Tongfei Chen

tongfeichen@microsoft.com tongfei.me

Education

2020 | **Ph.D. in Computer Science**
Johns Hopkins University, *Baltimore, MD, USA*
Center for Language and Speech Processing
Department of Computer Science
Advisor: Prof. Benjamin Van Durme
Thesis: *Ranking and retrieval under semantic relevance*

2014 | **B.Sc. in Computer Science**
Peking University, *Beijing, China*
Department of Computer Science and Technology
Advisor: Prof. Junfeng Hu
Thesis: *Large-scale unsupervised word segmentation for Classical Chinese: Research & system*

Work Experience

Jan. 2022 – *present* | **Senior Researcher**
Semantic Machines, *Microsoft, Bellevue, WA, USA*
Topics: Conversational AI; dialogue systems

Oct. 2020 – Dec. 2021 | **Senior Applied Scientist**
Project Turing, *Microsoft, Bellevue, WA, USA*
Topics: Microsoft Bing; Transformer model pre-training; question answering over tables; relevance ranking for tables

Aug. 2014 – Sept. 2020 | **Graduate Research Assistant**
Center for Language and Speech Processing, *Johns Hopkins University, Baltimore, MD, USA*
Advisor: Prof. Benjamin Van Durme
Topics: Information extraction; question answering; information retrieval; natural language inference; knowledge acquisition from text; scalable systems; approximate algorithms

May 2018 – Aug. 2018 | **Applied Scientist Intern**
Alexa AI, *Amazon, Seattle, WA, USA*
Host: Dr. Lambert Mathias
Topics: Amazon Alexa; dialogue context modeling; sequence transduction models

May 2017 – Aug. 2017 | **Research Intern**
IBM Thomas J. Watson Research Center, *Yorktown Heights, NY, USA*
Hosts: Dr. Jiří Navrátil, Dr. Bing Xiang
Topics: Confidence scoring; model calibration; meta-models

June 2012 – June 2014 | **Undergraduate Research Assistant**
Institute of Computational Linguistics, *Peking University, Beijing, China*
Advisor: Prof. Junfeng Hu
Topics: Word segmentation; graph & network analysis; ontology construction

Publications

Peer-reviewed papers

* Equal contributions.

[1] Luyu Gao, Zhuyun Dai, **Tongfei Chen**, Zhen Fan, Benjamin Van Durme, Jamie Callan (2021). Complement lexical retrieval model with residual semantic embeddings. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR; LNCS 12656)*. pp. 146–160. [IR] [NLP]

[2] Patrick Xia*, Guanghui Qin*, Siddharth Vashishtha, Yunmo Chen, **Tongfei Chen**, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, Benjamin Van Durme (2021). LOME: Large ontology multilingual extraction. In *Proceedings of the Software Demonstrations of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. pp. 149–159. [NLP]

[3] Yunmo Chen, **Tongfei Chen**, Benjamin Van Durme (2020). Joint modeling of arguments for event understanding. In *Proceedings of the First Workshop on Computational Approaches to Discourse (CODI@EMNLP)*. pp. 96–101. [NLP]

[4] Yunmo Chen, **Tongfei Chen**, Seth Ebner, Aaron Steven White, Benjamin Van Durme (2020). Reading the Manual: Event extraction as definition comprehension. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP (SPNLP@EMNLP)*. pp. 74–83. [NLP]

[5] **Tongfei Chen***, Zhengping Jiang*, Adam Poliak, Keisuke Sakaguchi, Benjamin Van Durme (2020). Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 8772–8779. [NLP]

[6] **Tongfei Chen**, Yunmo Chen, Benjamin Van Durme (2020). Hierarchical entity typing via multi-level learning to rank. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 8465–8475. [NLP]

[7] Yiming Wang, **Tongfei Chen**, Hainan Xu, Shuoyang Ding, Hang Lv, Yiwen Shao, Nanyun Peng, Lei Xie, Shinji Watanabe, Sanjeev Khudanpur (2019). Espresso: A fast end-to-end neural speech recognition toolkit. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. pp. 136–143. [SPEECH]

[8] Arya D. McCarthy, **Tongfei Chen**, Seth Ebner (2019). An exact No Free Lunch theorem for community detection. In *Proceedings of the 8th International Conference on Complex Networks and their Applications (Complex Networks)*. pp. 176–187. [GRAPH]

[9] Arya D. McCarthy, **Tongfei Chen**, Rachel Rudinger, David W. Matula (2019). Metrics matter in community detection. In *Proceedings of the 8th International Conference on Complex Networks and their Applications (Complex Networks)*. pp. 164–175. [GRAPH]

