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 is one of my major research interests *i.e.,* investigating 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 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 sequence-to-sequence learning, multi-task learning, subword level models in NLP, unsupervised learning in NLP and memory networks. I am also interested in exploring word and sentence representations.

Areas of Expertise

NLP: Multilingual Learning, Machine Translation, Machine Transliteration, NLP for Related languages, NLP for Indian languages, Distributed Representations for NLP, Grammar Correction, Information Extraction, Multiword Expressions, Crowdsourcing.

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
SUMMARY: Human languages are related to each other via evolutionary links, contact over a long period of time and typological as well as scriptural similarities. In my thesis, we investigated how relatedness among languages can be leveraged to build more accurate machine translation systems with lower resource requirements. The motivation for this is two-fold: (i) even after decades of research, it has been difficult to come with universal approaches to translation which can address the vast diversity in languages, (ii) most translation requirements are within a set of related languages or between a set of related languages and a *lingua franca* like English, Spanish, etc. We believe that a set of related languages provides the right level of abstraction to be able to learn translation systems which scale to a large number of useful and related language pairs using limited resources. Some directions of work we explored are: (i) unsupervised transliteration & translation incorporating linguistic similarity to aid the learning process, (ii) sub-word level translation to address data sparsity, (iii) translation using pivot languages so that multiple related languages can help each other (iv) joint, multilingual learning of translation and transliteration models.
• M.TECH in Computer Science & Engineering, *IIT Bombay.* 2006-2008. CPI: 9.21
ADVISOR: Prof. Om Damani.
THESIS: Compound Noun Multiword Expression Extraction
SUMMARY: Multiword experessions (MWE) are concepts which cross word boundaries, where the meaning of the entire unit is not completely composable from the meaning of the constituent words. Compound nouns like *green card*, idioms like *point a finger* are some examples of this phenomenon. They are widely prevalent in natural language and their accurate recognition can enhance the accuracy of machine translation and cross lingual information retrieval. MWEs result from institutionalized usage of compound words or due to idiomatic usage of constituent words. My efforts were directed towards: (i) Extracting multi-words from a free text corpus to compile a lexicon which can be useful for NLP applications using collocation detection based methods, (ii)

Explore the linguistic cues which can aid MWE extraction, (iii) Improve lexicographer productivity in the task of MWE lexicon creation.
- B.E in Computer Engineering, *University of Pune*. 2000-2004. % Marks: 65.53

Work Experience

- Senior Applied Researcher, Microsoft India (Machine Translation group), Feb 2018 to *present*.
- 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.

Invited Talks

1. **Tutorial** on *Multilingual Neural Machine Translation* at **COLING**, Barcelona, Spain, in September 2020 with Raj Dabre & Chenhui Chu *(upcoming)*.
2. **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.
3. **Keynote Talk** on *NLP for Indian Languages: A Language Relatedness Perspective Keynote Talk* at 5th WILDRE workshop (under LREC 2020) in May 2020.
4. **Invited Talk** on *Indic NLP: A Multilinguality and Language Relatedness Perspective* at Vaibhav Summit (Organized by MyGov, Govt. of India). October 2020.
5. **Invited Talk** on *NLP for Indian Languages: A Language Relatedness Perspective* at NASSCOM Data Science & AI - Center of Excellence, Bengaluru in August 2019.
6. **Tutorial** on *Multilingual Learning* at the Summer School on Machine Learning: Advances in Modern AI, IIIT Hyderabad, in July 2018.
7. **Lecture** on *Introduction to Neural MT* at CEP Worksop on course on Deep Learning for Natural Language Processing at IIT Patna, Jan 2020.
8. **Tutorial** on *Natural Language Processing - A Distributional Approach* at IIT Alumni Center Bengaluru AI Deep Dive Workshop, July 2019.
9. Tech Talk on *Machine Translation for related languages* at AXLE 2018 (Microsoft Academic Accelerator), Microsoft IDC Hyderabad, in May 2018.
10. 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)*
11. **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.
12. Lecture on *Machine Translation* at the IIIT-H Advanced Summer School on Natural Language Processing, IIIT Hyderabad, in June 2018.
13. 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.

14. 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.

15. **Tutorial** on *Translation and Transliteration between related languages* at International Conference on Natural Language Processing, Trivandrum, in December 2015 with Mitesh Khapra.

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

17. Lecture on *Introduction to SMT, NMT and Machine Translation between Related Languages* at BITS Pilani, Hyderabad Campus, in February 2019.

18. Talk 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

Journals

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

2. 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.

3. 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. 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.

2. 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.

3. 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.

4. 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.

5. 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.

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

7. 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.

8. 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.

9. 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)*

10. 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.

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

12. 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.

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

14. 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.

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

16. 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.

17. 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.

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

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

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.

Other Articles

1. Anoop Kunchukuttan, Pushpak Bhattacharyya. *Utilizing Language Relatedness to improve SMT: A Case Study on Languages of the Indian Subcontinent.* eprint arXiv:2003.08925. 2020.

2. Anoop Kunchukuttan, Munish Munia, Pushpak Bhattacharyya. *Multiword Expressions in the CLIA project.* Vishwabharat. Jan-June 2012.

3. Anoop Kunchukuttan. *The Reordering Problem in Statistical Machine Translation.* Survey Report. 2012.

Professional Activities

- Journal Reviewer: Computational Linguistics, ACM Transactions Asian & Low-Resource Language Information Processing (TALLIP), Natural Language Engineering (NLE).
- Conference Reviewer: ACL, EMNLP, NAACL, EACL, COLING, ICLR, AAAI, IJCAI, WMT, LREC, MT Summit, EAMT, CODS-CoMAD, ACML, ICON.
- Co-organizer for Workshop on Asian Language Translation shared task (2016-2020).
- Founder and Maintainer of *Indic NLP Catalog*, a catalog for indexing Indian language NLP resources.
- Founder and Member of *AI4Bharat-NLP Initiative*, a community effort to build 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`

Software

- *Indic NLP Library*: NLP library for Indian languages covering normalizer, transliterator, word segmenter, script information phonetic similarity, syllabification, etc.
- *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.
- *IndicBERT*: Multilingual ALBERT model for 11 Indian languages and English.
- 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.

Online Systems

- *Shata-Anuvaadak*: Automatic Statistical Machine Translation system for 110 Indian language pairs.
- *BrahmiNet*: Statistical Transliteration System for 306 language pairs of South Asia.

Resources

- IndicNLP Suite: largescale corpora, BERT embeddings, NLU datasets and NLU benchmark for Indian languages.
- 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

Relevant Coursework

Natural Language Processing, Statistical Foundations of Machine Learning, Probabilistic Graphical Models, Neural Networks, Statistical Relational Learning, Data Mining, Optimization Theory, Algorithms, Web Mining.

Technical & Language Skills

Programming Languages	Python, Java, C/C++, shell scripting, Perl, R
Technologies	TensorFlow, PyTorch, Moses MT, OpenNMT, GATE, UIMA, Lucene
Tools	Eclipse, profiling tools on Java/Windows, gdb, SVN, git, LaTeX
Platforms	Linux, Windows
Natural Languages	Malayalam (native), English (fluent), Hindi (fluent), Marathi (good working knowledge)

Last updated: 17th November 2020