

# Wei Xu

3330 Walnut St., University of Pennsylvania  
Philadelphia, PA 19104

✉ xwe@cis.upenn.edu ☎ (718) 612-9347

<http://www.cis.upenn.edu/~xwe/>

Last updated: May 24, 2016 (up-to-date version on my homepage)

---

<b>RESEARCH</b>	Machine Learning, Natural Language Processing, Social Media	
<b>CURRENT POSITION</b>	<b>Assistant Professor, Ohio State University</b> , Columbus, OH	Aug 2016 (expected) – Present
	<i>Department of Computer Science and Engineering</i>	
<b>EDUCATION</b>	<b>Ph.D. in Computer Science, New York University</b> , New York, NY	Jan 2014
	Thesis: Data-driven Approaches for Paraphrasing Across Language Variations	
	<b>B.S. and M.S. in Computer Science, Tsinghua University</b> , Beijing, CHINA	2004/2007
	Thesis: Event-relevance Based Summarization	
<b>EXPERIENCE</b>	<b>Postdoctoral Fellow, University of Pennsylvania</b> , Philadelphia, PA	Feb 2014 – Aug 2016
	<i>Computer Information and Science Department</i>	
	<b>Visiting Researcher, University of Washington</b> , Seattle, WA	Jan 2012 – Dec 2013
	<i>Computer Science and Engineering Department</i>	
	<b>Research Intern, Microsoft Research</b> , Seattle, WA	Jun – Sep 2011
	<i>Internet Services Research Center</i>	
	<b>Research Consultant, Educational Testing Service</b> , Princeton, NJ	Mar – Dec 2010
	<i>Natural Language and Speech Group</i>	
	<b>Software Engineer Intern, Amazon.com</b> , Seattle, WA	Jun – Aug 2010
	<i>Catalog Quality Group</i>	
	<b>Visiting Research Assistant, Hong Kong Polytechnic University</b>	Sep 2005 – Feb 2006
	<i>Department of Computing</i>	
<b>SERVICES</b>	<b>Area Chair:</b> EMNLP (2016)	
	<b>Publicity Chair:</b> NAACL (2016)	
	<b>Session Chair:</b> EMNLP (2015), NAACL (2015), AAAI (2015), ACL (2014)	
	<b>Organizer:</b>	
	– The 1st International Workshop on Noisy User-generatedText at ACL (2015)	
	– The 2nd International Workshop on Noisy User-generated Text at COLING (2016)	
	<a href="http://noisy-text.github.io/">http://noisy-text.github.io/</a>	
	– SemEval shared-task: Paraphrase and Semantic in Twitter (2015)	
	<b>(19 research teams worldwide participated)</b>	
	<a href="http://alt.qcri.org/semeval2015/task1/">http://alt.qcri.org/semeval2015/task1/</a>	
	– Mid-Atlantic Student Colloquium on Speech, Language and Learning (2016)	
	<a href="http://www.mascs11.org/">http://www.mascs11.org/</a>	

**Program Committee:**

ACL (2015, 2014 – Outstanding Reviewer Award, 2013), NAACL (2015), EMNLP (2015, 2014)  
 WWW (2016, 2015), AAAI (2016, 2015, 2012), KDD (2015), COLING (2014)  
 WWW Workshop on #Microposts (2016)  
 ACL Workshop on Social Factors in Natural Language Processing (2016)  
 EACL Workshop on Language Analysis in Social Media (2014)

**Steering Committee:** SemEval (2017)

**Thesis Committee:**

Maria Pershina, Ph.D. in Computer Science, New York University (2016)

**Journal Reviewer:**

Transactions of the Association for Computational (TACL)

**INVITED  
TALKS**

**Multiple-instance Learning from Unlimited Text**

University of Edinburgh, Edinburgh, United Kingdom	May 2016
Ohio State University, Columbus, OH	Apr 2016
University of North Carolina, Chapel Hill, NC	Apr 2016
Arizona State University, Tempe, AZ	Mar 2016
Vanderbilt University, Nashville, TN	Mar 2016
Imperial College London, London, United Kingdom	Mar 2016
University of Waterloo, Waterloo, ON, Canada (CS Seminar)	Mar 2016
Indiana University, Bloomington, IN (Computer Science Colloquium Series)	Feb 2016
Washington University, St Louis, MI (Computer Science & Engineering Colloquia Series)	Feb 2016
Simon Fraser University, Vancouver, BC, Canada	Feb 2016
University of Alberta, Edmonton, AB, Canada (Special Lecture)	Feb 2016
Yale University, New Haven, CT (CS Talk)	Feb 2016
University of Maryland, College Park, MD (CLIP Colloquium)	Oct 2015
Ohio State University, Columbus, OH (Clippers Seminar)	Oct 2015

