Prasanna Raj Noel Dabre | **Kyoto, Japan** | **DOB: 01/04/1989**
Email-Id: prajdabre@gmail.com | **Phone No: +818042446287**
Google Scholar | **LinkedIn** | **Last Updated: 23/10/2023**

Education

- Kyoto University - 04/2015 to 03/2018 - Ph.D. (Multilingual Low Resource Machine Translation)
- IIT Bombay, India - 07/2011 to 08/2014 - Masters in Technology (Indian Languages Machine Translation)
- St Francis Institute of Tech., India - 08/2006 to 05/2010 - Bachelors in Engineering (Computer Science)

Employment and Affiliation

- National Institute of Information and Communications Technology (NICT) - 05/2018 to Present - Researcher
- Indian Institute of Technology, Madras (IIT Madras), India - 01/2024 to Present - Adjunct Faculty
- St Francis Institute of Technology, India - 06/2010 to 05/2011 - Lecturer

Research Projects (NICT, Japan 2018-Present)

Efficient Multilingual Generation at Scale (May 2018-Present): I have focused on developing practical deep learning solutions that scale to many languages, especially in low-resource languages. A majority of my focus is on Indic and South-East Asian languages. I have leveraged unsupervised pre-training at scale to create and adapt LLMs, cross-lingual transfer learning and knowledge distillation in combination with heavy parameter sharing for compression.
Structured Document Translation (September 2021-Present): As a part of a collaborative research project with SAP, I have focused on structured document translation for Asian languages. Recent focus has been on retrieval augmented generation by leveraging in-context learning using LLMs, which is effective at handling long context as well as the rich structure present in web-documents.
Generation For Creoles (April 2022-Present): Creoles, being a group of neglected languages, do not have a well organized set of resources. My focus, as a part of a multi-organization collaboration, is on collecting data and implementing transfer learning methods specific to highly related languages to enable translation and generation for more than 26 Creoles.

Ph.D. Thesis (March 2018)

Exploiting Multilingualism and Transfer Learning for Low Resource Machine Translation: This thesis focuses on harnessing the power of multilingual parallel corpora for low resource machine translation. The focus was on domain adaptation, multilingual transfer learning and multisource translation. I showed the importance of linguistically similar languages when performing transfer learning.

Invited Talks and Teaching

1. Invited talk titled “Advances in Indic Natural Language Generation” in the OdiaGenAI workshop
2. Tutorial titled "Developing State-Of-The-Art Massively Multilingual Machine Translation Systems for Related Languages" at AACL-IJCNLP 2023 with Jay Gala and Pranjal Chitale (Also presented parts of this as a guest lecture titled “Multilingual Neural Machine Translation” in the NLP course taught in MBZUAI by Prof. Thamar Solrio and Prof. Alham Fikri)
3. Virtual talk titled "Efficiency in Deep Learning: An Application to Neural Machine Translation" at Google (2020).
4. Invited lecture titled "Generative Adversarial Networks" at Kyoto University (2018).
5. Tutorial titled "Multilingual Neural Machine Translation" at COLING 2020 with Anoop Kunchukuttan and Chenhui Chu. (Also presented parts of tutorial in invited talks at Kyoto University and IIT Bombay)
6. Tutorial titled "Neural Machine Translation: Basics, Practical Aspects and Recent Trends" at IJCNLP 2017 with Fabien Cromieres and Toshiaki Nakazawa

Patents (Japanese)

1. Multi-layer softmaxing of layers for flexible decoding (Accepted in 2023)
2. Multi-stage fine-tuning for low-resource machine translation (Filed in 2019)
3. BERTSeg: BERT Based Unsupervised Subword Segmentation for Neural Machine Translation (Filed in 2022)

Conference/Workshop Publications and Preprints

1. Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M Khapra, Raj Dabre, Rudra Murthy, Anoop Kunchukuttan. Airavata: Introducing Hindi Instruction-tuned LLM. In *CoRR*, volume abs/2401.15006, 2024. URL: https://arxiv.org/abs/2401.15006.

2. Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Ratish Puduppully, Anoop Kunchukuttan. RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models models via Romanization. In *CoRR*, volume abs/2401.14280, 2024. URL: https://arxiv.org/abs/2401.14280.

3. Pranjal A Chitale, Jay Gala, Varun Gumma, Mitesh M Khapra, Raj Dabre. An Empirical Analysis of In-context Learning Abilities of LLMs for MT. In *CoRR*, volume abs/2401.12097, 2024. URL: https://arxiv.org/abs/2401.12097.

4. Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, Raj Dabre. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. In *CoRR*, volume abs/2401.07078, 2024. URL: https://arxiv.org/abs/2401.07078.

5. Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, Doris Dippold. Natural Language Processing for Dialects of a Language: A Survey. In *CoRR*, volume abs/2401.05632, 2024. URL: https://arxiv.org/abs/2401.05632.

6. Wangjin Zhou, Zhengdong Yang, Chenhui Chu, Sheng Li, Raj Dabre, Yi Zhao, Kawahara Tatsuya. MOSFAD: Improving Fake Audio Detection Via Automatic Mean Opinion Score Prediction. In *CoRR*, volume abs/2401.13249, 2024. URL: https://arxiv.org/abs/2401.13249.

7. Nandini Mundra, Sumanth Doddapaneni, Raj Dabre, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra. A Comprehensive Analysis of Adapter Efficiency. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (CODS-COMAD24).*, pages 136-154, 2024. Publisher: Association for Computing Machinery. URL: https://dl.acm.org/doi/abs/10.1145/3632410.363246.

8. Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Hans Erik Heje, Diptesh Kanojia, Paul Belony, Marcel Bollmann, LoÃ¯c Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders SÃ¸gaard, Johannes Bjerva. CreoleVal: Multilingual Multitask Benchmarks for Creoles. In *CoRR*, volume abs/2310.19567, 2023. URL: https://arxiv.org/abs/2310.19567.

9. Raj Dabre, Diptesh Kanojia, Chinmay Sawant, and Eiichiro Sumita. YANMTT: Yet Another Neural Machine Translation Toolkit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 257-263, 2023. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2023.acl-demo.24.

10. Zhuoyuan Mao, Raj Dabre, Qianying Liu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. Exploring the Impact of Layer Normalization for Zero-shot Neural Machine Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1300-1316, 2023. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2023.acl-short.112.

11. Dominik Machácek, Peter Polak, Ondrej Bojar, and Raj Dabre. Robustness of Multi-Source MT to Transcription Errors. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3707-3723, 2023. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2023.findings-acl.228.

12. Ananya B. Sai, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. IndicMT Eval: A Dataset to Meta-Evaluate Machine Translation Metrics for Indian Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14210-14228, 2023. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2023.acl-long.795.

13. Varun Gumma, Raj Dabre, and Pratyush Kumar. An Empirical Study of Leveraging Knowledge Distillation for Compressing Multilingual Neural Machine Translation Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 103-114, 2023. Publisher: European Association for Machine Translation. URL: https://aclanthology.org/2023.eamt-1.11.

14. Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. SciCap+: A Knowledge Augmented Dataset to Study the Challenges of Scientific Figure Captioning. In *CoRR*, volume abs/2306.03491, 2023. URL: https://doi.org/10.48550/arXiv.2306.03491.

15. Dominik Machácek, Ondrej Bojar, and Raj Dabre. MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation. In *Proceedings of the 20th International Conference on Spoken Language Translation, IWSLT@ACL 2023, Toronto, Canada (in-person and online), 13-14 July, 2023*, pages 169-179, 2023. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2023.iwslt-1.12.

16. Zhuoyuan Mao, Haiyue Song, Raj Dabre, Chenhui Chu, and Sadao Kurohashi. Variable-length Neural Interlingua Representations for Zero-shot Neural Machine Translation. In *CoRR*, volume abs/2305.10190, 2023. URL: https://doi.org/10.48550/arXiv.2305.10190.

17. Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy F. Chen. Decomposed Prompting for Machine Translation Between Related Languages using Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.48550/arXiv.2305.13085.

18. Aswanth Kumar, Anoop Kunchukuttan, Ratish Puduppully, and Raj Dabre. In-context Example Selection for Machine Translation Using Multiple Features. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.48550/arXiv.2305.14105.

19. Raj Dabre, Bianka Buschbeck, Miriam Exel, and Hideki Tanaka. A Study on the Effectiveness of Large Language Models for Translation with Markup. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, Macau SAR, China*, pages 148-159, 2023. Publisher: Asia-Pacific Association for Machine Translation. URL: https://aclanthology.org/2023.mtsummit-research.13/.

20. Dominik Machácek, Raj Dabre, and Ondrej Bojar. Turning Whisper into Real-Time Transcription System. In *Proceedings of the 2023 The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, AACL-IJCNLP 2023, Bali, November 1-4, 2023*. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.48550/arXiv.2307.14743.

21. Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. IndicBART: A Pre-trained Model for Indic Natural Language Generation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1849-1863, 2022. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2022.findings-acl.145.

    Overview of the 9th Workshop on Asian Translation. In *Proceedings of the 9th Workshop on Asian Translation, WAT@COLING 2022, Gyeongju, Republic of Korea, October 17, 2022*, pages 1-36, 2022. Publisher: International Conference on Computational Linguistics. URL: https://aclanthology.org/2022.wat-1.1.

22. Raj Dabre. NICT's Submission to the WAT 2022 Structured Document Translation Task. In *Proceedings of the 9th Workshop on Asian Translation, WAT@COLING 2022, Gyeongju, Republic of Korea, October 17, 2022*, pages 64-67, 2022. Publisher: International Conference on Computational Linguistics. URL: https://aclanthology.org/2022.wat-1.6.

23. Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Hideki Tanaka, Masao Utiyama, and Eiichiro Sumita. FeatureBART: Feature Based Sequence-to-Sequence Pre-Training for Low-Resource NMT. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5014-5020, 2022. Publisher: International Committee on Computational Linguistics. URL: https://aclanthology.org/2022.coling-1.443.

24. Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. IndicNLG Benchmark: Multilingual Datasets for Diverse NLG Tasks in Indic Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5363-5394, 2022. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2022.emnlp-main.360.

25. Raj Dabre and Aneerav Sukhoo. KreolMorisienMT: A Dataset for Mauritian Creole Machine Translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Online only, November 20-23, 2022*, pages 22-29, 2022. Publisher: Association for Computational Linguistics. URL: https://aclanthology.org/2022.findings-aacl.3.

26. Haiyue Song, Raj Dabre, Zhuoyuan Mao, Chenhui Chu, and Sadao Kurohashi. BERTSeg: BERT Based Unsupervised Subword Segmentation for Neural Machine Translation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 85-94, 2022. Publisher: Association for Computational Linguistics. URL: https://aclanthology.org/2022.aacl-short.12.

27. Bianka Buschbeck, Raj Dabre, Miriam Exel, Matthias Huck, Patrick Huy, Raphael Rubino, and Hideki Tanaka. A Multilingual Multiway Evaluation Data Set for Structured Document Translation of Asian Languages. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Online only, November 20-23, 2022*, pages 237-245, 2022. Publisher: Association for Computational Linguistics. URL: https://aclanthology.org/2022.findings-aacl.23.

28. Zhengdong Yang, Wangjin Zhou, Chenhui Chu, Sheng Li, Raj Dabre, Raphael Rubino, and Yi Zhao. Fusion of Self-supervised Learned Models for MOS Prediction. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 5443-5447, 2022. Publisher: ISCA. URL: https://doi.org/10.21437/Interspeech.2022-10262.

29. Zhuoyuan Mao, Chenhui Chu, Raj Dabre, Haiyue Song, Zhen Wan, and Sadao Kurohashi. When do Contrastive Word Alignments Improve Many-to-many Neural Machine Translation? In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1766-1775, 2022. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2022.findings-naacl.134.

