Tao CHEN

Email: chentaokite@gmail.com
`http://www.cs.jhu.edu/~taochen`

1600 Amphitheatre Parkway ,
Mountain View, CA 94043, USA

Research Interest

My research spans in the fields of natural language processing, information retrieval, health informatics, social computing, multimedia, and applied machine learning.

Education

National University of Singapore (NUS), Singapore Aug 2010 - Apr 2016

- Ph.D in Computer Science, GPA 4.69/5.0
- Thesis: Analyzing Image Tweets in Microblogs
- Advisor: Associate Professor Min-Yen Kan

East China Normal University, Shanghai, China Sep 2006 - Jun 2010

- Bachelor in Software Engineering, GPA 3.73/4.0, Rank 2/161
- Awarded School Excellent Thesis and Shanghai Excellent Graduate

Experience

Senior Software Engineer, Google Research, USA Apr 2018 - Present

Working in a research team for user modeling, content understanding, and information retrieval. Some of most recent works:

- User modeling: Explore large language models for social media creator modeling from their multimodal posts and interactions, and build creator-aware retrieval and ranking system (eg, [1]).
- Document understanding: propose dynamic language models for evolving content (like social media posts) [7], and exploit document layout and multimodal content for document representation learning [8].
- Information retrieval: for first stage retrieval, propose a zero-shot hybrid system to combine lexical and deep retrieval model [6], and an end-to-end query term reweighting model to improve lexical retrieval [2]; for second stage re-ranking, propose a decoder-only architecture to speed up re-ranker inference ([4]) and augument re-ranker using external retrieval [3].

Postdoctoral Fellow, Johns Hopkins University, USA Nov 2016 - Mar 2018

- Advisor: Prof. Mark Dredze
- Worked on social media analysis for public health and natural language processing for clinical text using deep neural network.

Research Fellow, NUS, Singapore Jun 2016 - Oct 2016

- Advisor: Prof. Min-Yen Kan and Prof. Teck Khim Ng
- Worked on second language learning from reading the news.

Research Assistant, NUS, Singapore Sep 2014 - Mar 2016

- Advisor: Prof. Min-Yen Kan
- Worked on second language learning from reading the news.

Software Engineering Intern, Google, New York Jun 2013 - Sep 2013

- Worked in review summarization team and focused on sentiment analysis internationalization.

Software Engineering Intern, IBM, Shanghai, China Jun 2009 - Sep 2009

- Developed an IE extension that could bookmark, comment and share web pages within IBM internal social network.

Publications

[1] Spurthi Amba Hombaiah, **Tao Chen**, Mingyang Zhang, Michael Bendersky, Sergey Levi, Matt Colen, Vladimir Ofitserov, Marc Najork. Creator Context for Tweet Recommendation (under review).

[2] Karan Samel, Cheng Li, Weize Kong, **Tao Chen**, Mingyang Zhang, Shaleen Gupta, Swaraj Khadanga, Wensong Xu, Xingyu Wang, Kashyap Kolipaka, Michael Bendersky, Marc Najork. End-to-End Query Term Weighting (under review).

[3] Kai Hui, **Tao Chen**, Zhen Qin, Honglei Zhuang, Fernando Diaz, Mike Bendersky, Don Metzler. Retrieval Augmentation for T5 Re-ranker using External Sources. *arXiv:2210.05145 (arXiv'22).*

[4] Kai Hui, Honglei Zhuang, **Tao Chen**, Zhen Qin, Jing Lu, Dara Bahri, Ji Ma, Jai Gupta, Cicero Nogueira dos Santos, Yi Tay, Donald Metzler. ED2LM: Encoder-Decoder to Language Model for Faster Document Re-ranking Inference. *In Proceedings of the Findings of 60th Annual Meeting of the Association for Computational Linguistics (ACL'22).*

[5] Shuang Liu, Fan Zhang, Baiyang Zhao, Renjie Guo, **Tao Chen**, Meishan Zhang. APPCorp: A Corpus for Android Privacy Policy Document Structure Analysis. Frontiers of Computer Science. *Frontiers of Computer Science* (upcoming).

[6] **Tao Chen**, Mingyang Zhang, Jing Lu, Michael Bendersky, Marc Najork. Out-of-Domain Semantics to the Rescue! Zero-Shot Hybrid Retrieval Models. *In Proceedings of the 44th European Conference on Information Retrieval (ECIR'22).*

[7] Spurthi Amba Hombaiah, **Tao Chen**, Mingyang Zhang, Michael Bendersky, Marc Najork. Dynamic Language Models for Continuously Evolving Content. *In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'21).*

[8] Te-Lin Wu, Cheng Li, Mingyang Zhang, **Tao Chen**, Spurthi Amba Hombaiah, Michael Bendersky. LAMPRET: Layout-Aware Multimodal PreTraining for Document Understanding. *Visually Grounded Interaction and Language Workshop (NAACL'21).*

[9] Ansel MacLaughlin, **Tao Chen***, Burcu Karagol Ayan, Dan Roth. *In Proceedings of the 15 International Conference on Weblogs and Social Media (ICWSM'21).* (*corresponding author).

[10] **Tao Chen**, Mark Dredze, Jonathan P Weiner, Hadi Kharraz. Incorporating Contextual Information to Identify Geriatric Syndromes in Clinical Notes. *Journal of the American Medical Informatics Association (JAMIA)*, 2019; 26 (8-9): 787795. (Impact factor: 4.292).