**Large-scale Paraphrase Acquisition from Twitter**

DARPA's DEFT Project PI Meeting, Boulder, CO	May 2015
--	----------

**Learning and Generating Paraphrases from Twitter and Beyond**

Carnegie Mellon University, Pittsburgh, PA	Apr 2015
Columbia University, New York, NY (NLP Talk)	Apr 2015
Johns Hopkins University, Baltimore, MD (CLSP Colloquium)	Feb 2015

**Paraphrases in Twitter**

Twitter.com, San Francisco, CA	Feb 2015
--------------------------------	----------

**Modeling Lexically Divergent Paraphrases in Twitter (and Shakespeare!)**

The City University of New York, New York, NY (NLP Seminar)	Mar 2015
IBM Research - Almaden, San Jose, CA	Feb 2015
University of California, Berkeley, CA	Feb 2015
The University of Texas, Austin, TX (Forum for Artificial Intelligence)	Feb 2015
Yahoo!, New York, NY	Dec 2014
Carnegie Mellon University, Pittsburgh, PA (CL+NLP Lunch Seminar)	Nov 2014
Microsoft Research, Seattle, WA (Visiting Speaker Series)	Aug 2014

**Incremental Information Extraction**

Stanford Research Institute, Palo Alto, CA	Apr 2012
IARPA's KDD Project PI Meeting, San Diego, CA	May 2011

**Information Extraction Research**  
University of Washington, Seattle, WA

Jan 2011

**Event-based Summarization**  
Thomson Reuters, Eagan, Minnesota, MN  
France Telecom, Beijing, CHINA

Nov 2009  
Mar 2007

**PUBLICATIONS** Refereed Journal and Conference Papers (all have acceptance rates  $\approx 25\%$ )

(\* TACL is the new journal since 2013 that allows presentation at any ACL conferences of author's choice and has quickly become the most selective venue for natural language processing publications. It only accepts about 40 papers per year and does not accept submissions previously published or rejected from conferences.)

*Optimizing Statistical Machine Translation for Simplification*

**Wei Xu**, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch  
Transactions of the Association for Computational (TACL)\*, 2016  
presentation at ACL 2016 (expected)

*Problems in Current Text Simplification Research: New Data Can Help*

**Wei Xu**, Chris Callison-Burch, Courtney Napoles  
Transactions of the Association for Computational (TACL)\*, 2015  
oral presentation at EMNLP 2015

*Extracting Lexically Divergent Paraphrases from Twitter*

**Wei Xu**, Alan Ritter, Chris Callison-Burch, William B. Dolan, Yangfeng Ji  
Transactions of the Association for Computational (TACL)\*, 2014  
oral presentation at NAACL 2015

*Discovering User Attribute Stylistic Differences via Paraphrasing*

Daniel Preotiuc, **Wei Xu**, Lyle Ungar  
Proceedings of AAAI 2016

*Cost Optimization for Crowdsourcing Translation*

Mingkun Gao, **Wei Xu**, Chris Callison-Burch  
Proceedings of NAACL 2015

*Infusion of Labeled Data into Distant Supervision for Relation Extraction*

Maria Pershina, Bonan Min, **Wei Xu**, Ralph Grishman  
Proceedings of ACL 2014

*Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction*

**Wei Xu**, Raphael Hoffmann, Le Zhao, Ralph Grishman  
Proceedings of ACL 2013

*Paraphrasing for Style*

**Wei Xu**, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry  
Proceedings of COLING 2012

*Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-grams Models*

**Wei Xu**, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao  
Proceedings of EMNLP 2011

*Passage Retrieval for Information Extraction using Distant Supervision*

**Wei Xu**, Ralph Grishman, Le Zhao

Proceedings of IJCNLP 2011

*Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task*  
Kristen Parton, Kathleen R. McKeown, Bob Coyne, Mona T. Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gokhan Tur, **Wei Xu**, Sibel Yaman  
Proceedings of ACL-IJCNLP 2009

