Shi Feng - CURRICULUM VITAE

shif@uchicago.edu • +1 240 821 0347 • Google Scholar

Education

University of Chicago — Chicago, Illinois
Postdoc, Computer Science — *2021 –*
Advised Prof. Chenhao Tan.

University of Maryland — College Park, Maryland
PhD, Computer Science — *2016 – 2021*
Advised by Prof. Jordan Boyd-Graber.

Shanghai Jiao Tong University — Shanghai, China
B.S. in Computer Science — *2012 – 2016*
Member of the ACM Honor Class.

Publications

Machine Explanations and Human Understanding
TMLR and FaCCT — *2023*
Chacha Chen*, **Shi Feng***, Amit Sharma, Chenhao Tan

Learning Human-Compatible Representations for Case-Based Decision Support
ICLR 2023 — *2023*
Han Liu, Yizhou Tian, Chacha Chen, **Shi Feng** Yuxin Chen, Chenhao Tan

Learning to Explain Selectively
EMNLP 2022 — *2022*
Shi Feng, Jordan Boyd-Graber

Active Example Selection for In-Context Learning
EMNLP 2022 — *2022*
Yiming Zhang, **Shi Feng**, Chenhao Tan

Calibrate Before Use: Improving Few-shot Performance of Language Models
ICML 2021 — *2021*
Tony Z. Zhao*, Eric Wallace*, **Shi Feng**, Dan Klein, Sameer Singh

Concealed Data Poisoning Attacks on NLP Models
NAACL 2021 — *2020*
Eric Wallace*, Tony Z. Zhao*, **Shi Feng**, Sameer Singh

Quizbowl: The Case for Incremental Question Answering
JMLR 2021 — *2019*
Pedro Rodriguez, **Shi Feng**, Mohit Iyyer, He He, Jordan Boyd-Graber

What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play
IUI 2019, oral — *2019*
Shi Feng, Jordan Boyd-Graber

Universal Adversarial Triggers for Attacking and Analyzing NLP
EMNLP 2019, oral — *2019*
Eric Wallace, **Shi Feng**, Nikhil Kandpal, Matt Gardner, Sameer Singh

Misleading Failures of Partial-input Baselines
ACL 2019 — *2019*
Shi Feng, Eric Wallace, Jordan Boyd-Graber

Understanding Impacts of High-Order Loss Approximations and Features in Deep Learning Interpretation
ICML 2019 — *2019*
Sahil Singla, Eric Wallace, **Shi Feng**, Soheil Feizi

Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples for Question Answering
TACL 2019 *2019*
Eric Wallace, Pedro Rodriguez, **Shi Feng**, Jordan Boyd-Graber

Pathologies of Neural Models Make Interpretation Difficult
EMNLP 2018, oral *2018*
Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, Jordan Boyd-Graber

Improving Attention Modeling with Implicit Distortion and Fertility for Machine Translation
COLING 2016 *2016*
Shi Feng, Shujie Liu, Nan Yang, Mu Li, Ming Zhou, Kenny Q. Zhu

Knowledge-Based Semantic Embedding for Machine Translation
ACL 2016 *2016*
Chen Shi, Shujie Liu, Shuo Ren, **Shi Feng**, Mu Li, Ming Zhou, Xu Sun, Huofeng Wang

Working papers

Learning to Improve Spaced Repetition
In submission *2022*
Matthew Shu*, **Shi Feng***, Jordan Boyd-Graber

What Spurious Features Can Pretrained Language Models Combat?
In submission *2022*
Chenglei Si, Dan Friedman, Nitish Joshi, **Shi Feng**, Danqi Chen, He He

Workshop papers

How Pre-trained Word Representations Capture Commonsense Physical Comparisons
Commonsense Inference in NLP workshop *2019*
Pranav Goel, **Shi Feng**, Jordan Boyd-Graber

Interpreting Neural Networks with Nearest Neighbors
BlackboxNLP Workshop at EMNLP *2018*
Eric Wallace*, **Shi Feng***, Jordan Boyd-Graber

