Tae Yano, Ph.D.

tayano@microsoft.com | tyyano@gmail.com

Career Interests

Applied research and application development in the areas of natural language processing (NLP), including: statistical text analysis, machine learning for NLP, information extraction, topic modeling, computational social science, probabilistic graphical models, text-driven prediction, social media text, and application of NLP to real world problems.

Education

Carnegie Mellon University
Language Technologies Institute, School of Computer Science
Ph.D. in Language and Information Technologies
Concentration: natural language processing
Pittsburgh, Pennsylvania. Sept 2007 - Aug 2013

Columbia University in the City of New York
The Fu Foundation School of Engineering and Applied Science
M.S. in Computer Science
Concentration: natural language processing
Honors: Master of Science Teaching Assistance Fellowship
New York, New York. Sept 2005 - May 2007

Hunter College of the City University of New York
B.A. in Computer Science with honors
Honors: Magna Cum Laude, Dean's honoree
New York, New York. Sept 1998 - Jan 2001

Lewis and Clark College
B.A. in International Affairs
Honors: Dean's honoree, international student grant recipient
Portland, Oregon. Sept 1994 - Aug 1997

Industry Experience

Software Development Engineer (RSDE II)
Microsoft Bing
Bellevue, Washington. Aug 2013 – Present

Project: SMILE (Site Mining Initiative for Local Entity)

- Our team is responsible for the development and maintenance of the SMILE text mining pipeline, which continuously crawls and extracts business entity attributes from web sites, providing vital entity feeds for Bing's Local Entity index.

- I have designed and implemented various entity attribute (e.g., address, name) extraction algorithms for the en-US market. The work involves exploratory research, algorithm design, corpus engineering, machine learning model training and tuning, feature engineering, performance evaluation, model integration, and coordination of internal/external collaborations
- I have contributed to various parts of our extraction pipeline (using Scope), as well as porting our system to Microsoft Azure.
- I have supported external groups to develop markets and segment specific extraction solutions utilizing our extraction algorithms and resources.

Research Intern
Microsoft Research
Redmond, Washington. May 2012 - Aug 2012

Supervisors: Michael Gamon and Patrick Pantel
Project: Detecting salient entities in web documents.

- I have worked on research and development of the core prediction models, including model design, data collection, evaluation, and experiments.
- I am also a co-inventor of the subsequent patent derived from this research

Research Programmer
Center for Computational Learning Systems (CCLS) Columbia University
New York, New York. Jun 2007 - Aug 2007

Project: Building machine learning based system to predict electric power grid failure, in collaboration with the city of New York.

- Implemented various modules in python for model learning and integration

System Engineer
Software Engineering Group, Ricoh Corporation
West Caldwell, New Jersey. Sept 2001 - Sept 2005

- Design and development of large-scale embedded application suits in C/C++ for UNIX based document management devices
- Development and support of proprietary API for the device development;
- Education and training of 3rd party software developers for API

UNIX System Administrator and Application Programmer
Office of Instructional Computing and Information Technology (ICIT) and Department of Computer Science, Hunter College
New York, New York. Feb 1999 - Jan 2001

Research and Teaching Experience

Graduate Research Assistant
School of Computer Science, Carnegie Mellon University
Pittsburgh, Pennsylvania. Sept 2007 - Aug 2013

Advisors: Noah A. Smith, William Cohen
Projects: Statistical NLP and text analysis in political social media; Congressional Roll Call voting prediction from texts; Modeling and tracing of dynamic voting process;

Teaching Assistant

School of Computer Science, Carnegie Mellon University
Pittsburgh, Pennsylvania. Sept 2009 - May 2010

- Courses taught: Text Driven Forecasting (A graduate level course investigating the statistical techniques in text analysis and their use in predicting real world event); Language and Statistics II (A graduate level statistical NLP course); Advanced NLP Seminar.

