HAORAN ZHANG

Homepage GitHub ◇ haoranz6@illinois.edu

SUMMARY

Full stack researcher with 3-year experience in **Natural Language Processing**, with research interest in **Information Extraction**. Strong knowledge of Natural Language Processing and solid programming skills in Data Science. I am interested in how human and machine learn language.

EDUCATION

University of Illinois Urbana-Champaign *Aug 2019 - Present*
Master of Science, Information Management

Changsha University of Science & Technology *Jul 2014 - Jun 2018*
Bachelor of Science, Computer Science

RESEARCH EXPERIENCE

University of Illinois Urbana-Champaign *Sep 2019 - Present*
BLENDER Member, Supervised by Heng Ji.

Changsha University of Science & Technology *May 2018 - Sep 2019*
Research Assistant on Relation Extraction, Supervised by Daojian Zeng.

University of California, Los Angeles *Sep 2016 - Jun 2017*
Remote Summer Research on Computer Vision, Supervised by Yajia Yang.

SKILLS

Full Stack Researcher: Literature Review, New Ideas, Data Visualization, Experiments Design, Paper Writing, Presentation.

Tools: Python, Haskell, PyTorch, LaTeX, MySQL, MongoDB.

PROJECTS

Sequence-to-Unordered-Multi-Tree for Joint Extraction of Relations and Entities[2] *2020*

- Formulated the output sequence to unordered-multi-tree structure to mitigate the notorious exposure bias problem in the well-studied Seq2Seq model.
- Pointed out flaws of the widely-used NYT dataset, i.e. the models only memorize the appeared triplets rather than generalize to new entities.
- Implementing a toolkit containing [5 Models $\times$ 2 Datasets] to be open-sourced.

Sequence-to-Sequence for Joint Extraction of Relations and Entities[3] *2019*

- Figured out a linear algebra bug causing underfitting of training set in an ACL2018 paper.
- Based on theoretical analysis, added only one more non-linear layer to fix the bug.
- Yielded 14 and 31 (F1) absolute improvement over baseline on NYT and WebNLG dataset respectively.

Controlled Sequence-to-Sequence for Paraphrase Generation[4] *2018*

- Using only pairwise sentence training set, generated multiple paraphrases according to different keywords.

- The system was successfully deployed to both individual users scenario and data augmentation of models.

PAPERS AND MANUSCRIPTS

(* refers to equal contribution)

1. Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, David Liem, Ahmed Elsayed, Martha Palmer, Jasmine Rah, Cynthia Schneider, Boyan Onyshkevych. **COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation**. arXiv preprint. Retrieved from here.
2. Haoran Zhang*, Qianying Liu*, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara and Sadao Kurohashi, **Minimize Exposure Bias of Seq2Seq Models in Joint Entity and Relation Extraction.** EMNLP 2020 Findings. Retrieved from here.
3. Daojian Zeng*, Haoran Zhang*, Qianying Liu, **CopyMTL: Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning.** AAAI, 2020. Retrieved from here.
4. Daojian Zeng, Haoran Zhang, Lingyun Xiang, Jin Wang, Guoliang Ji, **User-Oriented Paraphrase Generation With Keywords Controlled Network**, in IEEE Access, vol. 7, pp. 80542-80551, 2019. doi: 10.1109/ACCESS.2019.2923057. Retrieved from here.

OTHER PROJECTS

Educational Data Mining on CS125 Assignments *2020*

- Using MongoDB to manage students' assignments without fixed schema.
- Cleaning the data and co-designing the annotation manual.

Distant Supervised Relation Extraction Dataset Visualization *2019*

- Visualizing data distribution via ggplot2.
- Visualizing long tail label problem of NYT dataset.
- Figure out label duplication between training and test set which may cause model overfitting.