Kabir **Ahuja**

Research Fellow, Microsoft Research

🌐 https://kabirahuja2431.github.io/ @ kabirahuja2431@gmail.com github.com/kabirahuja2431 Google Scholar
📍 Microsoft Research, #9 VIGYAN, Lavelle Road, Bangalore, KA, India 560001

Education

May 2019
Aug 2015
Birla Institute of Technology and Science (BITS) Pilani **Pilani, India**
B.E. (Hons.), Chemical Engineering
CGPA: 9.45/10 (Distinction)
Department Rank: 1
Thesis Title: Learning To Optimize Molecular Geometries Using Reinforcement Learning

Experience

Present
Aug 2021
Microsoft Research **Bangalore, India**
Research Fellow [🌐] | *Advisors: Dr. Sunayana Sitaram , Dr. Monojit Choudhury, Dr. Navin Goyal*
Working on building linguistically fair Multilingual Models covering different aspects around their performance, calibration, evaluation, interpretation and data collection. Also working on analyzing the properties of Self Attention in Transformers and the types of functions that can be expressed and learned by these networks.

Aug 2021
July 2020
Udaan.com **Bangalore, India**
Data Scientist | Mentor: Dr. Mohit Kumar
Worked on implementing and deploying Recommendation Systems to solve numerous use cases across various verticals of the company.

July 2020
Jan 2020
Microsoft Research **Bangalore, India**
Research Intern | Advisor: Dr. Navin Goyal
Worked on on analyzing and contrasting the learning capabilities of Recurrent Neural Networks and Transformers in recognizing Context Free Languages in finite precision.

Dec 2019
June 2019
Indian Institute of Science | MALL Lab [🌐] **Bangalore, India**
Research Assistant | Advisor: Dr. Partha Talukdar
Worked on controlled text generation for generating paraphrases of sentences while following a specified syntactic structure provided in form of an exemplar sentence.

Dec 2018
Aug 2018
Massachusetts Institute of Technology | Green Research Group [🌐] **Cambridge, MA, USA**
Visting Student Researcher | Advisors: Prof. William H. Green,Prof. Yi-Pei Li
Worked on devising Reinforcement Learning based methods for Molecular Geometry Optimization for efficiently finding the local minima of non-convex energy surfaces.

Publications

S=In Submission, C=Conference, W=Workshop, J=Journal

[C.6] **On Calibration of Massively Multilingual Language Models**
Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, Monojit Choudhury
The 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi **[EMNLP'22]**

[C.5] **Global Readiness of Language Technology for Healthcare: What would it Take to Combat the Next Pandemic?** [🔗]
Ishani Mondal*, Kabir Ahuja*, Mohit Jain, Jacki O'Neill, Kalika Bali, Monojit Choudhury
The 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea **[COLING'22]**

[C.4] **On the Economics of Multilingual Few-shot Learning: Modeling the Cost-Performance Trade-offs of Machine Translated and Manual Data** [🔗][code]
Kabir Ahuja, Monojit Choudhury, Sandipan Dandapat
2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Seattle **[NAACL'22]**

[C.3] **Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models** [🔗]
Kabir Ahuja*, Shanu Kumar*, Sandipan Dandapat, Monojit Choudhury
60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland **[ACL'22]**

[C.2] **On the Practical Ability of Recurrent Neural Networks to Recognize Hierarchical Languages** [🔗][code]
Satwik Bhattamishra, Kabir Ahuja, Navin Goyal
28th International Conference on Computational Linguistics **[Best Short Paper Award]** **[COLING'20]**

[C.1] **On the Ability and Limitations of Transformers to Recognize Formal Languages** [🔗][code]
Satwik Bhattamishra, Kabir Ahuja, Navin Goyal
2020 Conference on Empirical Methods in Natural Language Processing **[EMNLP'20]**

[J.2] **Learning to Optimize Molecular Geometries Using Reinforcement Learning** [🔗]
Kabir Ahuja, William H. Green, Yi-Pei Li
Journal of Chemical Theory and Computation [Impact Factor: 6.006] **[JCTC]**

