Simran Khanuja | simran-khanuja.github.io
CMU User ID: khanuja.simran7@gmail.com

Statement of Purpose | Carnegie Mellon University
Language Technologies Institute, PhD/MLT, Fall 2022

Today's data-driven Natural Language Processing (NLP) technologies fail to serve over ninety percent of the world's seven thousand languages. Drawing on my experiences as an intern at Microsoft Research (MSR) and a **Pre-Doctoral Researcher** at Google Research, I am inspired to make algorithmic advances in learning from **limited text** data. I have made initial contributions to advancing under-represented NLP by not only developing competent models and frameworks, but also by creating challenging evaluation sets, which are just as crucial for developing and deploying models. During my graduate studies, I'm keen on leveraging **non-text modalities** involving speech, images, videos or structured knowledge bases, in addition to text, to endow machines with language understanding capabilities. Research advances on this front would not only have a significant impact by being linguistically inclusive, but would also bridge the gap in modeling linguistic form to understanding linguistic meaning. Below, I detail how my experiences have prepared me to pursue this goal and my motivation to tackle some of these challenges in the future.

Modeling and Evaluation for Under-Represented NLP

The progress of any field is tightly coupled with its evaluation paradigm, something I first realized when surveying literature on contextual embedding models at MSR India. In multilingual Indian communities, *Code-Mixing* (CM) (the alternation of multiple languages in one utterance), is common, but there was no one-stop paradigm to evaluate a model's CM capabilities. Progress was impeded by scattered datasets covering few tasks. Inspired to address this gap, I spearheaded an initiative to create a benchmark for CM, alongside my mentors, Dr. Sunayana Sitaram and Dr. Monojit Choudhury. While we curated released datasets for most tasks, we built one for Natural Language Inference (NLI) from scratch using movie scripts. Each premise represented a multi-turn dialogue from which the model had to draw inferences. I laid out detailed annotation guidelines and conducted a two-phase pilot study in my lab, whose findings were leveraged to annotate a larger dataset (released at CALCS, **LREC 2020** [1]). This was a great learning experience, informing me of the processes involved in annotating data. With all the datasets in place, I evaluated multilingual models as baselines, and eventually presented our benchmark, *GLUECoS*, at **ACL 2020** [2] (open-sourced[1]).

When I was evaluating multilingual models for CM, I noticed a significant performance difference between English and other languages within the monolingual realm. While linguistic diversity and inclusion have evolved to be a pressing concern today, I realised that measures to quantify these phenomenon are lacking. Therefore, following discussions with my mentors, Dr. Partha Talukdar and Dr. Sebastian Ruder, I proposed an evaluation paradigm that measures the *inclusivity*, *equity*, and *accessibility* of a technology, to quantify the diversity of users it can serve (**Under Review: ACL 2022**) [3]. While we tailored pre-defined metrics for *inclusivity* and *accessibility*, I leveraged concepts I studied as part of my second major in **Economics**, and used the *Gini coefficient* (a well-established metric used for estimating societal wealth inequality) to measure *equity*. Using our paradigm, I highlighted the linguistic biases in models for Indian (IN) languages, concluding with novel approaches to model fine-tuning that mitigate these biases. This project was exciting and unique as I was exposed to the potential of interdisciplinary thinking and I wish to continue having this flavor in my research.

The reason why multilingual models exhibit performance gaps across languages is because they cannot equally represent the 100+ languages they are trained on, given their limited capacity. For the past year, I've been working with Dr. Partha Talukdar at Google Research India on modeling strategies for IN languages. As a first step, I built *MuRIL* [4], a multilingual model trained on a limited set of closely-related IN languages, to address deficiencies of massively multilingual models for IN. The **open-sourced** models performed exceptionally well with the base and large variants beating state-of-the-art models by **10%** and **3.3%** respectively. [**Press Coverage**: *Indian Express*[2], *Economic Times*[3]]. Not only was this effort well-received, but it also imparted me with critical technical skills involved in data processing and pre-training large language models (LMs), making this a wholesome experience. One promising direction to improve this model has been to train using phonetic transcriptions, capitalizing on the phonetic similarity amongst several IN languages, like Hindi and Urdu. This strengthens my inclination towards integrating information from multiple modalities to improve text LMs.

I recently integrated MuRIL into the semantic parsing module of Google Assistant for IN, using a teacher-student knowledge distillation framework named *MergeDistill* (**ACL 2021** Findings [5]) which I devised to merge multiple pre-trained LMs. This was an enriching experience as I got to tie in my research with a product, acquired competent engineering skills and worked with a diverse group of peers spanning geographies and product teams. An issue with these LMs is that they

[1] https://github.com/microsoft/GLUECoS
[2] https://www.newindianexpress.com/opinions/2020/dec/24/making-use-of-the-language-landscape-diversity-2240471.html
[3] https://economictimes.indiatimes.com/tech/technology/google-aims-to-help-researchers-startups-better-understand-indian-languages/articleshow/79773091.cms

aren't robust to Automatic Speech Recogniser (ASR) errors, and parsing speech directly is desirable. As a step towards this, I'm working with Dr. Alexis Conneau on retrieving text documents directly from speech queries using a pre-trained speech-text model.

