

# Shruti Rijhwani

<https://shrutirij.github.io>  
Google Scholar  
srijhwan@cs.cmu.edu

## EDUCATION

---

**Carnegie Mellon University** Pittsburgh, PA  
Ph.D. in Language Technologies, School of Computer Science 2018 – Aug 2022 (expected)  
– Advisor: Graham Neubig  
– Research Focus: Low-Resource Natural Language Processing  
\* Select publications: TACL 2021 [3], EMNLP 2020 [8], ACL 2020 [9], TACL 2020 [11], AACL 2019 [16]  
– Awards: Forbes 30 Under 30 in Science 2022 [link]; Bloomberg Ph.D. Fellowship

**Carnegie Mellon University** Pittsburgh, PA  
M.S. in Language Technologies, School of Computer Science 2016 – 2018

**Birla Institute of Technology and Science, Pilani** Pilani, India  
B.S. in Computer Science 2011 – 2015

## WORK EXPERIENCE

---

**Bloomberg AI** New York, NY  
Research Intern (*Mentors: Jing Wang, Daniel Preoțiuc-Pietro, Anju Kambadur*) May – August 2020  
– Project: Multilingual named entity recognition for noisy text outputs from OCR systems.  
– Skills: Python, PyTorch

**Bloomberg AI** New York, NY  
Research Intern (*Mentor: Daniel Preoțiuc-Pietro*) May – August 2019  
– Project: Temporally-aware named entity recognition.  
– Created a temporally distributed dataset and improved state-of-the-art neural network models for NER on Twitter data by incorporating temporal features; published at ACL 2020 [10]. Skills: Python, PyTorch

**Microsoft Research** India  
Research Fellow (*Mentors: Monojit Choudhury, Kalika Bali*) August 2015 – July 2016  
– Project: Code-switched language processing.  
– Word-level language identification, sentiment analysis, and speech synthesis of code-switched texts; published at ACL 2017 [20], EMNLP 2016 [23], and SSW 2016 [24]. Skills: C++, Python, C#

**Google Summer of Code, MIT Media Lab** Virtual  
Software Developer May – July 2015  
– Implemented features for MIT App Inventor, an open-source visual programming environment.  
– Skills: Java, JavaScript

**Microsoft Research** India  
Research Intern (*Applied Sciences*) January – May 2015  
– Developed scalable fuzzy search software for retrieval and deduplication of millions of misspelled entities.

**Microsoft** India  
Software Development Intern (*Bing Ads*) May – August 2014  
– Part-of-speech tagging for search queries using weakly-supervised Conditional Random Fields.  
– Skills: C#, Python

## AWARDS AND HONORS

---

- Forbes 30 Under 30 in Science [\[link\]](#) 2022
- Bloomberg Ph.D. Fellowship 2018–2021
- Honorable Mention, Society of Fellows in Critical Bibliography Essay Prize (awarded for [\[13\]](#)) 2021
- Carnegie Mellon University Graduate Research Fellowship 2016–2018
- Best Presentation Award, Student Research Symposium at Carnegie Mellon University 2018
- Best Poster Award, Machine Learning Project Symposium at Carnegie Mellon University 2016

## RESEARCH MENTORING

---

### Masters Students

- Sedrick Keh  
Project: Zero-Shot Language Modeling.
- Adithya Pratapa  
Project: Morphosyntactic Evaluation of Generated Texts; published at EMNLP 2021 [\[5\]](#).
- Rosaline Su  
Project: Dependency Induction with Visual Perception; published at CoNLL 2021 [\[6\]](#).
- Shuyan Zhou  
Project: Entity Linking for Low-Resource Languages; published at TACL 2020 [\[11\]](#) and DeepLo 2019 [\[18\]](#).
- Aman Madaan  
Project: Practical Comparable Data Collection for Low-Resource Languages via Images; published at PML4DC [\[15\]](#).
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang  
Project: Choosing Transfer Languages for Cross-Lingual Learning; published at ACL 2019 [\[17\]](#).

### Undergraduate Students

- Lindia Tjumatja  
Project: Transfer Learning for OCR Post-Correction; published at WiNLP 2021 [\[7\]](#).

