Simran **Khanuja**

Pre-Doctoral Researcher | Google Research

Website | E-mail | Github | Twitter | Google Scholar | LinkedIn

Education

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

Experience

Present | **Google Research | Natural Language Understanding** [🌐] — **Bangalore, India**
Aug 2020 | *Pre-Doctoral Researcher* | *Advisor: Dr. Partha Talukdar*
Projects: Multilingual Representations for Indian Languages (*MuRIL*), Merging Pre-trained Language Models using Distillation (*MergeDistill*), Multilingual Neural Semantic Parsing for Google Assistant (*mNSP*)

Jul 2020 | **Microsoft Research | Project Mélange** [🌐] — **Bangalore, India**
Jul 2019 | *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 | **Birla Institute of Technology and Science, Pilani** [🌐] — **Goa, India**
Aug 2018 | *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 | **International Institute of Information Technology | LTRC** [🌐] — **Hyderabad, India**
May 2018 | *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.1] **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
tfhub | huggingface [*Coverage*: Economic Times | Indian Express | Google AI Blog]
[Working Paper]

[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.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]
Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, Monojit Choudhury
code
International Workshop on Treebanks and Linguistic Theories **[TLT, SyntaxFest'19]**

Select Research Projects

Multilingual Representations for Indian Languages (MuRIL) Aug'20 - Present

Advisor: Dr. Partha Talukdar

- Built a multilingual model specifically focused on Indian languages, which is now open-sourced (**tfhub** | **huggingface**)
- Working on expanding language coverage, making the model robust to code-mixing and creating challenging evaluation test sets.

Merging Pre-trained Language Models using Distillation (MergeDistill) Nov'20 - Feb'21

Advisor: Dr. Partha Talukdar

- Experimented with merging multiple pre-trained language models in a teacher-student knowledge distillation framework, aptly called *MergeDistill*. [**ACL '21 Findings**]

Multilingual Neural Semantic Parsing for Google Assistant (mNSP) Dec'20 - Present

Advisor: Dr. Partha Talukdar

- Worked on building the neural semantic parser for *Google Assistant* which is being *launched* for eight Indian languages [*Bengali (BN), Gujarati (GU), Kannada (KN), Malayalam (ML), Marathi (MR), Tamil (TA), Telugu (TE), Urdu (UR)*].
- We merge MuRIL with the assistant model in production using MergeDistill to achieve an overall win-loss ratio of **1.51** over the affected **0.74%** of queries.

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

Advisors: 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**)

Talks and Interviews

"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)

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.

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.

Weekly Team Updates (NLU Team, Google Research) *Organiser* Sep '21 - Present

› Organise the weekly team updates for the NLU team at Google Research India.

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 and assisting in test corrections.

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

Miscellaneous

› Hosted the NLP networking session at IKDD 2021 where Dr. Monojit Choudhury was our guest speaker!
› Co-organising the Coffee Club program at Google India. This program enables women employees to seek mentorship from senior leaders across Google.
› Hosted a Fireside Chat with Jeff Dean on his virtual Google India visit!
› Core member and vocalist of the Music Society at BITS Goa.