Anoop Kunchukuttan

Microsoft India (R&D) Pvt. Ltd.
Microsoft Campus
Gachibowli
Hyderabad - 500032
Telangana, India

Date of Birth: November 21, 1982
Nationality: Indian
Phone: +91 9860999552
email: anoop.kunchukuttan@gmail.com
URL: http://anoopk.in

Research Interests

I am broadly interested in Natural Language Processing and Machine Learning. Multilingual NLP and self-supervised language learning are my major research interests with a goal to investigate techniques for making quality NLP solutions available to multiple languages economically and at scale. This is important to make available the benefits of the vast amount of knowledge to large sections of the population. In this context, I am interested in machine translation/ transliteration, joint multilingual learning of LLM/NLP models, cross-lingual NLP tasks, multilingual distributed representations, language typology and code mixing. I am very interested in looking at these problems in the context of related languages, particulary Indian languages; and building open-source software for Indian language NLP. To investigate these problems, I am generally interested in LLMs, sequence-to-sequence learning, multi-task learning, subword level models in NLP and unsupervised learning in NLP. I am also interested in exploring word and sentence representations as well as techniques for mining datasets at scale from large-scale corpora.

Areas of Expertise

NLP: Multilingual Learning, Self-supervised Learning, LLMs, Machine Translation, Machine Transliteration, NLP for Related languages, NLP for Indian languages, Distributed Representations for NLP, Large-scale Mining of NLP datasets.

Machine Learning: Supervised Classification, Unsupervised learning, Imbalanced dataset classification, optimizing performance metrics, prediction, sequence labelling, word alignment, string transduction and translation, sequence to sequence learning.

Education

- Ph.D in Computer Science & Engineering, *IIT Bombay*. 2012-2018.
Advisor: Prof. Pushpak Bhattacharyya.
Thesis: Machine Translation & Transliteration involving Related, Low-resource Languages

- M.Tech in Computer Science & Engineering, *IIT Bombay*. 2006-2008. CPI: 9.21
Advisor: Prof. Om Damani.
Thesis: Compound Noun Multiword Expression Extraction

- B.E in Computer Engineering, *University of Pune*. 2000-2004. % Marks: 65.53

Work Experience

- Prinicipal Applied Researcher, Microsoft India (Machine Translation group), Sep 2023 to *present*.
- Co-founder, Co-lead, AI4Bharat, July 2019 to *present*.
- Area Chair, ACL Rolling Review, Dec 2023 to *present*.
- Senior Applied Researcher, Microsoft India (Machine Translation group), Feb 2018 to Aug 2023.
- Adjunct Faculty, Department of Computer Science, IIT Madras, Aug 2022 to July 2024.
- Teaching Assistant (Project), Center for Indian Language Technology, IIT Bombay, responsible for research and mentoring teams in collaborative projects with Xerox Research, Crimson Interactive, Elsevier Publishing. Jan 2012-Dec 2017.

- Research Intern, Xerox Research Centre Europe, Jul-Oct 2012. *Mentors*: Dr. Nicola Cancedda & Dr. Sriram Venkatapathy.
- Research Engineer, Center for Indian Language Technology, IIT Bombay. Mar-Dec 2011.
- Team Lead, Persistent Systems, Jan 2011-Mar 2011.
- Module Lead, Persistent Systems, Sep 2008-Dec 2010.
- Member of Technical Staff, Persistent Systems, Jul 2004-Aug 2008.

Awards & Honours

- Outstanding Paper Award at ACL 2024 conference for IndicLLMSuite.
- Area Chair Award at ACL 2024 conference for RomanSetu.
- NASSCOM AI Gamechangers Award 2022 for IndicWav2Vec.
- NASSCOM AI Gamechangers Award 2021 for Samanantar.
- Best Thesis Talk on *Investigations into subword units for Statistical Machine Translation between related languages* at Research and Innovation Symposium in Computing, IIT Bombay. 2017.
- Outstanding Paper Award. Workshop on Subword and Character level models in NLP. 2017.

