Lan Sang

lan.sang@colorado.edu | 2300 Arapahoe Ave, Boulder, Colorado, USA

EDUCATION

University of Colorado Boulder Boulder, CO, United States
Ph.D. in Information Systems Expected 2026

University of Colorado Boulder Boulder, CO, United States
M.S. in Computational Linguistics May.2020

Nanjing Normal University Nanjing, China
B.A. in French and BBA in Business Administration Jun. 2018

PROFESSIONAL EXPERIENCE

Nuance Communications, Inc. Boulder, CO, United States
NLP Research Intern, Natural Language Understanding Team Jun. 2019-Aug. 2019

- Mapped annotated data from a Mandarin corpus in one annotation spec to another annotation spec
- Ran mapping rules of English data on Mandarin data to see how the rules performed and found out possible reasons that affected the accuracy of the mapping rules when ran on another language
- Presented to the team about key findings by the end of the internship

University of Colorado Boulder Boulder, CO, United States
Research Assistant, Computational Language and Education Research Center May. 2019-Jan. 2021

- Used html, python (flask framework) and MongoDB to build and maintain the website of United Verbs Index (uvi.colorado.edu), which is a system merges links and web pages from five different NLP projects: VerbNet, FrameNet, PropBank, OntoNotes and SynSemClass Lexicon
- Edited html and css files to improve the functions and user interface of the UVI website as well as updating VerbNet

RESEARCH EXPERIENCE & PROJECTS

SIGMORPHON 2020 Shared Task: Typologically Diverse Morphological Inflection Boulder, CO, United States
Course Project Mar.2020-May.2020

- Wrote grammar rules to generate lemmas, part-of-speech tags and word forms for three languages: Crimean Tatar (crh), Tagalog (tgl) and Livonian (liv)
- Built Finite-State Transducer (FST) using Foma by combining a lexicon-based model with a guesser to handle unseen lemmas and got accuracies of 96.38%(crh), 68.84%(liv) and 78.01%(tgl) on test datasets.

Topic Modeling Using Unsupervised Learning Models on User Review Dataset Boulder, CO, United States
Course Project Nov.2019

- Clustered customer reviews into groups and discovered the latent semantic structures using Python
- Preprocessed text dataset of Amazon user reviews by tokenization, stemming and extracted features by Term Frequency-Inverse Document Frequency (TF-IDF)
- Trained unsupervised machine learning models of K-Means Clustering and Latent Dirichlet Allocation (LDA)
- Visualized and analyzed the model training results

SKILLS

- Background: Computational Linguistics | Natural Language Processing | Machine Learning | Deep Learning
- Programming: Python | Java | SQL | JavaScript
- Tools: MySQL | Spark | Tensorflow | Keras | PyTorch | AWS | Tableau | Git | Praat | NLTK | Scikit-learn | NumPy
- Languages: English (Fluent) | Mandarin (Native) | French (Fluent)

PUBLICATIONS

- Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, **Lan Sang**, Mans Hulden et al(2020), "Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars." In *Proceedings of 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Pages 162 – 170, Association for Computational Linguistics