Nandan Thakur

thakur-nandan.github.io

Email: nandan.thakur@uwaterloo.ca

[Google Scholar] [Twitter] [GitHub] [LinkedIn]

EDUCATION

Univesity of Waterloo, Waterloo, ON, Canada — September 2021- Present
Ph.D. Student, David R. Cheriton of Computer Science
Advisor: Prof. Jimmy Lin

Birla Institute of Technology & Science, Pilani, Goa, India — August 2014 - May 2018
B.E. (Hons.) in Electronics & Instrumentation, Minor in Finance

PUBLICATIONS

[1] MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages.
Xinyu Zhang*, **Nandan Thakur***, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, Jimmy Lin. (* denotes equal contribution)
Transactions of the Association for Computational Linguistics (TACL), 2023.

[2] SPRINT: A Unified Toolkit for Evaluating and Demystifying Zero-shot Neural Sparse Retrieval.
Nandan Thakur, Kexin Wang, Iryna Gurevych, Jimmy Lin.
SIGIR 2023 - Resource Track.

[3] Injecting Domain Adaptation with Learning-to-hash for Effective and Efficient Zero-shot Dense Retrieval.
Nandan Thakur, Nils Reimers, Jimmy Lin.
ReNeuIR 2023 Oral Presentation.

[4] Evaluating Embedding APIs for Information Retrieval.
Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, **Nandan Thakur**, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Jimmy Lin.
ACL 2023 - Industry Track.

[5] Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages.
Xinyu Zhang*, **Nandan Thakur***, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, Jimmy Lin. (* denotes equal contribution)
WSDM Cup Competition, 2023.

[6] GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval.
Kexin Wang, **Nandan Thakur**, Nils Reimers, Iryna Gurevych.
NAACL-HLT 2022.

[7] BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.
Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych.
NeurIPS 2021 - Datasets and Benchmark Track.

[8] Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks.
Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych.
NAACL-HLT 2021.

PREPRINTS

[9] Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval.
Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Frederick Wieting, Jimmy Lin, Daniel Cer
Arxiv Preprint, 2023.

[10] HAGRID: A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution.
Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, **Nandan Thakur**, Jimmy Lin.
Arxiv Preprint, 2023.

[11] Resources for Brewing BEIR: Reproducible Reference Models and an Official Leaderboard.
Ehsan Kamalloo, **Nandan Thakur**, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, Jimmy Lin.
Arxiv Preprint, 2023.

[12] Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval.
Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, **Nandan Thakur**, Jheng-Hong Yang, Xinyu Zhang
Arxiv Preprint, 2023.

Research Experience

Univesity of Waterloo Sep 2021 - Present, Waterloo, ON, Canada
Ph.D. Student (Supervisor: Prof. Jimmy Lin)
Worked on multilingual information retrieval [1] [12] [9], data and model efficiency [3] and reproducibility [2] [11]. Recently focusing on retrieval-augmented generation with LMs [10] [4].

Google Research Sep 2022 - May 2023, California, USA
Student Researcher (Mentors: Daniel Cer, Jianmo Ni)
Worked on improving existing multilingual retrieval systems using PaLM 2 generated synthetic data, without expensive human-labeled training data for a wide variety of languages [9].

UKP Lab, Technical University of Darmstadt Nov 2019 - Aug 2021, Darmstadt, Germany
Research Assistant (Supervisors: Prof. Iryna Gurevych, Nils Reimers)
Worked on developing a benchmark to evaluate zero-shot out-of-domain (OOD) evaluation of retrieval systems [7] and worked on data-augmentation techniques to generate synthetic data for domain adaptation in pairwise sentence [8] and retrieval tasks [6].

(EMBL) European Molecular Biology Laboratory Jun - Aug 2018, Heidelberg, Germany
Research Trainee (Supervisors: Prof. Toby Gibson, Dr. Manjeet Kumar)
Worked on developing a prediction toolkit using machine learning to computationally predict kinase-substrate phosphorylation sites within (CAMK) protein sequences.

