SIJIA GE

+1(720) 530-7370 ◇ Boulder, CO ◇ sijiage007@gmail.com ◇ konic-nlp.github.io

SUMMARY

A self-driven NLPer with a proven ability to proficiently pre-process, and model diverse data. Experienced in **machine learning, deep learning, and software development**. Analytical problem solver with demonstrated skills in engineering innovative solutions. A supportive collaborator who works well internally and externally to gain consensus and implement solutions. Eager to master and enhance comprehensive skills with an interdisciplinary background in linguistics and computer science. Passionate about helping companies continue on a good path and increase profit margins.

Research experience includes **named entity recognition**, **text classification**, **semantic role labeling, and ontology application**; Industry experience involves **products design, UX/UI design**.
Urgently seeking internship & entry-level full-time job in **NLP engineer, Machine Learning Engineer, Data scientist, and Software engineer**.

EDUCATION

MS in Computational Linguistics, University of Colorado-Boulder |**GPA: 3.95/4.0** Aug 2021-Dec 2023
MA in Computational Linguistics, Nanjing Normal University |**GPA: 86.3/100** Sep 2016-Jun 2019
BA in Chinese Literature and Linguistics, Shanxi University |**GPA: 85.8/100** Sep 2012-Jul 2016

SKILLS

Technical Python, Keras, Pytorch, Java, NLTK, Scikit-learn, Transformers, Tensorflow, SQL, Django, Spacy, Flask, HTML, Axure RP, JavaScript, JQuery, Linux, Shell, SPSS, LaTex, Git, R, PlantUML, Neo4j, Tableau, Heroku, Numpy, Pandas, Scipy, Google Cloud Platform, Docker
Languages Mandarin Chinese, English

PROJECTS

Bike shared count Kaggle competition

- leveraged exploratory data analysis, feature engineering, feature transformation & normalization.
- applied XGBoost with parameter grid search into the model.
- ranked #3/52 in Public leaderboard with $R^2 = 0.935$ and #9/52 in Private Leaderboard with $R^2 = 0.939$.

A comparison in terms of image classification performance between a large model and a light-weighted model with denoising

- compared the performance of the image classification task on the CIFAR-10 dataset between a vanilla CNN and a CNN including ResNet-50 pre-trained models.
- the experiments showed that after preprocessing like Deep-Image-Prior(DIP) on the noisy data, the performance on a vanilla CNN can reach a similar level to that on a complicated CNN leveraging ResNet-50 without denoising. The vanilla CNN trains faster, with fewer weights.

UMR-writer: An efficient annotation tool for Uniform Meaning Representation(UMR)

- devise a typing interface, supporting frame lexicon display on the fly.
- add multiple events listeners and DOM operations that increased the annotation efficiency by **60%**.
- efficiently promote the UX and UI, covering multiple languages such as Arapahoe, Navajo, Kukama, Arabic, etc.

E-commerce online shopping website

- implemented the fundamental functions of an online shopping website, such as user account management, product display, edit cart, and checkout.
- built on Django 2.0, adopted JQuery & JavaScript to interact with the users at the front end.
- adopted SQLite as the model and built-in admin module as the backstage management system.

Viz-Wiz Visual Question & Answer Challenge: Answer Visual Questions from People Who Are Blind

- adopted the feature map on the next to last layer of the VGG-16 model as the image features.
- adopted the hidden state of the last layer of BERT as question features.

- adapted the challenge to a classification task and classified each image-question pair to the first 50000 frequent answers via concatenated features and got a score of 0.47.

Named Entity Recognition for Gene text

- initiated named entity recognition in the biomedical area with CRF, Bi-LSTM-CRF, and pre-trained language model(BioBERT and reached entity-based F-scores of 62%, 62%, 77%, and 85% respectively.
- leveraged **comet_ml** as a tool for training visualization

Music Store Simulation

- programmed the routine of a music store, including ordering new items, selling items, buying items, and so on.
- applied design patterns to the project, such as strategy, decorator, observer, singleton, etc.
- adopted Junit test module facilitated the robustness.

EXPERIENCE

Student Research Assistant May 2022 - present
CLEAR lab, University of Colorado-Boulder *Boulder, CO*

- compute the similarity across events whose participants and events are mapped to the **WikiData** knowledge graph using rule-based and machine-learning algorithms.
- promote the flask-based online annotation tool with a better user experience interface.

Product Manager Intern Mar 2021 - Jun 2021
Beijing Lingosail Tech Co., Ltd. *Beijing, China*

- designed the prototype of a product targeted for computer-aided translation users, including basic UI, function specifications, and interaction; arranged the overall progress of development as well as the profit model;
- tested the implemented specifications with black box testing.

PUBLICATIONS

- Sijia Ge. 2022. Integration of Named Entity Recognition and Sentence Segmentation on Ancient Chinese based on Siku-BERT. In Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, pages 167–173, Taipei, Taiwan. Association for Computational Linguistics.
- Ning Cheng, Bin Li, Liming Xiao, Changwei Xu, Sijia Ge, Xingyue Hao, and Minxuan Feng. 2020. Integration of Automatic Sentence Segmentation and Lexical Analysis of Ancient Chinese based on BiLSTM-CRF Model. In Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 52–58, Marseille, France. European Language Resources Association (ELRA).
- Song, L., Wen, Y., Ge, S., Li, B., Qu, W. (2020). An Easier and Efficient Framework to Annotate Semantic Roles: Evidence from the Chinese AMR Corpus. In: Hong, JF., Zhang, Y., Liu, P. (eds) Chinese Lexical Semantics. CLSW 2019. Lecture Notes in Computer Science, vol 11831. Springer, Cham.

EXTRA-CURRICULAR ACTIVITIES

- Peer mentor for international students and linguistic department first-year master students during the 2022-2023 academic year. offering help, answering questions, and directing them to the resources to help new students adapt to campus life.
- Course Manager for CSCI 5622 Machine Learning, mainly focus on facilitating writing auto-grading scripts, performing sanity checks, and canvas page management.
- volunteers for academic conferences, Boulder international film festival, etc. Tasks involve conference registration, attendees' hotel booking, Wifi support and stage production, overall coordination and organizing.