[10] **Tongfei Chen**, Chetan Naik, Hua He, Pushpendre Rastogi, Lambert Mathias (2019). Improving long distance slot carryover in spoken dialogue systems. In *Proceedings of the First Workshop of NLP for Conversational AI (NLP4ConvAI@ACL)*. pp. 96–105. **[best paper award]** [NLP]

[11] Zhongyang Li, **Tongfei Chen**, Benjamin Van Durme (2019). Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 4818–4823. [NLP]

[12] Pushpendre Rastogi, Arpit Gupta, **Tongfei Chen**, Lambert Mathias (2019). Scaling multi-domain dialogue state tracking via query reformulation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (NAACL)*. pp. 97–105. [NLP]

[13] J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, **Tongfei Chen**, Matt Post, Benjamin Van Durme (2019). Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL)*. pp. 839–850. [NLP]

[14] Yiming Wang, Xing Fan, I-Fan Chen, Yuzong Liu, **Tongfei Chen**, Björn Hoffmeister (2019). End-to-end anchored speech recognition. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 7090–7094. [SPEECH]

[15] **Tongfei Chen**, Jiří Navrátil, Vijay Iyengar, Karthikeyan Shanmugam (2019). Confidence scoring using whitebox meta-models with linear classifier probes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS); Proceedings of Machine Learning Research 89 (PMLR)*, pp. 1467–1475. [ML]

[16] Rashmi Sankepally, **Tongfei Chen**, Benjamin Van Durme, Douglas W. Oard (2018). A test collection for coreferent mention retrieval. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 1209–1212. [NLP]

[17] Hainan Xu, **Tongfei Chen**, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, Sanjeev Khudanpur (2018). A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5929–5933. [SPEECH] [NLP]

[18] **Tongfei Chen** (2017). Typesafe abstractions for tensor operations. In *Proceedings of the 8th ACM SIGPLAN International Symposium on Scala (SCALA@SPLASH)*. pp. 45–50. [PL]

[19] Benjamin Van Durme, Tom Lippincott, Kevin Duh, Deana Burchfield, Adam Poliak, Cash Costello, Tim Finin, Scott Miller, James Mayfield, Philipp Koehn, Craig Harman, Dawn Lawrie, Chandler May, Max Thomas, Annabelle Carrell, Julianne Chaloux, **Tongfei Chen**, Alex Comerford, Mark Dredze, Benjamin Glass, Shudong Hao, Patrick Martin, Pushpendre Rastogi, Rashmi Sankepally, Travis Wolfe, Ying-Ying Tran, Ted Zhang (2017). CADET: Computer Assisted Discovery Extraction and Translation. In *Proceedings of the 8th International Joint Conference on Natural Language Processing, System Demonstrations (IJCNLP)*, pp. 5–8. [NLP]

[20] **Tongfei Chen**, Benjamin Van Durme (2017). Discriminative information retrieval for question asnwering sentence selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2 (EACL)*, pp. 719–725. [NLP] [IR]

[21] Junhao Zhang, **Tongfei Chen**, Junfeng Hu (2015). On the relationship between Gaussian stochastic blockmodels and label propagation algorithms. *Journal of Statistical Mechanics: Theory and Experiment (J. Stat. Mech)*. 2015(3), P03009. [GRAPH]

[22] Ni Sun, **Tongfei Chen**, Liumingjing Xiao, Junfeng Hu (2014). Diachronic deviation features in continuous space word representations. In *Proceedings of the 13th China National Conference on Computational Linguistics (CCL; LNCS 8801)*, pp. 23–33. [NLP]

[23] **Tongfei Chen**, Xiaojun Zou, Weimeng Zhu, Junfeng Hu (2013). Human-computer interactive Chinese word segmentation: An adaptive Dirichlet process mixture model approach. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 1278–1284. [NLP]

[24] **Tongfei Chen**, Weimeng Zhu, Xueqiang Lv, Junfeng Hu (2013). A Kalman filter based human-computer interactive segmentation system for ancient Chinese texts. In *Proceedings of the 12th China National Conference on Computational Linguistics (CCL; LNCS 8202)*, pp. 25–35. [NLP]

Non-refereed system descriptions

[S1] Yunmo Chen, Seth Ebner, **Tongfei Chen**, Patrick Xia, Elias Stengel-Eskin, Tzu-Ray Su, J. Edward Hu, Nils Holzenberger, Ryan Culkin, Craig Harman, Max Thomas, Thomas Lippincott, Aaron Steven White, Kyle Rawlins, Benjamin Van Durme (2019). NIST TAC SM-KBP 2019 system description: JHU/UR framework. In *Proceedings of the Text Analysis Conference (TAC)*. [NLP]