30. Raj Dabre. NICT at MixMT 2022: Synthetic Code-Mixed Pre-training and Multi-way Fine-tuning for Hinglish-English Translation. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1122-1125, 2022. Publisher: Association for Computational Linguistics. URL: https://aclanthology.org/2022.wmt-1.111.

31. Raj Dabre and Atsushi Fujita. Investigating Softmax Tempering for Training Neural Machine Translation Models. In *Proceedings of the 18th Biennial Machine Translation Summit - Volume 1: Research Track, MTSummit 2021 Virtual, August 16-20, 2021*, pages 114-126, 2021. Publisher: Association for Machine Translation in the Americas. URL: https://aclanthology.org/2021.mtsummit-research.10.

32. Raj Dabre, Aizhan Imankulova, and Masahiro Kaneko. Studying The Impact Of Document-level Context On Simultaneous Neural Machine Translation. In *Proceedings of the 18th Biennial Machine Translation Summit - Volume 1: Research Track, MTSummit 2021 Virtual, August 16-20, 2021*, pages 202-214, 2021. Publisher: Association for Machine Translation in the Americas. URL: https://aclanthology.org/2021.mtsummit-research.17.

33. Raj Dabre, Aizhan Imankulova, Masahiro Kaneko, and Abhisek Chakrabarty. Simultaneous Multi-Pivot Neural Machine Translation. In *CoRR*, volume abs/2104.07410, 2021. URL: https://arxiv.org/abs/2104.07410.

34. Raj Dabre, Atsushi Fujita. Recurrent Stacking of Layers in Neural Networks: An Application to Neural Machine Translation. In *CoRR*, volume abs/2106.10002, 2021. URL: https://arxiv.org/abs/2106.10002.

35. Haiyue Song, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, Eiichiro Sumita. Pre-training via Leveraging Assisting Languages for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2020, Online, July 5-10, 2020*, pages 279-285, 2020. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2020.acl-srw.37.

36. Raj Dabre, Raphael Rubino, Atsushi Fujita. Balancing Cost and Benefit with Tied-Multi Transformers. In *Proceedings of the Fourth Workshop on Neural Generation and Translation, NGT@ACL 2020, Online, July 5-10, 2020*, pages 24-34, 2020. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2020.ngt-1.3.

37. Raj Dabre, Abhisek Chakrabarty. NICT's Submission To WAT 2020: How Effective Are Simple Many-To-Many Neural Machine Translation Models? In *Proceedings of the 7th Workshop on Asian Translation, WAT@AACL/IJCNLP 2020, Suzhou, China, December 4, 2020*, pages 98-102, 2020. Publisher: Association for Computational Linguistics. URL: https://aclanthology.org/2020.wat-1.9/.

38. Diptesh Kanojia, Raj Dabre, Shubham Dewangan, Pushpak Bhattacharyya, Gholamreza Haffari, Malhar Kulkarni. Harnessing Cross-lingual Features to Improve Cognate Detection for Low-resource Languages. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1384-1395, 2020. Publisher: International Committee on Computational Linguistics. URL: https://doi.org/10.18653/v1/2020.coling-main.119.

39. Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, Eiichiro Sumita. Improving Low-Resource NMT through Relevance Based Linguistic Features Incorporation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4263-4274, 2020. Publisher: International Committee on Computational Linguistics. URL: https://doi.org/10.18653/v1/2020.coling-main.376.

40. Haiyue Song, Raj Dabre, Atsushi Fujita, Sadao Kurohashi. Coursera Corpus Mining and Multistage Fine-Tuning for Improving Lectures Translation. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3640-3649, 2020. Publisher: European Language Resources Association. URL: https://aclanthology.org/2020.lrec-1.449/.

41. Zhuoyuan Mao, Fabien Cromierès, Raj Dabre, Haiyue Song, Sadao Kurohashi. JASS: Japanese-specific Sequence to Sequence Pre-training for Neural Machine Translation. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3683-3691, 2020. Publisher: European Language Resources Association. URL: https://aclanthology.org/2020.lrec-1.454/.

