Simran KHANUJA

linkedin.com/in/simran-khanuja-6b80b6144 @ khanuja.simran7@gmail.com
Bangalore, India github.com/SimranKhanuja simran-khanuja.github.io/

RESEARCH POSITIONS

Ongoing Aug' 20	Google Research, BANGALORE, India Pre-Doctoral Researcher, *Natural Language Understanding Team*
Aug' 2020 Aug' 2019	Microsoft Research, BANGALORE, India Research Intern, *Natural Language Processing Team*

EDUCATION

Aug' 2020	Birla Institute of Technology and Science, Pilani, GOA, India B.E. (Honors) Computer Science ; M.Sc.(Hons.) Economics, *CGPA : 8.81/10.00*
Aug' 2015	City International School, PUNE, India Central Board of Secondary Education, *CGPA : 95.4/100.0*
Aug' 2013	St. Mary's School, PUNE, India (National Topper) Indian Certificate of Secondary Education, *CGPA : 98.4/100.0*

RESEARCH GOAL

My research experience and interests revolve around **Natural Language Processing**. My primary research goal is to advance technologies in **low [text] resource** scenarios. It not only serves to create a significant **societal impact**, but also presents unique challenges of learning from **limited data** and encourages one to exploit **non-text modalities**, something a researcher striving to endow machines with "intelligence" must eventually confront.

PUBLICATIONS

- MergeDistill : Merging Pre-trained Language Models using Distillation
  Simran Khanuja, Melvin Johnson, Partha Talukdar
  Findings of 2021 Annual Conference of the Association for Computational Linguistics (ACL '21)
  pdf

- MuRIL : Multilingual Representations for Indian Languages
  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
  pdf | tfhub | huggingface
  Coverage : Economic Times | Indian Express | Google AI Blog

- GLUECoS : An Evaluation Benchmark for Code-Switched NLP
  2020 Annual Conference of the Association for Computational Linguistics (ACL '20)
  Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, Monojit Choudhury
  pdf | code | website

- A New Dataset for Natural Language Inference from Code-mixed Conversations
  CALCS Workshop : International Conference on Language Resources and Evaluation, 2020 (LREC '20)
  Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, Monojit Choudhury
  pdf | data

- Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities
  2019 International Conference on Natural Language Processing (ICON'19)
  Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury and Kalika Bali
  pdf

- Dependency Parser for Bengali-English Code-Mixed Data enhanced with a Synthetic Treebank
  18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)
  Urmi Ghosh, Dipti Misra Sharma, and Simran Khanuja
  pdf | code

Experience

Present
Aug' 2020

Google Research, Bangalore, India
Pre-Doctoral Researcher | Advisors : Dr. Partha Talukdar

- Built a multilingual model specifically focused on Indian languages named **MuRIL**, which is now open-sourced. (paper | tfhub | huggingface)
- Experimented with merging multiple pre-trained language models in a teacher-student knowledge distillation framework, aptly called **MergeDistill**. [**ACL '21 Findings**]
- Worked on building the neural semantic parser for the **Google Assistant** which is being **launched** for **eight Indian languages**. 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.

Jul' 2020
Jul' 2019

Microsoft Research, Bangalore, India
Research Intern | Advisors : Dr. Sunayana Sitaram, Dr. Monojit Choudhury

- Worked on experimenting with fine-tuning strategies for contextualized multilingual models using the XTREME benchmark as a test-bed.
- 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**]

Aug' 2018
May 2018

MT-NLP Lab, IIIT Hyderabad, India
Summer Research Intern | Advisors : Dr. Dipti Misra Sharma [Github Link]

- Worked on generating valid Hindi-English code-mixed data to improve the accuracy of a code-mixed language model. Similar methodology of generation implemented for Bengali-English. [**TLT@ SyntaxFest '19**]
- Mainly used rule-based approaches wherein we chunk parallel sentences consistently and perform a post interleaving of the chunks based on head matching.
- Tools worked with include the Stanford Parser, LTRC Hindi Parser and the GIZA++ word alignment tool.

Nov' 2018
Aug 2018

BITS Pilani, K.K. Birla Goa Campus, India
Undergraduate Research Project | Supervisor : Dr. Sreejith V. [Github Link] [Report Link]

- 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.

Jul' 2017
May 2017

Liveweaver India Pvt. Ltd., Pune, India
Summer Intern | NLP for Voice Recognition

- Performed basic NLP taks including, morphological segmentation, coreference resolution, sentiment analysis and named entity recognition. Also worked with Elasticsearch for clustering, indexing paragraphs and calculating scores of similarity between the paragraphs.

Teaching and Mentoring

- Conducted a **hands-on Tensorflow Tutorial** session at the CVIT Summer School, 2021, attended by **100+** members from Academia.
- Spoke about my journey to Google Research to a cohort of **160+ potential pre-doctoral candidates**.
- Shared my research path with members from the Rotaract Club at BITS Hyderabad. The interview can be found here.
- Teaching Assistant for the course, "Financial Management" at BITS Goa
- Teaching Assistant for the course, "Applied Econometrics" at BITS Goa. Conducted hands-on tutorials for Data Analysis in Python

Honors and Awards

2013	Secured an All India (National) Rank 1 in the 10th board examination, ICSE
2016	Received the institute's merit scholarship for the first semester

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.

Skills

Languages	Python, C++, Java, HTML
Frameworks	Tensorflow, NLTK
Tools	Visual Studio, Git, GIZA++, Stanford Parser, Elasticsearch
Courses	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