Xi Victoria Lin

Contact Information

Ph.D. Student
Department of Computer Science and Engineering
University of Washington
AC486 Paul G. Allen Center
185 Stevens Way
Seattle, WA 98195-2350, USA

Mobile: +1-206-519-8086
E-mail:
xilin@cs.washington.edu
Website:
http://victorialin.net

Research Interests

Natural language understanding, machine learning and knowledge representation.

Education

University of Washington, Seattle, WA
Ph.D. candidate, Computer Science and Engineering, June 2013 - *present*
- Adviser: Professor Luke Zettlemoyer (November 2013 - *present*)
  Professor Benjamin Taskar (June 2013 - November 2013)
- Research Interest: semantic parsing, natural language understanding.

University of Pennsylvania, Philadelphia, PA
Ph.D. graduated as M.Sc., Computer and Information Science, May 2013
- Cumulative GPA: 3.86/4.0
- Adviser: Professor Benjamin Taskar
- Area of Study: Natural Language Processing and Machine Learning

University of Oxford, Kellogg College, Oxford, UK
M.Sc., Computer Science (w. Distinction), September 2011
- Thesis Topic: *Fine-grained Named Entity Classification in Machine Reading*
- Adviser: Professor Stephen Pulman
- Area of Study: Computer Science focused on Logic and Semantics

The Hong Kong Polytechnic University, Kowloon, HK
B.S., Electronic and Information Engineering, August 2010
- Cumulative GPA: 4.0/4.0
- Thesis Topic: *Robust Real-time Face Detection*
- Adviser: Professor Kenneth K. M. Lam
- Minor in Computer Science

Xi'an Jiaotong University, Xi'an, China
Special Class for the Gifted Young, January 2008
- University preparatory program emphasizing math and science training for 9th-grade students selected from all over mainland China.

Research Projects

Program Synthesis from Natural Language Using Neural Networks.
Demo: http://tellina.rocks
Joint work with Chenglong Wang, Kevin Vu, Dr. Michael Ernst, Dr. Luke Zettlemoyer and Deric Pang.
Even if a competent programmer knows what she wants to do and can describe it in English, it can still be difficult to write code to achieve the goal. Existing resources, such as question-and-answer websites, tabulate specific operations that someone has wanted to perform in the past, but they are not effective in generalizing to new tasks, to compound tasks that require combining previous questions, or sometimes even to variations of listed tasks. Our goal is to make programming easier and more productive by letting programmers use their own words and concepts to express the intended operation, rather than forcing them to accommodate the machine by memorizing its grammar. We have built a system, Tellina, that lets a programmer describe a desired operation in natural language, then automatically translates it to a programming

language for review and approval by the programmer. In a controlled study, programmers who had access to Tellina outperformed those who did not, even when Tellina's predictions were not completely correct, to a statistically significant degree. We are continuingly improving Tellina's network structure and training data to make it more accurate and cover more programming languages. (Summer 2016 - *present*)

Cross-lingual Distant Supervision for Relation Extraction.
Joint work with Dr. Luke Zettlemoyer and Dr. Sameer Singh
Distant supervision is the paradigm that creates its own training data by heuristically matching the contents of a database to corresponding text. Such paradigm can be naturally extended to a multi-lingual setting since the same fact can be expressed in multiple languages. The widely used common-sense knowledge base, *Freebase*, stores multi-lingual aliases for its entities. A naive approach is hence to use the aliases as anchor texts for mapping relation tuples to text, and train a relation extractor for each language separately. However, for a long tail of Freebase entities, either the foreign translation is missing or they are never mentioned in a foreign language, which makes the distant supervision signal much weaker in many other languages than in English. In this project, we propose bilingual projection for strengthening the distant supervision signal in the secondary language, Chinese. The assumption is that text patterns which are translations to each other should express the same set of relations. On the other hand, patterns that express the same relations are likely to be translations of each other. We are currently investigating the effect of adding the above assumptions to the distant supervision model. (Spring 2014 - *present*)

Multi-label Learning with Posterior Regularization.
Joint work with Dr. Sameer Singh, Luheng He, Dr. Benjamin Taskar and Dr. Luke Zettlemoyer
In many multi-label learning problems, especially as the number of labels grow, it is challenging to gather completely annotated data. In this work, we present a new approach for multi-label learning from incomplete annotations. Specifically, it uses the posterior regularization technique to enforce soft constraints on the output of inductive classifiers, regularizing them to prefer sparse and low-rank (SLR) predictions. By avoiding strict low-rank constraints we enable a learned trade-off leading to better generalization of the classifiers. The resulting joint objective can be optimized efficiently using EM and stochastic gradient descent. Experiments in both image and text domains demonstrate the contributions of each modeling assumption and show that the approach achieves state-of-the-art performance on a number of challenging datasets. (Fall 2013 to Spring 2014)

Research Internships

Microsoft Research, Redmond, WA. (Summer 2015)
Efficient Relation Extraction from Embedded Knowledge Graph and Text.
Joint work with Dr. Kristina Toutanova, Dr. Wen-tau Yih, Dr. Hoifung Poon and Dr. Chris Quirk
Modeling relation paths has offered significant gains in embedding models for knowledge base (KB) completion. However, enumerating paths between two entities is very expensive, and existing approaches typically resort to approximation with a sampled subset. This problem is particularly acute when text is jointly modeled with KB relations and used to provide direct evidence for facts mentioned in it. In this project, we proposed the first exact dynamic programming algorithm which enables efficient incorporation of all relation paths of bounded length, while modeling both relation types and intermediate nodes in the compositional path representations. We conduct a theoretical analysis of the efficiency gain from the approach. Experiments on two datasets show that it addresses representational limitations in prior approaches and improves accuracy in KB completion.