42. Sheng Li, Xugang Lu, Raj Dabre, Peng Shen, Hisashi Kawai. Joint Training End-to-End Speech Recognition Systems with Speaker Attributes. In *Odyssey 2020: The Speaker and Language Recognition Workshop, 1-5 November 2020, Tokyo, Japan*, pages 385-390, 2020. Publisher: ISCA. URL: https://doi.org/10.21437/Odyssey.2020-54.

43. Raj Dabre, Atsushi Fujita. Combining Sequence Distillation and Transfer Learning for Efficient Low-Resource Neural Machine Translation Models. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 492-502, 2020. Publisher: Association for Computational Linguistics. URL: https://aclanthology.org/2020.wmt-1.61/.

44. Raj Dabre, Atsushi Fujita. Recurrent Stacking of Layers for Compact Neural Machine Translation Models. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6292-6299, 2019. Publisher: AAAI Press. URL: https://doi.org/10.1609/aaai.v33i01.33016292.

45. Raj Dabre, Eiichiro Sumita. NICT's participation to WAT 2019: Multilingualism and Multi-step Fine-Tuning for Low Resource NMT. In *Proceedings of the 6th Workshop on Asian Translation, WAT@EMNLP-IJCNLP 2019, Hong Kong, China, November 4, 2019*, pages 76-80, 2019. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/D19-5207.

46. Raj Dabre, Atsushi Fujita, Chenhui Chu. Exploiting Multilingualism through Multistage Fine-Tuning for Low-Resource Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1410-1416, 2019. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/D19-1146.

47. Sheng Li, Raj Dabre, Xugang Lu, Peng Shen, Tatsuya Kawahara, Hisashi Kawai. Improving Transformer-Based Speech Recognition Systems with Compressed Structure and Speech Attributes Augmentation. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 4400-4404, 2019. Publisher: ISCA. URL: https://doi.org/10.21437/Interspeech.2019-2112.

48. Aizhan Imankulova, Raj Dabre, Atsushi Fujita, Kenji Imamura. Exploiting Out-of-Domain Parallel Data through Multilingual Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*, pages 128-139, 2019. Publisher: European Association for Machine Translation. URL: https://aclanthology.org/W19-6613/.

49. Raj Dabre, Kehai Chen, Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita. NICT's Supervised Neural Machine Translation Systems for the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 168-174, 2019. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/w19-5313.

50. Benjamin Marie, Raj Dabre, Atsushi Fujita. NICT's Machine Translation Systems for the WMT19 Similar Language Translation Task. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 208-212, 2019. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/w19-5428.

51. Raj Dabre and Eiichiro Sumita. NICT's Supervised Neural Machine Translation Systems for the WMT19 Translation Robustness Task. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 533-536, 2019. Publisher: Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/w19-5362.

52. Chenhui Chu and Raj Dabre. Multilingual Multi-Domain Adaptation Approaches for Neural Machine Translation. In *CoRR*, volume abs/1906.07978, 2019. URL: http://arxiv.org/abs/1906.07978.

53. Raj Dabre, Anoop Kunchukuttan, Atsushi Fujita, and Eiichiro Sumita. NICT's Participation in WAT 2018: Approaches Using Multilingualism and Recurrently Stacked Layers. In: *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation, WAT@PACLIC 2018, Hong Kong, December 1-3, 2018*, Publisher: Association for Computational Linguistics, 2018. URL: https://aclanthology.org/Y18-3003/.

54. Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, Publisher: Association for Computational Linguistics, 2017. URL: https://doi.org/10.18653/v1/P17-2061.

55. Raj Dabre, Fabien Cromierès, and Sadao Kurohashi. Kyoto University MT System Description for IWSLT 2017. In: *Proceedings of the 14th International Conference on Spoken Language Translation, IWSLT 2017, Tokyo, Japan, December 14-15, 2017*, Publisher: International Workshop on Spoken Language Translation, 2017. URL: https://aclanthology.org/2017.iwslt-1.8.