Supervised Individual Research and
Department Research Assistant (DRA)

Computer Science Department, School of Engineering and Applied Science,
Columbia University
New York, New York. Sept 2006 - May 2007; May 2006 - Aug 2006

Advisor: Rebecca J. Passonneau
Projects: CLiMB 2 - Computational Linguistics for Metadata Building

- Information extraction and text mining from textbooks in the Art History domain.
- Design and implementation of annotation ontology, annotation study, and machine learning experiments
- Co-authored the internal and external publications and presentations.

Master of Science Teaching Assistant Fellow

Computer Science Department, School of Engineering and Applied Science,
Columbia University
New York, New York. Jan 2005 - May 2007

- Courses Taught: Advanced Programming (3 semesters); Data Structure in C/C++ (2 semesters); Programming Languages and Translator (2 semesters); Discrete Mathematics (1 semester)

Skills

Programming Language:

- C#, C and Python: extensive, hands on knowledge (5+ years)
- Scope programming for Cosmos (4+ years)
- Solid understanding of OOP concepts and practice
- Experienced in working with large scale data on distributed development platform
- Working knowledge in C++ and Java (1 – 2 years)
- Other programming experience: LISP, PL/SQL, Matlab, R, Perl, awk, various shell scripting, and assembly (8086).

Languages:

- Fluent Japanese and English,
- College Level Spanish

Patent

Identifying salient items in documents

Michael Gamon, Patrick Pantel, Xinying Song, Tae Yano, Johnson Tan Apacible
U. S. Patent 9251473 B2
February 02, 2016

Publications

Ph.D. Thesis:

Text as Actuator: Text-Driven Response Modeling and Prediction in Politics

Carnegie Mellon University, School of Computer Science Technical Report CMU-LTI-13-006. Pittsburgh, PA. July 2013

Thesis Committee: Noah A. Smith (University of Washington), William W. Cohen (Carnegie Mellon University), Jason I. Hong (Carnegie Mellon University), Philip Resnik (University of Maryland).

Refereed Conferences (with presentation):

Identifying salient entities in web pages

Michael Gamon, Tae Yano, Xinying Song, Johnson Apacible and Patrick Pantel
In proceedings of the 22nd International Conference on Information and Knowledge Management (CIKM). Burlingame, CA. Oct 2013

Exploring venue-based city-to-city similarity measures

Daniell Preoiuc-Pietro, Justin Cranshaw, and Tae Yano
In proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. Chicago IL. Aug 2013

A Penny for your Tweets: Campaign Contributions and Capitol Hill Microblogs

Tae Yano, Dani Yogatama, and Noah A. Smith
In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM 2013). Boston, MA. July 2013

Textual Predictors of Bill Survival in Congressional Committees

Tae Yano, John Wilkerson and Noah A. Smith
In Proceedings of the North American Association for Computational Linguistics Human Language Technologies Conference (NAACL). Montreal, Quebec. June 2012

Bayesian Nonparametric for Information Extraction

Jacob Eisenstein, Tae Yano, William W. Cohen, Noah A. Smith, and Eric Xing
In Proceedings of the EMNLP Workshop on Unsupervised Learning in NLP. Edinburgh, UK. July 2011

Seeing a Home away from the Home: Distilling proto-Neighborhood from Incidental data with Topic Modeling

Justin Cranshaw and Tae Yano
In Proceedings of the Workshop on Computational Social Science and the Wisdom of Crowds, Annual Conference on Neural Information Processing System (NIPS). Vancouver, B.C., Canada. Dec 2010

Shedding (a Thousand Points of) Light on Biased Language

Tae Yano, Philip Resnik, and Noah A. Smith

In Proceedings of the NAACL-HLT Workshop on Creating Speech and Language Data With Mechanical Turk. Los Angeles, CA. June 2010

What's Worthy of Comment? Content and Comment Volume in Political Blogs with Topic Models

Tae Yano and Noah A. Smith
In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM). Washington, D.C. May 2010

Predicting Response to Political Blog Posts with Topic Models

Tae Yano, William W. Cohen, and Noah A. Smith
In Proceedings of the North American Association for Computational Linguistics Human Language Technologies Conference (NAACL). Boulder, CO. May/June 2009