[11] **Tao Chen**, Mark Dredze, Jonathan P Weiner, Leilani Hernandez, Joe Kimura, Hadi Kharraz. Extraction of Geriatric Syndromes From Electronic Health Record Clinical Notes: Assessment of Statistical Natural Language Processing Methods. *JMIR Medical Informatics (JMI)* , 2019;7(1): e13039.

[12] Yuki Lama, **Tao Chen***, Mark Dredze, Amelia Jamison, Sandra Crouse Quinn, David A Broniatowski (2018). Discordance between HPV Twitter Images and Disparities in HPV Risk and Disease: Mixed Methods Analysis. *Journal of Medical Internet Research (JMIR)*, 2018; 20(9): e10244. (Impact factor: 4.945; *corresponding author)

[13] David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, **Tao**

Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health (AJPH)*, 2018, e1-e7. (Impact factor: 5.381; Covered by over 250 news articles; Ranked #37 of 12.2 million research outputs by Altmetric and in the top-20 for 2018)

[14] **Tao Chen** and Mark Dredze (2018). Vaccine Images on Twitter: Analysis of What Images are Shared. *Journal of Medical Internet Research (JMIR),* 2018; 20(4): e130. (Impact factor: 4.945)

[15] Francesco Gelli, Xiangnan He, **Tao Chen** and Tat-Seng Chua (2017). How Personality Affects our Likes: Towards a Better Understanding of Actionable Image. *In Proceedings of the 25th ACM International Conference on Multimedia (MM'17).*

[16] Kang Hong Jin, **Tao Chen**, Muthu Kumar Chandrasekaran and Min-Yen Kan (2016). A Comparison of Word Embeddings for English and Cross-Lingual Chinese Word Sense Disambiguation. *In Proceedings of COLING Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'16).*

[17] **Tao Chen**, Xiangnan He and Min-Yen Kan (2016). Context-aware Image Tweet Modelling and Recommendation. *In Proceedings of the 24th ACM International Conference on Multimedia (MM'16).*

[18] Xiangnan He, **Tao Chen**, Min-Yen Kan and Xiao Chen (2015). Review-aware Explainable Recommendation by Modeling Aspects. *In Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15).*

[19] Bang Hui Lim, Dongyuan Lu, **Tao Chen** and Min-Yen Kan (2015). #mytweet via Instagram: Exploring User Behaviour across Multiple Social Networks. *In Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15).*

[20] **Tao Chen**, Naijia Zheng, Yue Zhao, Muthu Chandrasekaran and Min-Yen Kan (2015). Interactive Second Language Learning from News Websites. *In Proceedings of ACL Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA'15).*

[21] **Tao Chen**, Hany Salaheldeen, Xiangnan He, Min-Yen Kan and Dongyuan Lu (2015). VELDA: Relating an Image Tweet's Text and Images. *In Proceedings of the 29st AAAI Conference on Artificial Intelligence (AAAI'15).*

[22] Jing Wu, Wei Guo, Chenxi Luo and **Tao Chen** (2015). Local Governments' Preference on Housing Market Interventions:Measurement and Analysis on Determinant Factors (in Chinese). *Finance & Trade Economics*, No.12, 2015.

[23] **Tao Chen**, Dongyuan Lu, Min-Yen Kan and Peng Cui (2013). Understanding and Classifying Image Tweets. *In Proceedings of the 21st ACM International Conference on Multimedia (MM'13).*

[24] **Tao Chen** and Min-Yen Kan (2013). Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus. *Language Resources and Evaluation*, 47(2)(2013). (Impact Factor: 1.029)

[25] Aobo Wang, **Tao Chen** and Min-Yen Kan (2012). Re-tweeting from a Linguistic Perspective. *In Proceedings of the NAACL-HLT 2012 Workshop on Language in Social Media (LSM'12).*

Services

Journal Reviewer:

- Information Retrieval Journal (IRJ)
- Language Resources and Evaluation (LRE)
- Social Network Analysis and Mining (SNAM)
- IEEE Transactions on Multimedia (TMM)
- Journal of Medical Internet Research (JMIR)
- JMIR Public Health and Surveillance (JPH)
- Journal of the American Medical Informatics Association (JAMIA)
- Digital Communications and Networks

Conference Program Committee Member:

- ACL (2017-2023), EMNLP (2016-2022), SIGIR (2018-2022), The Web Conference (2021-2023), AAAI (2020-2022), CIKM (2021-2022), IJCAI (2021), NAACL (2021), ICME (2016-2018), ASONAM (2016-2020), ECIR (2022-2023), IJCNLP (2017)

Session Chair: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining
Volunteer: The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015), 2012 Machine Learning Summer Schools (MLSS)

Awards & Scholarships

Ann E. Nolte Writing Award for paper [13]	Feb 2019
Validating Startup Concept (VaSCo) Award from NUS (S$10K)	Sep 2015
AAAI 2015 Student Travel Scholarship	Jan 2015
Google Anita Borg Memorial Scholarship: Asia Pacific (28 recipients in Asia Pacific)	Sep 2014
1st Place in Elsevier CodeForScience Singapore Competition	Sep 2012
Machine Learning Summer School (MLSS) Singapore Scholarship	Jul 2011
National University of Singapore Research Scholarship	2010 - 2014
IBM Chinese Excellent Student Scholarship	Sep 2009
National Inspirational Scholarship	Oct 2007
First-class Scholarship in East China Normal University	2007 - 2010

Skills

Programming: Python, C++, Java, Ruby, SQL, JSP, HTML, Javascript
Language: English (fluent), Chinese (native)

Last updated: Mar 14, 2023