Xiang Dai

Research scientist at
CSIRO Data61

Email: dai.xiang.au@gmail.com

Summary

I am currently a research scientist at CSIRO Data61. My research interests lie in natural language processing, deep learning, and digital health. Some topics I focus on include extracting entities and their relations from the unstructured text; pre-training domain-specific language representation models; building document-level models that can handle long sequence inputs; and, applying NLP techniques to the medical domain.

Education

University of Sydney
Ph.D., Natural Language Processing, 2017 - 2021
Supervisors: Sarvnaz Karimi, Ben Hachey, Cecile Paris, Joachim Gudmundsson
Thesis: Recognising Biomedical Names: Challenges and Solutions

University of Sydney
Master's Degree, Information Technology and Information Technology Management, 2015 - 2017
Supervisors: Joachim Gudmundsson, Michael Horton

South China University of Technology
Bachelor's Degree, Computer Science, 2003-2007

Experience

CSIRO Data61 [Sydney, Australia]
Research scientist, Aug 2023 - Present
Postdoc, Dec 2021 - Aug 2023

University of Copenhagen [Copenhagen, Denmark]
Postdoc, Feb 2021 - Dec 2021

Saarland University [Saarbrücken, Germany]
Visiting Researcher, Aug 2020 - Jan 2021, hosted by Dietrich Klakow

Bosch Center for Artificial Intelligence [Stuttgart, Germany]
Research Intern, Mar - Jul 2020

University of Sydney [Sydney, Australia]
Teaching Assistant, 2017 - 2019

CSIRO Data61 [Sydney, Australia]
Research Assistant, Feb - Jul 2017
Summer Scholar, Nov 2016 - Feb 2017

Lanke Technology Co., Ltd. [Guangzhou, China]
Engineer, 2010 - 2014

Qiaoqing Technology Co., Ltd. [Guangzhou, China]
Engineer, 2008 - 2010

Peer-Reviewed Publications

Journal

2 Clinician-facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI
Hubert D. Zając, Dana Li, **Xiang Dai**, Jonathan F. Carlsen, Finn Kensing, Tariq Q. Andersen
ACM Transactions on Computer-Human Interaction (TOCHI), 2023

1 Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis
Maciej Rybinski, **Xiang Dai**, Sonit Singh, Sarvnaz Karimi, Anthony Nguyen
JMIR Medical Informatics, 2021

Conference

10 Rethinking the Role of Entity Type in Relation Classification
Xiang Dai, Stephen Wan and Sarvnaz Karimi
The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL-IJCNLP 2023)

9 MultiFin: A Dataset for Multilingual Financial NLP
Rasmus Kær Jørgensen, Oliver Brandt, Mareike Hartmann, **Xiang Dai**, Christian Igel, Desmond Elliott
Findings of the Association for Computational Linguistics: EACL 2023

8 Revisiting Transformer-based Models for Long Document Classification
Xiang Dai, Ilias Chalkidis, Sune Darkner, Desmond Elliott
Findings of the Association for Computational Linguistics: EMNLP 2022

7 mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model
Rasmus Kær Jørgensen, Mareike Hartmann, **Xiang Dai**, Desmond Elliott
Findings of the Association for Computational Linguistics: EMNLP 2021

6 SearchEHR: A Family History Search System for Clinical Decision Support
Xiang Dai, Maciej Rybinski, Sarvnaz Karimi
International Conference on Information and Knowledge Management (CIKM 2021)

5 An Analysis of Simple Data Augmentation for Named Entity Recognition
Xiang Dai, Heike Adel
International Conference on Computational Linguistics (COLING 2020)

4 Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media
Xiang Dai, Sarvnaz Karimi, Ben Hachey, Cecile Paris
Findings of the Association for Computational Linguistics: EMNLP 2020

3 An Effective Transition-based Model for Discontinuous NER
Xiang Dai, Sarvnaz Karimi, Ben Hachey, Cecile Paris
Annual Meeting of the Association for Computational Linguistics (ACL 2020)

2 NNE: A Dataset for Nested Named Entity Recognition in English Newswire
Nicky Ringland, **Xiang Dai**, Ben Hachey, Sarvnaz Karimi, Cecile Paris, James R. Curran
Annual Meeting of the Association for Computational Linguistics (ACL 2019)

1 Using Similarity Measures to Select Pretraining Data for NER
Xiang Dai, Sarvnaz Karimi, Ben Hachey, Cecile Paris
Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)

Workshop

8 CSIRO Data61 Team at BioLaySumm Task 1: Lay Summarisation of Biomedical Research Articles Using Generative Models
Mong Yuan Sim, **Xiang Dai**, Maciej Rybinski and Sarvnaz Karimi
ACL Workshop on Biomedical Natural Language Processing (BioNLP 2023)
Shared task system paper: our system ranked 3rd for task 1 of BioLaySumm 2023 shared task among all 21 teams.

7 Can Social Media Inform Dietary Approaches for Health Management? A Dataset and Benchmark for Low-Carb Diet
Skyler Zou, **Xiang Dai**, Sarvnaz Karimi, Pennie Taylor and Grant Brinkworth
ACL Workshop on Biomedical Natural Language Processing (BioNLP 2023)

6 Detecting Entities in the Astrophysics Literature: A Comparison of Word-based and Span-based Entity Recognition Methods
Xiang Dai, Sarvnaz Karimi
AACL-IJCNLP Workshop on Information Extraction from Scientific Publications (WIESP 2022)
Shared task system paper: our system ranked 3rd for the DEAL shared task among all 12 teams.

