
Education

- 2022-Present **PhD Computer Science**, *Tokyo Institute of Technology*, Tokyo, Japan
- Advisor: Dr. Naoaki Okazaki
 - MEXT Scholar (文部科学省奨学金)
- 2016-2018 **MS Computer Science**, *Yonsei University*, Seoul, South Korea
- Advisor: Dr. Yo-Sub Han
 - Thesis: *Efficient Algorithms for Two Parsing Problems on Probabilistic Finite Automata*
 - Outstanding International Student Scholarship recipient.
- 2011-2015 **BS Discrete Mathematics**, *Georgia Institute of Technology*, Atlanta, GA
- Minor in Korean
 - Advisor: Dr. Anton Leykin
 - Thesis: *Straight Line Programs and Automatic Differentiation in Python*

Work Experience

- 05/22-Present **PhD Student Researcher**, *Google*, Tokyo, Japan
- Researcher focusing on language modeling and federated analytics on the Gboard team.
- 08/19-04/22 **Software Engineer**, *Google*, Mountain View, CA
- Software engineer on the Gboard team.
 - Developed compressed statistical language models for mobile keyboard input.
 - Developed TensorFlow Federated infrastructure and compressed sketching models for privacy-preserving federated analytics.
 - TA for Google Tech Exchange Applied Data Structures and Algorithms (2020, 2021).
- 02/19-05/19 **Software Engineering Intern**, *Google*, New York City, NY
- Software engineering intern in the Speech and Language Algorithms research group.
 - Developed a compression scheme for finite-state transducers used in keyboard language models which led to a first-author publication.
 - Compression algorithm reduced the space requirements of the Gboard keyboard lexicon data structure by 90% compared to the uncompressed version and 58% compared to the previous production compression scheme.
- 09/16-12/18 **Graduate Teaching Assistant**, *Yonsei University*, Seoul, South Korea
- TA for CSI2103 Data Structures, CSI3108 Algorithm Analysis, CSI3109 Automata and Formal Languages, and CSI6512 Graduate Analysis of Algorithms.
- 01/16-05/16 **Upper School Computer Science Teacher**, *Maclay School*, Tallahassee, FL
- 05/15-08/15 **Data Science Intern**, *AirSage Inc.*, Atlanta, GA
- 05/12-08/12 **Software Development Intern**, *AirSage Inc.*, Atlanta, GA

Publications (*denotes primary authorship)

- Distributional Properties of Subword Regularization**
 - *Marco Cognetta**, Vilém Zouhar, Naoaki Okazaki. EMNLP 2024
- The bread emoji Team's Submission to the IEEE BigData 2024 Cup: Predicting Chess Puzzle Difficulty Challenge**
 - Tyler Woodruff, Oleg Filatov, *Marco Cognetta*. IEEE Big Data 2024
 - Our submission won 1st place and a \$1000 USD prize.
- Two Counterexamples to *Tokenization and the Noiseless Channel***
 - *Marco Cognetta**, Vilém Zouhar, Sangwhan Moon, Naoaki Okazaki. LREC-COLING 2024
- Parameter-Efficient Korean Character-Level Language Modeling**
 - *Marco Cognetta**, Sangwhan Moon, Lawrence Wolf-Sonkin, Naoaki Okazaki. EACL 2023
- SoftRegex: Generating Regex from Natural Language Descriptions using Softened Regex Equivalence**
 - Jun-U Park, Sang-Ki Ko, *Marco Cognetta*, Yo-Sub Han. EMNLP 2019
- On the Compression of Lexicon Transducers**
 - *Marco Cognetta**, Cyril Allauzen, Michael Riley. FSMNLP 2019

7. **Online Infix Probability Computation for Probabilistic Finite Automata**
 - *Marco Cagnetta**, Yo-Sub Han, Soon Chan Kwon. ACL 2019
8. **Incremental Computation of Infix Probabilities for Probabilistic Finite Automata**
 - *Marco Cagnetta**, Yo-Sub Han, Soon Chan Kwon. EMNLP 2018
9. **Online Stochastic Pattern Matching**
 - *Marco Cagnetta**, Yo-Sub Han. CIAA 2018

Preprints (*denotes primary authorship)

1. **Tokenization as Finite-State Transduction**
 - *Marco Cagnetta**, Naoaki Okazaki.
 - <https://arxiv.org/abs/2410.15696>
2. **An Analysis of BPE Vocabulary Trimming in Neural Machine Translation**
 - *Marco Cagnetta**, Tatsuya Hiraoka, Naoaki Okazaki, Rico Sennrich, Yuval Pinter.
 - <https://arxiv.org/abs/2404.00397>

Conference Talks

1. **LotteryTickets.jl: Sparsify Your Flux Models**
 - JuliaCon 2023 (Boston, USA)
 - <https://www.youtube.com/watch?v=ZmcaUyZLi4Q>

Invited Talks

1. **The Tokenization Landscape**
 - Plenary Meeting of the National Institute of Advanced Industrial Science and Technology Artificial Intelligence Research Center's Knowledge and Information Research Team (AIST AIRC-KIRT) (March 2024, Tokyo, Japan)

Service

- **The Gradient** (<https://thegradients.pub/>) - Editorial Board (2021 - Present)
- **Seminars on Formal Languages and Neural Networks (FLaNN)** (<https://flann.super.site/>) - Co-organizer (2022 - Present)
- **Workshop for Natural Language Processing Open Source Software (NLP-OSS)** - Programme Committee (2023)
- **FSU ACM Programming Contest** - Question Writer (2020, 2021^(x2), 2022)
- **Hackbright Academy** - Volunteer Mentor (2020^(x2), 2021)
- **ACM International Collegiate Programming Contest (Korea Regional)** - Question Writer (2017, 2018)

Advising

1. **Gordon Lichtstein** (2024) - *Esperanto Morphological Tokenization*
 - High School Extracurricular Senior Project
2. **Junyoung Lee** (2023) - *Jamo-Level BPE in Korean Machine Translation*
 - Nanyang Technological University (NTU Singapore) Bachelor's Thesis
3. **Emil Hukic** (2022) - *FST Tokenization for NLP*
 - Young Science and Engineering Researchers Program (YSEP) Final Project
4. **Kosuke Endo** (2022) - 画像キャプション生成におけるJPEG圧縮への頑健性の改善
 - English Title: *Improved Robustness to JPEG Compression in Image Caption Generation*
 - Tokyo Institute of Technology Bachelor's Thesis (co-advised with Zhishen Yang)
 - Presented at the Japanese Association for Natural Language Processing conference (NLP2023)
5. **Haksu Kim, Yumin Lim, Myeongjang Pyeon** (2018) *Solving k-MPS using Probabilistic Finite-State Automata*
 - Yonsei University Capstone Project

Skills

- **Programming Languages:** Python, C++, Julia
- **Human Languages:** English, Korean, Esperanto