Future Interests and Graduate Studies

Learnings from my past experiences lead me to believe that model building and evaluation should be approached **holistically**, where multimodal agents learn via interactions, and evaluation assesses learning capabilities. In *GLUECoS*, the best models performed only slightly over chance for CM NLI. Built over conversational data, this dataset tests phenomena that are an integral part of daily human communication, but models trained on **static web data** lack these capabilities. While working on Assistant, I also realized that since CM is predominantly a social, spoken phenomenon, web-text evaluation sets don't capture speech-related phenomena, like disfluencies, and also cannot measure how well a model deals with varying degrees of CM based on users and context.

During my graduate studies, I'm keen on devising modeling techniques that aren't solely dependent on volumes of text to learn language. Heavy reliance on text also excludes non-literate people and speakers of languages without written systems to benefit from technology. From a cognitive perspective, humans understand the meaning of textual language not in isolation, but within the larger context of a rich **multimodal** environment leveraging **knowledge** they've gathered through their lifetime. A lack of grounding in current approaches have made models adept at detecting higher-order compositional patterns but they are unable to reason why they compose a certain response. In *Metaphors we live by*, Lakoff and Johnson also note that most abstract concepts seem to be grounded in basic physical metaphors giving us leverage to bootstrap language use for an array of experiences. This tight coupling between language and the world we live in is what I look forward to exploiting in pursuit of endowing machines with language understanding capabilities.

Multimodality may offer unique solutions to problems in multilinguality that I have tackled in the past. From a **modeling** perspective, while languages are represented using different scripts at the surface, the underlying information we wish to convey remains consistent across its manifestations. Therefore, instead of jointly modeling form, can we leverage this consistency and use grounding as a mechanism to learn multiple languages? From an **evaluation** perspective, designing culturally relevant tests in text is challenging and probing multilingual representations using images would reveal biases along a multi-dimensional axis. A simple input of *'a person wearing clothes'* in multiple languages would reveal different perceptions of gender, skin tone and clothing across several cultures. Finally, from an **application** perspective, cross-modal modeling can empower non-literate people to use technology and also improve overall user-experience.

Why a PhD at CMU: Drawing on my experiences, I'm excited to pursue in-depth, focused research as a PhD student. Owing to its brilliant faculty and student community, I believe that Carnegie Mellon University (CMU) is an ideal place for me to accomplish my goals. At CMU, I am particularly interested in working with [redacted], whose work on low-resource languages and multilingual NLP closely aligns with my past work and future interests. His recent work on utility estimation formed the basis of my work on quantifying the linguistic diversity of NLP technologies [3]. I would also be fortunate to pursue multimodal research with [redacted], given my interests in leveraging non-text modalities to ground language and build interpretable, linguistically inclusive systems. I would like to work with [redacted] since his interests in language grounding and interactive systems is in sync with what I intend to pursue. [redacted] work on robust and efficient NLP is also inspiring and applicable in resource-lean scenarios, that I'm often faced with in my research on under-represented NLP. I would like to analogously learn from esteemed faculty in the sciences, since utilising knowledge from these disciplines is imperative for me to pursue my goals. I can bestow models with inductive biases grounded in linguistic theory to build generalizable systems, drawing from *Linguistics*; build inclusive systems with a user-first approach, drawing from *Economics*; and leverage non-text modalities to build grounded social agents, drawing from *Cognition*. I look forward to collaborating with exceptional graduate students at CMU and am inspired by the works of students in similar areas like, Cindy Wang, Paul Pu Liang, Shruti Rijhwani and Volkan Cirik, to name a few. Overall, pursuing a PhD at CMU would be an ideal way to lay the groundwork for a research career in academia.

[1] S. Khanuja, S. Dandapat, S. Sitaram, and M. Choudhury, "A new dataset for natural language inference from code-mixed conversations," CALCS, LREC, 2020.

[2] S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram, and M. Choudhury, "GLUECoS: An Evaluation Benchmark for Code-Switched NLP," ACL, 2020.

[3] S. Khanuja, S. Ruder, and P. Talukdar, "Evaluating Inclusivity, Equity, and Accessibility of Nlp Technology: A Case Study for Indian Languages," 2021.

[4] S. Khanuja, D. Bansal, S. Mehtani, et al., "MuRIL: Multilingual Representations for Indian Languages," 2021.

[5] S. Khanuja, M. Johnson, and P. Talukdar, "MergeDistill: Merging language models using Pre-trained Distillation," ACL Findings, 2021.