56. Raj Dabre, Fabien Cromières, and Sadao Kurohashi. Enabling Multi-Source Neural Machine Translation By Concatenating Source Sentences In Multiple Languages. In: *Proceedings of Machine Translation Summit XVI, Volume 1: Research Track, MTSummit 2017, September 18-22, 2017, Nagoya, Aichi, Japan*, Publisher: International Workshop on Spoken Language Translation, 2017. URL: https://aclanthology.org/2017.mtsummit-papers.8.

57. Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation. In: *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation, PACLIC 2018, Cebu City, Philippines, November 16-18, 2017*, Publisher: The National University (Phillippines), 2017. URL: https://aclanthology.org/Y17-1038/.

58. Chenhui Chu, Raj Dabre, and Sadao Kurohashi. Parallel Sentence Extraction from Comparable Corpora with Neural Network Features. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, PortoroÅ$\frac{3}{4}$, Slovenia, May 23-28, 2016, Publisher: European Language Resources Association (ELRA), 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/363.html.

59. Raj Dabre, Yevgeniy Puzikov, Fabien CromierÃ¨s, and Sadao Kurohashi. The Kyoto University Cross-Lingual Pronoun Translation System. In: *Proceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany, August 11-12, 2016. Publisher: The Association for Computer Linguistics, 2016. URL: https://doi.org/10.18653/v1/w16-2349. DOI: 10.18653/v1/w16-2349.

60. Diptesh Kanojia, Raj Dabre, and Pushpak Bhattacharyya. Sophisticated Lexical Databases - Simplified Usage: Mobile Applications and Browser Plugins For Wordnets. In: *Proceedings of the 8th Global WordNet Conference (GWC 2016)*, Bucharest, Romania, January 27-30, 2016, Publisher: Global Wordnet Association, 2016. URL: https://aclanthology.org/2016.gwc-1.22/.

61. John Richardson, Raj Dabre, Chenhui Chu, Fabien CromierÃ¨s, Toshiaki Nakazawa, and Sadao Kurohashi. KyotoEBMT System Description for the 2nd Workshop on Asian Translation. In: *Proceedings of the 2nd Workshop on Asian Translation (WAT 2015)*, Kyoto, Japan, October 16, 2015, Publisher: Workshop on Asian Translation, 2015. URL: https://aclanthology.org/W15-5006/.

62. Rohit More, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Raj Dabre. Augmenting Pivot-based SMT with Word Segmentation. In: *Proceedings of the 12th International Conference on Natural Language Processing (ICON 2015)*, Trivandrum, India, December 11-14, 2015, Publisher: NLP Association of India, 2015. URL: https://aclanthology.org/W15-5944/.

63. Raj Dabre, Fabien CromierÃ¨s, Sadao Kurohashi, and Pushpak Bhattacharyya. Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages. In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, USA, May 31 - June 5, 2015, Publisher: The Association for Computational Linguistics, 2015. URL: https://doi.org/10.3115/v1/n15-1125. DOI: 10.3115/v1/n15-1125.

64. Raj Dabre, Chenhui Chu, Fabien CromierÃ¨s, Toshiaki Nakazawa, and Sadao Kurohashi. Large-scale Dictionary Construction via Pivot-based Statistical Machine Translation with Significance Pruning and Neural Network Features. In: *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29)*, Shanghai, China, October 30 - November 1, 2015. Publisher: ACL, 2015. URL: https://aclanthology.org/Y15-1033/.

65. Raj Dabre, Jyotesh Choudhari, and Pushpak Bhattacharyya. Tackling Close Cousins: Experiences in Developing Statistical Machine Translation Systems for Marathi and Hindi. In: *Proceedings of the 11th International Conference on Natural Language Processing (ICON 2014)*, Goa, India, December 18-21, 2014, Publisher: NLP Association of India, 2014. URL: https://aclanthology.org/W14-5103/.

