Statement of Purpose

Amirhossein Kazemnejad
ah.kazemnejad@gmail.com

.

My objective is to pursue a P.h.D. in Computer Science with a focus on Natural Language Processing. Specifically, my research interests include **controlled and conditional text generation**, **machine comprehension**, and **question answering systems**, and I am currently working with Dr. Mohammad Taher Pilehvar, investigating the surprisal effects of text anomalies on language models. Having worked in two research groups and being involved in the industry for almost two years put me in a unique position to offer this program a strong computer science background accompanied by solid engineering skills and two years of research experience.

During my first research experience, I worked at Advanced BigData Analysis Lab (ABDAL) on a data mining project, which was defined by HamrahAvval (a local wireless operator). We were set to find abnormal and unfair utilization of the company's services by using their provided dataset. Although I am unable to elaborate upon the exact details of the dataset due to HamrahAvval confidentiality, I could say that it was basically an encrypted time-series dataset without any annotations. The main challenge was the dataset's immense size (100M+ records) and noisy nature, which made it difficult to infer any useful information. Therefore, we designed and constructed a graph representation of the dataset from its raw records as a higher-level source of information. We then used a combination of data mining techniques and graph-based features to detect unusual patterns and nodes. In this study, I was responsible for designing the feature sets of each node and edge. Also, I developed a program using Apache Spark to query a subset of the graph, which was stored on multiple servers. One of the main challenges of this project was the dataset's colossal size. Even the abstract graph version had 3M nodes and 17M edges. Moreover, the amount of incompatibility and noise in records' values, which are often present in real-world datasets, made us to invent specific mechanisms to verify each value. Currently, our system has replaced the company's legacy rule-based system, and helped them to ban much more fraudulent users with higher accuracy.

Although the project above paved my way into data science and AI, the progress of state-of-the-art dialog systems at that time made me heavily interested in NLP, and in fact, changed my direction toward it. With Dr. Mahdieh Soleymani at Machine Learning Lab, I led a research in improving the retrieve-and-edit framework. This two-step framework augments the text generation models with a non-parametric memory which is filled with training samples. Traditional sequence-to-sequence models, especially for small datasets, suffer from the lack of diversity and complexity in their outputs; therefore, introducing training samples to such neural networks enables them to use training instances as a prototype for generating high-quality outputs. Previous research papers in this field have extensively studied the first part—the retrieving procedure—but mostly neglected the second step. I, however, concentrated on the editing step by introducing *Micro Edit Vectors* (MEV), which changed the way that these models incorporate the training set in their generation process.

The MEV concept allowed the model to have granular control over its usage of the training corpus' instances. Although this improvement benefited all text-generation tasks (such as dialog generation and machine translation), our evaluation was focused because of obvious resource and time limitations. We not only showed the superiority of the method in the paraphrase generation but also suggested clear improvements in data augmentation for text (which is currently followed by one of the lab members). Furthermore, it had an unexpected promising results in Text Style Transfer because of its ability to combine two sentences in a granular manner. For this study, I submitted a first-authored [1] paper to AAAI 2020, which has one of the lowest acceptance rates among the AI conferences, but it got rejected with an average score of 7 (from 8, 7, and 6 scores; detailed reviewers' comments are available upon request). Nonetheless, I am preparing it for resubmitting to ACL 2020. Leading a re-

search project and taking its full responsibility, although it is often challenging for an undergraduate student, had valuable experiences for me such as giving task to a research team and coming out of research failures. Moreover, lab's computational resources were so limited that I had to design a pipeline of free online services to run my experiments. This pipeline not only ended up being used by other lab members, but also my article explaining it on Medium [2] got published by one of its largest publications. As another outcome of this project, I have also publish a blog post [3] describing the positional encoding in Google's Transformer architecture.

Recently, I've been working with Dr. Pivehvar in order to investigate the reaction of language models to semantical text anomalies, comparing it to the human brain. Take this sentence as an example: "Celine will come to the party. She ought to bring skyscrapers." This sentence, due to its apparent semantic violation, puts a more cognitive load on the reader's mind. Hence, we seek to find whether models trained on human language corpora exhibit similar degrees of surprise as humans. Although this project is in its early stages, we believe that it can help us to understand how well Deep Learning models understand semantics and how we can improve them. Also, it encourages better computational representations of words and contexts as models will need to capture more world knowledge. Our aim is to create a semantic-based benchmark dataset similar to WIC [4].

Before beginning my career in research, I had worked for a few companies as a software engineer. These industry experiences helped me to maintain a solid programming skills so that I am now an open-source contributor at Google's Tensorflow 2.0 project [5]. My programming skills also came in handy in my second research project as it enabled me to quickly evaluate 40+ prototypes of my model and conveniently modify other papers' code in order to create baselines for my experiments. Additionally, the lessons that I learned from leading the technical team at my startup, assisted me in managing my second research project and my university's course projects.

In summary, I have experienced leading a research project, teaching as a lecture assistant, working as software engineer, and even running my own startup. By comparing these experiences, which I am glad that I grained them during my bachelor's degree, I become more and more confident about my choice of pursuing a P.h.D.. Through these experiences, I found myself most engaged when I am working on projects with no well-specified description and, more importantly, no well-defined solutions. These problems although are more challenging, their forthcoming reward does absolutely worth it. Moreover researching in universities, due to the open nature of scientific research, has the potential of impacting more people, which, however, may not be the case in private companies. Therefore, my ultimate goal is to be a professor at a research-focused university.

Though I am open to a wide variety of research within NLP, my experience working with semantic-intensive projects has inspired an interest in building systems that demand broad semantic knowledge such as machine comprehension. I intend to work on improving the model's representation of information they retrieve from available sources—such as the Internet. Also, because of experience working with text generation models, making their output quality comparable to those of humans is exciting for me. Moreover, I am curious to explore conditioning the output of NLP models to other parameters such as the persona of users, which is particularly useful in dialog systems.

.

References

(1) **Kazemnejad, Amirhossein**, Mohammadreza Salehi, and Mahdie Soleymani Baghshah. "Paraphrase Generation by Learning How to Edit from Samples" *Submitted to ACL 2020*

(2) **Kazemnejad, Amirhossein**. "How to Do Deep Learning Research with Absolutely No GPUs — Part 1." Medium, October 18, 2019. https://medium.com/swlh/how-to-do-deep-learning-research-with-absolutely-no-gpus-part-1-1517450d4010.

(3) **Kazemnejad, Amirhossein**. "Transformer Architecture: The Positional Encoding." Medium, September 30, 2019. https://medium.com/@kazemnejad/transformer-architecture-the-positional-encoding-d961dbd952e0.

(4) Pilehvar, Mohammad Taher, and Jose Camacho-Collados. "WiC: The Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations." *In Proceedings of the NAACL 2019*.

(5) GitHub. "Kazemnejad - Overview." Accessed November 17, 2019. https://github.com/kazemnejad.