RUI DONG

85 Franklin Street, Apt 3, Brookline, MA 02445 ⋄ (617) · 992 · 4676 ⋄ dongrui@ccs.neu.edu

EDUCATION

Northeastern University *2014.09-present*
Ph.D. candidate, College of Computer and Information Science
Fudan University *2011.09-2014.07*
M.sc., School of Computer Science
East China Normal University *2007.09-2011.07*
B.Eng., School of Information Science and Technology

RESEARCH INTEREST

My research interests are in natural language processing and machine learning. I am currently focusing on text error correction and noisy language modeling based on neural network models.

PUBLICATION

okralact – a multi-engine Open Source OCR training system
Konstantin Baierer, **Rui Dong**, Clemens Neudecker. *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing* (HIP 2019), September 2019, Sydney, Australia.

Noisy Neural Language Modeling for Typing Prediction in BCI Communication.
Rui Dong, David A Smith, Shiran Dudy, Steven Bedrick. *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies* (SLPAT 2019), June 2019, Minneapolis, United States.

Multi-Input Attention for Unsupervised OCR Correction.
Rui Dong, David A. Smith. *Association for Computational Linguistics* (ACL 2018), July 2018, Melbourne, Australia.

Weakly-Guided User Stance Prediction via Joint Modeling of Content and Social Interaction.
Rui Dong, Yizhou Sun, Lu Wang, Yupeng Gu, Yuan Zhong. *International Conference on Information and Knowledge Management* (CIKM 2017), November 2017, Singapore, Singapore.

A Probabilistic Approach to Latent Cluster Analysis.
Zhipeng Xie, **Rui Dong**, Zhengheng Deng, Zhenying He, Weidong Yang. International Joint Conference on Artificial Intelligence (IJCAI 2013), August 2013, Beijing, China.

RESEARCH EXPERIENCE

Table-based Fact Verification *2019-present*
I am working with Prof. David A. Smith to design a model for table-based fact verification. We aimed at improving the verification accuracy for the statements involving numerical reasoning over table cells.

Predictive Typing for Brain-Computer Interface System *2018-2019*
I worked with Prof. David A. Smith to design a neural language model for the typing module of a Brian-Computer Interface designed for patients with Lock-In Syndrome. We investigated how to improve the performance of language models on typing prediction task given the noisy output from BCI systems. This work has been accepted by SLPAT, 2019.

Optical Character Recognition Post-correction *2017-2018*

I worked with Prof. David A. Smith to design an unsupervised model for OCR post-correction that exploits repeated texts in large corpus. An attention-based Seq2Seq model is applied as the correction model and different attention combination strategies are introduced to jointly align, correct and vote among duplicate texts. This work has been accepted by ACL, 2018.

Argument Clustering *2016-2017*

I worked with Prof. Lu Wang to design a deep clustering model to uncover the argumentative facets of different topics according to online discussions. A hierarchical attention-based model is applied to learn the representation for each sentence and a dictionary-based clustering model is designed to group sentences into different facets.

Understanding User Stance via Comment Board *2015-2016*

I worked with Prof. Yizhou Sun and Prof. Lu Wang to design a unified model to understand users' positions in different topics according to their comments on the news websites. We propose to combine the modeling of the comments and the interactions between the users to predict their stances. This work has been accepted by CIKM, 2017.

User Characteristic Analysis *2014-2015*

I worked with Prof. Yizhou Sun to design a model to learn users' personalities according to their behavior, e.g., interactions with the dialog boxes, recorded in the games.

Cluster Ensemble *2012-2014*

I worked with Prof. Zhipeng Xie to design an algorithm to utilize the clustering results generated by different algorithms to find a consensus clustering result. This work has been accepted by IJCAI, 2013.

WORK EXPERIENCE

Amazon.com, Inc 2020.06 - 2020.08
Applied Scientist Intern *Cambridge, MA*

· Improve the accuracy of semantic parsing on small datasets via multi-domain training, transfer learning and few-shot learning.

Onboard Data, Inc 2019.09 - 2019.12
Research Intern *Cambridge, MA*

· Build a system for table detection, table recognition and table classification on scanned images of tables.

Digital Humanities Lab, Leipzig University 2018.12 - 2019.08
Researcher *Leipzig, Germany*

· Implement a client/server architecture system for harmonizing the input data, parameterization and provenance tracking of training different OCR engines.
· Compare the performance of OCR models on texts with different fonts and languages.
· Create a large dataset for training OCR models via unsupervised method on historical German text.

Ancestry.com, LLC 2016.06 - 2016.08
Data Scientist Intern *San Francisco, CA*

· Implement an information extraction system from the noisy OCR results of historical newspaper via entity recognition, entity linking and entity relation extraction.

TECHNICAL STRENGTHS

Computer Languages Python, Tensorflow, PyTorch, C, C++, C#, Matlab