66. Raj Dabre, Aneerav Sukhoo, and Pushpak Bhattacharyya. Anou Tradir: Experiences in Building Statistical Machine Translation Systems for Mauritian Languages - Creole, English, French. In: *Proceedings of the 11th International Conference on Natural Language Processing (ICON 2014)*, Goa, India, December 18-21, 2014, Publisher: NLP Association of India, 2014. URL: https://aclanthology.org/W14-5113/.

67. Diptesh Kanojia, Manish Shrivastava, Raj Dabre, and Pushpak Bhattacharyya. PaCMan: Parallel Corpus Management Workbench. In: *Proceedings of the 11th International Conference on Natural Language Processing (ICON 2014)*, Goa, India, December 18-21, 2014, Publisher: NLP Association of India, 2014. URL: https://aclanthology.org/W14-5126/.

68. Diptesh Kanojia, Pushpak Bhattacharyya, Raj Dabre, Siddhartha Gunti, and Manish Shrivastava. Do not do processing when you can look up: Towards a Discrimination Net for WSD. In: *Proceedings of the Seventh Global Wordnet Conference (GWC 2014)*, Tartu, Estonia, January 25-29, 2014, Publisher: University of Tartu Press, 2014. URL: https://aclanthology.org/W14-0126/.

69. Raj Dabre, Archana Amberkar, and Pushpak Bhattacharyya. Morphological Analyzer for Affix Stacking Languages: A Case Study of Marathi. In: *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, Publisher: Indian Institute of Technology Bombay, 2012. URL: https://aclanthology.org/C12-2023/.

Journal Publications

1. Jay P. 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, and Anoop Kunchukuttan. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. In *Transactions on Machine Learning Research*, November 2023. URL: https://openreview.net/pdf?id=vfT4YuzAYA.

2. Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. Low-resource Multilingual Neural Translation Using Linguistic Feature-based Relevance Mechanisms. In *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, volume 22, number 7, pages 191:1-191:36, 2023. URL: https://doi.org/10.1145/3594631.

3. Haiyue Song, Raj Dabre, Chenhui Chu, Sadao Kurohashi, and Eiichiro Sumita. SelfSeg: A Self-Supervised Sub-Word Segmentation Method for Neural Machine Translation. In *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, volume 22, number 8, article 215, August 2023. Publisher: Association for Computing Machinery. URL: https://doi.org/10.1145/3610611.

4. Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A Survey of Multilingual Neural Machine Translation. *ACM Comput. Surv.*, 53(5):99:1–99:38, 2021. URL: https://doi.org/10.1145/3406095.

5. Raphael Rubino, Benjamin Marie, Raj Dabre, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita. Extremely low-resource neural machine translation for Asian languages. In *Machine Translation*, volume 34, number 4, pages 347-382, 2020. URL: https://doi.org/10.1007/s10590-020-09258-6.

6. Raj Dabre, Fabien Cromierès, and Sadao Kurohashi. Exploiting Multilingual Corpora Simply and Efficiently in Neural Machine Translation. *J. Inf. Process.*, 26:406–415, 2018. URL: https://doi.org/10.2197/ipsjjip.26.406.

7. Chenhui Chu, Raj Dabre, and Sadao Kurohashi. A Comprehensive Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. *J. Inf. Process.*, 26:529–538, 2018. URL: https://doi.org/10.2197/ipsjjip.26.529.

Software and Resources created

1. **Airavata:** A Hindi Instruction-tuned LLM See: https://ai4bharat.github.io/airavata/

2. **IndicTrans2 and BPCC:** The current state-of-the-art machine translation model covering 22 Indic languages along with the world's largest dataset for said languages. See: https://github.com/AI4Bharat/IndicTrans2

3. **YANMTT:** Yet another neural machine translation toolkit is built on top of HuggingFace transformers. This can be used for distributed/multi-node massively multilingual pre-training and efficient transfer learning via sequence-to-sequence models. It is currently used for IndicBART and IndicNLG projects. See: https://github.com/prajdabre/yanmtt (**Published at ACL 2023 as a demo paper**)

4. **Indic MT Evaluation Suite:** A meta-evaluation dataset to evaluate machine translation evaluation metrics for 5 Indian languages. See: https://github.com/AI4Bharat/IndicMT-Eval