## INVITED TALKS

---

- AI4Bharat Talk Series 2022  
“No Language Left Behind: Unlocking Text Data for Under-Resourced Languages”
- Microsoft Research India NLP Talk Series 2022  
“No Language Left Behind: Unlocking Text Data for Under-Resourced Languages”
- NeuralSpace Platform Launch Event 2022  
“Cross-lingual Transfer for Low-Resource NLP”
- Featured Session, Grace Hopper Conference 2021  
“Digitizing Endangered Language Texts: How NLP Can Help Language Revitalization”
- SIGTYP Lecture Series 2021  
“Cross-Lingual Entity Linking for Low-Resource Languages” [\[video\]](#)
- George Mason NLP Research Group 2021  
“OCR Post-Correction for Endangered Language Texts”
- Language Technologies Institute Colloquium at Carnegie Mellon University 2020  
“Zero-shot Neural Transfer for Cross-lingual Entity Linking” [\[video\]](#)
- University of Utah Data Science Seminar 2020  
“Entity Linking for Low-Resource Languages” [\[video\]](#)
- Microsoft Research India Podcast 2020  
“Building a Career in Research Through the MSR India Research Fellow Program” [\[podcast\]](#)
- Natural Language, Dialog, and Speech Symposium at the New York Academy of Sciences 2019  
“Temporally-Aware Named Entity Recognition”

## TEACHING AND ACADEMIC SERVICE

---

- Teaching Assistant at Carnegie Mellon University  
Machine Translation and Sequence-to-Sequence Models (11-731) Fall 2019  
Search Engines (08-710) Spring 2018  
Data Mining (08-711) Spring 2018
- LTI Diversity, Equity, and Inclusion Committee at CMU, 2021-2022
- Organizer, Workshop on Computational Methods for Endangered Languages at ACL 2022
- Chair, Student Research Workshop at ACL 2020
- Organizer, CMU SCS Graduate Application Support Program, 2020
- Diversity and Inclusion Committee at NAACL 2019
- Mentoring  
CMU Language Technologies Mentoring Program (for new graduate students; 2021), CMU Graduate Application Support Mentor (2020, 2021), CMU AI Mentoring Program (for undergraduates; 2019, 2020, 2021)
- Area Chair  
EMNLP 2022 (Track: Multilinguality)
- Reviewing  
EAAMO 2022, ACL 2022, AAAI 2022, AAAI 2021, ARR 2021, EACL 2021, NAACL 2021, ACL 2021, AmericasNLP 2021, AAAI 2020, HAMLETS 2020, LREC 2020, EMNLP 2020, \*SEM 2020, AACL SRW 2020, AfricaNLP2020, TALLIP 2019, CALCS 2018

## PUBLICATIONS

---

- [1] “Developing Optical Character Recognition for Kwak’wala”  
D. Rosenblum, **S. Rijhwani**, M. King, A. Anastasopoulos, G. Neubig  
*Workshop on Computational Methods for Endangered Languages (ComputEL): Special Session, 2022.*
- [2] “Reading it Right: A Case Study in Pintupi-Luritja”  
S. Disbray, B. Foley, **S. Rijhwani**, M. Mistica  
*Digital Approaches to Multilingual Text Analysis, 2022.*
- [3] “Lexically-Aware Semi-Supervised Learning for OCR Post-Correction” [\[pdf\]](#)  
**S. Rijhwani**, D. Rosenblum, A. Anastasopoulos, G. Neubig  
*Transactions of the Association for Computational Linguistics (TACL), 2021.*
- [4] “MasakhaNER: Named Entity Recognition for African Languages” [\[pdf\]](#)  
D. I. Adelani *et al.*, including **S. Rijhwani**  
*Transactions of the Association for Computational Linguistics (TACL), 2021.*
- [5] “Evaluating the Morphosyntactic Well-formedness of Generated Texts” [\[pdf\]](#)  
A. Pratapa, A. Anastasopoulos, **S. Rijhwani**, A. Chaudhary *et al.*  
*Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.*
- [6] “Dependency Induction Through the Lens of Visual Perception” [\[pdf\]](#)  
R. Su, **S. Rijhwani**, H. Zhu, J. He, X. Wang, Y. Bisk, G. Neubig  
*Conference on Computational Natural Language Learning (CoNLL), 2021.*
- [7] “Explorations in Transfer Learning for OCR Post-Correction” [\[pdf\]](#)  
L. Tjuatja, **S. Rijhwani**, G. Neubig  
*Fifth Widening Natural Language Processing Workshop (WiNLP), 2021.*
- [8] “OCR Post-Correction for Endangered Language Texts” [\[pdf\]](#)  
**S. Rijhwani**, A. Anastasopoulos, G. Neubig  
*Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.*