[J.1] **Syntax-Guided Controlled Generation of Paraphrases** [🔗][code]
Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, Partha Talukdar
Transactions of the Association for Computational Linguistics **[TACL]**

[W.1] **Beyond Static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages** [🔗]
Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram, Monojit Choudhury
NLP Power! The First Workshop on Efficient Benchmarking in NLP, ACL, Dublin **[NLPPower@ACL'22]**

Select Research Projects

Calibration of Multilingual Models Feb'22 - July'22
Advisors: Dr. Sunayana Sitaram, Dr. Monojit Choudhury, Dr. Sandipan Dandapat

- Empirically investigated that Multilingual PLMs were massively mis-calibrated in a zero-shot cross lingual setup across multiple tasks and languages.
- Particularly observed strong correlations between Expected Calibration Error (ECE) and amount of pre-training data in the target language as well as its typological relatedness with the language used during fine-tuning.
- Showed that with Temperature Scaling and Label Smoothing along with collecting a few instances of labelled data in target language, one could reduce the calibration errors significantly, achieving the errors close to those observed for English for all languages tested.
- Accepted as a short paper at EMNLP 2022.

Performance and Cost Trade-Offs of Machine Translated and Manual Data Oct'21 - Feb'22
Advisors: Dr. Monojit Choudhury, Dr. Sandipan Dandapat

- Adapted the Production Function theory from microeconomics to introduce Performance Functions that were used to measure the cost and performance trade-offs of using Machine Translated vs Manually Annotated data for fine-tuning Multilingual Models.
- Through a case study on TyDiQA benchmark we demonstrated that manual data is both more sample efficient for fine-tuning compared to translated data as well as essential to reach higher levels of performance.
- Also showed that Translated data can be helpful in low resource settings where there is less amount of data in any language to begin with and can act as a form of Data Augmentation.
- Accepted as a long paper at NAACL 2022.

Predicting Performance of Multilingual Models Across Languages Aug'22 - December'22
Advisors: Dr. Monojit Choudhury, Dr. Sandipan Dandapat

- Formulated Multi-Task regression approaches to predict Zero-Shot performance of Multilingual models across different languages without explicitly evaluating them on test data, using model-specific, linguistic and tokenizer features.
- Demonstrated that our proposed approach out-performed existing single-task and averaging baselines significantly for low-resource tasks and languages.
- Interpreted the factors influencing zero-shot performance by analyzing the SHAP values of the regression models and demonstrated importance of Tokenizer Quality for tasks involving token-level predictions.
- Accepted as a long paper at ACL 2022. Coverage/Mentions - ruder.io/acl2022

Analysis of Neural Sequence Models on Formal Languages Jan'20 - July'20

Advisor: Dr. Navin Goyal

- Worked on analyzing and contrasting the learning capabilities of Recurrent Neural Networks and Transformers in recognizing Formal Languages in finite precision.
- Showed that encoder only Transformers can learn to recognize certain counter languages but are limited in their capabilities to recognize regular languages.
- Investigated the extent of generalization exhibited by these models in acquiring the underlying rules of a formal language and how it is impacted by architectural choices in a model like depth, number of attention heads, the type of Positional Encodings used etc. This work got accepted at EMNLP'20
- Also showed theoretically and empirically that RNNs can recognize context-free languages with bounded depth.
- Developed probing tasks to interpret the intermediate representations and validate the findings. This work was accepted at COLING'20.

Syntax Controlled Paraphrase Generation June'19 - Dec'19

Advisor: Dr. Partha Talukdar

- Worked on developing a sequence to sequence model for **Syntactically Controlled Paraphrase Generation** composed of Bidirectional LSTMs, Tree LSTM and Pointer Generator Networks.
- Incorporated the desired syntax using Constituency Parse Trees and implemented a gated mechanism to selectively choose leaf node representations while decoding.
- Trained on Quora Question Pairs (QQP) and ParaNMT datasets, our proposed model outperformed existing methods by a significant margin on metrics like BLEU, ROUGE, METEOR and Tree-Edit Distance as well as in Human Evaluation.
- Published in TACL (Volume 8, 2020)