Invited Talks & Teaching

1. **Tutorial** on *Data and Model Centric Approaches for Expansion of Large Language Models to New languages* at **EMNLP**, Suzhou, China, in December 2025 *(upcoming)* with Raj Dabre, Rudramurthy V, Mohammed Safi Ur Rahman & Thanmay Jayakumar.
2. **Tutorial** on *Extending English Large Language Models to New Languages* at summer schools in **IIIT Hyderabad** (July 2024), **IIIT Delhi** (May 2024) and course lecture (**IIT Hyderabad**, March 2024). Upcoming **tutorial at EMNLP 2025**.
3. **Tutorial** on *Multilingual Neural Machine Translation* at **COLING**, Barcelona, Spain, in September 2020 with Raj Dabre & Chenhui Chu.
4. **Tutorial** on *Statistical Machine Translation between related languages* at **North American Chapter of the Association for Computational Linguistics (NAACL)**, San Diego, United States, in June 2016 with Prof. Pushpak Bhattacharyya & Mitesh Khapra.
5. **Tutorial** on *Multilingual Learning* at the Summer School on Machine Learning: Advances in Modern AI, IIIT Hyderabad, in July 2018.
6. **Tutorial** on *the AI4Bharat Initiative* at the ICON 2021, in Dec 2021 with Mitesh Khapra and Pratyush Kumar.
7. **Tutorial** on *Machine Learning for Machine Translation* at International Conference on Natural Language Processing, Delhi, in December 2013 with Prof. Pushpak Bhattacharyya, Piyush Dungarwal and Shubham Gautam.
8. **Tutorial** on *Translation and Transliteration between related languages* at International Conference on Natural Language Processing, Trivandrum, in December 2015 with Mitesh Khapra.
9. **Tutorial** on *Machine Translation* at the ACM India Summer School on Natural Language Processing, IIIT Hyderabad, in July 2021.
10. **Tutorial** on *Machine Translation* at the IIIT-H Advanced Summer School on Natural Language Processing, IIIT Hyderabad, in June 2018.
11. **Keynote Talk** on *NLP for Indian Languages: A Language Relatedness Perspective Keynote Talk* at 5th WILDRE workshop (under LREC 2020) in May 2020.
12. **Invited Talk** on *Mining Datasets at scale for Building High-quality NLP Models* at IIT Hyderabad. January 2023.

13. **Invited Talk** on *Indic NLP: A Multilinguality and Language Relatedness Perspective* at Vaibhav Summit (Organized by MyGov, Govt. of India). October 2020.

14. **Invited Talk** on *NLP for Indian Languages: A Language Relatedness Perspective* at NASSCOM Data Science & AI - Center of Excellence, Bengaluru in August 2019.

15. **Mini Course** on *Natural Language Processing - A Deep Learning Approach* at IIT Alumni Center Bengaluru AI Deep Dive Workshop, July 2019.

16. **Course Lecture** on *RNN, LSTM, Language Modeling and Sequence to Sequence Modeling* at IIT Bombay for course on Deep Learning for NLP. 2021, 2022.

17. **Course Lecture** on *Machine Translation* at IIT Hyderabad (2020), IIIT Lucknow (2022), BITS Pilani-Hyd (2019) for course on NLP.

18. **Course Lecture** on *Introduction to Neural MT* at CEP Worksop on course on Deep Learning for Natural Language Processing at IIT Patna, Jan 2020.

19. **Course Lecture** on *Understanding the Indian Languages: Challenges & Opportunities* at Atal FDP on AI in NLP at KIIT University, Bhubaneshwar, Oct 2020.

20. **Course Lecture** on *Detection of Controversies, Polarization and Fake Information* at IIM Visakhapatnam in July 2018.

21. **Invited Talk** on *Multilingual Learning and Mining Datasets for Building High-quality NLP Models* at IISER Bhopal in Apr 2022.

22. Tech Talk on *Machine Translation for related languages* at AXLE 2018 (Microsoft Academic Accelerator), Microsoft IDC Hyderabad, in May 2018.

