Statement of Purpose

Ashim Gupta (MSc by thesis applicant for Fall—2019)

My Objectives – I, Ashim Gupta, am presenting my statement of purpose in applying to the **University of Alberta** for admission to the MSc. program in Computer Science. My research primarily focuses on **Natural Language Processing (NLP)**, and **Machine Learning**.

My motivation for pursuing a masters degree in natural language processing stems from my positive first-hand experience working as an NLP researcher with **Prof. Pawan Goyal** [1] and **Prof. Sudeshna Sarkar** [2] at IIT Kharagpur. During my graduate study, I expect to do high-quality research, teach undergraduate and graduate classes to refine my knowledge.

NLP Research at IIT Kharagpur – While researching at IIT Kharagpur, I have been intrigued by the research in information extraction with a focus on biomedical textual data. Particularly, in one project, I have been responsible for developing specialized NER systems for biomedical domain. I was tasked with development of CRF based drug tagger making use of morphological, orthographic, and linguistic features. Although the system worked well, I realized the approach involving designing of hand-engineered features for each of the different entity types was not scalable to the number of biomedical entity types we were interested in. This led me to investigate approaches involving LSTM-CRFs, that automatically learn feature representations and have performed exceptionally for general english domain. After preliminary experiments, I discovered that these neural methods fail to address problems specific to biomedical domain.

On further analysis, I identified two particular cases specific to scientific medical text where the systems needed immediate improvement: 1) the generous usage of abbreviations/acronyms in biomedical text without explicit mention of their full forms, and 2) the words with different meaning in biomedical and general text i.e. *polysemy*. We hypothesized that making use of the context more effectively could ameliorate these issues. We adopted two solutions for this: First, making use of contextualized word representations obtained from a bi-directional language model. Second, replacing a fixed transition matrix with a context conditioned matrix modeled with a neural network. Consequently, the system surpassed the performance of two very competitive baselines by at least 3 % absolute F-score on 4 prominent biomedical datasets. Performing further analysis of the results, I discovered that our system was also superior for larger entities (spanning more than 5 tokens). Our work has been accepted to **41st European Conference on Information Retrieval (ECIR) 2019** [5]. This work has not only taught me the intuition and importance of defining research hypotheses out of preliminary observations, but also the discipline to systematically conduct experiments to validate these hypotheses.

Motivated by my earlier experience of teaching at a school in Mirzapur near Varanasi, where I realized the unavailability of standard text books in native Indian languages, I have undertaken a machine translation project for morphologically rich low-resource indic languages with **Prof. Pawan Goyal**. We are currently exploring completely unsupervised approaches as well as those with limited supervision. Our preliminary experiments have informed us of the superiority of statistical translation approaches (phrase-based SMT) over the neural network methods for distant and low-resource language pairs. Our current approach involving the use of a bilingual lexicon for initial phrase table induction and iterative back translation has achieved a BLEU score of 7.0 for English-Hindi language pair. Analysis of results have revealed the shortcomings of this PBSMT model which, although provides acceptable translations for short phrases, fails to produce completely fluent full sentence translations. We hypothesize that using a Neural MT model, which takes into account a larger context, initialized with synthetic data generated from PBSMT can help alleviate this issue. We are also attempting to simultaneously exploit the lexical similarities shared between many Indian languages (Hindi, Marathi etc.). Working on this project, I am realizing how recent advances in AI coupled with increased outreach of these systems have a very vital role to play in future of developing countries like India.

Self-motivated Research – I believe it is imperative for young researchers to be able to think and work independently. As researchers, we have ideas and I believe it is crucial to explore these ideas to maintain a strong sense of intellectual curiosity. During my time at IIT Kharagpur, I have independently assisted a master's research student in which I proposed to formulate metaphor detection as a sequence tagging problem. I was responsible for implementing the neural LSTM architecture along with carrying out extensive experiments. Results demonstrated the efficacy of using a sequence prediction module with carefully chosen set of input features. This work has been published in *FigLang Workshop* collocated with **NAACL, 2018** [6].

My intellectual curiosity has also led me to independently explore the problem of scientific text classification. I implemented a hierarchical neural network architecture, in which I modeled the implicit sequential structure between sentences in a scientific abstract using CRFs. I extended this approach by modifying the CRF to more effectively model the global context in an abstract by using a neural network based transition matrix, achieving an improvement of 0.4 % over the current state-of-the-art. During my analysis, I found that using attention primarily helped for sentences greater than 60 tokens in length. I am currently exploring the use of attention statistics pooling based approach, recently introduced at INTERSPEECH, 2018 to further increase the system's performance.

Background in Computer Science – My exposure to computer science and programming began after I entered the **IIT-BHU**. Although I did struggle initially, the subject had already sparked enough interest in me to pursue it beyond my curriculum. To gain practical development experience, I have worked as an Android Development Intern after my sophomore year and as a Software Engineer after graduating from college. Courses such as *Algorithms and Data Structures*, *Computer Systems*, *Numerical Methods*, *Linear Control Systems*, and *Digital Control Systems* have provided me with a strong mathematical and computational background. My interest in control theory prompted me to undertake my Bachelor's thesis on Multi-Sensor Data fusion using Kalman Filter under **Prof. D.N. Vishwakarma** [4], in which I integrated the noisy measurements from a low-cost inertial measurement unit (IMU) with global positioning sensor (GPS) data. The experience was enriching as it gave me an opportunity to apply my theoretical knowledge of linear algebra, calculus, and probability.

I had my first experience with research when I collaborated with **Prof. Rajeev Srivastava** [3], in Department of Computer Science, during my junior year to work on Fuzzy c-means clustering based Image Retrieval system. I also implemented a relevance feedback mechanism using a Support Vector Machine (SVM). Impressed by problem-solving skills, the professor motivated me to submit a research proposal to Design and Innovation Hub (DiH) on application of image retrieval to mammograms suspicious of breast cancer. I developed a specialized region growing algorithm to segment suspicious regions from these mammograms. We experimented and found that extracting gabor and wavelet features coupled with Euclidean distance based metric works surprisingly well in our case. More details of our work, which was accepted to **IEEE UPCON, 2015**, can be found here [7]. While working on this project, I fostered a deep interest in academic research and the aptitude for problem solving that it requires.

The MSc. program offered at the University of Alberta with its research focus would be the perfect platform to prepare me for PhD. After having studied some of the research papers from these professors, I have a strong feeling that Alberta is a great match for my research interests. Indeed, it would be a privilege to do my masters studies at the **Department of Computing Science, University of Alberta**.

Thank you very much for taking the time to read.

References

[1] http://cse.iitkgp.ac.in/~pawang/

[2] http://cse.iitkgp.ac.in/~sudeshna/

[3] https://www.iitbhu.ac.in/dept/cse/people/rajeevcse

[4] http://www.old.iitbhu.ac.in/eee/index.php/peo/faculty/23-dnv.html

[5] https://ashim95.github.io/docs/ecir_paper.pdf

[6] https://ashim95.github.io/docs/naacl_paper.pdf

[7] https://ashim95.github.io/docs/ieee_paper.pdf