EMILY CHENG

emilyshana.cheng@upf.edu

EDUCATION

Universitat Pompeu Fabra September 2022-
PhD Candidate in Linguistics
Representational Compression in Neural Language Models
Supervisor: Marco Baroni

Massachusetts Institute of Technology December 2021
Master of Engineering in Computer Science (2021) GPA: 4.7/5.0
Supervisors: Boris Katz and Andrei Barbu
Bachelor of Science in Computer Science and Engineering (2020)
Bachelor of Science in Mathematics (2020)

PUBLICATIONS

1. **Emily Cheng**, Corentin Kervadec, Marco Baroni. Bridging Information-Theoretic and Geometric Compression in Language Models. *Under review*
2. **Emily Cheng**, Mathieu Rita, Thierry Poibeau. On the Correspondence between Compositionality and Imitation in Emergent Neural Communication. *In Findings of ACL 2023.*
3. **Emily Cheng**, Yen-Ling Kuo, Josefina Correa, Ignacio Cases, Boris Katz, and Andrei Barbu. Quantifying the Emergence of Symbolic Communication. *In Proceedings of CogSci 2022.*
4. **Emily Cheng**, Yen-Ling Kuo, Ignacio Cases, Boris Katz, and Andrei Barbu. Towards Modeling the Emergence of Symbolic Communication. Poster in *Proceedings of the ICRA-2021 Social Intelligence Workshop.*

EXPERIENCE

Intrinsic Dimensionality & Representational Compression in Neural Language Models Fall 2022-
Doctoral thesis *Barcelona, Spain*

· Explore relationship between linguistic structure, information-theoretic compression, and intrinsic dimensionality in neural language models.

Compositionality and Imitation Learning in Artificial Emergent Language Spring-Fall 2022
Visiting researcher at ENS Ulm, CNRS *Paris, France*

· Supervised by Thierry Poibeau. Supported by Paris AI Research Institute.
· Explore relationship between compositionality of emergent languages and ease of imitation learning.

Emergent Symbolic Communication in Humans and Machines Fall 2020 - Fall 2021
Master's Research: MIT Infolab *Cambridge, MA*

· Thesis: *Understanding Symbolic Communication*
· Supervised by Boris Katz and Andrei Barbu.
· Characterized the transition from sub-symbolic to symbolic communication between human players, and later machine players via a communication game.

Few-Shot Text Classification with Meta-Learning Spring 2020
MIT Undergraduate Research: Natural Language Processing Group *Cambridge, MA*

· Supervised by Regina Barzilay.
· Extended pipeline for few-shot documentation topic classification in PyTorch to include zero-shot classification baselines.

Reverse-engineering Nanophotonic Systems with BNNs Fall 2018 - Spring 2019
MIT Undergraduate Research: Soljacic Group *Cambridge, MA*

· Supervised by Marin Soljacic.
· Implemented a Bayesian neural network with multiplicative normalizing flows to reverse-design the hyperparameters of nanophotonic systems.

INDUSTRY EXPERIENCE

Palantir Technologies Summer 2020
Software Engineering Intern *Remote*

· Developed insurance risk models using PySpark in Palantir Foundry in collaboration with external clients
· Architected map visualization backend.

Two Sigma Investments Summer 2019
Quant Research Intern: News Team *New York, NY*

· Designed and evaluated alpha models in Python and Groovy to forecast equity and options returns with news data.

Virtu Financial January 2019
Algo Quant Research Intern *New York, NY*

· Developed cross-asset market impact models using Python for cash equities execution.

Goldman Sachs Summer 2018
Securities Research Intern: Equities Flow Vol, FICC SMM Execution Services *New York, NY*

· Developed alpha models in Python to forecast realized volatility for trading single stock options that is in production.
· Designed and built an order fill model for systematic trading simulation in Java to integrate submitted orders with historical market simulation data.

SELECTED PROJECTS

L2 Acquisition and Language Convergence in Neural Language Models Fall 2020
9.190 Group Project

· Conducted cross-lingual transfer and contact language experiments between monolingual French and English LSTM models using toy dataset.
· Found that utterances of monolingual models do not converge, but rather become mutually intelligible.

Cross-Lingual Text-to-Speech Transfer Learning for Low-Resource Languages Spring 2020
6.864 Group Project

· Performed cross-lingual transfer learning on text-to-speech synthesis using German to English single-speaker datasets.
· Designed and conducted Mean Opinion Score tests, finding evidence for optimal periods of transfer from partially trained systems.

Automatic Image Colorization with Semantic Prior Fall 2017
6.867 Group Project

· Created a Keras/TensorFlow machine learning pipeline that predicts a colorized output image given grayscale input and a semantic tag.
· Designed, implemented, and trained the scene classifer and automatic colorizer CNNs, including data preprocessing and postprocessing.

AWARDS

Fulbright France Open Research Grant Semifinalist 2021
Meta-learning in low-resource multilingual generalization *Paris, France*

In collaboration with LATTICE at CNRS and École Normale Supérieure.

TEACHING

6.031 Software Construction Fall 2020
Graduate Teaching Assistant *Cambridge, MA*

· Held lab hours, graded assignments for students in MIT's intermediate Java software class.

Global Teaching Labs January 2020
Instructor *Grenoble, France*

· Taught middle school, high school, and preparatory school students concepts in math, physics, and computer science as part of an MIT STEM outreach program in Grenoble, France.
· Created and carried out lesson plans in both French and English for students aged 8th grade to prépa.

MIT Math Learning Center Fall 2018 - Spring 2019
Teaching Assistant *Cambridge, MA*

· Held twice-weekly office hours for students in the math department taking Differential Equations (18.03), Linear Algebra (18.06), Probability and Random Variables (18.600), Physics (8.01/2) and Calculus (18.01/2).
· Reviewed lecture material and helped students with problem sets and code implementation.

MIT Math Department Fall 2017, Fall 2018
Grader *Cambridge, MA*

· Provided weekly feedback to students on assignments and exams for Probability and Random Variables (18.600) and Statistics (18.650).

ACTIVITIES & OUTREACH

Reviewer
EMNLP, NeurIPS 2023; NeurIPS 2022; NeurIPS, ICLR 2021

COURSEWORK

6.867 Machine Learning (G)	6.860 Statistical Learning Theory (G)	6.337 Numerical Methods (G)
6.435 Bayesian Inference (G)	6.031 Software Construction	18.615 Stochastic Processes (G)
6.864 Natural Language Processing (G)	6.046 Design & Analysis of Algorithms	6.436 Probability Theory (G)
6.884 Sensorimotor Learning (G)	24.933 Semantics & Pragmatics (G)	9.190 Comp. Linguistics (G)

SKILLS

Computer Languages	Python, Java, C/C++
Software & Tools	Pandas/Numpy/Scipy, PyTorch/Keras/TensorFlow, Git, Linux, AWS

LANGUAGES

English (native), Mandarin (fluent), French (C1), Spanish (B1)