[S2] Mozhi Zhang, Jordan Boyd-Graber, Michelle Yuan, C. Anton Rytting, Weiwei Yang, Philip Resnik, Ting Hua, Adam Poliak, Adam Teichert, **Tongfei Chen**, Xu Han, Linghao Jin, João Sedoc, Benjamin Van Durme (2019). LoReHLT19 System Description UMD-JHU. [NLP]

[S3] Patrick Xia, Elias Stengel-Eskin, **Tongfei Chen**, Seth Ebner, Nils Holzenberger, Ryan Culkin, Pushpendre Rastogi, Xutai Ma, Benjamin Van Durme (2018). NIST TAC SM-KBP 2018 system description: JHU/UR pipeline. In *Proceedings of the Text Analysis Conference (TAC).* [NLP]

Presentation and Talks

Aug. 1, 2019	Improving long distance slot carryover in spoken dialogue systems *NLP4ConvAI@ACL 2019: Best paper talk, Florence, Tuscany, Italy*
May 29, 2019	Uncertain natural language inference *DARPA LORELEI/AIDA Site Visit, JHU*
Mar. 25, 2019	Uncertain natural language inference *Center for Language and Speech Processing Student Seminar, JHU*
Nov. 16, 2018	Towards typesafe deep learning in Scala *Scale by the Bay 2018, San Francisco, CA, USA*
Mar. 18, 2018	Towards typesafe deep learning in Scala *Northeast Scala Symposium 2018, Cambridge, MA, USA*
Oct. 23, 2017	Typesafe abstractions for tensor operations *SCALA@SPLASH 2017, Vancouver, BC, Canada*
Feb. 15, 2017	Discriminative information retrieval for knowledge discovery *DARPA DEFT/LORELEI Site Visit, JHU*
Oct. 25, 2016	Discriminative information retrieval for knowledge discovery *Center for Language and Speech Processing Student Seminar, JHU*
Oct. 17, 2013	HCI Chinese word segmentation: An adaptive Dirichlet process mixture model approach *IJCNLP 2013, Nagoya, Aichi, Japan*
Oct. 11, 2013	A Kalman filter based HCI segmentation system for ancient Chinese texts *CCL 2013, Suzhou, Jiangsu, China*

Honors and Awards

July 2019	Best paper award (2/25) *The 1st Workshop on NLP for Conversational AI @ ACL 2019*
June 2014	Outstanding undergraduate thesis *Peking University*

Open Source

- Espresso: Fast end-to-end automatic speech recognition based on fairseq. (800+ stars on GitHub) https://github.com/freewym/espresso (contributor)
- Nexus: Experimental typesafe tensors and deep learning in Scala. (200+ stars on GitHub) https://github.com/ctongfei/nexus

- Progressbar: A terminal-based progress bar for JVM. (700+ stars on GitHub)
  https://github.com/ctongfei/progressbar

Service

- Senior program committee member:
  - AAAI 2022
- Reviewer / Program committee member:
  - ACL Rolling Review 2021
  - AAAI 2021, 2020
  - AACL 2020
  - ACL 2021, 2020
  - CCL 2017
  - CL 2021
  - DI@KDD 2021
  - EMNLP 2021, 2020, 2019, 2018
  - KG4IR@SIGIR 2017
  - NAACL 2021, 2019
  - RCQA@AAAI 2020
  - TADGM@ICML 2018
- Secondary reviewer:
  - ACL 2019, 2018, 2017, 2015, 2014
  - ACL Demo Track 2017
  - EACL 2017
  - EMNLP 2017, 2014
  - IJCNLP 2017
  - NAACL 2015
  - TACL 2017, 2015
  - WWW 2015
- PhD recruitment committee 2018–2020, Johns Hopkins University
- North American Computational Linguistics Olympiad (NACLO) organizing committee 2016

Skills

- Programming languages:
  - Scala *(expert)*
  - Python *(proficient)*
  - Java *(proficient)*
  - C/C++ *(proficient)*
  - C# *(proficient)*
  - Haskell *(intermediate)*
- Natural languages:
  - Mandarin Chinese *(native)*
    (Standard & Sichuanese)
  - English *(proficient)*
  - Japanese *(intermediate)*
- Libraries and tools:
  - Deep learning: PyTorch, TensorFlow
  - Information retrieval: Lucene, FAISS
  - Data serialization: Thrift, ProtoBuf
  - Data visualization: Gephi
  - Distributed computing: Spark
  - Workflow orchestration: Ducttape
  - Scala ecosystem: Cats, Shapeless
  - Typesetting: LaTeX