Yelugam Pranay Kumar

(+413) 2754654 | pranay.yelugam@gmail.com | linkedin.com/pranayyelugam | github.com/pranayyelugam

Research Interests

Natural Language Processing, Common Sense Reasoning, Question Answering, Artificial Intelligence.

Education

University of Massachusetts — Amherst, MA
Master of Science in Computer Science (GPA: 3.96) — *Feb 2021 – May 2022*
Relevant course work: Advanced Natural Language Processing, Advanced Machine Learning, Algorithms for Data Science, Statistics, Applied Information Retrieval, Distributed Systems, and Intelligent Visual Computing

Indian Institute of Information Technology — Allahabad, India
Bachelor of Technology in Computer Science — *Aug 2014 – July 2018*
Relevant course work: Data Structures and Algorithms, Operating Systems, Data Mining, Machine Learning, Artificial Intelligence

Experience

Member of Technical Staff, Eightfold.ai — Santa Clara, CA
Core Infrastructure Team — *Feb 2022 - Present*

- **Data-Lake and Data-Warehouse:** Working on solutions for reliable data transfer in the Data-Lake pipeline.
  - Developed a model to detect outliers in the ingestion of data from Redshift tables to QuickSight spice based data tables.
  - Created Data Refresh Layer to allow for seamless refresh of data from Athena and Redshift based data sources to QuickSight dashboards.
  - Improved the Candidate matching algorithm by caching the profiles and retrieving the Vector representations efficiently.

Research Intern — **Advisors: Prof. Andrew McCallum, Prof. Mohit Iyyer & Xiang Lorraine Li** — Amherst, MA
Information Extraction and Synthesis Laboratory [link] — *May 2021 - Present*

- **Commonsense Frame Completion:** Working on generative Common Sense and its evaluation.
  - Proposed commonsense frame completion (CFC), a new generative task which evaluates common sense via multiple open-ended generations. Created a dataset for this task using AMR representation of commonsense context sentences.
  - We also propose a method of probabilistic evaluation for this task which strongly correlates with human judgements.
  - This evaluation approach aligns answers for a question into clusters and measures the KL divergence between distributions of the answer clusters. We justify this approach by establishing similarity between the human answers.
- **GLM+KG:** Generative Language Models (GLM) and Knowledge Graphs(KG) for commonsense question answering.
  - Created models with sequential of the Language Models and Knowledge Graphs for generative commonsense QA.
  - Used ConceptNet to extract a subgraph for a question-answer pair in the ProtoQA dataset and generated additional data from the subgraph to finetune a generative LM (GPT2).
  - Used Nucleus Sampling as our decoding approach to generate a variety of responses from the finetuned model for a question and evaluated the responses using ProtoQA-Evaluator [link].

Graduate Research Assistant — **Advisors: Prof. Andrew McCallum & Neha Kennard** — Amherst, MA
Information Extraction and Synthesis Laboratory [link] — *Feb 2021 - May 2021*

- **Discourse Structure:** Worked on problems in discourse structure at the document level. [link]
  - Developed and annotated a large dataset of scientific peer review text to highlight discourse structure.
  - Developed classification and span selection models to automatically detect the discourse structure using this dataset.

Software Engineer, Samsung — Delhi, India
Unified Metadata Team — *July 2018 - Dec 2020*

- Implemented a CNN and Longest Common Subsequence based solution to infer content similarity from content providers.
- Implemented *Simhash* for clustering duplicate content from the content description.
- Designed an ETL pipeline that ingests real time TV programs and schedules from different providers to Samsung TV.

Publications

- Probabilistic evaluation of a novel Generative Common Sense Question Answering task — in review at a NLP conference.
- Neha Kennard, Tim O'Gorman, Rajarshi Das, **Pranay Kumar Yelugam**, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Hamed Zamani, and Andrew McCallum. DISAPERE: A Dataset for Discourse Structure in Peer Review Discussions. Annual Conference of the North American Chapter of the Association for Computational Linguistics (**NAACL**), 2022.
- Bharadwaju, **Pranay Kumar Yelugam**, K. Anudeep, A. Vamshi Krishna, Bakshi Rohit Prasad, Sonali Agarwal "Real time mining of ego networks for exploring social associations". **CICT**, 2017

Skills

- **Technical Skills**: PyTorch, sckit-learn, pandas, numpy, NLTK, matplotlib, C/C++, Java, Scala, JavaScript
- **Industry Knowledge**: NLP, Machine Learning, Natural Language Processing, Deep Learning, AI