Jia Xu

RESEARCH INTERESTS Artificial Intelligence, Machine Learning, Natural Language Processing

EMPLOYMENT

Assistant Professor
Stevens Institute of Technology

Assistant Professor
Graduate Center & Hunter College, City University of New York

Associate Professor and Ph.D. Supervisor
Head of the Center for Machine Intelligence (CMI)
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS)

Assistant Professor and Ph.D. Supervisor
Head of the Center for Machine Intelligence (CMI)
IIIS, Tsinghua University

Senior Researcher
German Institute for Artificial Intelligence (DFKI)

OTHER AFFILIATIONS

Research Fellow
I-DSLA (Institute for Data Science Learning and Applications)
Rutgers University - Newark

EDUCATION

Department 6 of Computer Science (Informatik 6), RWTH-Aachen University
Ph.D. in Computer Science
Thesis: Sequence Segmentation for Statistical Machine Translation
Advisor: Hermann Ney

Department of Computer Science, TU-Berlin
Diploma in Computer Science (MSc&BSc)
(5-year program completed in 3.5 years)
Thesis: A Computational Model for Human Auditory System
Advisor: Werner Hemmert

PUBLICATIONS

(RECEIVED ABOUT 700 CITATIONS)

- Abdul Rafae Khan, Jia Xu, and Weiwei Sun. Coding Textual Inputs Boosts the Accuracy of Neural Networks. In Proceedings of EMNLP, 2020.
- Abdul Rafae Khan, Asim Karim, Hassan Sajjad, Faisal Karmiran, and Jia Xu. A Clustering Framework for Lexical Normalization of Roman Urdu. In Journal of Natural Language Engineering, March 2019.

  [Top Journal; Impact Factor: **1.065**]

- Weimin Lyu, Sheng Huang, Abdul Rafae Khan, Shengqiang Zhang, Weiwei Sun and Jia Xu. CUNY-PKU Parser at SemEval-2019 Task 1: Cross-Lingual Semantic Parsing with UCCA. In Proceedings of the 2019 International Workshop on Semantic Evaluation at NAACL, March 2019.

- Tianyuan Sun, Yongcai Wang, Deying Li, Zhaoquan Gu, and Jia Xu. WCS: Robust Network Localization by Weighted Component Stitching. In Proceedings of IEEE/ACM Transactions on Networking, 2018.

  [Top Journal; Impact Factor: **3.11**]

- Abdul Rafae Khan, Subhadarshi Panda, Jia Xu, and Lampros Flokas. Hunter NMT System for WMT18 Biomedical Translation Task: Transfer Learning in Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation at EMNLP, October 2018.

- Hoang Cuong and Jia Xu. Assessing Quality Estimation Models for Sentence-Level Prediction. In Proceedings of COLING, August 2018.

  [Top NLP Conference; Acceptance Rate: **27%** in '12]

- Zhixian Lei, Xuehan Ye, Yongcai Wang, Deying Li, and Jia Xu. On the Efficient Online Model Adaptation by Incremental Simplex Tableau. In Proceedings of AAAI, June 2017.

  [Top AI Conference; Acceptance Rate: **24.6%**]

- J. Xu, Kuang Y.Z., Baijoo S., Lee J.H., Shahzad U., Ahmed M., Lancaster M., Carlan C.. Hunter MT: A Course for Young Researchers in WMT17. In Proceedings of the Second Conference on Machine Translation (WMT17) at EMNLP, Copenhagen, Denmark, Sep. 2017.

- Z. Lei, X. Ye, Y. Wang, D. Li, J. Xu. On the Efficient Online Model Adaptation by Incremental Simplex Tableau. In Proceedings of AAAI, San Francisco, CA, Feb. 2017.

  [Top AI Conference; Acceptance Rate: **24.6%**]

- P. A. Papakonstantinou, J. Xu, G. Yang (*all authors alphabetically ordered – paper on the mathematics of Machine Learning*). On the Power and Limits of Distance-Based Learning. In Proceedings of ICML, NY, NY, Jun. 2016.