*Extractive Summarization using Inter- and Intra- Event Relevance*  
Wenjie Li, **Wei Xu**, Mingli Wu, Chunfa Yuan, Qin Lu  
Proceedings of COLING-ACL 2006

### **Refereed Workshop Papers (all peer-reviewed)**

*Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition*  
Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, **Wei Xu**  
Proceedings of ACL 2015 Workshop on Noisy User-generated Text (WNUT)

*SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT)*  
**Wei Xu**, Chris Callison-Burch, William B. Dolan  
Proceedings of SemEval 2015

*Poetry of the Crowd: A Human Computation Algorithm to Convert Prose into Rhyming Verse*  
Quanze Chen, Chenyang Lei, **Wei Xu**, Ellie Pavlick, Chris Callison-Burch  
Proceedings of HCOMP 2014

*Gathering and Generating Paraphrases from Twitter with Application to Normalization*  
**Wei Xu**, Alan Ritter, Ralph Grishman  
Proceedings of ACL 2013 Workshop on Building and Using Comparable Corpora (BUCC)

*A Preliminary Study of Tweet Summarization using Information Extraction*  
**Wei Xu**, Ralph Grishman, Adam Meyers, Alan Ritter  
Proceedings of NAACL 2013 Workshop on Language Analysis in Social Media (LASM)

*New York University 2011 System for KBP (Knowledge Base Population) Slot Filling*  
Ang Sun, Ralph Grishman, **Wei Xu**, Bonan Min  
Proceedings of TAC 2011 (best performance system in NIST KBP-2011 Evaluation slot filling track)

*A Parse-and-Trim Approach with Information Significance for Chinese Sentence Compression*  
**Wei Xu**, Ralph Grishman  
Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG)

*Transducing Logical Relations from Automatic and Manual Annotation*  
Adam Meyers, Michiko Kosaka, Heng Ji, Nianwen Xue, Mary Harper, Ang Sun, **Wei Xu**, Shasha Liao  
Proceedings of ACL-IJNLP Workshop on Linguistic Annotation 2009

*Automatic Recognition of Logical Relations for English, Chinese and Japanese in the GLARF Framework*  
Adam Meyers, Michiko Kosaka, Nianwen Xue, Heng Ji, Ang Sun, Shasha Liao, **Wei Xu**  
Proceedings of NAACL-HLT Workshop on Semantic Evaluations 2009

*Using Non-Local Features to Improve Named Entity Recognition Recall*  
Xinnian Mao, **Wei Xu**, Yuan Dong, Haila Wang  
Proceedings of PACLIC 2007

*Domain Extension of Chinese Named Entity Recognition*

**Wei Xu**, Bin Fu, Liu Liu, Chunfa Yuan, Wenjie Li  
Frontiers of Content Computing (FCC) 2007

*Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis*

**Wei Xu**, Wenjie Li, Mingli Wu, Wei Li, Chunfa Yuan  
Proceedings of CICLing 2006

*Building Document Graph for Text Summarization: An Event-based Approach*

**Wei Xu**, Wenjie Li, Mingli Wu, Wei Li, Chunfa Yuan  
Proceedings of ICCPOL 2006

*The THU/PolyU System at MSE 2006: An Event-relevance based Approach*

**Wei Xu**, Chunfa Yuan, Mingling Wu, Wenjie Li  
Proceedings of MSE 2006

## Theses

*Data-driven Approaches for Paraphrasing Across Language Variations*

**Wei Xu**

Ph.D. Thesis, Department of Computer Science, New York University

## TEACHING

**Instructor**, New Brunswick, NJ

Summer 2016 (expected)

**North American Summer School on Logic, Language, and Information (NASSLLI)**

Course: *Social Media and Text Analytics* (interdisciplinary)

<http://rucss.rutgers.edu/nasslli2016/>

**Instructor, University of Pennsylvania**, Philadelphia, PA

Summer 2015

Course: *Social Media and Text Analytics* (graduate-level)

<http://socialmedia-class.org/>

Designed and taught a new course, covering from basic to state-of-the-art machine learning algorithms for text processing and social media analysis. 20 students attended.