5 NLNDE at CANTEMIST: Neural Sequence Labeling and Parsing Approaches for Clinical Concept Extraction
Lukas Lange, **Xiang Dai**, Heike Adel, Jannik Strötgen
Iberian Languages Evaluation Forum (IberLEF 2020)
Shared task system paper: our team ranked 4th for task 1; 5th for task 2; 4th for task 3 of the CANTEMIST shared task among all 24 teams.

4 Shot Or Not: Comparison of NLP Approaches for Vaccination Behaviour Detection
Aditya Joshi, **Xiang Dai**, Sarvnaz Karimi, Ross Sparks, Cecile Paris, C Raina MacIntyre
EMNLP Workshop on Social Media Mining for Health Applications (SMM4H 2018)
Shared task system paper: our system ranked 1st for task 4 of the SMM4H 2018 shared task among all 17 systems.

3 Recognizing Complex Entity Mentions: A Review and Future Directions
Xiang Dai
ACL Student Research Workshop (ACL-SRW 2018)

2 Medication and Adverse Event Extraction from Noisy Text
Xiang Dai, Sarvnaz Karimi, Cecile Paris
Australasian Language Technology Association Workshop (ALTA 2017)

1 Automatic Diagnosis Coding of Radiology Reports: A Comparison of Deep Learning and Conventional Classification Methods
Sarvnaz Karimi, **Xiang Dai**, Hamed Hassanzadeh, Anthony Nguyen
ACL Workshop on Biomedical Natural Language Processing (BioNLP 2017)

Talks

1 Challenges and Opportunities in AI4Xray
  2021-Oct-08 at Department of Diagnostic Radiology, Rigshospitalet, Denmark

Other Academic Activities

Supervision

1 Farhana (Hanna) Pethan PhD (Feb 2022-present), University of Sydney, co-advised with Adam Dunn, Heiko Spallek, Mike Conway and Sarvnaz Karimi

3 Mong Yuan Sim PhD (March 2024-present), University of Adelaide, co-advised with Wei (Emma) Zhang

5 Biaoyan (Byron) Fang Postdoc (Oct 2023-present), CSIRO, co-advised with Sarvnaz Karimi

7 Ritvik Dinesh Intern (Nov 2023-Feb 2024), University of Sydney

6 Harrison Bailey Intern (Nov 2023-Feb 2024), Australian National University

4 Skyler Zou Intern (Dec 2022-Feb 2023), University of New South Wales

3 Mong Yuan Sim Honours (July 2023-Nov 2023), University of Adelaide, co-advised with Wei (Emma) Zhang and Cecile Paris

3 Mong Yuan Sim Intern (Nov 2022-Feb 2023), University of Adelaide

2 Anqi Wu Intern (Nov 2022-Feb 2023), Monash University

Conference PC (reviewing)

1 ACL Annual Meeting of the Association for Computational Linguistics 2020; 2021; 2023

2 EMNLP Conference on Empirical Methods in Natural Language Processing 2020 (Outstanding Reviewer); 2021; 2022; 2023

3 ACL Rolling Review 2021; 2022; 2023; 2024 (Action Editor)

4 NAACL Conference of the North American Chapter of the Association for Computational Linguistics 2021

5 EACL Conference of the European Chapter of the Association for Computational Linguistics 2021 (Outstanding Reviewer); 2023

6 NeurIPS Conference on Neural Information Processing Systems 2021; 2022; 2023

7 ICLR International Conference on Learning Representations 2022; 2024

8 AAAI Conference on Artificial Intelligence 2023; 2024

9 AACL Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics 2023 (Session chair)

10 ICML International Conference on Machine Learning 2023; 2024

11 COLING International Conference on Computational Linguistics 2024

12 COLM Conference on Language Modeling 2024

Journal reviewing

1 ACM Transactions on Intelligent Systems and Technology 2018

2 International Journal of Medical Informatics 2018,2021,2023

3 Language Resources and Evaluation 2020

4 Computer Speech & Language 2021

5 Informatics in Medicine Unlocked 2022

6 Computers in Industry 2022

7 npj Digital Medicine 2022

8 IEEE/ACM Transactions on Audio, Speech, and Language Processing 2022

9 Artificial Intelligence In Medicine 2024

10 Information Processing & Management 2024

Workshop reviewing

1 ALTA Annual Workshop of the Australasian Language Technology Association 2019; 2020; 2021; 2022; 2023

2 Student Research Workshop at ACL (2020, 2021, 2022, 2023); AACL (2020, 2022); NAACL (2022, 2024)

3 Workshop on NLP for COVID-19 at EMNLP 2020

4 Workshop on Weakly Supervised Learning at ICLR 2021

5 Workshop on Information Extraction from Scientific Publications at AACL 2022

Membership

ALTA (Australasian Language Technology Association) executive committee, 2019-2020

ACL (Association for Computational Linguistics), 2018-present

Honors and Awards

Postgraduate Research Support Scheme
USYD, 2019 and 2020

Engineering and Information Technologies Research Scholarship
AUD 27K per annum, 3.5 years
Faculty of Engineering and Information Technologies, USYD, 2018 - 2021

Data61 top-up Scholarship
AUD 10K per annum, 3.5 years
Data61, CSIRO, 2018 - 2021

SIT Summer Scholarships
AUD 6K
School of Information Technologies, USYD, November 2016 - February 2017

Postgraduate Coursework High Honour Roll
School of Information Technologies, USYD, April 2016, May 2017 and June 2018

MIT and MITM Half-Fee Scholarship
School of Information Technologies, USYD, April 2016