Taehee **Jung**

Minneapolis, MA 55414

412-403-8832 | taj41@pitt.edu | theejung.github.io | Google Scholar

Education

University of Pittsburgh — *Pittsburgh, PA, USA*

Ph.D., Department of Statistics — *Aug 2018 - Present*

- Advisor: Lucas Mentch

University of California, Berkeley — *Berkeley, CA, USA*

Master of Arts, Department of Statistics — *Aug 2016 - May 2017*

Korea Advanced Institute of Science and Technology (KAIST) — *Daejeon, Korea*

Bachelor of Arts, Department of Management Science — *2007 - 2011*

- Academic excellence scholarship (2007 - 2011)

Professional Experience

Amazon, Alexa AI — *Seattle, WA*

Applied Scientist Intern — *Summer 2021*

- Collaborators: Sungjin Lee, Joo-kyung Kim
- Applying Calibration method to rescale poor predicted probabilities on Extreme Multi Label Classification Task.

Amazon, Alexa AI — *Seattle, WA*

Applied Scientist Intern — *Summer 2020*

- Collaborators: Liyuan Zhang, Tommy Powers
- Generated user's synthetic utterances for Alexa model training, using bi-directional transformer language models (BERT) and variational autoencoder (VAE).

3M HIS M*Modal, Natural Language Understanding Machine Learning team — *Pittsburgh, PA*

Research NLU Intern — *Summer 2019*

- Collaborators: Thomas Schaaf
- Developed a part-of-speech tagger for medical documents, using recent deep learning techniques (FLAIR).
- Interviewed with 3M Career website with respect to the internship experience [link].

University of Pittsburgh — *Pittsburgh, PA*

Graduate Research Assistant — *Dec 2018 - Present*

(1) Posterior calibration training for classification tasks — *Fall 2019 - Present*

- Collaborators: Thomas Schaaf (3M M*Modal, CMU LTI), Lucas Mentch (UPitt Stats), Dongyeop Kang (CMU LTI)
- Developed a posterior calibration optimization methods for sentence classification tasks . This work was presented in [ACL 2020].
- Currently working for mathematical proof of the proposed calibration method and its extension to the other NLP tasks such as generation and sequence tagging

(2) Analyzing corpus and system biases in current summarization — *Spring 2019*

- Collaborators: Eduard Hovy (CMU LTI), Lucas Mentch (UPitt Stats), Dongyeop Kang (CMU LTI)
- Analyzed how current corpora or summarization systems are biased toward certain sub-aspects (i.e., position of the sentences, importance, and meaning diversity). This work was presented in [EMNLP 2019].

(3) Geofitting: Constraining Relational Semantics into Word Embeddings *Fall 2018*

- Collaborators: Eduard Hovy (CMU LTI), Dongyeop Kang (CMU LTI)
- Anayzed geometric tendencies (e.g., slope, distance) on different set of inter-world relations on existing word vectors (i.e. Word2Vec, Glove).
- Developed an optimization technique called *geo-fitting* that minimizes the geometric differences but keeps the properties from the original word vectors, and showed that the geometry constrained vectors help many downstream NLP tasks (e.g., semantic similarity, textual entailment). This work is submitted to [CoNLL 2020].

University of California, Berkeley *Berkeley, CA*

(1) Empirical Examination on Generative Adversarial Models, ML class project [report] *Feb-May 2017*

- Studied/implemented different types of adversarial generative models: vanilla, deep convolutional, Wassertein, and Cycle GAN.
- Compared them on different tasks: image generation, interpolation, projection, arithmetic operations, translation, and completion.
- Won the best credit among Berkeley's Machine Learning course projects.

(2) S&P 500 Stock Portfolio with Sentiment Analysis, Capstone Project [report] [ppt] *Jan - May 2017*

- Analyzed sentiment of tweets predicting volatility and developed autoregressive predictive models
- Constructed a stock portfolio using Markowitz objective function with l1 penalty

(3) Creating R package for MFA, class project of Statistical Computing [demo] *Nov - Dec 2016*

- Developed R package for Multiple Factor Analysis to perform MFA, bootstrapping and visualized the results on Shiny app

Korea Credit Bureau (KCB) *Seoul, South Korea*

Data Scientist *Dec 2010 - May 2016*

(1) Big, Public Data Analysis, funded by Ministry of Government Administration & Jeonju City *Sep - Dec 2015*

- Analyzed and visualized boarding histories, local floating population, petition documents, and home to office O/D information
- Constructed new strategies for Jeonju city's bus rearrangement plan [article] [report] (written in Korean)

(2) Fraud Detection System Development, funded by Woori Bank *Dec 2014 - Aug 2015*

- Segmented customers' transaction histories and developed predictive ensemble models on regression, neural net, and decision tree
- Compared with previous FDS, overall detection rate increases more than 20% [article] (written in Korean)

(3) Migrant Analysis for Magok New Town, funded by LG Economic Research Institute *Jun - Sep 2014*

- Developed demographic database of migrants and classified them using decision tree for potential cultural demand to the new town

(4) Location based Decision Support System, funded by Korea Credit Bureau & R114 *May - Aug 2012*

- Developed an integrated database (e.g., location, real estate, residential's demographic and financial data) called K-atlas

Publication

WORK in Submission or Progress

Don't Treat Every Relation the Same! Constraining Word Vectors to Geometric Properties of Inter-Word Relations, Taehee Jung*, Dongyeop Kang*, Eduard Hovy (*equal contribution), Submitted to CoNLL 2020

Published Papers

ACL'20 **Posterior Calibrated Training on Sentence Classification Tasks**, Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, Thomas Schaaf, The 58th Annual Meeting of the Association for Computational Linguistics, 2020

EMNLP'19 **Earlier Isn't Always Better: Sub-aspect Analysis on Corpus and System Biases in Summarization**, Taehee Jung*, Dongyeop Kang*, Lucas Mentch, Eduard Hovy (*equal contribution), Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019

Academic Experience

Facebook Women in Research Lean In. *Menlo Park, CA*

Invited for women PhD candidates and researchers *Sep 11 - 13, 2019*

Teaching Assistant at University of Pittsburgh *Pittsburgh, PA*

Statistics & Probability for Business Management *Spring 2019, Spring 2020*

Applied Statistical Methods *Fall 2018, Fall 2020*

Service

2019 **External Reviewer**, Annual Meeting of the Association for Computational Linguistics (ACL)

2019 - 2020 **Lead Organizer**, Statistics and Machine Learning Group, University of Pittsburgh *Pittsburgh, PA*

Honor

2019 **Student Scholarship**, Conference on Empirical Methods in Natural Language Processing (EMNLP) *Hong Kong, China*

Skills

Programming Python, Java, R, SAS, SQL, Unix/Linux, Tableau, QGIS

Language English (Fluent), Korean (Native), Spanish (Beginner)