Shubham Toshniwal

Senior Research Scientist, NVIDIA

shtoshni@gmail.com shtoshni shtoshni.github.io Google Scholar

Education

Toyota Technological Institute at Chicago (TTIC)
Ph.D., Computer Science, 2017–2022
Title: Efficient and Interpretable Neural Models for Entity Tracking
Advisors: Kevin Gimpel, Karen Livescu

M.S., Computer Science, 2015–2017
Advisor: Karen Livescu

Indian Institute of Technology Kanpur (IITK)
B.Tech., Computer Science, 2009–2013

Industry

NVIDIA, New York — Senior Research Scientist, Sep 2023–Current
Math Reasoning with Large Language Models

Fundamental AI Research (FAIR), Meta AI, New York — Research Scientist, Jan 2022–Aug 2023
Reasoning with Large Language Models

Google Research, San Francisco — Software Engineering Intern, June–Sep 2018
Image Grounded Language Representation Learning — *Daniel Gillick, Alessandro Presta*

Google Research, New York — Software Engineering Intern, June–Sep 2017
Multilingual Speech Recognition — *Tara Sainath, Ron Weiss*

IBM Research, New Delhi — Software Engineer, 2013–2015
Dialoging with Watson — *Jitendra Ajmera, Sachindra Joshi*

Preprints

OpenMathInstruct-2: Accelerating AI for Math with Massive Open-Source Instruction Data
Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, Igor Gitman

Nemotron-4 340B Technical Report
NVIDIA

Code Pretraining Improves Entity Tracking Abilities of Language Models
Najoung Kim, Sebastian Schuster, *Shubham Toshniwal*

Publications

OpenMathInstruct-1: A 1.8 Million Math Instruction Tuning Dataset
Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, Igor Gitman
NeurIPS Datasets and Benchmarks Track (*Oral*)

Major Entity Identification: A Generalizable Alternative to Coreference Resolution
Kawshik Manikantan, *Shubham Toshniwal*, Makarand Tapaswi, Vineet Gandhi
EMNLP 2024

Learning to Reason and Memorize with Self-Notes
Jack Lanchantin*, *Shubham Toshniwal**, Jason Weston, Arthur Szlam, Sainbayar Sukhbaatar
NeurIPS 2023

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models
Srivastava et al.
TMLR 2023

Adapting Pretrained Text-to-Text Models for Long Text Sequences
Wenhan Xiong, Anchit Gupta, *Shubham Toshniwal*, Yashar Mehdad, Wen-tau Yih
Findings of EMNLP 2023

Baked-in State Probing
Shubham Toshniwal, Sam Wiseman, Karen Livescu, Kevin Gimpel
Findings of EMNLP (short) 2022

Chess as a Testbed for Language Model State Tracking
Shubham Toshniwal, Sam Wiseman, Karen Livescu, Kevin Gimpel
AAAI 2022

On Generalization in Coreference Resolution
Shubham Toshniwal, Sam Wiseman, Karen Livescu, Kevin Gimpel
CRAC@EMNLP 2021 (*Best Short Paper*)

Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks
Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, Kevin Gimpel
EMNLP 2020 (short)

PeTra: A Sparsely Supervised Memory Model for People Tracking
Shubham Toshniwal, Allyson Ettinger, Kevin Gimpel, Karen Livescu
ACL 2020

A Cross-Task Analysis of Text Span Representations
Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, Kevin Gimpel
RepL4NLP 2020

Pre-trained Text Embeddings for Enhanced Text-to-Speech Synthesis
Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, *Shubham Toshniwal*, Karen Livescu
Interspeech 2019

Parsing Speech: A Neural Approach to Integrating Lexical and Acoustic-Prosodic Information
Trang Tran*, *Shubham Toshniwal**, Mohit Bansal, Kevin Gimpel, Karen Livescu, Mari Ostendorf
NAACL HLT 2018 *(Oral)*

A Comparison of Techniques for Language Model Integration in Encoder-Decoder Speech Recognition
Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, Karen Livescu
SLT 2018

Multilingual Speech Recognition With A Single End-To-End Model
Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, Kanishka Rao
ICASSP 2018 *(Oral)*

Multitask Learning with Low-Level Auxiliary Tasks for Encoder-Decoder Based Speech Recognition
Shubham Toshniwal, Hao Tang, Liang Lu, Karen Livescu
Interspeech 2017 *(Oral)*

Jointly Learning to Align and Convert Graphemes to Phonemes with Neural Attention Models
Shubham Toshniwal, Karen Livescu
SLT 2016

Patents

System and method for cognitive filtering of audio in noisy environments
Jitendra Ajmera, Nitendra Rajput, Saurabh Srivastava, *Shubham Toshniwal*
US Patent No. 10,187,738 B2, issued January 22, 2019

Generating natural language dialog using a questions corpus
Jitendra Ajmera, Ajay K. Gupta, Sachindra Joshi, *Shubham Toshniwal*
US Patent No. 10,049,152 B2, issued August 14, 2018

Visual Information Processing Allocation between a Mobile Device and a Network
Anirban Majumder, Samik Datta, Sharad Jaiswal, Nisheeth Shrivastava, Sreedal Menon, *Shubham Toshniwal*
US Patent No. 8,913,838 B2, issued December 16, 2014

Awards

Best short paper at CRAC@EMNLP 2021

Developed *Usher*, an intelligent museum guide mobile application, that won several internal IBM awards

All India Rank 13 in IIT-JEE 2009 among 400,000 candidates

All India Rank 1 in UPTU-SEE 2009 among 300,000 candidates

Media

Work on multilingual speech models has been covered in several Google AI blogs [1, 2]

Featured in the first episode of PyTorch Lightning Community Spotlight to discuss work on chess LMs

Work on Usher, an intelligent museum guide application, featured in **MIT Technology Review**

Invited Talks

Microsoft India, 2023

IBM Research India, 2023

TTIC 20th Anniversary Workshop, Chicago, 2023

Services

Reviewer: ICLR 2025, EMNLP 2024, ACL 2024, TACL 2024, EMNLP 2023, ACL 2023, MemARI@NeurIPS 2022, EMNLP 2022, SLT 2022, ICLR 2021, SLT 2020, RepL4NLP 2020, ICLR 2019, NeurIPS 2018, EMNLP 2018, RepL4NLP 2018, CoNLL 2017

Co-organizer of Speech & Language Reading group at TTIC (SLATTIC) from 2018-2021

Co-organizer of Student Workshop 2019 at TTIC

Teaching

Teaching Assistant, Fall 2017
TTIC 31020, Introduction to Statistical Machine Learning

Coursework

Speech Technologies, Natural Language Processing, Probabilistic Graphical Models, Statistical Machine Learning, Advanced Natural Language Processing, Unsupervised Learning and Data Analysis, Dynamical Systems with Applications