging Pre-trained Language Models using Distillation** [PDF]
Simran Khanuja, Melvin Johnson, Partha Talukdar
Annual Conference of the Association for Computational Linguistics (Virtual) **[Findings of ACL'21]**

[C.2] **GLUECoS: An Evaluation Benchmark for Code-Switched NLP** [PDF]
Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, Monojit Choudhury
code | website
Annual Conference of the Association for Computational Linguistics (Virtual) **[ACL'20]**

[C.1] **Unsung Challenges of Building and Deploying Language Technologies for LRL Communities** [PDF]
Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, Kalika Bali
International Conference on Natural Language Processing, Hyderabad, India [**ICON'19**]

[W.3] **FLEURS: Few-Shot Learning Evaluation of Universal Representations of Speech** [PDF]
Alexis Conneau*, Min Ma*, Simran Khanuja*, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, Ankur Bapna
IEEE Spoken Language Technology Workshop [**SLT 2022**]

[W.2] **A New Dataset for Natural Language Inference from Code-mixed Conversations** [PDF]
Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, Monojit Choudhury
data
International Conference on Language Resources and Evaluation [**CALCS, LREC'20**]

[W.1] **Dependency Parser for Bengali-English Code-Mixed Data enhanced with a Synthetic Treebank** [PDF]
Urmi Ghosh, Simran Khanuja, Dipti Misra Sharma
code
International Workshop on Treebanks and Linguistic Theories [**TLT, SyntaxFest'19**]

Select Research Projects

Multilingual Representations for Indian Languages (MuRIL) Aug'20 - Present
Advisors: Dr. Partha Talukdar, Dr. Sebastian Ruder

- Built a multilingual model specifically focused on Indian languages, which is now open-sourced (**large** | **base**). The base variant beats mBERT by **10%** and the larger one beats XLM-R by **3.3%**. [**Press Coverage - 1, 2, 3**].
- The large model has been used by **all the winning teams** at the Question-Answering challenge for Hindi and Tamil, hosted by Google India.
- Evaluated the inclusivity, equity and accessibility of multiple multilingual models for all Indian languages, hence quantifying the diversity of users each can serve. Devised a purely empirical approach to fine-tuning that maximizes metrics across all three dimensions. [**Under Review**]
- Working on expanding language coverage, making the model robust to code-mixing and creating challenging evaluation test sets that cover all 22 scheduled Indian languages.

Cross-Lingual Text Speech Embeddings (XTeSE) Aug'21 - Present
Advisor: Dr. Alexis Conneau

- Built a cross-modal speech-text retrieval model by fine-tuning a pre-trained speech-text model on parallel pairs using a contrastive ranking loss.
- Currently doing large-scale inference and analysis of the trained models and extending it to be multilingual.

Merging Pre-trained Language Models using Distillation (MergeDistill) Nov'20 - Feb'21
Advisor: Dr. Partha Talukdar

- Devised a framework named *MergeDistill* to merge multiple pre-trained language models using knowledge distillation in a task-agnostic manner. [**ACL '21 Findings**]
- Applied this technique in neural semantic parsing for Google Assistant with positive results.

General Language Understanding and Evaluation for Code-Switching (GLUECoS) Jul'19 - Mar'20
Advisors: Dr. Sunayana Sitaram, Dr. Monojit Choudhury

- Experimented with cross lingual word embeddings and multilingual generalized language models on a variety of downstream NLP tasks for code-mixed data. Eventually built a benchmark for the evaluation of models/methods that process code-mixed data, which is now open-sourced (**code** | **website**) [**ACL '20**]
- Proposed and oversaw the creation of a new dataset for the task of conversation entailment in code-mixed data, which is now open-sourced (**data**) [**CALCS@LREC '20**]

Multi-modal Emotion Aware Connected Healthcare (EACH) Jul'19 - Mar'20
Advisor: Dr. Sreejith V.

- Implemented an Emotion Recognition System to recognize the emotional state of a patient. The final state is a weighted average of several parameters including facial expressions, speech signals and physiological signals including heart rate and breathing rate (**github** | **report**)

Generating Synthetic Code-Mixed data using Syntactic Parse Trees May'18 - Aug'18

Advisor: Dr. Dipti Misra Sharma

- Worked on generating valid code-mixed data to improve the accuracy of a code-mixed language model [**TLT@ SyntaxFest '19**]
- Used rule-based approaches wherein we chunk parallel sentences consistently and perform a post inter-leaving of the chunks based on head matching. Tools worked with include the Stanford Parser, LTRC Hindi Parser and the GIZA++ word alignment tool (**github**)

Academic Service

Reviewer	MRL@EMNLP'21, TALLIP'20
Sub-Reviewer	EMNLP'21, EMNLP'20, ACL'20

Teaching and Leadership

An Introduction to (Modern) TensorFlow *Co-Instructor* Aug '21

- Conducted a hands-on Tensorflow tutorial session attended by **100+** members from Academia.

Coffee Club Mentorship Program, Google *Co-Lead* July '21 - Present

- Co-Lead of the Coffee Club program at Google India. This program enables women employees to seek mentorship from senior executives across Google in pursuit of encouraging diversity at higher levels within the organisation.

BITS Alumni Mentorship Program *Mentor* Aug '20 - Present

- Mentorship for undergraduate students looking to secure research theses opportunities. My mentee secured a year-long thesis at Microsoft Research, India!

NLP Reading Group, Google Research, Bangalore, India *Participant* Aug '20 - Present

- Active participant in our weekly reading group where I regularly present research papers and engage in discussions.

IKDD NLP Session *Host* Aug '21

- Hosted the NLP networking session at IKDD 2021 where Dr. Monojit Choudhury was our guest speaker!

Fireside Chat with Jeff Dean *Host* Sep '20

- Hosted a Fireside Chat with Jeff Dean on his virtual Google India visit!

Music Society, BITS Goa *Core Member* Aug '17 - Aug '18

- Managed events of the Music Society, BITS Goa as one of the five core members in the organizational team and gave performances as a vocalist.

Applied Econometrics (F342) *Teaching Assistant* Jan'19 - May'19

- Conducted hands-on tutorials for Data Analysis in Python and R.

Financial Management (F315) *Teaching Assistant* Aug'17 - Dec'17

- Helping students understand concepts, clearing doubts and assisting in test corrections.

Talks and Interviews

"Decode with Google"

- Speaker List | Talk (2:28:00 onwards) (registration required) [🌐] August 2022

"An Introduction to (Modern) TensorFlow"

- CVIT Summer School, IIIT Hyderabad [🌐] August 2021 (Remote)

"Journey into Research"

- Rotaract Club, BITS Hyderabad [🌐] January 2021 (Remote)
- Google Research, India December 2020 (Remote)

"ICSE National Topper"

- India Times | Times of India | Indian Express May 2013 (Pune, India)

Skills

Languages	Python, C++, Java, HTML
Frameworks	Tensorflow, NLTK
Tools	Visual Studio, Git, GIZA++, Stanford Parser, Elasticsearch
Relevant Coursework	Machine Learning, Information Retrieval, Neural Networks, Data Structures and Algorithms, Design and Analysis of Algorithms, Object Oriented Programming, Probability and Statistics, Applied Econometrics, Mathematics and Statistics