Sebastin **Santy**

AI Center Fellow, Microsoft Research

sebast.in @ hi@sebast.in github.com/sebastinsanty Google Scholar

1023, Microsoft Research, #9 VIGYAN, Lavelle Road, Bangalore, KA, India 560001

Education

Dec 2018 | **Birla Institute of Technology and Science (BITS) Pilani** — **Goa, India**
Aug 2015 | B.E. (Hons.), Electronics & Instrumentation (Graduated Early)

Experience

Present / **Jul 2019** | **Microsoft Research | Center for Societal Impact through Cloud & AI** [🌐] — **Bangalore, India**
AI Center Fellow | Advisors: Dr. Kalika Bali, Dr. Monojit Choudhury, Tanuja Ganu
Worked on "Language Technologies for All" – on understanding the disparity between research and deployment especially for low resource languages and developing technologies to mitigate this gap.

Jan 2019 | *Research Intern | Advisors: Dr. Kalika Bali, Dr. Monojit Choudhury*
Developed Interactive Neural Machine Translation which assists translators with suggestions on-the-fly with the aim to bring forth a synergy between the pros of human and machine translation.

Dec 2018 / **Jul 2018** | **Carnegie Mellon University | School of Computer Science** — **Pittsburgh, USA**
Research Intern (Bachelor Thesis) | Advisor: Prof. David Touretzky
Worked on real-time obstacle collision avoidance and doorway navigation with limited sensor capacity of monocular vision and low computational constraints.

Jun 2018 / **May 2018** | **University College London | Web Intelligence Group** [🌐] — **London, UK**
Summer Research Intern | Advisors: Prof. Emine Yilmaz, Dr. Rishabh Mehrotra
Worked on user tasks and needs understanding specifically on understanding how incorporating task information can lead to better, more relevant search/retrieval.

Aug 2018 / **May 2018** | **NumFOCUS | Julia Computing** [🌐] — **Remote**
Google Summer of Code 2018 | Mentor: Dr. Lyndon White
Developed DataDepsGenerators.jl, a provenance-preserving data preprocessing tool to generate executable metadata for published datasets. Added support to popular data repositories such as FigShare, Zenodo.

Aug 2017 / **May 2017** | **Mozilla Inc. | Bugzilla Team** [🌐] — **Remote/Austin,USA**
Google Summer of Code 2017 | Mentor: Dylan Hardison
Worked on Bugzilla, a bug-tracking tool to introduce a new single page bug filing interface enter_bug.cgi. This replaced the previous interface which was based on conditional forms that required navigation through multiple pages often making the overall process cumbersome.

Publications

BERTologiCoMix: Probing the Capabilities of Multilingual BERT through Code-Mixing
Sebastin Santy, Anirudh Srinivasan, Monojit Choudhury
[Under Review] [EACL'21]

The State and Fate of Linguistic Diversity and Inclusion in the NLP World [🔗]
Sebastin Santy*, Pratik Joshi*, Amar Budhiraja*, Kalika Bali, Monojit Choudhury (* = Equal Contribution)
Annual Conference of the Association for Computational Linguistics, Seattle, USA [ACL'20]

Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi [🔗]
Sebastin Santy*, Devansh Mehta*, Ramaravind Mothilal, Brij Mohan Lal Srivastava, Alok Sharma, Anurag Shukla, Vishnu Prasad, Venkanna U, Amit Sharma and Kalika Bali (* = Equal Contribution)
International Conference on Language Resources and Evaluation, Marseille, France [LREC'20]

Deploying Language Technologies for Underserved Communities [Invited Poster] [🔗]
Kalika Bali, Monojit Choudhury, Sunayana Sitaram, Sebastin Santy
UNESCO International Conference on Language Technologies for All, Paris, France [LT4All]

Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities [🔗]
P. Joshi, C. Barnes, Sebastin Santy, S. Khanuja, S. Shah, A. Srinivasan, S. Bhattamishra, S. Sitaram, M. Choudhury, K. Bali
16th International Conference on Natural Language Processing, Hyderabad, India [ICON'19]