Industry Experience

KNOLSKAPE Sep 2018 - Oct 2019, Bengaluru, India
Data Scientist (Manager: Chaithanya Yambari)
Worked on developing Krawler.ai, an enterprise product for effectively searching a company's large messy content libraries (pdf, xlsx, docx, etc.) with multimodal search. Implemented search functionality using Elasticsearch and backend data ingestion using Flask, Apache-Airflow and MongoDB.

Belong.co Jul - Dec 2017, Bengaluru, India
Data Science Intern (Manager: Vinodh K. Ravindranath)
Worked on topic modeling for clustering millions of candidate resumes. Extracted keywords using FlashText and automatically clustered candidates using GuidedLDA, a semi-supervised LDA algorithm.

Honor and Awards

University of Waterloo (UW) Graduate Scholarship 2021 - Present
BEIR benchmark included in CS224U teaching material at Stanford University. 2021
Created both the ELLIS NLP 2021 and SustaiNLP 2021 workshop websites. 2021
Got Selected as a speaker for PyCon Italia in 2020 (Cancelled due to Covid-19) 2020
Finalists in Technology Premier League (TPL) held by CIO & Leader, IT Next. 2019
Received a **fully-funded Machine Learning (ML) Fellowship** in EMBL Heidelberg 2018

Teaching Experience

Teaching Assistant, University of Waterloo — 2021 - Present

- CS 135 (Designing Functional Programs) - Fall 2021
- CS 136 (Elementary Algorithm Design and Data Abstraction) - Winter 2022, Spring 2023
- CS 241 (Foundations of Sequential Programs) - Spring 2022
- CS 479/679 (Introduction to Artificial Intelligence) - Winter 2023
- CS 370 (Numerical Computation) - Fall 2023

Services

Competition Lead Organizer: WSDM Cup 2023.
Shared-task Lead Organizer: TREC RAG 2024 (Upcoming)
Reviewer (*CL/NLP conferences): ACL Rolling Review: Oct-Nov (2021), Jan-Apr (2022)
Reviewer (ML conferences): NeurIPS 2023
Reviewer (IR conferences): SIGIR 2023, ECIR 2024.

Invited Talks

Koç University: Advanced Information Retrieval (Tutorial) — Virtual, June 2023
Stanford University: Heterogenous Benchmarking in IR Research — USA, November 2022
OpenNLP Meetup: BEIR, An Open-Source Benchmark for IR Systems — Virtual, June 2021

Coursework

University of Waterloo: CS 680: Introduction to Machine Learning (Ongoing), CS 889: Data Sources for Emerging Tech, CS 886: Graph Neural Networks, CS 886: Robustness of Machine Learning, CS 679: Neural Networks, CS 848: Information Retrieval, CS 649: Human-Computer Interaction, CS 854: Experimental Performance Evaluation. 2021-Present

BITS Pilani: Machine Learning, Neural Networks & Fuzzy Logic, Data Structures & Algorithms, Probability & Statistics, Linear Algebra, Econometric Methods, Discrete Mathematics. 2014-2018

Press and Media

Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages, WSDM Cup Competition 2023 — February 2023
Domain Adaptation with Generative Pseudo-Labeling (GPL) (Pinecone.ai) — August 2022
Extending Neural Retrieval Models to New Domains and Languages (Transformers-at-Work, Zeta Alpha) — December 2021
BEIR benchmark as a helpful ML library (ML News by Yannic Kilcher) — August 2021
Making the Most of Data: Augmentation with BERT (Pinecone.ai) — March 2021
Advance BERT model via transferring knowledge from Cross-Encoders to Bi-Encoders (Towards Data Science) — May 2020

Competences

Languages Bengali (*native*), English (*fluent*, TOEFL 110), Hindi (*fluent*), German (*elementary*, A2)
Programming Python, JavaScript, ReactJS, R, C++, HTML, CSS, Excel, MATLAB, Racket, LaTeX.
Libraries and Services Pytorch, JAX, Tensorflow, Flask, Django, SQL, MongoDB, Docker, Elasticsearch, Redis, RabbitMq, Apache-Airflow, Postman.

Co-Curricular

Mime Club Coordinator, BITS Pilani — 2014-2018
Led a team of 30 student performers in one of the most popular clubs in college. Involved in acting, sound mixing, designing slides and creating stories for more than 10 shows over a span of 4 years.