The UMD Neural Machine Translation Systems at WMT17 Bandit Learning Task
The Second Conference on Machine Translation *2017*
Amr Sharaf, **Shi Feng**, Khanh Nguyen, Kianté Brantley, Hal Daumé III

Deployed Projects

KAR^{3}L: Spaced repetition meets representation learning karl.qanta.org

We want to see if representation learning can help us improve human memorization, and more generally, human learning. We implemented this flashcard app as a testbed for this idea. In traditional spaced repetition learning systems, all flashcards are treated as equal, so are all the users. This over-simplified model ignores useful signals that can help us infer the state of the user's memory: if the user correctly answers a question about Mozart, this should tell us something about the his/her knowledge about classical music, and in turn the probability of correctly answering a question about Beethoven. Our proposed algorithm uses representation learning to exploit connections like this, and is currently deployed on this interface. We have a paper in preparation for this project.

Play With QANTA: Human-computer Cooperative QA

We want to see if post-hoc explanations improves human-AI cooperation, and built this online interface is a testbed for this idea. Each human player on the interface is assisted with a human-level question answering AI, where the AI communicated its predictions via several post-hoc explanations. Our **IUI'19** paper is based on experiments conducted using this interface. We have an on-going work that investigates whether we could adapt to each user and intelligently select which explanation to show in order to maximize the human-AI team performance.

QANTA: Human-level *Quizbowl* System github.com/pinafore/qb

At HSNCT’17 we beat *top* human players for the first time (video). I’m mainly responsible for the *buzzer* of QANTA, which controls when to buzz and when to wait. The buzzer was trained with reinforcement learning using game history collected from Protobowl. This RL buzzer was first introduced to the system for HSNCT’17 and turned out to be crucial to the victory against human.

Talks

Pragmatic Interpretability, USC NLG Seminars	Nov 17, 2022
Pragmatic Interprteability, NEC Labs Europe	Nov 7, 2022
NAACL 22 Tutorial on Human-centered Evaluations of Explanations	July 10, 2022
NLP Highlights Podcast: Pathologies of Neural Models Make Interpretation Difficult	Apr 25 2019
Pathologies of Neural Models Make Interpretation Difficult, UPenn	Mar 25 2019
Pathologies of Neural Models Make Interpretation Difficult, UCSD	Mar 19 2019
Pathologies of Neural Models Make Interpretation Difficult, UCI	Mar 18 2019

Awards and Service

Best reviewer award, EMNLP’18, NeurIPS’20

Reviewer: EMNLP’18’19’20’21’22, ACL’19’20’21, AAAI’20, CoNLL’20, NeurIPS’20’21, ICLR’21’22, NAACL’21

Work Experience

Salesforce Research, *Research Intern* 2020.6 – 2020.8

- Advised by Bryan McCann
- Pretrain poisoning. We show that malicious unlabeled data during pretraining can lead to biases and backdoors in the downstream model based on the pretrained representations. In particular, we show that injecting a few thousand sentences to GPT-2’s unlabeled training set exacerbates the gender bias of a sentiment classifier based on GPT-2 by 10% absolute difference. The security implications of this vulnerability is immense: unlabeled data for pretraining is almost always collected from the web with very minimal data cleaning. Poisoning against the unlabeled data is thus easy to carry out by an adversary without a lot of resources. We have a paper in preparation for this work.

Microsoft Research, *Research Intern* 2018.6 – 2018.8

- Health AI team
- Domain adaptation for machine translation. A boosting approach to safely select in-domain data to adapt a general translation system to the medical domain.

Microsoft Research Asia, *Research Intern* 2015.8 – 2016.2

- Natural Language Computing Group
- Built the first neural machine translation system with Theano for NLC group.
- Improved the attention mechanism, results published at COLING’16.
- Experimented sequence-to-sequence for many other tasks, including pos tagging, parsing, and Chinese couplet completion (link).