INMT: Interactive Neural Machine Translation [🔗]
Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, Kalika Bali
Conference on Empirical Methods in Natural Language Processing, Hong Kong [Systems Demo] [EMNLP'19]

CoSSAT: Code-Switched Speech Annotation Tool [🔗]
Sanket Shah, Pratik Joshi, Sebastin Santy, Sunayana Sitaram
Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP, EMNLP, Hong Kong [AnnoNLP@EMNLP'19]

Towards Task Understanding in Visual Settings [🔗]
Sebastin Santy, Wazeer Zulfikar, Rishabh Mehrotra, Emine Yilmaz.
33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA [Student Abstract] [AAAI'19]

DataDepsGenerators.jl: Automatic generation of DataDeps.jl registration code [🔗]
Lyndon White, Sebastin Santy
Journal for Open Source Software [JOSS'18]

BITS Darshini: A Modular, Concurrent Protocol Analyzer Workbench [🔗]
Prasad Talasila, Mihir Kakrambe, Anurag Rai, Sebastin Santy, Neena Goveas, Bharat Deshpande
19th International Conference on Distributed Computing and Networking, Varanasi, India [ACM ICDCN'18]

Select Research Projects

Interactive Neural Machine Translation (INMT) Jan'19 - Present
Advisors: Dr. Kalika Bali, Dr. Monojit Choudhury, Dr. Sandipan Dandapat, Tanuja Ganu

- Worked on understanding how translators can be assisted with suggestions from a machine translation system. On basis of the insights gathered, developed an interactive translation interface to make the translation process quicker and better in terms of quality. [🌐] [EMNLP'19] [Working Paper]
- Engaging with non-profits Translators without Borders, Pratham Books' Story Weaver and CGNet Swara (covered by LiveMint) looking at possible solutions for deploying INMT for low resource languages.
- Developing new interfaces to tailor to specific use-cases of translation such as document translation and web-page localization (using browser extension) and offline translation (INMT lite).

State of Language Technologies for Low-Resource Languages Jan'19 - Present
Advisors: Dr. Kalika Bali, Dr. Monojit Choudhury

- Conducted a quantitative analysis on disparity of language resources being used at NLP conferences. Created data model of publications (authors, papers, institutions) and used statistical measures as well as entity embeddings to classify and track the progress of representation of different languages over conference iterations. [🌐] [ACL'20]
- Understand and track the impact of different language-based technological interventions that were carried out by CGNet Swara in the Gond Community of Chattisgarh. Gondi is a language spoken by 3 million people however is a severely low-resourced language mainly attributed to non-existence of its own script. [LREC'20] [LT4All]
- Surveyed the challenges faced for deployment of language technologies to marginalized communities. [ICON'19]
- Coverage/Mentions - ACL'20: Quartz, NLP Newsletter, ruder.io/nlp-beyond-english, SIGTYP Newsletter | LREC'20: Times of India, Hindustan Times, ETV

Behaviour of Transformer Language Model Attention Heads for Different Stimuli Aug'20 - Present
Advisor: Dr. Monojit Choudhury

- Analyzed how the attention heads of BERT change with fine-tuning. Here, we introduce Code-Mixing to the models and judge the *responsivity* of different heads to them being provided as input. [Under Review]
- Analyzing how different attention heads behave to stimuli that is based on a diverse set of linguistic tasks provided as input to BERT. [Working Paper]

User Tasks and Needs Understanding Jan'18 - Present
Advisors: Prof. Emine Yilmaz, Dr. Rishabh Mehrotra, Prof. Prasanta Bhattacharya

- Designed an ontology of common tasks carried out on a day-to-day basis. These tasks were derived from Wikihow which is collection of "How-To" questions that can be modelled as users seeking answers on how to solve a particular task.
- Showed the use of such a task ontology to produce more coherent image captioning. [AAAI'19]
- In order to adapt to increasing number of new tasks, developed a naive method to dynamically enrich this task ontology using the principles of Bayesian Rose Trees and Chinese Restaurant Processes. [🌐]
- Analyzed which tasks are of interest to different communities around the world. [Working Paper]