Reinforcement Learning for Molecular Geometry Optimization Aug'18 - Dec'18

Advisors: Advisors: Prof. William H. Green,Prof. Yi-Pei Li

- Designed a Reinforcement Learning based approach to find the local minima of Potential Energy Surfaces of different molecules while minimizing the number of optimization steps.
- Utilized Self Attention to handle variable size state and action spaces while implementing Policy and Value Networks and trained them using Proximal Policy Optimization.
- Achieved better performance (30% fewer steps on average) than the state of the art optimization schemes for geometry optimization like BFGS, L-BFGS, FIRE etc, on both in-domain and out of domain optimization environments.
- Published in Journal of Chemical Theory and Computation (2021)

Honours and Awards

Best Short Paper Award in COLING, 2020 [🌐] For our paper: On the Practical Ability of Recurrent Neural Networks to Recognize Hierarchical Languages

BITS Merit Cum Need Scholarship, 2015-2019 Recipient of university's Scholarship awarded to top 2% students of a batch.

Winner of Turing's Large Scale Models for Inclusion Hackathon Challenge, 2022 Implemented Inclusivity Toolkit to diagnose the biases of language models across various dimensions by bringing together numerous bias detection method in literature.

Microsoft Global Hackathon Award Winner, 2021 [🌐] Implemented a healthcare chat-bot to assess patients blood test reports and answer queries based on that.

Flipkart Grid Challenge [🌐] Appeared among top 10 finalists out of 6000 participating teams in the Flipkart Grid Challange 2019 involving an Object Localisation problem for E-Commerec products.

Teaching and Volunteering Roles

SNLP Reading Group, MSR India *Organizer* Jun'22 - Present

- Organize weekly meet-ups discussing recent papers and trends in Natural Language Processing as well arrange invited talks in the related areas.

Natural Language Processing, Plaksha University *Teaching Assistant* Jan'22 - Apr'22

- Designed Assignments covering different topics in NLP from Bag Of Word Models to Massively Multilingual Language Models.
- Mentored students for their course projects.

Neural Networks and Fuzzy Logic, BITS Pilani *Teaching Assistant* Jan'19 - May'19

- Helped the course be more hands-on by introducing programming assignments for the course encompassing topics like K-nearest neighbours, Logistic Regression and Deep Autoencoders.
- Hosted a Kaggle competition where students were asked to implement a Recommendation System using Deep Auto-Encoders.
- Conducted workshops on Python programming, Neural Networks in practice and Deep Learning frameworks to provide a practical flavour to the concepts taught in class.

National Service Scheme, BITS Pilani *School Volunteer and Class Head* Aug'15 - May'17

- Taught under-privileged students high-school level Physics, Chemistry and Mathematics.
- Prepared assignments and conducted tests to prepare them for High-School and College Entrance Exams.

Relevant Coursework

Computer Science: Data Structures and Algorithms, Computer Programming, Object Oriented Programming, Machine Learning, Neural Networks and Fuzzy Logic, Artifical Intelligence

Mathematics: Mathematics-1 (Multivariate Calculus), Mathematics-2 (Linear Algebra, Complex Variables and Calculus), Mathematics-3 (Differential Equations), Probability and Statistics, Discrete Mathematics, Non-Linear Optimization, Numerical Methods for Chemical Engineering

Skills

Programming Languages:	Python, C, C++, Java, Matlab
Libraries and Frameworks:	Pytorch, Hugging Face Transformers, Weights and Biases, Numpy, Tensorflow

Academic Service

Workshop Co-Organizer	SumEval Workshop at AACL'22
Peer Reviewer	AAAI'23, EMNLP'22, NAACL'22, ACL'22

References

- Dr. Monojit Choudhury .... *Principal Data and Applied Scientist, Microsoft Turing, India* [🌐]
- Dr. Navin Goyal .... *Principal Researcher, Microsoft Research, India* [🌐]
- Dr. Sunayana Sitaram .... *Senior Researcher, Microsoft Research, India* [🌐]