23. Talk on *Investigations into subword units for Statistical Machine Translation between related languages* at Research and Innovation Symposium in Computing, IIT Bombay, in April 2017. *(Best Thesis Talk Award)*

24. Invited Poster on *Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent* at CODS-COMAD 2018, in Goa, January 2018.

25. Talk on *Orthographic Syllable as basic unit for SMT between Related Languages* at Inter-Research-Institute Student Seminar in Computer Science, ACM India, in Kolkata, January 2017.

26. Introductory Talks on *Introduction to Machine Translation & Transliteration*

    - Faculty Development Programme, Dharamsihn Desai Institute of Technology, Nadiad, in June 2018.
    - Machine Learning Summer School, Vidyalankar Institute of Technology, Mumbai, in June 2017.
    - Viva College of Engineering, Mumbai, in June 2016.
    - Cummins College of Engineering, Pune, in August 2015.

Publications

Preprints

1. Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, Ratish Puduppully. *RomanLens: The Role Of Latent Romanization In Multilinguality In LLMs.* ArXiv preprint arXiv:2502.07424. 2025. *(under review)*

2. Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop Kunchukuttan, Mitesh M Khapra. *Cross-Lingual Auto Evaluation for Assessing Multilingual LLMs.* ArXiv preprint arXiv:2410.13394. 2024. *(under review)*

3. Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, Raj Dabre. *BhasaAnuvaad: A Speech Translation Dataset for 13 Indian Languages.* ArXiv preprint arXiv:2411.04699. 2024. *(under review)*

4. Sanjay Suryanarayanan, Haiyue Song, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, Raj Dabre. *Pralekha: An Indic Document Alignment Evaluation Benchmark.* ArXiv preprint arXiv:2401.2411.19096. 2024. *(under review)*

Books

1. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Machine Translation and Transliteration involving Related, Low-resource Languages.* CRC Press. 2021.

Journals

1. Jay Gala, Pranjal A Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, Anoop Kunchukuttan. *IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages.* Transactions on Machine Learning Research **(TMLR)**. 2023.

2. Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh Khapra and others. *Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages.* Transactions of the Association for Computational Linguistics **(TACL)**. 2022.

3. Raj Dabre, Chenhui Chu, Anoop Kunchukuttan. *A Comprehensive Survey of Multilingual Neural Machine Translation.* ACM Computing Surveys **(CSUR)**. 2020.

4. Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, Bamdev Mishra. *Learning Multilingual Word Embeddings in Latent Metric Space: A Geometric Approach.* Transactions of Association of Computational Linguistics **(TACL)**. 2019.

5. Anoop Kunchukuttan, Mitesh Khapra, Gurneet Singh, Pushpak Bhattacharyya. *Utilizing Orthographic Similarity for Multilingual Neural Machine Transliteration.* Transactions of the Association for Computational Linguistics **(TACL)**. 2018.

Conferences

1. Nandini Mundra, Aditya Nanda Kishore, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M Khapra. *An Empirical Comparison of Vocabulary Expansion and Initialization Approaches for Language Models.* **CoNLL**. 2024.

2. Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, Mitesh M. Khapra. *IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages.* **ACL**. 2024. *(Outstanding Paper Award)*

3. Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Ratish Puduppully, Anoop Kunchukuttan. *RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization.* **ACL**. 2024. *(Area Chair Award)*

4. Anushka Singh, Ananya B. Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M Khapra. *How Good is Zero-Shot MT Evaluation for Low Resource Indian Languages?.* ACL. 2024.

5. Kaushal Maurya, Rahul Kejriwal, Maunendra Desarkar, Anoop Kunchukuttan. *CharSpan: Utilizing Lexical Similarity to Enable Zero-Shot Machine Translation for Extremely Low-resource Languages.* **EACL**. 2024.

6. Sanjana Soni Kartik, Anoop Kunchukuttan, Tanmoy Chakraborty, Md. Shad Akhtar. *Synthetic Data Generation and Joint Learning for Robust Code-Mixed Translation.* **COLING-LREC**. 2024.

7. Nandini Mundra, Sumanth Doddapaneni, Raj Dabre, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra. *A Comprehensive Analysis of Adapter Efficiency.* CoDS-COMAD. 2024.

8. Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, Nancy F Chen. *DecoMT: Decomposed Prompting for Machine Translation Between Related Languages using Large Language Models.* **EMNLP**. 2023.

9. Aswanth Kumar, Ratish Puduppully, Raj Dabre, Anoop Kunchukuttan. *In-context Example Selection for Machine Translation Using Multiple Features.* **EMNLP Findings**. 2023.

10. Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul NC, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra. *Aksharantar: Open Indic-language Transliteration datasets and models for the Next Billion Users.* **EMNLP Findings**. 2023.

11. Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, Pratyush Kumar. *Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages.* **ACL**. 2023.

12. Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M Khapra, Pratyush Kumar, Rudra Murthy V, Anoop Kunchukuttan *Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages.* **ACL**. 2023.

13. Yash Madhani, Mitesh M Khapra, Anoop Kunchukuttan. *Bhasha-Abhijnaanam: Native-script and romanized Language Identification for 22 Indic languages.* ACL. 2023.

14. Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre. *IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation Metrics for Indian Languages.* **ACL**. 2023.

15. Divyanshu Aggarwal, Vivek Gupta, Anoop Kunchukuttan. *IndicXNLI: Evaluating Multilingual Inference for Indian Languages.* **EMNLP**. 2022.

16. Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M Khapra, Pratyush Kumar. *IndicNLG Suite: Multilingual Datasets for Diverse NLG Tasks in Indic Languages.* **EMNLP**. 2022.

17. Chaitanya Agarwal, Vivek Gupta, Anoop Kunchukuttan, Manish Shrivastava. *Bilingual Tabular Inference: A Case Study on Indic Languages.* Annual Conference of the North American Chapter of the Association for Computational Linguistics **(NAACL 2022)**. 2022.

18. Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, Pratyush Kumar. *IndicBART: A Pre-trained Model for Natural Language Generation of Indic Languages.* Findings of the Association for Computational Linguistics: ACL 2022 **(ACL-Findings 2022)**. 2022.

19. Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra. *Towards Building ASR Systems for the Next Billion Users.* Proceedings of the AAAI Conference on Artificial Intelligence **(AAAI 2022)**. 2022.

20. Anoop Kunchukuttan, Siddharth Jain, Rahul Kejriwal. *A Large-scale Evaluation of Neural Machine Transliteration for Indic Languages.* Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics **(EACL 2021)**. 2021.

21. Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, Pratyush Kumar. *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages.* Findings of EMNLP **(EMNLP-Findings**). 2020.

22. Pratik Jawanpuria, Satya Dev N T V, Anoop Kunchukuttan, Bamdev Mishra. *Learning Geometric Word Meta-Embeddings.* Proceedings of the 5th Workshop on Representation Learning for NLP. 2020.

23. Rudra Murthy V, Anoop Kunchukuttan, Pushpak Bhattacharyya. *Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages.* Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies **(NAACL 2019)**. 2019.

24. Mayank Meghwanshi, Pratik Jawanpuria, Anoop Kunchukuttan, Hiroyuki Kasai, Bamdev Mishra. *McTorch, a manifold optimization library for deep learning.* The ACM India Joint International Conference on Data Science and Management of Data **(CODS-COMAD 2019)**. 2019.

25. Rudramurthy V, Anoop Kunchukuttan, Pushpak Bhattacharyya. *Judicious Selection of Training Data in Assisting Language for Multilingual Neural NER.* Conference of Association of Computational Linguistics **(ACL 2018)**. 2018.

26. Anoop Kunchukuttan, Pratik Mehta, Pushpak Bhattacharyya. *The IIT Bombay English-Hindi Parallel Corpus.* Language Resources and Evaluation Conference **(LREC 2018)**. 2018.

27. Mayank Meghwanshi, Pratik Jawanpuria, Anoop Kunchukuttan, Hiroyuki Kasai, Bamdev Mishra. *McTorch, a manifold optimization library for deep learning.* Workshop on Machine Learning Open Source Software **(MLOSS 2018, co-located with NIPS)**. 2018.

28. Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, Pushpak Bhattacharyya. *Utilizing Lexical Similarity between Related, Low-resource Languages for Pivot-based SMT.* International Joint Conference on Natural Language Processing **(IJCNLP 2017)**. 2017.

29. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Learning variable length units for SMT between related languages via Byte Pair Encoding.* 1st Workshop on Subword and Character level models in NLP **(SCLeM 2017, co-located with EMNLP)**. 2017. *(Outstanding Paper Award)*

30. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Orthographic Syllable as basic unit for SMT between Related Languages.* Conference on Empirical Methods in Natural Language Processing **(EMNLP 2016)**. 2016.

31. Anoop Kunchukuttan, Mitesh Khapra, Pushpak Bhattacharyya. *Substring-based unsupervised transliteration with phonetic and contextual knowledge.* Conference on Natural Language Learning **(CoNLL 2016)**. 2016.

32. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Faster decoding for subword level Phrase-based SMT between related languages.* Third Workshop on NLP for Similar Languages, Varieties and Dialects **(VarDial 2016, co-located with COLING)**. 2016.

33. Rohit More, Anoop Kunchukuttan, Raj Dabre, Pushpak Bhattacharyya. *Augmenting Pivot based SMT with word segmentation.* International Conference on Natural Language Processing **(ICON 2015)**. 2015.

34. Pratik Mehta, Anoop Kunchukuttan, Pushpak Bhattacharyya. *Investigating the potential of postordering SMT output to improve translation quality.* International Conference on Natural Language Processing **(ICON 2015)**. 2015.

35. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Addressing Class Imbalance in Grammatical Error Detection with Evaluation Metric Optimization* . International Conference on Natural Language Processing **(ICON 2015)**. 2015.

36. Aditya Joshi, Anoop Kunchukuttan, Pushpak Bhattacharyya, Mark J Carman. *SarcasmBot¿: An open-source sarcasm-generation module for chatbots.* KDD Workshop on Issues of Sentiment Discovery and Opinion Mining **(WISDOM 2015, co-located with KDD)**. 2015.

37. Anoop Kunchukuttan, Ratish Puduppully , Pushpak Bhattacharyya. *Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent.* Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: System Demonstrations **(NAACL 2015)**. 2015.

38. Rajen Chatterjee, Anoop Kunchukuttan, Pushpak Bhattacharyya. *Supertag Based Pre-ordering in Machine Translation* . International Conference on Natural Language Processing **(ICON 2014)**. 2014.

39. Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya. *Shata-Anuvadak: Tackling Multiway Translation of Indian Languages* . Language and Resources and Evaluation Conference **(LREC 2014)**. 2014.

40. Mitesh M. Khapra, Ananthakrishnan Ramanathan, Anoop Kunchukuttan, Karthik Visweswariah, Pushpak Bhattacharyya. *When Transliteration Met Crowdsourcing : An Empirical Study of Transliteration via Crowdsourcing using Efficient, Non-redundant and Fair Quality Control.* Language and Resources and Evaluation Conference **(LREC 2014)**. 2014.

41. Anoop Kunchukuttan, Rajen Chatterjee, Shourya Roy, Abhijit Mishra and Pushpak Bhattacharyya. *Trans-Doop: A Map-Reduce based Crowdsourced Translation for Complex Domain* . Proceedings of the Association of Computational Linguistics: System Demonstrations **(ACL 2013)**. 2013.

42. Anoop Kunchukuttan, Shourya Roy, Pratik Patel, Somya Gupta, Kushal Ladha, Mitesh Khapra, Pushpak Bhattacharyya. *Experiences in Resource Generation for Machine Translation through Crowdsourcing* . Language and Resources and Evaluation Conference **(LREC 2012)**. 2012.

43. Anoop Kunchukuttan, Shourya Roy, Pratik Patel, Somya Gupta, Kushal Ladha, Mitesh Khapra, Pushpak Bhattacharyya. *Experiences in Resource Generation for Machine Translation through Crowdsourcing.* CrowdConf. 2011.

44. Anoop Kunchukuttan and Om P.Damani. *A System for Compound Noun Multiword Expression Extraction for Hindi.* International Conference on Natural Language Processing **(ICON 2008)**. 2008.