Talks

"The State and Fate of Linguistic Diversity in the NLP World"

- NLP with Friends [🌐] [📹] November 2020 (Remote)
- ACL 2020 Presentation [🌐] [📹] July 2020 (Remote)
- ML Invited Authors Series, Ada Support June 2020 (Remote)
- Lab Sabha, Microsoft Research India May 2020 (Remote)

"Repeatable Data Setup for Repeatable Science"

- PyData New York City (NYC) 2018 [🌐] [📹] October 2018 (Microsoft, 11 Times Square, NYC, USA)

Honours and Awards

BITS Alumni Association (BITSAA) Travel Grant, 2019 [🌐] For attending AAAI'19 held in Honolulu, HI, USA.

BITS International Programmes Scholarship, 2018 Awarded for pursuing Summer Internship at UCL, London, UK.

Mozilla All-Hands, 2017 [🌐] Invited as one of 100 contributors to attend this employee meet held at Austin, TX, USA.

Iken Scientifica, 2009 | Student Icon of India [📹] [🌐] Adjudged as Student Icon of India among ∼ 200k participants. Invited for a tête-à-tête with the former President of India, and renowned rocket scientist, Dr. APJ Abdul Kalam at his residence. The entire competition was aired on The National Geographic Channel and was supported by the Ministry of Science And Technology, Government Of India.

National Standard Examination in Chemistry 2015 | State Topper Prelims to International Chemistry Olympiad.

C.Subramaniam Award for Character Excellence, 2013 Among Secondary School students across the school system.

International Computer Fair And Seminar (CoFAS), 2012 | Runner's Up Web design and implementation Round.

Teaching and Leadership Roles

SNLP Reading Group, MSR India *Organizer* Jul'19 - Present

- Organize a weekly lab-wide Reading Group focused on research taking place in the area of Speech and Natural Language Processing (SNLP). We read recent and classical papers as well as arrange for invited talks in related areas.

Open Source Development (OSD) Labs, BITS Goa *Founding Member* [🐙] Aug'16 - Apr'18

- Initiative to foster the open source culture at BITS Goa. Hosted several informative and technical sessions.
- Setup new open source organizations for different academic divisions such as Academic Registration and Counselling Division [🐙], Student Welfare Division [🐙] to enable students to actively contribute to university softwares.

Introduction to Computer Programming (CS F111) *Teaching Assistant* Jan'18 - May'18

- Responsibilities included evaluating labs, and helping students with the coursework and home/lab assignments.

Introduction to Programming, Center for Technical Education (CTE), BITS Goa *Co-Instructor* Dec'17 - Aug'17

- Voluntarily teaching students to object oriented concepts and functional programming using Python.

Waves, Quark, BITSMUN (Festivals at BITS Goa) *Lead Developer* Dec'15 - Dec'17

- Responsible for building and maintaining websites of Waves (cultural fest), Quark (technical fest), BITS Model UN.

Software and Open Source Contributions

- **Microsoft**: Interactive Neural Machine Translation - *commits*, *webpage*
- **Mozilla**: Bugzilla - *commits*, Firefox (Gecko Engine) - *commits*, TreeHerder - *commits*
- **Other External**: scikit-learn - *commits*, DDGenerators.jl - *commits*, Virtual ACL 2020 - *commits*, UserContext - *github*
- **@BITS**: BITS-Darshini - *commits*, SWD - *commits*, ARC - *commits*, Quark 2017 - *commits*, Waves 2016 - *commits*

Academic Service

Reviewer	EACL'21, EMNLP'20 (Demo), MLADS'20, ICON'20
Sub-Reviewer	EMNLP'20, ACL'20, LREC'20, CODS-COMAD'20, CoNLL'19, Interspeech'19, ICON'19
Volunteer	ACL'20, AAAI'19, Panini Linguistics Olympiad (PLO) '19/'20