Deepak Nathani   *https://deepakn97.github.io*   deepakn1019@gmail.com

Statement of Purpose

of Deepak Nathani (Ph.D. applicant at University of California, Santa Barbara for Fall–2022)

My primary research interests are in NLP, specifically in Natural Language Generation; examples include Summarization, Controlled Text Generation, and Dialogue Generation. In the future, I want to lead a research group at a university, advance the field and train the next generation of researchers. Hence, I believe that pursuing a Doctoral Degree is the next step in this journey. During my time as a graduate student, I plan to make progress towards addressing some of the problems in natural language generation, some of which are mentioned below:

- *Controlled Text Generation:* How can we control the text generated using Large Language models? Some examples include controlling the hallucinations and stylized generation.

- *Summarization and Question answering:* How can we design systems which can reliably extract meaningful information from a plethora of text available on the internet?

- *Dialogue Generation:* How can we create better open domain and goal-oriented dialogue models which are diverse and grounded in real-world information?

I have spent the past year working with **Dr. Partha Talukdar** as a **Pre-Doctoral Researcher** at **Google** on the intersection of Dialogue and Controlled Text Generation. Prior to this, I have worked on various research topics in Machine Learning such as Topological Data Analysis, Node Classification, and Knowledge Graph Completion. Below, I talk about my introduction to research and some of the projects that have significantly impacted my research approach.

Controllable Text Generation and Conversational Agents   The release of general-purpose Large Language Models (LLMs) like GPT-3, BERT has proven to be a key development in the field of NLP. We find ourselves reliant on these LLMs for various tasks, Natural Language Generation generation being one of them. While these models are capable of human-like text generation, they also suffer from various issues like hallucination and degeneration - misrepresenting information, which is potentially worse than generating less fluent text. Since these models are trained on large amounts of unstructured texts available on the internet, it can be challenging to steer the generation towards specific attributes such as politeness, formality.

Motivated by this, in 2020, I joined Google Research to work on Controllable Text Generation and Conversational agents under the guidance of Dr. Partha Talukdar. While at Google, I have worked towards creating a Conversational Health Assistant. This project provided me with a unique opportunity to work on an open-ended research problem from the ground up. I brought structure to the project, which required understanding the existing literature on goal-oriented dialogue agents by identifying the core research areas: understanding the user input, a reinforcement-learning based decision module, and a text generation system to generate personalized dialogue.

While most of the recent research on Text Style Transfer assumes access to extensive style-labeled data, recent work [6] has attempted "few-shot" style transfer. In a related work [3], we pushed the state-of-the-art for few-shot style transfer with a new method modeling the stylistic difference between paraphrases. Moreover, this newly proposed method is better able to control the amount of style transfer using a scalar knob. We also designed an automatic evaluation suite for low-resource languages with no pre-existing evaluations sets using Adapter based Transformers [5].

During this work, we found that most style transfer systems today suffer from the exact copying problem – the model copies the input verbatim without changing the style. One possible reason for this could be that we train the neural model to recreate the input sentence from an encoded vector representation. This training paradigm works well, but it can also be the cause of limited diversity in the generated texts. While we try to alleviate the verbatim copying problem to some extent in this work, I believe that there is much work to be done to create diverse style transfer systems while preserving the

content. To that extent, I am currently exploring methods to improve diversity and use the label-free text style transfer work for personalizing dialogue agents.

I also led an effort to collect a TyDi QA style question answering dataset for two Indian languages – ChAII (Challenge in AI for India), which involved collaborating with annotators, preparing and maintaining the data collection pipeline, and launching a baseline QA model. We recently launched a challenge on Kaggle to motivate ML practitioners to work on question answering for low-resource Indian languages.

Prior Research Experience During my undergraduate studies at IIT, Hyderabad, I was introduced to the field of Machine Learning in my junior year. I joined the Krama lab led by **Dr. Manohar Kaul** to explore this interest. As my first project, we worked on a method to solve the partial assignment problem and applied it to image matching using point clouds. Throughout this project, I learned how researchers approach an unsolved problem, do a literature survey, work on a method to address specific gaps in the previous solution, devise a new approach, and most importantly, design experiments that can convey the motivation and methodology succinctly. This work was accepted at **ICML 2018**. [7].

I enjoyed working in a research environment and continued to work with Dr. Manohar. In my junior year, I was presented with an opportunity to intern at IBM Research under Dr. Sumit Bhatia and Dr. Bapi Chatterjee to work on another interesting problem of node classification using Topological Data Analysis.During the internship, I applied my previous experience, getting familiar with the research topic and formulating the method. Through this experience, I realized the importance of being up to date with the literature to come up with original research ideas. This work was later accepted at **Complex Networks, 2020**. [1].

While working at IBM, I came across the ongoing research on attention mechanisms in neural networks and related work on node classification – Graph Attention Networks [8]. I explored the possibility of applying the concept of attention in the Knowledge Graph Completion problem. Subsequently, we proposed the first graph attention-based neural network for the same. Apart from contributing towards the method formulation, I contributed significantly towards designing experimentation and academic writing. We were able to publish our work at **ACL, 2019** [4].

Following this, we worked on the less explored Few-Shot graph classification problem. We proposed a novel clustering-based solution, and the work was accepted at **ICLR, 2020**. [2].

The **ACL, 2019** [4] work, coupled with the interest in LLMs such as GPT-2 at the time, motivated me to explore the attention-based NLP models and their applications further.

Why get a Ph.D.? My long-term research goal is to bridge the language barrier between humans and machines by creating natural language systems grounded in real life and factually correct. Moreover, these systems should be personalized. While I am open to pursuing interesting research ideas within the broader field, I am particularly interested in improving the controllability in language generation systems and using these controlled systems for tasks such as dialogue generation.

My future goals are directed towards academia, where I can work on progressing the field further and get an opportunity to advise the next generation of researchers. I believe that a Ph.D. program will provide me with the necessary structure to conduct the research I am enthusiastic about and simultaneously help me hone my teaching and research advising skills. This career choice is influenced by my positive experiences as the lead of a Programming Club during my undergraduate. I enjoyed teaching and advising younger students.

At UC Santa Barbara, I am particularly interested in the research directions explored by **William Wang** and **Lei Li**. Specifically, **William Wang's** work on model-agnostic explanations for dialogue response generation is exciting and relevant to my own research goals. Following the work done within UCSB NLP group, I am confident that the Doctoral program offered at UC Santa Barbara is an excellent fit for me. Moreover, I believe that the large NLP and ML community at University of California, Santa Barbara will help me gain varied perspectives and valuable experience as a researcher and a teacher.

References

[1] Sumit Kaur Bhatia, Bapi Chatterjee, Deepak Nathani, and Manohar Kaul. A persistent homology perspective to the link prediction problem. In *COMPLEX NETWORKS*, 2019.

[2] Jatin Chauhan, Deepak Nathani, and Manohar Kaul. Few-shot learning on graphs via super-classes based on graph spectral measures, 2020.

[3] Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. Few-shot controllable style transfer for low-resource settings: A study in indian languages, 2021.

[4] Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based embeddings for relation prediction in knowledge graphs, 2019.

[5] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online, October 2020. Association for Computational Linguistics.

[6] Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. Textsettr: Few-shot text style extraction and tunable targeted restyling, 2021.

[7] Charu Sharma, Deepak Nathani, and Manohar Kaul. Solving partial assignment problems using random clique complexes. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4586–4595. PMLR, 10–15 Jul 2018.

[8] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.