  [Top Machine Learning Conference; Acceptance Rate: **24.0%**]

- J. Xu. System Description of ICT/Dublin NIST-15 Machine Translation System. In Workshop of NIST Open Machine Translation, Jun. 2015.

- P. A. Papakonstantinou, J. Xu, C. Zhu (*first two authors alphabetically ordered*). Bagging by Design (On the Sub-optimality of Bagging). In Proceedings of AAAI, Quebec City, Quebec, Canada, Jul. 2014.

- M. Dong, Y. Cheng, Y. Liu, J. Xu (*corresponding author*), M. Sun. Query Lattice for Translation Retrieval. In Proceedings of COLING, Dublin Ireland, Aug. 2014,

- C. Gan, Z. Qin, J. Xu, T. Wan. Salient Object Detection in Image Sequences via Spatial-Temporal Cue. In Proceedings of Visual Communications and Image Processing (VCIP), Sarawak, Malaysia, Nov. 2013.

- J. Xu, C. Kennington, C. Przywara and L. Wanzare. Comparable Corpora in Wikipedia Text for Machine Translation. In Proceedings of the 6th NIC Symposium 2012: 25 Years HLRZ/NIC (Book Section), ISBN: 9783893367580. Jülich, Germany. Feb. 2012.

- J. Xu and W. Sun. Generating Virtual Parallel Corpus: A Compatibility Centric Method. In Proceedings of MT-Summit, Xiamen, China. Sep. 2011.

- W. Sun and J. Xu. Enhancing Chinese Word Segmentation Using Unlabeled Data. In Proceedings of the EMNLP 2011. Edinburgh, UK, Jul. 2011.

- J. Xu, J. Gao, K. Toutanova and H. Ney: Synchronous Learning of Chinese Word Segmentation and Word Alignment. In the Handbook of Natural Language Processing and Machine Translation (Book Chapter), Springer, New York. 2011.
- A. Eisele and J. Xu: Improving Machine Translation Performance Using Comparable Corpora. In Proceedings of the LREC Workshop on Building and Using Comparable Corpora, Malta, May 2010.
- C. Federmann, A. Eisele, Y. Chen, S. Hunsicker, J. Xu and H. Uszkoreit: Further Experiments with Shallow Hybrid MT Systems. In Proceedings of the ACL Workshop on Statistical Machine Translation, pages 77-81, Uppsala, Sweden, Jul. 2010.
- J. Xu, J. Gao, and K. Toutanova: Character-based Chinese-English Statistical Machine Translation. Technical Report at RWTH-Aachen University. Aachen, Germany, Oct. 2009.
- J. Xu, J. Gao, K. Toutanova and H. Ney: Bayesian Semi-Supervised Chinese Word Segmentation for Statistical Machine Translation. In Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK, Aug. 2008.
- Y. Deng, J. Xu, and Y. Gao: Phrase Table Training for Precision and Recall: What Makes a Good Phrase and a Good Phrase Pair? In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, OH, Jun. 2008.
- J. Xu, Y. Deng, Y. Gao and H. Ney: Domain Dependent Machine Translation. In Proceedings of the Machine Translation Summit XI, Copenhagen, Danmark, Sep. 2007.
- J. Xu, R. Zens, and H. Ney: Partitioning Parallel Documents Using Binary Segmentation. In Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Statistical Machine Translation, pp. 78-85, New York City, NY, Jun. 2006.
- D. Vilar, J. Xu, L. F. D'Haro and H. Ney: Error Analysis of Statistical Machine Translation Output. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 697-702, Genova, Italy, May 2006.
- J. Xu, E. Matusov, R. Zens, and H. Ney: Integrated Chinese Word Segmentation in Statistical Machine Translation. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT), Pittsburgh, PA, Oct. 2005.
- R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney: The RWTH Phrase-based Statistical Machine Translation System. InProceedings of the International Workshop on Spoken Language Translation (IWSLT), Pittsburgh, PA, Oct. 2005.
- J. Xu, R.Zens, and H. Ney: Sentence Segmentation Using IBM Word Alignment Model 1. In Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 2005), pp. 280-287, Budapest, Hungary, May 2005.
- J. Xu, R. Zens, and H. Ney: Do We Need Chinese Word Segmentation for Statistical Machine Translation? In Proceedings of the Third SIGHAN Workshop on Chinese Language Learning, pp. 122-128, Barcelona, Spain, Jul. 2004.

