Shashank Siripragada

Boston, MA 414-544-4714 siripragada.s@northeastern.edu https://shashanksiripragada.github.io/

Education

Northeastern University **Expected May 2023**
Masters in Information Systems, (GPA : 4.0) *Boston, MA*
Coursework: Data Science, Application Engineering

International Institute of Information Technology **July 2017**
Bachelors in Electronics and Communication Engineering *Hyderabad, India*

Technical Skills

Languages: Python, SQL, C++, C
ML/Vision: PyTorch, OpenCV, scikit-learn, Pandas
Tools: Qlik, Tableau, Qt, AzureML Studio, SLURM, Shell

Experience

International Institute of Information Technology **May 2019 – July 2021**
Research Fellow *Hyderabad, India*

- Research focused on developing Neural Machine Translation (NMT) systems and Multilingual datasets for 11 Indian Languages.
- Released *cvit-pib, mkb* one of the largest Multilingual parallel corpora for training NMT systems on Indian languages.
- The work done as a part of this project was published in WAT 2019, LREC 2020, CODS COMAD 2021 and was featured in premier translation forums WMT, WAT 2020.

Primera Medical Technologies **June 2017 – May 2019**
Data Scientist *Hyderabad, India*

- Built predictive models for early detection and intervention in patients at risk of C.Difficile, hospital overstay, SNF placement to assist hospital staff in patient logistics.
- Designed comprehensive Qlik dashboards using EDI 835&837 data for monitoring Insurance Claims & Denials at enterprise scale.

Hyundai Motor India Engineering **June 2016 – July 2016**
Intern *Hyderabad, India*

- Developed an application to calculate Aperture Ratio from an image of a speaker grill using OpenCV/C++ & Qt.

Publications

- **Revisiting Low Resource Status of Indian Languages in Machine Translation** CODS COMAD, India, 2021
- **A Multilingual Parallel Corpora Collection Effort for Indian Languages** LREC, France, 2020

Projects

Large Scale Parallel Corpus from The Web | *Python, PyTorch, flask* **Jan 2021**

- Developed and released a flask web application to extract large-scale parallel corpus from news sources.
- This application pipeline contains efficient translation, document retrieval and sentence alignment modules enabling users to work at scale.
- Demonstrated improvements in corpus size and quality with iterative improvements in machine translation and document retrieval performance.

Research Paper Miner | *Python* **Dec 2016**

- Implemented a tool to extract algorithm names from research papers to help users navigate scientific research by specific domains.
- The workflow consists of pdf-to-text conversion, tokenization, named entity recognition (NER) and employs cosine similarity on word2vec vectors to determine relevant algorithm names and domains.

Image Captioning | *Python, PyTorch* **Apr 2017**

- Implemented an encoder-decoder framework to generate natural language descriptions given an image, experimenting with Vanilla RNNs, GRU and LSTM networks in PyTorch.