Articles/Technical Reports

1. Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, Pratyush Kumar. *A Primer on Pretrained Multilingual Language Models.* Arxiv preprint 2107.00676. 2021.
2. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Utilizing Language Relatedness to improve SMT: A Case Study on Languages of the Indian Subcontinent.* Arxiv preprint arXiv:2003.08925. 2020.
3. Anoop Kunchukuttan. *Indic NLP Library: A unified approach to NLP for Indian languages.* Technical Report. 2020.
4. Anoop Kunchukuttan. *The IndoWordnet Parallel Corpus.* Technical Report. 2020.
5. Anoop Kunchukuttan, Munish Munia, Pushpak Bhattacharyya. *Multiword Expressions in the CLIA project.* Vishwabharat. Jan-June 2012.
6. Anoop Kunchukuttan. *The Reordering Problem in Statistical Machine Translation.* Survey Report. 2012.

Shared Tasks

1. Vikrant Goyal, Anoop Kunchukuttan, Rahul Kejriwal, Siddharth Jain, Amit Bhagwat. *Contact Relatedness can help improve multilingual NMT: Microsoft STCI-MT @ WMT20.* Conference on Machine Translation (WMT 2020) . 2020.
2. Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Win Pa Pa, Isao Goto, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Sadao Kurohashi. *Overview of the 6th Workshop on Asian Translation.* 6th Workshop on Asian Language Translation **(WAT 2019, co-located with EMNLP)**. 2019.
3. Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh, Sadao Kurohashi. *Overview of the 5th Workshop on Asian Translation.* 5th Workshop on Asian Language Translation **(WAT 2018, co-located with PACLIC)**. 2018.
4. Raj Dabre, Anoop Kunchukuttan, Atsushi Fujita and Eiichiro Sumita. *NICT's Participation in WAT 2018: Approaches Using Multilingualism and Recurrently Stacked Layers.* 5th Workshop on Asian Language Translation **(WAT 2018, co-located with PACLIC)**. 2018.
5. Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan, Pushpak Bhattacharyya. *Comparing Recurrent and Convolutional Architectures for English-Hindi Neural Machine Translation.* 4th Workshop on Asian Language Translation **(WAT 2017, co-located with IJCNLP)**. 2017.
6. Sandhya Singh, Anoop Kunchukuttan, Pushpak Bhattacharyya. *Integrating Neural Probabilistic Language Models with SMT for English-Indonesian Translation.* 3rd Workshop on Asian Language Translation **(WAT 2016, co-located with COLING)**. 2016.
7. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Data representation methods and use of mined corpora for Indian language transliteration* . Named Entities Workshop: Shared Task **(NEWS 2015, co-located with ACL)**. 2015.
8. Anoop Kunchukuttan, Ratish Pudupully, Rajen Chatterjee, Abhijit Mishra, Pushpak Bhattacharyya. *The IIT Bombay SMT System for ICON 2014 Tools Contest.* NLP Tools Contest at ICON 2014 **(ICON 2014)**. 2014. *(3^{rd} position)*
9. Anoop Kunchukuttan, Sriram Chaudhury, Pushpak Bhattacharyya. *Tuning a Grammar Correction System for Increased Precision.* Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task **(CoNLL 2014)**. 2014.
10. Piyush Dungarwal, Rajen Chatterjee, Abhijit Mishra, Anoop Kunchukuttan, Ritesh Shah, Pushpak Bhattacharyya. *The IIT Bombay Hindi,English Translation System at WMT 2014.* Workshop on Machine Translation **(WMT 2014)**. 2014.
11. Anoop Kunchukuttan, Ritesh Shah, Pushpak Bhattacharyya. *IITB System for CoNLL 2013 Shared Task: A Hybrid Approach to Grammatical Error Correction* . Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task **(CoNLL 2013)**. 2013.
12. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Partially modelling word reordering as a sequence labelling problem.* First Workshop on Reordering for Statistical Machine Translation co-located with Computational Linguistics Conference **(RMT 2012, co-located with COLING)**. 2012.

