Dhruv Kumar

ddhruvkr@gmail.com | +1 226-978-3751 | Google Scholar| Website | LinkedIn | GitHub

EXPERIENCE

Grammarly, Vancouver, Canada

Senior Applied Research Scientist **Jul 2021 - Present**

- Working on RAG and integrating tools into LLMs such as calendar, internal documents, email APIs, etc
- Led a team of five, including linguists and ML engineers, to build the core text-generation model behind Voice Composer, a voice-to-text application, currently being used by millions of users. Improved the model on multiple aspects (such as robustness, fluency, coverage, naturalness, etc) by over 10% points.
- Led a team to build a paraphrasing model and service that is being used by over 100k users per week.
- Developed and open-sourced CoEdIT - a series of instruction-tuned models for text revision and editing ranging from 770M to 11B parameters. The models perform competitively to the biggest LLMs available and have been downloaded over 100k times on HuggingFace. Extended the work to **multilingual models**.
- Mentored (and co-mentored) four research interns on:
  - Iterative text editing
  - LLMs for Self-contradictions in long documents
  - Personalization in LLMs
- Published 8+ papers and a couple of patents (Google Scholar).
- Taken over 75 interviews for full-time and intern ML researcher/engineer positions.

Borealis AI, Toronto, Canada

Machine Learning Research Intern **Sep 2020 - Dec 2020**

Turing - text-to-SQL semantic parsing

- Showed how explicit schema linking and regularization techniques can improve cross-domain generalizability and performance of a state-of-the-art semantic parser.
- Showed the efficacy of data-dependent initialization for transformers in improving generalization and for training deeper models on small datasets. [Code]

University of Waterloo, Waterloo, Canada

Graduate Research Assistant **Sep 2018 - Aug 2020**

GEM Benchmark

- Worked with Prof. Wei Xu, Sebastian Gehrmann, Mounica Maddela, and Prof. Ondrej Dusek to build an evaluation framework for the Generation, evaluation and metrics (GEM) workshops. [Code]

Unsupervised Sentence Simplification

- Designed an edit-based algorithm for unsupervised sentence simplification. The model is controllable and interpretable and achieves SARI scores competitive with those of supervised models. [Code]
- Extended the above algorithm to include a sequence-to-sequence generative model for more complex edits, thus combining the advantages of edit-based and generative approaches. [Code]

Music-conditioned lyrics generation

- Designed bimodal (text and audio) neural network models based on variational autoencoders (VAE) to generate lyric lines for instrumental pieces of music.
- Extended the above approach to align the learned latent spaces of audio and text representations using generative adversarial networks (GAN) and conditional variational autoencoders (CVAE).

Consentio: Managing Consent to Data Access using Permissioned Blockchains

- Designed a consent management system based on permissioned blockchains that can handle up to 6000 transactions per second. [Code]

Arcesium (DE Shaw Group), Hyderabad, India

Software Engineer, Fund and Investor Accounting **Jul 2016 - May 2018**

- Worked as a full-stack engineer to enhance and maintain the post-trade automation platform (written in Java) for funds operated by J.P. Morgan and D.E. Shaw.

Citigroup, Pune, India

Software Engineering Intern, Equities **May 2015 - Jul 2015**

- Designed and implemented the first prototype (MEAN stack) of the Trading Controls application. Received full-time offer.

EDUCATION

University of Waterloo, School of Computer Science, Waterloo, Canada **Sep 2018 - Aug 2020**
Master of Mathematics (Thesis), Computer Science
Thesis: Iterative Edit-based Unsupervised Sentence Simplification

Indian Institute of Information Technology, Allahabad, India **Jul 2012 - Jun 2016**
Bachelor of Technology (Hons.), Information Technology
Thesis: Compressed Knowledge transfer via Factorization Models in Recommender Systems

- Proposed a Joint Matrix Factorization algorithm for Music Recommendation that utilizes geographical and time-based tagging information of artists, in addition to implicit user feedback (user clicks). [Code]
- Developed an algorithm to incorporate metadata in Factorization Machines, lowering RMSE value to 0.836 as compared to 0.853 when using a Joint Matrix Factorization method on the MovieLens 1M dataset. [Code]

TECHNICAL SKILLS

- **Interests**: Natural Language Processing, Machine Learning
- **Programming Languages and Frameworks**: Python, Java, SQL, C/C++, Pytorch, HuggingFace, Scikit, NLTK

ACADEMIC SERVICE

- Reviewer: 1) ARR (ACL rolling reviews), 2) GEM: Natural Language Generation, Evaluation, and Metrics workshop, 3) In2Writing: Workshop on Intelligent and Interactive Writing Assistants, 4) AISG (AI for Social Good) Workshop

SELECTED PUBLICATIONS

- **mEdIT, Multilingual Text Editing via Instruction Tuning** — *Under Review* — V Raheja, D Alikaniotis, V Kulkarni, B Alhafni, **D Kumar**
- **ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models** — *Under Review* — J Li, V Raheja, **D Kumar**
- **Personalized Text Generation with Fine-Grained Linguistic Control** — *PERSONALIZE@EACL 2024* — B Alhafni, V Kulkarni, **D Kumar**, V Raheja
- **Speakerly™: A Voice-based Writing Assistant for Text Composition** — *EMNLP Industry Track 2023* — **D Kumar***, V Raheja*, A Kaiser-Schatzlein, R Perry, A Joshi, J Hugues-Nuger, S Lou, N Chowdhury
- **CoEdIT: Text Editing by Task-Specific Instruction Tuning** — *EMNLP 2023* — V Raheja, **D Kumar,** R Koo, D Kang
- **Improving iterative text revision by learning where to edit from other revision tasks** — *EMNLP 2022* — ZM Kim, W Du, V Raheja, **D Kumar**, D Kang
- **Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision** — *In2Writing@ACL 2022* **(Best Paper Award)** — W Du, ZM Kim, V Raheja, **D Kumar**, D Kang
- **Understanding Iterative Revision from Human-Written Text** — *ACL 2022* — W Du, V Raheja, **D Kumar**, ZM Kim, M Lopez, D Kang
- **GRS: Combining Generation and Revision in Unsupervised Sentence Simplification** — *ACL 2022* — M Dehghan, **D Kumar**, L Golab
- **LyricJam: A system for generating lyrics for live instrumental music** — *ICCC 2021* — O Vechtomova, G Sahu, **D Kumar**
- **Optimizing Deeper Transformers on Small Datasets** — *ACL 2021* — P. Xu, **D. Kumar**, W. Yang, W. Zi, K. Tang, C. Huang, J.C.K. Cheung, S.J.D. Prince, Y. Cao
- **The gem benchmark: Natural language generation, its evaluation and metrics** — *ArXiv 2021* — S. Gehrmann et al.
- **Generation of lyrics lines conditioned on music audio clips** — *NLP4MusA@ISMIR 2020* — O Vechtomova, G Sahu, **D Kumar**
- **Iterative edit-based unsupervised sentence simplification** — *ACL 2020* — **D. Kumar**, L. Mou, L. Golab, O. Vechtomova
- **Consentio: Managing consent to data access using permissioned blockchains** — *ICBC 2020* — R Agarwal*, **D Kumar***, L Golab, S Keshav