COMPETITION AWARDS

I won leading ranks in the following machine translation evaluations:

- *SemEval 2019*, representing CUNY and Peking University, "French-20K-open track" in Task 1: Cross-lingual Semantic Parsing with UCCA, Rank **1st** (in one track), **leader** [This competition have 72 teams registered, and 7 teams submitted results, worldwide. Team member Abdul Rafae Khan and Weimin Lv received student scholarship of SemEval'19.]

- *SemEval 2019*, representing CUNY and Peking University, "English-Wiki-open" and "English-20K-open" and "German-20K-closed" and "German-20K-open track" in Task 1: Cross-lingual Semantic Parsing with UCCA, Rank **2nd** (in four tracks), **leader** [This competition has 72 teams registered, and 7 teams submitted results, worldwide.]
- *WMT 2018 (biggest machine translation of the year)*, representing Hunter College, CUNY, English-French (Bio-medical), Rank **1st**, **leader**
- *WMT 2017 (biggest machine translation of the year)*, representing Hunter College, CUNY, Finnish-English (News), Rank **1st** (in BLEU) and Rank **2nd** (human evaluation), **leader**
  First undergraduate team to participate in an MT competition
- *NIST 2015*, representing ICT/CAS, Chinese-English, Rank 4th (among all) and Rank 1st (among academic affiliations), **leader**
- *WMT 2011*, representing DFKI, English-German, Rank 1st (in BLEU), **leader**
- *NIST 2008*, representing MSR Redmond, Chinese-English, Rank 1st
- *NIST 2006*, representing RWTH-Aachen, Chinese-English, Rank 4th, **leader**
- *NIST 2005*, representing RWTH-Aachen, Chinese-English, Rank 4th
- *NIST 2004*, representing RWTH-Aachen, Chinese-English, Rank 2nd
- *GALE 2008*, representing RWTH-Aachen, Chinese-English, Rank 2nd (as NightInGale)
- *GALE 2007*, representing RWTH-Aachen, Chinese-English, Rank 2nd (as NightInGale)
- *GALE 2006*, representing RWTH-Aachen, Chinese-English, Rank 2nd (as NightInGale)
- *TC-Star 2007*, representing RWTH-Aachen, Chinese-English, Rank 1st
- *TC-Star 2006*, representing RWTH-Aachen, Chinese-English, Rank 1st
- *TC-Star 2005*, representing RWTH-Aachen, Chinese-English, Rank 1st

US Patent

- Jianfeng Gao, Kristina Nikolova Toutanova and Jia Xu:
  Unsupervised Chinese Word Segmentation for Statistical Machine Translation.
  Application No. 12/163.119, (ASM 8310) MS 323635.01
  Redmond, Washington, Jun. 2008.

Administration & Service

Academic year 2012-2013, served full-time as:

- Chair for Elite Tsinghua Computer Science Program
- Chair of student activities and leadership development of IIIS, Tsinghua

Postdoc Supervision

- Cuong Huang, Oct. 2017 - Oct. 2018

Completed PhD Thesis

- Abdul Kahn, Jun.2019, Robust Neural Machine Translation.

