Yunyan Duan

https://www.linkedin.com/in/yunyan-duan/

Email : yyduan.pku@gmail.com
Mobile : (847)-596-1503

Education

Northwestern University Evanston, IL
PhD candidate, ***Linguistics*** *Sep. 2015 – Dec. 2020 (expected)*

- Research area: computational linguistics, psycholinguistics, reading, eye tracking

Peking University Beijing, China
Bachelor of Science, ***Statistics****; Bachelor of Science,* ***Psychology*** *(Double degree)* *Jul. 2013*

Work Experience

Tencent Remote
Internship, Data Scientist *Jul. – Aug. 2020*

- Implemented text correction pipelines with language models and neural network models to correct errors in Mandarin speech-to-text transcriptions
- Analyzed common speech-to-text errors in 1800 documents with 8 million tokens by annotating character-level and word-level errors, and examined error patterns in different genres
- Summarized error patterns in terms of phonetic rules, and evaluated the performance of different pipelines
- Wrote a report with solid background introduction to research papers on text correction and detailed analyses of the dataset, providing insights for reducing speech-to-text models' word error rate

Google Los Angeles, CA
Internship, Analytical Linguist *Jun. – Sep. 2019*

- Researched user-task interactions to identify meaningful signals predictive of the quality of crowd-sourcing labels
- Collaborated with software engineers and data scientists to design data processing pipeline
- Analyzed 100+ tasks with a wide variety of content, each task consisting of 10-200 questions and completed by 100-500 raters, to identify tracking events that predict label quality and consistency across raters
- Developed an R package and Python scripts to automatically query data from database, extract predictive metrics, visualize typical event trajectories, and generate an analytical report for every task
- Presented key findings to the team, which came into a product that impacts internal uses of crowd-sourcing labels

Kellogg School of Management, Northwestern University Evanston, IL
Research Assistant *Aug. 2018 – Feb. 2019*

- Part-time research assistant collaborated with professors to help data collection for research projects
- Wrote Python scripts to crawl sports competition results from 5 websites and organized the crawled data
- Wrote Python scripts to extract text data from pdf files and integrate datasets using fuzzy string matching

Academic Journal Publications

1. Chang, W., **Duan, Y.**, Qian, J., Wu, F., Jiang, X., & Zhou, X. (2020). Gender interference in processing Chinese compound reflexive: Evidence from reading eye-tracking. *Language, Cognition and Neuroscience.* 1-16.
2. **Duan, Y.**, & Bicknell, K. (2019). A rational model of word skipping in reading: Ideal integration of visual and linguistic information. In *Proceedings of the 41th Annual Conference of the Cognitive Science Society*: 275-281. ***Winner of best Computational Modeling paper in Perception & Action.***
3. **Duan, Y.**, & Bicknell, K. (2017). Refixations gather new visual information rationally. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*: 301-306.
4. Yu, H., **Duan, Y.**, & Zhou, X. (2017). Guilt in the eyes: Eye movement and physiological evidence for guilt-induced social avoidance. *Journal of Experimental Social Psychology, 71*, 128-137.
5. ***Duan, Y.**, & *Wu, O. (2016). Learning with auxiliary less-noisy labels. *IEEE Transactions on Neural Networks and Learning System, 28(7)*, 1716-1721. (* indicates equal contributions.)
6. Luo, Y., **Duan, Y.**, & Zhou, X. (2015). Processing rhythmic pattern during Chinese sentence reading: an eye movement study. *Frontiers in Psychology, 6*, 1881.
7. Wang, L., **Duan, Y.**, Theeuwes, J., & Zhou, X. (2014). Reward breaks through the inhibitory region around attentional focus. *Journal of Vision, 14(12):2*, 1-7.

Conference Presentations (excluding those with proceedings)

1. Sheng, Y., **Duan, Y.**, & Wu, F. Corpus-based analysis of complement coercion in Mandarin Chinese. Poster presentation at the 25th annual conference on Architecture and Mechanisms for Language Processing (AMLaP), Moscow, Russia, 6-8 September 2019.
2. **Duan, Y.**, Berzak, Y., Bicknell, K., & Levy, R. Inferring sentence comprehension from eye movements in reading. Poster presentation at the 32nd annual CUNY Conference on Human Sentence Processing, University of Colorado, Boulder, Colorado, 29-31 March 2019.
3. **Duan, Y.**, & Bicknell, K. (2019). A rational model of word skipping in reading: ideal integration of visual and linguistic information. Poster presentation at the 32nd annual CUNY Conference on Human Sentence Processing, University of Colorado, Boulder, Colorado, 29-31 March 2019.
4. **Duan, Y.**, & Bicknell, K. (2016). Word identification in reading is constructive: Refixations seek new visual information. Poster presentation at the 22nd annual conference on Architecture and Mechanisms for Language Processing (AMLaP), Bilbao, Spain, 1-3 September 2016.
5. Hu, J., Liu, J., **Duan, Y.**, Zhao, C., Gong, X., Xiang, Y., Jiang, C., & Zhou, X. (2014). Resting-state functional connectivity indexes emotion recognition bias. Poster presentation at the 20th Annual Meeting of the Organization for Human Brain Mapping (OHBM), Hamburg, Germany, 8-12 June 2014.
6. **Duan, Y.**, Yu, H., & Zhou, X. (2014). Avoiding eyes reveals guilty heart: An eye movement study on interpersonal guilt. Poster presentation at the 6th Chinese International Conference on Eye Movements (CICEM), Beijing, China, 5-9 May 2014.
7. *Feng, W., ***Duan, Y.**, Luo, Y., & Zhou, X. (2013). When language hurts you: Aggression provoked by rhetorical questions. Poster presentation at the 1st Brain Research Symposium by PKU-IDG/McGovern Institute, Beijing, China, 20-21 August 2013.

Research Projects

Rational models of eye movements in reading

Research Assistant at Department of Linguistics, Northwestern University, Evanston, IL *Sep. 2015 – present*

- To understand how eye movement control and word identification work in reading, we use rational analysis with computational modeling to study how different sources of information (visual, linguistic, and contextual) act interactively to identify words and influence eye movement decisions.
- Techniques: *Bayesian inference, language models, reinforcement learning*

Inferring sentence comprehension from eye movements in reading

Visiting Student at Department of Brain and Cognitive Sciences, MIT, Cambridge, MA *Aug. 2018 – Dec. 2018*

- Although psycholinguistic experiments suggest a strong eye-mind link, the evidence is mostly based on aggregated data. We predict readers' comprehension of individual sentences from their eye movements using machine learning and neural networks, and evaluate how well the predictive models generalize to new readers and new sentences.
- Techniques: *Machine learning, convolutional neural network (CNN), language models, model selection*

Natural language processing for predicting readmission in pediatric ICU

Research Assistant at Feinberg School of Medicine, Northwestern University, Chicago, IL *Sep. 2016 – Aug. 2017*

- Useful information for predicting readmission after pediatric ICU hospitalization can come from electronic health records, especially social workers' notes. We formalize the idea by extracting text features from this unstructured data and implementing classifiers to predict readmission probability.
- Techniques: *Classification with unbalanced classes, natural language processing (bag-of-words model, sentiment analysis)*

Learning with auxiliary less-noisy labels

Research Intern at Institute of Automation, Chinese Academy of Sciences, Beijing, China *Apr. – Sep. 2014*

- We propose a learning method that considers both noisy labels and auxiliary less-noisy labels available in a small portion of the training data. Our method outperforms traditional classifiers that do not explicitly consider the auxiliary less-noisy labels.
- Techniques: *Matlab, ExpectationMaximization algorithm, logistic regression, crowd-sourcing*

Skills

Programming languages: Python (5+ years), R (5+ years), SQL (2+ years)
Machine learning: Supervised learning (e.g. classification, regression), unsupervised learning (e.g. clustering)
Natural language processing: Text analysis, sentiment analysis, neural network models
Python Packages & tools: Scikit-learn, Numpy, Pandas, NLTK, Stanford CoreNLP
Other computer skills: Git, command line, AWS (EC2)
Linguistics: Praat, IPA, phonetics and phonology, syntax, part-of-speech annotation, semantics
Statistics: Linear-mixed models, logistic regression, cluster analysis, Bayesian inference, etc.

Online Learning

Coursera https://www.coursera.org/
Certificate: ***Reinforcement Learning*** *Specialization* *Mar. 2020*

- Learned the space of RL algorithms, how to build a RL system for sequential decision making, how to formalize a task as a Reinforcement Learning problem, and how to begin implementing a solution.

Coursera https://www.coursera.org/
Certificate: ***Deep Learning*** *Specialization* *Nov. 2018*

- Through 5 courses, developed a profound knowledge of deep learning from its foundations (neural networks) to its advanced techniques and industry applications (convolutional neural networks, recurrent neural networks, etc.).

Stanford Online https://lagunita.stanford.edu/
Statement of Accomplishment: ***Mining Massive Datasets*** *Jul. 2017*

- This course covers “big-data” algorithms, including PageRank, stream algorithms, clustering, social-network graph analysis, large-scale machine learning, recommendation systems, computational advertising, etc.

Awards & Fellowships

Successful Participants of Mathematical Contest in Modeling (MCM), 2013
First-class Award in Beijing district in China Undergraduate Mathematical Contest in Modeling, 2011
Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment (CURE), 2011
Second-class Freshman Scholarship, Peking University, 2009

Part-time Projects

My personal Github page: https://yyd27.github.io/
Individual contributor *Apr. 2017 – present*

- Build a personal static website using Jekyll
- Maintain the website, update website regularly, share blog posts, and keep up with security updates

Architecture highlights in Shanghai Shanghai library open data challenge
Team lead *May – Aug. 2019*

- Led a team of 7 to develop a website featuring architectures of historical importance in Shanghai
- Developed back-end code (implemented in Python/Django) to categorize architectures based on text descriptions
- Managed weekly updates, participated in discussion of product design, and prepared final presentation

Word evolution in ancient Chinese poems Shanghai library open data challenge
Individual contributor *May – Aug. 2018*

- Developed a website aiming to help researchers gain insights into word evolution, style change, and social evolution reflected in ancient Chinese poems over hundreds of years
- Independently came up with the idea, designed features, developed applications, and wrote documentation
- Implemented website using Python/Django and visualized data patterns using R Shiny