Academic Services and Activities

- Area Chair, ACL Rolling Review.
- Co-Founder and Co-Lead of *AI4Bharat*, a research group at IIT Madras working to advance Indian language NLP.
- Journal Reviewer: Computational Linguistics (CL), Transactions of ACL (TACL), Transaction of Machine Learning Research (TMLR), ACM Transactions Asian & Low-Resource Language Information Processing (TALLIP), Natural Language Engineering (NLE).
- Conference Reviewer: ACL, EMNLP, NAACL, NeurIPS, EACL, COLING, ICLR, AAAI, IJCAI, COLM, WMT, LREC, MT Summit, EAMT, CODS-CoMAD, ACML, ICON.
- Demo Chair: CoDS-COMAD 2023.
- Co-organizer for Workshop on Asian Language Translation shared task (2016-2024).
- Founder and Maintainer of *Indic NLP Catalog*, a catalog for indexing Indian language NLP resources.
- Organizing Committee and Workshop Chair, COLING 2012.
- Co-advising students.
- Masters' and Ph.D theses reviews.

Selected Projects

- Multilingual Neural Machine Translation for Indian languages *(at Microsoft from Feb 2018 to present)*
- Neural Machine Transliteration for Indian languages *(at Microsoft from Feb 2018 to present)*
- Multi-stage crowdsourcing system for collecting translations for a complex domain like legal documents *(at IIT Bombay from Dec 2011 to May 2012)*
- Extracting information on Illicit Drug, Alcohol and Substance abuse history from free-text medical records *(at Persistent Systems from Mar 2010 to Dec 2010)*
- A platform for building information extraction solutions for medical reports *(at Persistent Systems from Mar 2010 – Mar 2011)*
- High performance, scalable document de-identification (anonymization) and indexing system for free text medical records. *(at Persistent Systems from Aug 2009 to Feb 2010)*
- Automatic Deidentification (anonymization) of Medical Records to confirm to US HIPAA guidelines. *(at Persistent Systems from Apr 2009 to Feb 2010)*
- Information Extraction on surgical pathology reports to extract test results. *(at Persistent Systems from Sep 2008 to Mar 2009)*

Software & Resources created

Source Code on **Github**: `https://github.com/anoopkunchukuttan`, `https://github.com/AI4Bharat`

Software

- *Indic NLP Library*: NLP library for Indian languages covering normalizer, transliterator, word segmenter, script information phonetic similarity, syllabification, etc.
- *Models from AI4Bharat*: Airavat LLM, IndicBERT, IndicBART, IndicTrans, IndicXlit, IndicWav2Vec, IndicFT.
- *GEOMM*: Geometry-Aware Multilingual Mapping, a toolkit for learning multilingual word embeddings. More generally, it can be used to learn mappings between different high dimensional spaces.
- IIT Bombay Unsupervised Transliterator: Unsupervised transliteration system which uses phonetic features to define transliteration priors. This is an EM based method which builds on Ravi and Knight's 2009 work.

- Multilingual Neural Machine Translation System: A multilingual Neural Machine Translation system written in Tensorflow.
- *METEOR-Indic*: MT evaluation tool extended for 18 Indian languages.
- *Moses Job Scripts*: A simple experiment management system for Moses.
- *CFILT Pre-order (Maintainer)*: Source-side pre-ordering of English for English to Indian language translation.
- *McTorch*: A manifold optimization library for deep learning.

Resources

- *Datasets from AI4Bharat*: IndicLLMSuite, IndicXtreme, IndicLID, IndicMT-Eval, IndicCorp, Samanantar, Aksharantar, IndicGLUE, IndicNLGSuite, ShrutiLipi, IndicWav2Vec.
- IIT Bombay English-Hindi Parallel Corpus
- Indian Language NLP Resources Catalog
- Mined Transliteration corpora for 110 Indian language pairs
- Transliteration corpora for English-Hindi gathered through crowdsourcing
- GEOMM Multilingual Word Embeddings for Indian languages
- Translation Resources for 110 Indian language pairs

Last updated: 8$^{\text{th}}$ March 2025