Completed MSc Theses

- Geliang Chen (Dec. 2016, Tsinghua Univeristy):
  Phrase-based Language Model for Statistical Machine Translation
- Xiaojun Zhang (Nov. 2011, Univeristy of Saarland):
  Two-level Parallel Text Extraction from Comparable Corpora
  Co-supervised with Hans Uszkoreit

Completed Bachelor Theses

- Shun Zheng (May 2014, Beijing University of Posts and Telecommunications): Improvements on Word Alignment Models in Statistical Machine Translation
- Zhengping Che (Jun. 2013, Elite Tsinghua Computer Science Program, Tsinghua University): Dirichlet Process Model for Phrase-based Machine Translation
- Yulong Zeng (Jun. 2013, Elite Tsinghua Computer Science Program, Tsinghua University): A Comparative Study of Generative Model and Discriminative Model
- Zhibo Zhang (Jun. 2013, Elite Tsinghua Computer Science Program, Tsinghua University): Machine Learning-based Crime Prediction
- Suiqian Luo (Jun. 2013, Elite Tsinghua Computer Science Program, Tsinghua University): Improving Training Models in Statistical Machine Translation
- Geliang Chen (Jun. 2013, Peking University): A Novel Approach for Language Model
- Yong Cheng (Jun. 2012, Beijing Jiao Tong University): Analysis of User Behaviors in Social Network

Current Grants

- **208,400 USD**, Industry Sponsor, 2019-2021, **PI**, Missing Links in News Summarization, Pending.
- 299,888 USD, **NSF** grant, 2018-2023, Co-PI (PI of Subcontract), CNS 1747728 IUCRC Phase I Rutgers, Newark: Center for Accelerated Real Time Analytics (CARTA), Grant No. CNS-1747728.
- 660,000 RMB (100,000 USD), NSFC (NSF-China) grant, 2017-2019, Co-PI, Key Problems for Tightly-coupled, Multi-signal Fusion based Simultaneously Locating and Mapping.
- 200,000 RMB (33,000 USD), KLIIP grant, 2015-2016, Principal Investigator, Novel machine learning methods, Grant No. 20156020
- 500,000 RMB (83,000 USD), ICT grant (Innovation subjects), 2015-2017, Principal Investigator, Ensemble learning in machine translation, Grant No. 20156020
- 660,000RMB (100,000 USD), NSFC grant, 2014-2017, *co-PI*,
  New Approaches to the Limits of Efficient Propositional Reasoning: Algorithms, Approximations and Foundations, Grant No. 20131351464

Completed EU and DARPA Grants

- ACCURAT, EU grant, 2010-2012, *project leader at DFKI and package leader*
- EuroMatrix-Plus, EU grant, 2010-2012, *member*
- Quaero, EU grant, 2009-2010, *member*
- GALE, DARPA grant, 2006-2008, *member*
- TC-Star, EU grant, 2004-2006, *member*

Teaching

I designed and instructed the following courses:

- *Natural Language Processing C-SC 74040* (graduate course)
  Graduate Center, CUNY, Fall 2018.
  Average score in teacher's evaluation: 6 out of 7 - **Excellent**
- *Supervised Research CSCI 49600* (undergraduate course),
  Hunter College, CUNY, Fall 2018.

- *Language Technology CSCI 49362* (undergraduate course),
  Hunter College, CUNY, Fall 2018.
  Average score in teacher's evaluation: 5.1 out of 7 - **Very Good**
- *Individual Study/Research Project C-SC 79000 35822* (graduate course),
  Graduate Center, CUNY, Fall 2018.
- *Language Technology CSCI 49362* (undergraduate course),
  Hunter College, CUNY, Spring 2018.
  Average score in teacher's evaluation: 6.2 out of 7 - **Excellent**
- *Supervised Research CSCI 49600* (undergraduate course),
  Hunter College, CUNY, Spring 2018.
- *Individual Study/Research Project C-SC 79000 35822* (graduate course),
  Graduate Center, CUNY, Spring 2018.
- *Software Design and Analysis III CSCI 33500* (undergraduate course),
  Hunter College, CUNY, Fall 2017.
- *Language Technology CSCI 49362* (undergraduate course),
  Hunter College, CUNY, Fall 2017.
- *Individual Study/Research Project C-SC 79000 35822* (graduate course),
  Graduate Center, CUNY, Fall 2017.
- *Supervised Research CSCI 49600* (undergraduate course),
  Hunter College, CUNY, Spring 2017.
- *Natural Language Processing C-SC 74040* (graduate course)
  Graduate Center, CUNY, Fall 2016.
- *Language Technology LING 83600 32100* (graduate course)
  Graduate Center, CUNY, Fall 2016.
- *Software Analysis and Design I CSCI 13500* (undergraduate course)
  Hunter College, CUNY, Fall 2016.
- *Machine Learning and Machine Translation* (graduate course)
  Tsinghua University, Fall 2014.
- *Speech Communication of Human and Machine* (undergraduate Yao-class course)
  Tsinghua University, Fall 2013, co-taught with P.C. Ching, Tan Lee, Helen Meng and William S.-Y. Wang
- *Machine Learning and Pattern Recognition* (undergraduate Yao-class course)
  Tsinghua University, Spring 2013.
- *Machine Learning and Machine Translation* (graduate course)
  Tsinghua University, Fall 2012
- *Language Technology* (graduate course)
  University of Saarland, Spring 2011, co-taught with Martin Kay and Hans Uszkoreit