**Guest Lecturer, Rutgers University**, New Brunswick, NJ

Spring 2015

Course: *Natural Language Processing*

**Guest Lecturer, The Ohio State University**, Columbus, OH

Fall 2014

Course: *Natural Language Processing and Information Extraction for the Social Web*

**Guest Lecturer, University of Pennsylvania**, Philadelphia, PA

Spring 2014

Course: *Advanced Topics in Natural Language Processing*

## AWARDS

**Co-author, NSF EAGER: Simplification as Machine Translation**

2014

## AND

## GRANTS

**MacCracken Fellowship** (Award Value: \$230,000)

2007 – 2012

**Travel Grants: Google, Grace Hopper, CRA-Women**

## STUDENT RESEARCH ADVISING

Quanze Chen (undergraduate student at UPenn) [HCOMP 2014]

Bin Fu (undergraduate student at Tsinghua, then Ph.D. at CMU, now Google NYC) [FCC 2007]

Chenyang Lei (undergraduate student at UPenn) [HCOMP 2014]

Mingkun Gao (masters student at UPenn, now a PhD student at UIUC) [NAACL 2015]

Maria Pershina (Ph.D. student at NYU) [ACL 2014]

Siyu Qiu (masters student at UPenn) [paper in submission]

<b>OPEN SOURCE</b>	<i>Multiple-instance Learning Paraphrase Model</i>	Dec 2014
<b>CODE / DATA</b>	<a href="https://www.cis.upenn.edu/~xwe/multip/">https://www.cis.upenn.edu/~xwe/multip/</a>	
	<i>Twitter Paraphrase Corpus</i> (shared-task at SemEval-2015)	Oct 2014
	<a href="http://alt.qcri.org/semEval2015/task1/">http://alt.qcri.org/semEval2015/task1/</a>	
	<i>Event-based Twitter Summarization System</i>	Nov 2013
	<a href="https://github.com/cocoxu/twittersummarization">https://github.com/cocoxu/twittersummarization</a>	
	<i>Twitter Translation Phrase Table</i>	Oct 2014
	<a href="https://github.com/cocoxu/twitterparaphrase">https://github.com/cocoxu/twitterparaphrase</a>	
	<i>Parallel Shakespeare Corpus and Model</i>	Jul 2012
	<a href="https://github.com/cocoxu/Shakespeare">https://github.com/cocoxu/Shakespeare</a>	

<b>PRESS</b>	<i>One Per Cent</i> New Scientist	Oct 29, 2014
	<i>Government Aims to Build a “Data Eye in the Sky”</i> New York Times	Oct 11, 2011

**REFERENCES**

**Chris Callison-Burch**  
**Assistant Professor, Computer and Information Science Department**  
**University of Pennsylvania**, Philadelphia, PA  
 Sloan Research Fellow (2014), Chair of the NAACL Executive Board (2012-2013)  
 ☎ (267) 909-2668  
 ✉ ccb@cis.upenn.edu  
<http://www.cis.upenn.edu/~ccb/>

**Bill Dolan**  
**Research Manager/Principal Researcher, Microsoft Research**, Redmond, WA  
 Manager of the Natural Language Processing Group  
 ☎ (425) 706-3709  
 ✉ billdol@microsoft.com  
<http://research.microsoft.com/en-us/people/billdol/>

**Raymond J. Mooney**  
**Professor, Computer Science Department, University of Texas**, Austin, TX  
 AAAI fellow (2005), ACM fellow (2010), ACL fellow (2014)  
 ☎ (512) 471-9558  
 ✉ mooney@cs.utexas.edu  
<https://www.cs.utexas.edu/~mooney/>

**Ralph Grishman**  
**Professor, Computer Science Department, New York University**, New York, NY  
 Department Chair (1986-1988), President of ACL (1991)  
 ☎ (212) 998-3497  
 ✉ grishman@cs.nyu.edu  
<http://cs.nyu.edu/grishman/>

**Joel Tetreault**  
**Senior Research Scientist, Yahoo! Labs**, New York, NY  
 Treasurer of NAACL (2013-2015)  
 ☎ (310) 612-0315  
 ✉ tetreaul@yahoo-inc.com  
<http://www.cs.rochester.edu/~tetreaul/academic.html>