SARANG SHRIVASTAVA

in sarangshrivastava24 sarang.github.io @ sarang24s@gmail.com +1-669-342-9495 Menlo Park, USA

EXPERIENCE

5+ years of experience in designing, developing and fielding NLP/ML solutions into production in the financial domain. Currently focusing on improving the data quality for the firm in the Legal Entities and Products space using Generative AI and Large Language Models (LLMs).

I have also worked on a variety of problems in the space of DocumentAI ranging across, but not limited to, document structure understanding, representation learning, entity recognition, relation extraction, learning to rank, layout-aware language models and tabular structure understanding.

VP - Senior Applied NLP/ML Engineer

Goldman Sachs

Jan 2022 – Jan 2023 Menlo Park, California, USA

Focus Areas: *LLMs, Generative AI, T0pp, Flan, SetFit, Sentence Transformers, Approximate nearest neighbour search, Search and Retrieval*

- **Record Linking across large databases**
  - Trained a sentence transformer on database records, followed by an approximate nearest neighbour search on the embedding space of records. This step focused on achieving high recall and reduced the search space from **10 million records to 100**.
  - Used **LLMs (T0pp and Flan)** for bootstrapping the training data required to fine-tune a high-precision discriminative model (SetFit).
  - Fine-tuned **Setfit** and trained an ensemble of LLMs and Setfit as the final model to link records.
  - Improved the linking accuracy by more than **10%** above the baseline.

- **Record De-Duplication in a database**
  - Curated a dataset of duplicate records by working closely with the business team using Prodigy as the choice of the annotation tool
  - Fine-tuned a **BERT** based classification model to filter non-human users who were not relevant to the business
  - Leveraged **sentence transformers** to map the textual descriptions of the remaining records into an embedding space
  - Performed an **approximate nearest neighbour search** on the embedding space to create a graph-based representation of the entire database. The nodes of the graph represent the record and the edges represent the distance between them in their embedding space
  - Used Breadth First search on the graph to form the clusters of duplicate records. Augmented the stopping condition of the algorithm by leveraging **LLMs (Flan)**
  - Fine-tuned Setfit to trim down the clusters created in the previous step to remove the false positives introduced. Achieved an accuracy of **92%** for this problem

EDUCATION

- Bachelor of Technology in Computer Science and Engineering
  MNNIT, Allahabad - 2016 (G.P.A : 8.53/10)

SKILLS

Python C++ SQL NLTK **prodigy** **PyTorch** **TensorFlow** Keras **spaCy** **scikit-learn** **sentence-transformers** **Huggingface** NumPy DeepSpeed PAWLS Flask FastAPI Kubernetes

PUBLICATIONS/PREPRINTS

- Hitkul, **Sarang Shrivastava**, Eliot Brenner, **"Layout-Aware Neural Model for Resolving Hierarchical Table Structure"**, Submited to ACL Rolling Review, November 2021, [Link]
- **Sarang Shrivastava**, Afreen Shaikh, Shivani Shrivastava, Chung Ming Ho, Pradeep Reddy, Vijay Saraswat, **"Handling tree structured text: parsing directory pages"**, Available on Arxiv [Link]
- Priyanja Singh, **Sarang Shrivastava**, **"Privacy of Organization in Online Social Networks"**, in Springer proceedings of the International Conference on Advanced Computing, Networking and Informatics, ICACNI 2017 [Link]
- **Sarang Shrivastava**, Priyanja Singh, Ranvijay **"Impostor Detection through Chat Analysis"**, in International Multi-Conference on Information processing, IMCIP 2016 [Link]

SERVICE/OUTREACH

Mentored an undergraduate student for a period of 6 months in the field of Finance and Natural Language Processing as part of the Women in Tech Programme, April 2021 - October 2021

Gave a talk on **"Knowledge extraction for professionals from financial documents"** at GIDS 2020

Associate - Applied NLP/ML Engineer

Goldman Sachs

Jan 2020 – Dec 2021 Bengaluru, India

- **Joint NER and Relation extraction on Layout Rich Documents**
  Focus Areas: Language models v/s Layout aware Language models
  - Curated two datasets for NER and Relation extraction tasks on visually rich documents - Cover pages in Credit Agreements and Directory pages in Prospectuses
  - Fine-tuned **BERT, ROBERTA, LayoutLMv1, LayoutLMv2** in various settings and showed that joint training of NER and relation extraction tasks using layout-aware language models on layout-rich documents outperforms standard language models
- **Semantic Understanding of Tabular structures - Identifying Table components like Column headers, Row headers, Captions etc and hierarchy between them**
  Focus Areas: Layout aware Language Models, Tabular Data, Classification, Relation Extraction
  - Curated a dataset of table components(column headers, row headers, captions content cells) and hierarchy between column headers
  - Fine-tuned **BERT and LayoutLMv1** for the component identification and hierarchy detection tasks and showed that incorporating layout information in language models helps in tasks where visual structure and layout of textual data are important
- **Ternary Relation extraction in Prospectuses - Answering questions like "Who is the Legal advisor of Amazon for this fund offering?"**
  Focus Areas: Relation Extraction, BERT, Random Forest
  - Curated a **ternary relation extraction** dataset amongst organizations, person names and roles
  - Developed a **BERT** based model leveraging entity markers to enrich entity embeddings sent to the relation head
  - Reduced review time of cases from **2 hours to less than 1 minute** when a new onboarding of a client takes place
- **Reading order detection on Directory pages**
  Focus Areas: Heuristics, Random Forests
  - Curated a dataset depicting the reading order for the Directory page documents
  - Proposed a novel set of features and trained a random forest model for identifying Directory pages(usually a couple of pages in a 300 page long document) in Prospectuses
  - Proposed and implemented parsing of text present in directory pages in a tree-structured format. This enabled downstream Relation extraction tasks to be performed with ease

Analyst - Applied NLP Engineer

Goldman Sachs

Jan 2018 – Dec 2019 Bengaluru, India

Focus Areas: Python, Kafka, HBase, Java

- **Document Processing pipeline for NLP extractors**
  - Designed, built and integrated a document processing pipeline for running various financial extractors on the fly
- **Lex: Firm's strategic Document Storage system**
  - Reduced the transformation/indexing time of documents in our Document Storage system from 2 hours to 1 second. This involved breaking down a monolithic service into multiple scalable microservices

Mentored an undergraduate student for the successful completion and implementation of statistical plugins for Apache SpamAssassin during the **Google Summer of Code** programme, May 2019 - July 2019

Taught Data Structures and Algorithms to a batch of **180 students** as part of the Computer Club classes during my undergrad school

EXPERIENCE: NON-ML

Software Engineer

Arista Networks

Jul 2016 – Dec 2017 Bengaluru, India

- Implemented the Virtual IP support for CVX clusters that gives customers a single point of contact within the cluster. The virtual Ip actively follows the master node of the cluster
  Focus Areas: Virtual IP, Data centers, Networking
- Implemented Directory level replication service that internally uses rsync and inotify Linux utilities to figure out when a particular file is changed and then synchronizes it across the cluster using rsync
  Focus Areas: Operating Systems, Replication, Rsync

ACHIEVEMENTS

KVPY (Kishore Vigyan Protsahan Yojana) scholar (Govt of India)

All India rank of 3873 out of 20 million students who appeared in the JEE-MAINS exam.

Participated and won 4 robotics competitions during my undergrad - Built a Hand Gesture Controlled Bot, Line followers, Maze solvers with shortest path finding, prototype for Autonomous self-control car etc