- [9] “Soft Gazetteers for Low-Resource Named Entity Recognition” [\[pdf\]](#)  
**S. Rijhwani**, S. Zhou, G. Neubig, J. Carbonell  
*Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- [10] “Temporally-Informed Analysis of Named Entity Recognition” [\[pdf\]](#)  
**S. Rijhwani** and D. Preotiuc-Pietro  
*Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- [11] “Improving Candidate Generation for Low-resource Cross-lingual Entity Linking” [\[pdf\]](#)  
S. Zhou, **S. Rijhwani**, J. Wieting, J. Carbonell, G. Neubig  
*Transactions of the Association for Computational Linguistics (TACL)*, 2020.
- [12] “AlloVera: A Multilingual Allophone Database” [\[pdf\]](#)  
D. R. Mortensen, X. Li, P. Littell, A. Michaud, **S. Rijhwani**, A. Anastasopoulos, *et al.*  
*Language Resources and Evaluation Conference (LREC)*, 2020.
- [13] “Damaged Type and Areopagitica’s Clandestine Printers” [\[pdf\]](#) [\[press coverage\]](#)  
C. N. Warren, P. Williams, **S. Rijhwani**, M. G’Sell  
*Milton Studies*, 2020.
- [14] “A Summary of the First Workshop on Language Technology for Language Documentation and Revitalization” [\[pdf\]](#)  
G. Neubig, **S. Rijhwani**, A. Palmer, J. MacKenzie, H. Cruz, X. Li, M. Lee *et al.*  
*First Joint SLTU and CCURL Workshop*, 2020.
- [15] “Practical Comparable Data Collection for Low-Resource Languages via Images” [\[pdf\]](#)  
A. Madaan, **S. Rijhwani**, A. Anastasopoulos, Y. Yang, G. Neubig  
*Practical Machine Learning for Developing Countries Workshop (PML4DC)*, 2020.
- [16] “Zero-shot Neural Transfer for Cross-lingual Entity Linking” [\[pdf\]](#)  
**S. Rijhwani**, J. Xie, G. Neubig, J. Carbonell  
*Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019.
- [17] “Choosing Transfer Languages for Cross-Lingual Learning” [\[pdf\]](#)  
Y. Lin, C. Chen, J. Lee, Z. Li, Y. Zhang, M. Xia, **S. Rijhwani**, J. He *et al.*  
*Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- [18] “Towards Zero-resource Cross-lingual Entity Linking” [\[pdf\]](#)  
S. Zhou, **S. Rijhwani**, G. Neubig  
*Second Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo)*, 2019.
- [19] “Parser Combinators for Tigrinya and Oromo Morphology” [\[pdf\]](#)  
P. Littell, T. McCoy, N. Han, **S. Rijhwani**, Z. Sheikh, D. Mortensen, T. Mitamura, L. Levin  
*Language Resources and Evaluation Conference (LREC)*, 2018.
- [20] “Estimating Code-Switching on Twitter with a Novel Generalized Word-Level Language Detection Technique” [\[pdf\]](#)  
**S. Rijhwani**, R. Sequiera, M. Choudhury, K. Bali, C. S. Maddila  
*Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.
- [21] “Does the Geometry of Word Embeddings Help Document Classification? A Case Study on Persistent Homology-Based Representations” [\[pdf\]](#)  
P. Michel\*, A. Ravichander\*, **S. Rijhwani**\*  
*Second Workshop on Representation Learning for NLP*, 2017.
- [22] “Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages” [\[pdf\]](#)  
M. Yoder, **S. Rijhwani**, C. Rosé, L. Levin  
*Second Workshop on NLP and Computational Social Science*, 2017.

- [23] “Understanding Language Preference for Expression of Opinion and Sentiment: What do Hindi-English Speakers do on Twitter?” [\[pdf\]](#)  
K. Rudra, **S. Rijhwani**, R. Begum, K. Bali, M. Choudhury, N. Ganguly  
*Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- [24] “Experiments with Cross-lingual Systems for Synthesis of Code-Mixed Text” [\[pdf\]](#)  
S. Sitaram, S. K. Rallabandi, **S. Rijhwani**, A. W. Black  
*Ninth ISCA Speech Synthesis Workshop (SSW)*, 2016.
- [25] “Translating Codemixed Tweets: A Language Detection Based System”  
**S. Rijhwani**, R. Sequiera, M. Choudhury, K. Bali  
*Third Workshop on Indian Language Data Resource and Evaluation: System Demonstrations*, 2016.