SIJIA GE

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

SUMMARY

Proactive NLPer with a strong track record in adept data preprocessing and modeling. Proficient in **machine learning, deep learning, and software development**. A collaborative team player who drives consensus and impactful solution implementation, thriving in diverse environments. My interdisciplinary background in linguistics and computer science fuels my eagerness to enhance my skill set, making a positive impact on companies' growth and profitability.
I bring hands-on expertise in **named entity recognition, text classification, semantic role labeling, and ontology application**, backed by industry insights in **product design and UX/UI design**.
Currently, I'm actively pursuing internships and entry-level full-time roles like **NLP Engineer, Machine Learning Engineer, Data Scientist, or Software Engineer**.

EDUCATION

MS in Computational Linguistics, University of Colorado-Boulder |**GPA: 3.969/4.0** 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

- Employed exploratory data analysis, meticulous feature engineering, and effective feature transformation and normalization techniques. Implemented XGBoost with a comprehensive parameter grid search, optimizing model performance.
- Achieved a commendable rank of 3/52 on the Public Leaderboard with an impressive $R^2 = 0.935$, and secured a competitive 9/52 position on the Private Leaderboard with a solid $R^2 = 0.939$.

Comparative Study: Image Classification Performance of Large vs. Lightweight Models with Denoising

- Conducted an in-depth analysis of image classification performance using the CIFAR-10 dataset, contrasting a standard Convolutional Neural Network (CNN) with a CNN enhanced by a pre-trained ResNet-50 model.
- Demonstrated that by employing preprocessing techniques such as Deep-Image-Prior (DIP) on noisy data, the performance of a vanilla CNN—characterized by faster training speed and fewer parameters—can attain a comparable level to that of a complex ResNet-50 model without denoising.

UMR-Writer: An Efficient Annotation Tool for Uniform Meaning Representation (UMR)

- Devised an innovative keyboard interface featuring real-time display of frame lexicons, streamlining node addition from five steps to a single editing command, resulting in a remarkable **60% increase in annotation efficiency.**
- Implemented multiple event listeners and DOM operations to significantly enhance the user experience (UX) and UI.
- Successfully extended tool functionality to encompass various languages including Arapahoe, Navajo, Kukama, Arabic, among others. Additionally, the tool demonstrated its utility in annotating clinical data (THYME).

E-commerce Online Shopping Website

- Successfully developed core functionalities of an online shopping platform, encompassing user account administration, product presentation, cart editing, and secure checkout mechanisms.
- Created the platform using Django 2.0, integrating JQuery and JavaScript to facilitate engaging user interactions on the frontend. Implemented SQLite as the underlying database model and leveraged the built-in admin module for efficient backend management.

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

- Utilized feature maps from the penultimate layer of the VGG-16 model to extract image features; employed the hidden state of the final layer of BERT to capture question features.

- Transformed into a classification task by categorizing each image-question pair among the top 50,000 frequently occurring answers. This was achieved by combining the extracted features, resulting in a score of 0.47.

Advanced Named Entity Recognition for Genetic Text

- Spearheaded the implementation of named entity recognition techniques within the biomedical domain, employing CRF, Bi-LSTM-CRF, and a pre-trained language model—BioBERT, which was pretrained on PubMed data. Achieved impressive entity-based F-scores of 62

Music Store Simulation

- Developed a comprehensive music store simulation, encompassing tasks like inventory management, order processing, sales transactions, and more.
- Applied advanced design patterns—strategy, decorator, observer, singleton—effectively enhancing the project's architectural robustness and maintainability. used Junit testing to ensure the reliability and quality of the simulation.

EXPERIENCE

Student Research Assistant May 2022 - Present
CLEAR Lab, University of Colorado-Boulder *Boulder, CO*

- Conducting research to compute event similarity across participant and event mappings in the **WikiData** knowledge graph. Utilizing a linear programming algorithm to align likely matched pairs and employing machine learning models to determine match significance (Ongoing).
- Enhancing the flask-based online annotation tool by redesigning the user experience interface. Reduced the annotation process from multiple steps to a single command, catering to multiple languages.

CSCI 5622 Machine Learning 23S Course Manager Jan 2023 - May 2023
Department of Computer Science, University of Colorado-Boulder *Boulder, CO*

- Assisted Teaching Assistants (TAs) in script development for autograding and code testing. Created assignment rubrics for the course.
- Managed the gradebook, organized and posted student grades, and addressed daily logistical inquiries.
- Provided timely responses to student queries, coordinated canvas page management with instructors.

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

- Designed the prototype for a computer-aided translation product, covering UI, functional specifications, and user interactions as well as the testing. Led development progress and devised the profit model. Achived 40% profit and 20% active users higher than the expectation.

PUBLICATIONS

- Sijia Ge, Jin Zhao, Kristin Wright-bettner, Skatje Myers, Nianwen Xue, and Martha Palmer. 2023. UMR-Writer 2.0: Incorporating a New Keyboard Interface and Workflow into UMR-Writer. In Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 211–219, Toronto, Canada. Association for Computational Linguistics.
- 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

- Served as a peer mentor for international students and first-year master's students throughout the 2022-2023 academic year. Volunteering for CCL 2017 in Nanjing, DSNA 24 in Boulder, Boulder International Film Festival, and the 61st ACL conference, etc.