Academic Exchange and Long-term Visits

- Visiting Researcher, Jul. 2013 - Jan. 2014
  MSRA Star-Track Visiting Young Faculty Program

- Ph.D. Intern, Feb. 2007 - May 2007
  IBM T. J. Watson Research Center, Yorktown Heights, USA
  Research topic: Phrase extraction and domain adaptation
  Practical work: large-scale Chinese-English statistical machine translation system
  Manager: Yuqing Gao, Mentor: Yonggang Deng
  One ACL Publication and one MT-Summit during this internship.

- Ph.D. Intern, Oct. 2007 - Feb. 2008
  Microsoft Research NLP group, Redmond, USA
  Research topic: Chinese word segmentation for statistical machine translation
  Practical work: NIST MT evaluation campaign 2008
  Manager: Bill Dolan, Mentor: Jianfeng Gao
  One Coling Publication during this internship and one US patent.

INVITED TALKS

- Rutgers University, Newark, Aug. 2016
- Google Research, NYC, Dec. 2015
- University of Columbia, New York City, Nov. 2015
- University of Washington, Seattle, Nov. 2015
- USC, Los Angeles, Nov. 2015
- CWMT, Tutorial "Ensemble Learning for Machine Translation", Macau, Nov. 2014
- DFKI, Berlin, Germany, Oct. 2014
- RWTH-Aachen University, Germany, Oct. 2014
- Aarhus University, Denmark, Oct. 2014
- Stanford University, USA, Apr. 2014
- Facebook, USA, Mar. 2014
- MSR-Asia, Star-Track Research Talk, Beijing, Jun. 2013
- MSR-Asia, Natural Language Computing Group, Beijing, Jul. 2013
- MSR-Asia, "MSRA-IIIS" Bilateral Seminar Series, Beijing, Dec. 2012
- CUHK, Machine Learning Workshop, Hong Kong, Jul. 2012
- MSR, Natural Language Processing Group, Redmond (USA), Sep. 2011
- SRI International, Menlo Park (USA), Sep. 2011
- CAS, National Laboratory of Pattern Recognition Group, Beijing, Nov. 2010
- MSR-Asia, Natural Language Computing Group, Beijing, Nov. 2010
- Systran Software, Paris, Dec. 2009
- Limsi, Spoken Language Processing Group, Orsay (France), 2009
- DFKI, Language Technology Group, Saarbruecken (Germany), Oct. 2009
- IBM, Speech Technology Group, Watson (USA), Feb. 2008

REVIEW OF PH.D. THESIS

- External reviewer for: Aji Prasetya Wibawa, Advanced Javanese-to-Indonesian statistical machine translation, University of South Australia, Aug. 2013.
- Hongyu Liang, Efficient Approximation Algorithms for Graph Optimization Problems: Through a Combinatorial Lens, Tsinghua University, Jun. 2013.
- Jing He, Community Finding and Information Propagation in Social Networks, Tsinghua University, Jun. 2012.

WORKSHOP ORGANIZER

- CUHK, Machine Learning Workshop, Hong Kong, Jul. 2012

LANGUAGES

- Fluent in English
  and German
- Native Speaker in Chinese