Haejin Cho

(438) 722-0368 | haejin2909@gmail.com |https://www.linkedin.com/in/haejin-cho-220a3a1b2/ |
https://jinjin-tonic.github.io/ | Piano & Video game & Travel | Montreal, QC

EDUCATION

University of British Columbia Vancouver, BC
Master of Data Science - Computational Linguistics **EXPECTED** 07/2021
Relevant Coursework: NLP; ML; Neural Network; Data Visualization; Corpus Linguistics; Algorithms; Statistics

Yonsei University Seoul, South Korea
Bachelor of Arts - Korean Language and Literature **GRADUATED** 08/2019
GPA: 4.22/4.3; Highest Honours 2018, 2019;
Relevant Coursework: Syntax; Semantics; Phonology; Morphology; Corpus Linguistics

EXPERIENCE

Yonsei Institute of Language and Information Studies Seoul, South Korea
Research Associate 03/2016 – 09/2017

- Participated in publishing the 8th edition of Dong-a Yonsei's Elementary Korean Dictionary
- Analyzed vocabulary frequency used in elementary textbooks to select lexical items for the dictionary
  Advanced skills: Lexicology, Corpus linguistics, MS excel

Susinsa Project Seoul, South Korea
Research Associate 01/2017 – 11/2018

- Contributed to the creation of a database of entities from late 19th century historical Korean/Japanese texts
- Analyzed named entities in texts and established a database of semantic relations between those entities
  Advanced skills: named entity recognition, semantics, XML

PERSONAL PROJECT

COVID-19 Vaccine Sentiment Analysis [Project Link] [GitHub] 12/2020-02/2021

- Achieved an accuracy of 78% with the best model among the trials of different ML models such as LGBM, Logistic Regression, XGBoost and deep learning architectures.
- Deployed an LSTM model to predict sentiments of COVID-19 vaccines related tweets using Heroku and Flask
- Scraped tweets using Tweepy API, manually annotated sentiments of tweets, and preprocessed raw text data using tools in NLTK to perform a supervised learning task
  Advanced skills: FNN, NLTK, LSTM, PyTorch, Tweepy, Sentiment Analysis, Text Annotation, Flask, Heroku

ACADEMIC PROJECTS

Building a Multilingual Parallel Corpus using TED transcripts [Blog Link] 03/2021

- Scraped transcripts from TED talks to create a parallel multilingual corpus using BeautifulSoup
- Using spaCy, automatically extracted named entities in English, Chinese corpora and improved annotation quality by using Amazon Mechanical Turk
- Deployed the parallel corpora with an interface for users using HTML, JS, CSS, and FastAPI
  Advanced skills: Web Scraping, Python, Corpus Linguistics, spaCy, NER, AMT, FastAPI, Docker, Git

High/Medium-Resource Machine Translation Project: French to English, Danish to English

- Achieved 42.85 BLEU-4 score with a seq2seq with additive attention model in the French to English task.
- Preprocessed raw text data using TorchText and spaCy tokenizers, and trained the model with pyTorch.
- Tried different architectures such as transformers via openNMT.
  Advanced skills: Seq2Seq, Machine Translation, spaCy, TorchText, pyTorch, openNMT,

Creating a POS tagger using HMM algorithm 01/2021

- Implemented a semi-supervised POS tagger from scratch using HMM and Viterbi algorithm, achieved 79 % accuracy
- Visualized the result statistics with Altair
  Advanced skills: Semi-supervised Learning, HMM, Viterbi, Python, POS tagging, Data Visualization

ADDITIONAL INFORMATION

- **Technical Skills**: **Python**, Git, R, **PyTorch**, Pandas, Altair, **NLTK**, **spaCy**, MS Excel, OpenNMT, Jupyter Notebook
- **Languages**: Fluent in Korean, English; Conversational Proficiency in Japanese; Elementary Proficiency in French
- **Extra Accredited Courses**: Linear Algebra, Introduction to Calculus at Athabasca University
- **Online Courses and Certificates:** Japanese Language Proficiency Test N1