Allen Institute for Artificial Inteliigence, Seattle, WA. (Summer 2014)
Knowledge Extraction for Elementary Physics Reading Comprehension.
Joint work with Dr. Tom Kwiatkowski
Elementary physics reading comprehension posts significant challenges for text understanding and knowledge inference. Compared to newswire and common-sense reading comprehension, a piece of physics text could describe 1) relations between abstract concepts such as force, mass, speed etc.; 2) physical interactions between real world objects; 3) mathematical calculation of quantities. We thus formulated the reading comprehension problem in a knowledge extraction and inference pipeline, and focused on designing a knowledge representation which supports efficient inference. By the end of this project (which lasted for only 9 weeks) we concluded that hand-engineering the domain-specific knowledge schema does not scale and caught us in the knowledge representation explosion faced by most non-statistical AI techniques. This project left two difficult open problems: 1) can we develop a symbolic language to describe the operation and inference required to solve physics problems? 2) can we train machines to deduce and execute highly abstract inference procedures in an end-to-end set-up? (The solution to 1 probably lies in WolframAlpha.)

CONFERENCE PUBLICATIONS

[1] Lin, X. V., Wang, C., Pang, D., Vu, K., Zettlemoyer, L., Ernst, M. Program Synthesis from Natural Language Using Recurrent Neural Networks. *In submission.*

[2] Toutanova, K., Lin, X. V., Yih, W., Poon, H., Quirk, C. Compositional Learning of Embeddings for Relation Paths in Knowledge Bases and Text. In *processings of the 54th Anunual Meeting of the Association for Computational Linguistics (ACL)*, August 7-12, 2016, Berlin, Germany.

OTHER PUBLICATIONS

[3] Lin, X.V., Singh, S., He, L., Taskar, B., and Zettlemoyer, L. Multi-label Learning with Posterior Regularization. In *NIPS Workshop on Modern Machine Learning and Natural Language Processing*, December 08–12, 2014, Montreal, Canada.

[4] Lin, X.V. Fine-grained Named Entity Classification in Machine Reading. *M.Sc. thesis.* Oxford University. 2011.

OTHER PROJECTS

Aspect-Oriented Event Summarization based on Wiki-Templates.
Supervised by Dr. Benjamin Taskar.
(Winter 2013 to Summer 2013).

Named Entity Linking with Dual Decomposition.
Supervised by Dr. Benjamin Taskar.
(Fall 2012).

What Were They Talking About: Summarizing Quotations in News.
Supervised by Dr. Benjamin Taskar.
(Summer 2012).

Bullet Points: Structured Summarization for News Events.
Supervised by Dr. Benjamin Taskar.
(Spring 2012 to Summer 2012).

Fine-grained Named Entity Classification in Machine Reading.
Supervised by Dr. Stephen Pulman.
Psycholinguistic study suggests that our lexical memory organizes named entities and common nouns in a taxonomic hierarchy. The task of fine-grained named entity classification intends to classify the set of named entities in text into entry-level categories. By "entry-level categories" we mean the categories used by humans for

understanding the article. For example, besides knowing an entity appeared in text is a *person*, readers are more interested in knowing whether it is a *politician*, a *journalist* or an *actor*. We focus specifically on the *person* domain, where we found that in news reports, a single document is often enough for deciding the right type. We formulate the problem as searching for "is-a" relations between a person and an occupation, and develop a bootstrapping-style classifier which jointly discovers new person/occupation entities as well as new "is-a" relation tuples. The classifier performs competitively on manually annotated sentences. (Spring 2011 to Summer 2011).

Teaching Experience

University of Pennsylvania, Philadelphia, PA

Teaching Assistant

- CIS520: Machine Learning Fall 2012
  - Making exam questions; answering Piazza questions; office hours; grading.

Honours and Awards

University of Pennsylvania

- CIS Department Doctorate Fellowship, 2011–2013.

The Hong Kong Polytechnic University

- Best Academic Performance Award, EIE Department, 2009–2010
- Hong Kong SAR Government Scholarship, 2009-2010
- Hong Kong Polytechnic University Post-entry Scholarship, 2008–2009
- Hong Kong & Kowloon Electrical Appliances Merchants Association Scholarship, 2008–2009
- Apple Inc. WWDC Student Scholarship, 2009

References Available to Contact

Dr. Luke Zettlemoyer (e-mail: lsz@cs.washington.edu)

⋆ *Ph.D. research adviser*

- Assistant Professor , Computer Science and Engineering, University of Washington
  - ◇ 185 Stevens Way, Seattle, WA 98195-2350

Dr. Benjamin Taskar (1977-2013)

⋆ *former Ph.D. research adviser*

- Boeing Associate Professor , Computer Science and Engineering, University of Washington

Dr. Stephen Pulman (e-mail: sgp@clg.ox.ac.uk; phone: +44-186-561-0800)

⋆ *M.Sc. thesis advisor*

- FBA Professor, Computational Linguistics, University of Oxford
  - ◇ Wolfson Building, Parks Road, Oxford OX1 3QD

Dr. Kenneth K.M. Lam (e-mail: enkmlam@polyu.edu.hk; phone: +852-2766-6207)

⋆ *B.Eng. thesis advisor*

- Professor, Electronic and Information Engineering, The Hong Kong Polytechnic University
  - ◇ Room DE503c, EIE Department, HKPU, Hung Hom, Kowloon, Hong Kong.