

Sumanth Doddapaneni

PhD Student, IIT Madras

[Website](#) [Github](#) [Google Scholar](#) [Email](#)

Education

Jan 2022 Ongoing	Indian Institute of Technology (IIT), Madras Ph.D, Computer Science & Engineering Advisor: Mitesh M. Khapra	Chennai, India
Aug 2016 May 2020	Indian Institute on Information Technology (IIIT), Sri City B.Tech., Electronics & Communications Engineering	Sri City, India

Experience

Oct 2021 Present	AI4Bharat, IIT Madras PhD Researcher Advisors: <i>Dr. Mitesh M. Khapra, Dr. Anoop Kunchukuttan, Dr. Pratyush Kumar</i> Working on Multilingual Language modeling and Machine Translation, with a focus on low-resource Indian languages	Chennai, India
Nov 2022 Mar 2023	Mila - Quebec AI Institute Visiting Researcher Host: <i>Dr. Rahul Aralikkatte, Dr. Jackie Chi Kit Cheung</i> Exploring pretraining methods to develop better multilingual summarization models.	Remote
Oct 2020 Sep 2021	Robert Bosch Centre for Data Science and AI, IIT Madras Post Baccalaureate Fellow Advisors: <i>Dr. Mitesh M. Khapra, Dr. Anoop Kunchukuttan, Dr. Pratyush Kumar</i> Built SOTA models for Machine Translation (IndicTrans) and Automatic Speech Recognition (IndicWav2Vec) for Indian Languages	Chennai, India

Select Publications

P=Preprints, C=Conference, W=Workshop, J=Journal, *=Equal Contribution

- [J.1] **Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages** [\[Code\]](#)
Gowtham Ramesh*, Sumanth Doddapaneni*, et. al
Transactions of the Association for Computational Linguistics [TACL 2022]
- [J.2] **A Survey in Adversarial Defences and Robustness in NLP** [\[Code\]](#)
Shreya Goyal, Sumanth Doddapaneni, Mitesh M Khapra, Balaraman Ravindran
ACM Computing Surveys [CSUR 2023]
- [C.1] **Towards Building ASR Systems For The Next Billion Users** [\[Code\]](#)
Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra
36th AAAI Conference on Artificial Intelligence [AAAI 2022]
- [C.2] **Effectiveness of Mining Audio and Text Pairs from Public Data for Improving ASR Systems for Low-Resource Languages** [\[Code\]](#)
Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra
48th IEEE International Conference on Acoustics, Speech, and Signal Processing [ICASSP 2023]
- [P.1] **A Primer on Pretrained Multilingual Language Models** [\[Code\]](#)
Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, Pratyush Kumar
Pre-Print [ArXiv 2021]
- [P.2] **IndicXTREME: A Multi-Task Benchmark For Evaluating Indic Languages** [\[Code\]](#)
Sumanth Doddapaneni, Rahul Aralikkatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, Pratyush Kumar
Pre-Print [ArXiv 2022]
- [P.3] **Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages** [\[Code\]](#)
Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy V, Anoop Kunchukuttan
Pre-Print [ArXiv 2022]

Select Research Projects

Multilingual Summarisation

Nov'22 - Present

Advisors: [Dr. Rahul Aralrikatte](#), [Dr. Jackie Chi Kit Cheung](#)

- > Studying the effect of pretraining in multilingual summarisation
- > Pretraining and fine-tuning generative models for abstractive summarisation

Multilingual Language Modeling

Oct'21 - Present

Advisors: [Dr. Mitesh M. Khapra](#), [Dr. Anoop Kunchukuttan](#), [Dr. Pratyush Kumar](#) [[🔗](#)][Code]

- > Building Large scale monolingual corpora and Language Models for Indian Languages
- > Building strong testsets to evaluate the zero-shot generalisation of the models
- > Studying the various objectives and data regimes that improve zero-shot generalisation of the models
- > Understanding the efficacy of adapters in zero-shot generalisation of the LMs

Speech Recognition

Jun'21 - Feb'22

Advisors: [Dr. Mitesh M. Khapra](#), [Dr. Anoop Kunchukuttan](#), [Dr. Pratyush Kumar](#) [[Try the Model](#)][Code]

- > Building state-of-the-art models for Automatic Speech Recognition for Indian languages
- > Understanding the various effects of LM on downstream ASR task
- > Accepted as a long paper at AAAI 2022

Machine Translation

Feb'21 - Oct'21

Advisors: [Dr. Mitesh M. Khapra](#), [Dr. Anoop Kunchukuttan](#), [Dr. Pratyush Kumar](#) [[Try the model](#)][Code]

- > Built state of the art translation model for Indian languages to English and *vice versa*
- > Created the largest bilingual corpora (~50M pairs) across 11 languages for NMT training.
- > Work published in TACL (Volume 10, 2022)

Academic Service

Program Committee MRL @EMNLP'21

Volunteer EACL'21, ICML'21, NeurIPS'21, EMNLP'21, ACL'22, EMNLP'22

Reviewer ACL'23; ACL Rolling Review 2022, 2023

Honours and Grants

Google Research Selected to attend the Google Research Week 2023

Naver Labs and Univ. Grenoble Alpes Selected to attend the ALPS Winter School 2022

Robert Bosch Centre Received the Post Baccalaureate Fellowship to work on interdisciplinary AI

TFRC Grant Received TPU Research credits to carry out work on LMs

MSR Travel Grant Received Microsoft Research Travel grant to attend ACL 2022

Talks & Volunteering Roles

Invited Talk, Google Research India *Speaker*

Feb 2023

- > Talk on "Building Natural Language Understanding (NLU) capabilities for Indic languages"

Invited Talk, Swiggy Data Science *Speaker*

Jan 2022

- > Talk on "Towards Building ASR Systems for the Next Billion Users"

Organizer at NLP with Friends *Organizer*

Jan 2023 - Present

NLP Reading Group, AI4Bharat *Organizer*

Jan 2022 - Dec 2022

References

- > Dr. Mitesh M. Khapra Associate Professor, IIT Madras, India [[🔗](#)]
- > Dr. Anoop Kunchukuttan Senior Applied Researcher, Microsoft AI and Research, India [[🔗](#)]
- > Dr. Pratyush Kumar Senior Researcher, Microsoft Research, India [[🔗](#)]
- > Dr. Raj Dabre Researcher, NICT, Japan [[🔗](#)]
- > Dr. Rahul Aralrikatte Postdoctoral Fellow, MILA & Visiting Researcher, Google [[🔗](#)]