Relation between Agreement Measures on Human Labeling and Machine Learning Performance: Results from an Art History Domain

Rebecca Passonneau, Tom Lippincott, Tae Yano, and Judith Klavans
In Proceedings of the International Conference on Language Resources and Evaluation (LREC). Marrakesh, Morroco. May/Jun 2008

Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access

Judith Klavans, Carolyn Sheffield, Eileen Abels, Joan Beaudoin, Laura Jenemann, Jimmy Lin, Tom Lippincott, Rebecca Passonneau, Tandeep Sidhu, Dagobert Soergel, and Tae Yano
In proceedings of the LREC Workshop on Language Resources for Content-Based Image Retrieval (OntoImage 2008). Marrakesh, Morroco. May/Jun 2008

Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning

Rebecca Passonneau, Tae Yano, and Judith Klavans
In Proceedings of the Workshop on Metadata Mining for Image Understanding, International Conference on Computer Vision Theory and Applications (VISAPP). Funchal, Portugal. Jan 2008

Internal Reports and Collaborations:

Understanding Document Aboutness - Step One: Identifying Salient Entities

Michael Gamon, Tae Yano, Xinying Song, Johnson Apacible and Patrick Pantel
Microsoft Research Technical Report MSR-TR-2013-73

Emotional Convergence Among Members of Online Social Networks

Doug Pierce, David P. Redlawsk, William W. Cohen, Tae Yano, Ramnath Balasubramanyan
APSA 2012 Annual Meeting Presentation. New Orleans, LA. Sept 2012

Assessing the effects of emotion-laden messages in a social network

David Redlawsk, Doug Pierce, William W. Cohen, Tae Yano, and Ramnath Balasubramanyan
APSA 2011 Annual Meeting Presentation. Seattle WA. Sept 2011

Experiments on Non-Topical Paragraph Classification of Art History Textbooks

Term project for Search Engine Technology, Columbia University
New York, New York. Fall 2006

KP: A knitting language
Term project for Programming Languages and Translator, Columbia University
New York, New York. Fall 2004

Presentations (without publication)

Location Summit 2015
Microsoft internal conference
Presentation: SMILE – Chain Business Entity Extraction for Bing Local
Redmond, Washington. July 2015

Practice of Machine Learning Conference (PMLC) 2014
Microsoft internal conference
Presentation: SMILE – The Never Ending Story of Address Extraction
Redmond, Washington. March 06, 2014

New Directions in Analyzing Text as Data 2012
Presentation: Textual Predictors of Bill Survival in Congressional Committees
Harvard University, Institute for Quantitative Social Science
Cambridge, Massachusetts. Oct 05 - 06, 2012

Social Computational Systems (SoCS) Ph.D Symposium
Presentation: Predicting Voter Behavior from the Observables
University of Minnesota, Twin Cities
Minneapolis, Minnesota. Jun10 - 12, 2011

Text-as-Data Conference 2011
Presentation: Observing Motivated Reasoners and Affective Tipping Point in Political Decision Making
Northwestern University, Kellogg School of Management
Evanston, Illinois. Mar 10 - 12, 2011

Text-as-Data Conference 2010
Northwestern University, Kellogg School of Management
Evanston, Illinois. Mar 08 - 10, 2010

IBM Statistical Machine Learning and Its Application (SMiLe)
Presentation: Modeling Political Blogs with Response
IBM Thomas J. Watson Research Center
Yorktown Heights, New York. Oct 08 - 09, 2009

Conference Reviewing

- International AAAI Conference on Weblogs and Social Media (ICWSM)
- The European Chapter of the Association for Computational Linguistic (EACL)
- Association for Computational Linguistic (ACL)
- International World Wide Web conference (WWW)

- The North American Chapter of the Association for Computational Linguistics (NAACL)
- Conference on Empirical Methods in Natural Language Processing (EMNLP)
- Association for the Advancement of Artificial Intelligence Conference (AAAI)
- The International Conference on Language Resources and Evaluation (LREC)