Sam Thomson | Curriculum Vitae

(312)-315-3526 • sthomson@cs.cmu.edu

www.samthomson.com

Education

Carnegie Mellon University, School of Computer Science **Pittsburgh, PA**
Ph.D. in Language and Information Technology (expected) *Aug 2014–Jan 2019*
Proposed thesis: Encoding and Decoding Graph Representations of Natural Language

Carnegie Mellon University, School of Computer Science **Pittsburgh, PA**
Master of Language Technologies *Aug 2012–Jul 2014*

Cornell University, School of Arts and Sciences **Ithaca, NY**
BA cum laude in Mathematics (Computer Science minor) *Aug 2000–May 2004*

Research Experience

Carnegie Mellon University **Pittsburgh, PA**
Graduate Research Assistant, Advisor: Noah A. Smith *Aug 2012–present*

- Visiting Ph.D. Student at University of Washington, Seattle, Sep 2015–present

- Advanced the state of the art in semantic dependency parsing, abstract meaning representation parsing, semantic role labeling (FrameNet and PropBank), coreference resolution, and scene graph parsing.

- Developed novel algorithms for maximum spanning connected subgraph solving, prize-collecting Steiner tree solving, marginalizing softmax-margin SegRNNs, and backpropagating through structured argmaxes.

- Improved accuracy, runtime, and usability of the frame-semantic parser SEMAFOR (1,000+ downloads). Parallelized, increasing speed by a factor of 7, and reduced memory-usage by a factor of 3. Demo: `http://demo.ark.cs.cmu.edu/parse/`.

- Supervised undergraduate researchers, summer 2014 and fall 2017.

Professional Experience

Knewton **New York, NY**
Software Engineer, Adaptive Learning Team *Feb 2012–Jul 2012*

- