WEI ZHANG

wynnzh@gmail.com
+1 412-737-3288
Acton, Massachusetts USA
Google Scholar Profile
Homepage

EDUCATION

Master of Science (Research Track, Full scholarship) | ***Language Technologies*** Aug. 2012 – Aug. 2014
Language Technologies Institute, SCS, Carnegie Mellon University Pittsburgh, PA, USA

Bachelor of Science | ***Computer Science and Engineering*** Sep. 2003 – July 2007
Harbin Institute of Technology Harbin, Heilongjiang, China

WORK EXPERIENCE

Principal Machine Learning Engineer Sep 2022 - March 2023
Saks.com LLC New York City, NY, USA (remote)

With deep learning, large language models, multi-modal encoders and reinforcement learning, as a technical leader and impact maker, I helped saks.com transform its onsite experience of Search, Recommendations and Personalization to improve customer satisfaction and business success measured by conversion, GMV and profit. I also led the team to identify significant opportunity in product search and discovery and being hands-on for implementation. My research background on DL and NLP research enables me to think critically, while being realistic and mindful about the timelines and trade offs to achieve business goals. I also built strong collaboration with partner teams.

- **Semantic Search with CLIP**: For Saks.com, I created vector-based semantic product Search with large multi-modal encoders such as CLIP, which is fine-tuned with E-commerce customer behavior data
- **Product Information Generation using GPT:** I fine-tuned GPT2 model to generate and predict user search terms from product title and descriptions to bridge the query-product language gap for BM25 search.
- **Search Query attribute extraction with Neural Language Models:** Applying GPT2 to extract product brands, types, sizes, colors in a generative manner.
- **Multi-objective Search Ranking:** Working closely with Professor Charles Elkan to design product search ranking algorithm that considers both user likelihood of click and revenue. The new algorithm achieved 9% conversion rate lift for search.

Senior Staff Data Scientist May 2021 - Oct. 2022
Search and Recommendations, Wayfair LLC Boston, MA, USA

I worked with stakeholders (data scientists, ML engineers) to apply cutting-edge deep learning approaches to upgrade Wayfair's search and recommendation services. I made hands-on contributions and supervised junior members to implement DL models, drove the roadmap for greenfield project. I also help Wayfair gain external exposure by publishing papers.

- **Semantic Search**: Lead the team to conduct research on semantic search. Online A/B test demonstrated an estimated 70M incremental GRS gain. Paper published on ACL 2022 workshop.
- **Query Knowledge Graph**: Initiated and drove the roadmap of Wayfair's knowledge graph leveraging contextual representation, named entity recognition and approximated nearest neighbour search
- **Wayfair's Language Model**: Learning query / product representations using from user click behaviors.

Research Scientist, Challenge Lead, Tech Lead Sept. 2014 – Apr. 2021
IBM T. J. Watson Research Center, MIT-IBM Watson Research, IBM Watson MA/NY,USA

Within IBM Research and IBM Watson, I have served as tech lead for the early efforts of neural machine comprehension back in 2015/2016, and a research lead on language model finetuning and explanability work in IBM Research, as well as solid contributor to RL and HCI work that led to best paper awards, top-tier papers and patents.

- **Explainablility of Large Fundational NLP models**: 1. Lead/first-author of the work on explainable sample-based explanation of black-box NLP models that enhances the interpretability by pin-pointing critical text segments as explanations; paper submitted to *ACL 2021*; 2. Multi-granular BERT: making intrinsically explainable BERT by attention regularization from the variational inference perspective, a paper submitted to *ACL 2021*. 3. co-authored the BlackBox NLP 2019 paper on applying n-gram representation to create intrinsically explainable contextual embedding.
- **Evaluation of Transformer-based Language Models**: First author of AAAI 2021 paper studying the intrinsic evaluation of language models, and showed that BERT family outperforms static embedding to a large margin.

- **Neural Machine Reading Comprehension with Attention Mechanism:** Supervised Interns on studying multi-passage reading comprehension using reinforcement learning (Wang et al., AAAI 2017) and an simpler approach by concatenating passages (Wang et al., ICLR 2018).
- **Collaborative Human-agent and Agent-agent Interaction with Reinforcement Learning**: Created the AI agents for IBM GuessTheWord!, which incorporates word game agents trained with multi-agent reinforcement learning, using Transformers, static embeddings, and word evocation dataset, and found out that multi-agent training leads to serious semantic drift in the setting. Co-authored papers on studying human behavior in agent-human interactions, one of which won the **2020 HCI best paper award**. The study discovered that humans collaborates better with AI when they know their partners are AI and can adapt to their behaviors.
- **Reinforcement Learning for AI-AI Interaction**: Created a method called *Random Action Replay* for stabilizing multi-agent RL training by re-sampling actions from policy in experience replay. US. Patent filed 2020.
- **Speech Emotion Recognition with Area Attention and Data Augmentation:** Supervised students on using area-attention for recognizing emotions using IEMOCAP dataset. Paper accepted to ICASSP 2021.
- **Time-series Financial Forecasting:** Major contributor of project of predicting Analysts' company projected earnings change using a Transformer-based self-attention model with task-oriented reinforcement learning training algorithm. US Patent 2020.
- **NLP for Finance**: Lead on using Company Earning Calls to predict Financial Analysts' consensus rating change of the company; Hierarchical embeddings of Financial Entities (Sectors, Companies, overall) for LSTM-based entity-news attention model.

- **Neural Machine Reading Comprehension:** Team lead for question answering/machine reading comprehension, an IBM core NLP problem. Was one of the first team around the world working on Stanford question answering dataset (SQuAD) and led on the leaderboard for some time at debut. Supervised Interns on studying multi-passage reading comprehension using reinforcement learning (Wang et al., AAAI 2017). Supervised students from Univ. of Montreal and Collaborated with Yoshua Bengio's team on reading comprehension.
- **Zero-shot Spell Checking with Token-based language models:** Lead on unsupervised spell correction using statistical n-gram language models with Bayesian Models built on Wikipedia dataset. Created a dataset for spell checking ground truth from Wikipedia Edit History.
- **Attention Models for Knowledge Extraction**: Used common-sense knowledge to interrogate contextual embedding; Supervised interns on applying Graph Neural Nets on encoding both time-series and graph information in a temporal setting; Co-authored a paper of KG relation classification with multi-granular Attention-based language model.
- **Neural Turing Machines/Memory Networks: NeurIPS 2015** Workshop paper on studying the memory Hieraracy of Neural Turing Machines, which inspired predecessor of the attention mechanism of Transformers, and showed superior performance than NTM on sequence memorization and bAbI QA task.
- **Auto-associative Memory:** A new improved neural model of auto-associative memory that can learn to store memory elements. We showed the memorization was improved for many problems: long-sequence memorization, image classification and can serve as base model of machine comprehension.

Research Assistant Oct. 2012 – Aug. 2014

Carnegie Mellon University, School of Computer Science, Language Technologies Institute Pittsburgh, PA, USA

- **Location Identification/Disambiguation in Social Media:** Recognizing location expressions in social media data such as twitter messages, and applying preference learning method to location disambiguation using features such as geo-codes, context words as well as user profile. Advisor is Dr. Judith Gelernter and visited NIST with her.
- **Multi-lingual named entity recognition:** Studied and compared methods for multi-lingual named entity recognition: translation-based v.s. language-specific NER models and found out that translation-based worked better in limited resource scenarios.

Intern May 2013 – Aug. 2013
PNC Bank Pittsburgh, PA, USA

- Worked on bank advertisement solicitation problem using a continuous value HMM model.

Research Engineer Sep. 2007 – Dec. 2011
Beijing Document Service Beijing, China

- Search Engine Optimization using improved ranking algorithms and User Logs
- Recommendation and Personalization of User profiles in Scientific Library
- Automatic Knowledge Graph construction using Textual Documents of Scientific Papers
- Chinese Named Entity Recognition

Research Assistant Sep. 2006 – May 2007
IR Lab, Harbin Institute of Technology Harbin, Heilongjiang, China

- At information Retrieval (IR) lab, I worked on Chinese Word Sense Disambiguation using effective features and SVM, and as a team member won the ACL SemEval 2007 Task 11 No. 1.

Selected Publications

1. **[IEEE SLT 2022]** Xiaoming Zhang, Fan Zhang, Xiaodong Cui, *Wei Zhang*. Speech Emotion Recognition With Complementary Acoustic Representations. 2022 IEEE Spoken Language Technology Workshop

2. **[ECNLP @ ACL 2022]** Zheng Liu, *Wei Zhang*, Yan Chen, Weiyi Sun, Michael Du, Benjamin Schroeder. On Generalization of Semantic Product Search. ACL 2022 NLP for E-commerce NLP Workshop.

3. **[Arxiv]** Zixuan Yuan, Yada Zhu, *Wei Zhang*, Ziming Huang, Guangnan Ye, Hui Xiong. Multi-Domain Transformer-Based Counterfactual Augmentation for Earnings Call Analysis.

4. **[ACL 2021]** *Wei Zhang*, Ziming Huang, Yada Zhu, Guangnan Ye, Xiaodong Cui, Fan Zhang. On Sample Based Explanation Methods for NLP: Faithfulness, Efficiency and Semantic Evaluation.

5. **[AAAI 2021]** *Wei Zhang*, Murray Campbell, Yang Yu, Sadhana Kumaravel. Circles are like Ellipses, or Ellipses are like Circles? Measuring the Degree of Asymmetry of Static and Contextual Word Embeddings and the Implications to Representation Learning. *35th AAAI conference on Artificial Intelligence* (**AAAI 2021**)

6. **[ICASSP 2021]** Mingke Xu, Fan Zhang, Xiaodong Cui, *Wei Zhang*. Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation. *45th International Conference on Acoustics, Speech, and Signal Processing* (**ICASSP** 2021)

7. **[CHI 2021]** Zahra Ashktorab, Casey Dugan, James Johnson, Qian Pan, *Wei Zhang*, Sadhana Kumaravel, Murray Campbell. Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction. **ACM CHI Conference 2021**

8. **[CSCW 2021]** Zahra Ashktorab, Q Vera Liao, Casey Dugan, James Johnson, Qian Pan, *Wei Zhang*, Sadhana Kumaravel, Murray Campbell. Human-ai collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction.* ***CSCW2***

9. **[NeurIPS 2020]** Zahra Ashktorab, Casey Dugan, J. Johnson, Qian Pan, *Wei Zhang*. The Design and Development of Games with a Purpose for AI Systems. *NeurIPS 2020 Workshop of Human And Machine in-the-Loop Evaluation and Learning Strategies.*

10. **[CHI 2020 Best Paper Award]** Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, Sadhana Kumaravel, *Wei Zhang*. Mental models of ai agents in a cooperative game setting. *Proceedings of the* ***2020 CHI*** *Conference on Human Factors in Computing Systems*

11. **[ICEI 2020]** Sihao Xu, *Wei Zhang*, Fan Zhang. Multi-Granular BERT: An Interpretable Model Applicable to Internet-of-Thing devices. *2020 IEEE International Conference on Energy Internet (ICEI)*
12. **[NeurIPS 2019]** Katy Gero, Maria Ruis, Zahra Ashktorab, J Johnson, Sadhana Kumaravel, *Wei Zhang*, Qian Pan, Murray Campbell, Casey Dugan, David Millen, Sarah Miller, Werner, Geyer. Passcode: A cooperative word guessing game between a human and AI agent. **NeurIPS 2019 Demonstration.**
13. **[ACL 2019 Workshop]** Zhiguo Wang, Yue Zhang, Mo Yu, *Wei Zhang*, Lin Pan, Lingfeng Song, Kun Xu, Yousef El-Kurdi. Multi-Granular Text Encoding for Self-Explaining Categorization. ***ACL 2019 BlackboxNLP***
14. Yang Yu, Kazi Saidul Hasan, Mo Yu, *Wei Zhang*, Zhiguo Wang. Knowledge base relation detection via multi-view matching. *European Conference on Advances in Databases and Information Systems. 2019*
15. **[AAAI 2018]** Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, *Wei Zhang*, Shiyu Chang, Gerry Tesauro, Bowen Zhou, Jing Jiang. R3: Reinforced Ranker-Reader for Open-Domain Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence.* ***AAAI 2018***
16. **[ICLR 2018]** Shuohang Wang, Mo Yu, Jing Jiang, *Wei Zhang*, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, Murray Campbell. Evidence Aggregation for Answer Re-Ranking in Open-Domain Question Answering. *International Conference on Learning Representations.* ***ICLR 2018***
17. *Wei Zhang*, Bowen Zhou. Learning to update Auto-associative Memory in Recurrent Neural Networks for Improving Sequence Memorization. (arXiv preprint arXiv:1709.06493).
18. Yang Yu, *Wei Zhang*, Kazi Hasan, Mo Yu, Bing Xiang, Bowen Zhou. End-to-end answer chunk extraction and ranking for reading comprehension. arXiv preprint arXiv:1610.09996.
19. **[NeurIPS 2015]** *Wei Zhang*. Yang Yu. Bowen Zhou. Structured Memory for Neural Turing Machines. ***NeurIPS 2015*** *workshop of Reasoning, Memory and Attention.*
20. *Wei Zhang*. Judith Gelernter. Exploring Metaphorical Senses and Word Representations for Identifying Metonyms. *Arxiv: arXiv:1508.04515 [cs.CL].. 2015.*
21. Yang Yu, *Wei Zhang*. Chang-wei Hang, Bing Xiang, Bowen Zhou. Empirical Study on Deep Learning Models for Question Answering. *Arxiv: arXiv:1510.07526 [cs.CL]. 2015.*
22. *Wei Zhang*. Judith Gelernter. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science 9 (1), 37-70. 2013.*
23. Judith Gelernter. *Wei Zhang*. Cross-lingual geo-parsing for non-structured data. *Proceedings of the 7th Workshop on Geographic Information Retrieval., 64-71 2013.*
24. **[ACM SigSatial 2013]** Judith Gelernter, Gautam Ganesh, Hamsini Krishnakumar, *Wei Zhang*. Automatic gazetteer enrichment with user-geocoded data. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, GEOCROWD 2013.*
25. (**[SemEval 2007 Task 11 Winner]**) Yuhang Guo, Wanxiang Che, Yuxuan Hu, *Wei Zhang*, and Ting Liu. HIT-IR-WSD: A wsd system for english lexical sample task. *Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics* ***SemEval 2007***.

Honors and Awards

CHI Best Paper Award 2020

As the co-author, I created the RL-trained AI game agent used to study the human behaviors.

Tuition Waiver + Stipend 2012-2014

Supported by Language Technologies Institute for the most of the two years study plus monthly stipend

No. 1 in ACL SemEval 2007 Word Sense Disambiguation Task 2007
As a team member won the WSD Task 11 Challenge

Renmin Scholarship 2005
Won the scholarship while in Harbin Institute of Technology.

Community Involvement and Talks

Program Committee Members : INTERSPEECH 2023, ECNLP@ACL 2022, KDD '22, ACL '21 '20, NeurIPS 2020, EMNLP '20, '19 , NAACL '20, EACL '20, SLT '20

Skills

Languages: Chinese (Native), English (Fluent)
Programming and Frameworks: Python, Java, SQL, C, Tensorflow, Pytorch, Torch, Theano, NumPy, SciPy, Huggingface, Terraform, Triton, Docker, MLflow, Flask, Django, Jupyter Notebook, Eclipse, Pycharm, Snowflake etc.