5. **IndicBART:** A pre-trained encoder-decoder model for 11 Indic languages and English. This has been widely used for Indic languages natural language generation. See: https://huggingface.co/ai4bharat/IndicBART

6. **IndicNLG Benchmark and Models:** A benchmark containing training and evaluation datasets for natural language generation of 11 languages and 5 generation tasks. Models for the same have been released. See: https://ai4bharat.iitm.ac.in/indicnlg-suite/

7. **SciCap+:** An enhanced version of the scientific figure captioning dataset, SciCap, with retrieved paragraphs to enhance caption quality. See: https://github.com/zhishenyang/scientific_figure_captioning_dataset

8. **Coursera Parallel Corpus:** A parallel corpus for Chinese-Japanese-English machine translation mined from Coursera lectures. See: https://github.com/shyyhs/CourseraParallelCorpusMining

9. **JaRuNC:** A news commentary dataset for Japanese-Russian machine translation represnting an exremely low-resource dataset. See: https://github.com/aizhanti/JaRuNC

Community Presence

- **AI4Bharat**: As a visiting researcher since 2021, adjunct faculty at IIT Madras since January 2024, I am one of the leads behind the Indic natural language understanding and generation efforts.
- **CFILT**: As a visiting researcher since 2023, I am guiding students on linguistics centric machine translation, language understanding and pragmatics. My application for adjunct faculty is under process.
- **Workshop Organization**: I have been a co-organizer for the Workshop on Asian Translation since 2018, focusing on the Indic MT tasks. Since 2023, I have been co-organizing the M3Oriental workshop, which focuses on speech technologies for Asian languages.

- **Reviewing and Chairing:** Reviewer/committee member since 2012 for venues such as **ACL, NAACL, IJCAI, EMNLP, CoNLL, WMT, WAT, MT Summit, IJCNLP, ALR, TALLIP, TASLP and CSL**. Area chair for **ACL, NAACL, EMNLP and EACL** since 2022.

Academic and Professional Achievements

- Recipient of **MEXT Scholarship** via University Recommendation (2014-2018)
- Achieved **AIR 110** among **136027 candidates** in **GATE 2011**
- Secured **Department 1st rank** in St. Francis Institute of Technology in the B.E. Programme
- Received **TATA scholarship 2 years** in a row for securing distinction in B.E. examinations

Technical Skills

- **Programming Languages**: C, C++, Java
- **Scripting Languages**: Python (working with TensorFlow and PyTorch), MATLAB
- **Web Technologies**: HTML, AJAX, JQuery, PHP (basic), JSP (basic)
- **Operating Systems**: Linux, Windows
- **Database**: MySQL, Oracle
- **Tools**: VS Code, LaTeX, Sublime

Collaborative Activities

- **Kyoto University/ Tokyo Metropolitan University/ Tokyo Institute of Technology/ IIT Madras/ IIT Bombay/ Charles University/ Cape Town University**: Guiding students.
- **SAP**: Managing the research and development collaboration between NICT and SAP.

Internships

Improving GNMT for Low Resource Languages: Google's deep neural machine translation system (GNMT) relies on a large amount of data for optimal translation quality. The objective was to apply a combination of various existing ideas to leverage resource rich languages leading to significant improvements in translation quality for the resource poor languages.
*(Host: **Tetsuji Nakagawa**, Organization: **Google Inc., Japan**, Aug'16 - Oct'16)*

Other Voluntary Responsibilities

- Organising Member of COLING 2012 (December 2012) and WAT 2018-2022, Department Coordinator of TechConnect 2013 (January 2013), System Administrator of CFILT, IIT Bombay (July 2011-July'2014)

Extra Curricular Activities

- Working with international relations groups as Japanese-English interpreter
- Hobbies: **Astronomy**, Reading Fiction and Literature, Watching Anime and Manga, Cooking
- Languages: English (native), Hindi/Urdu (native), Marathi (native), Japanese (fluent), Chinese (beginner)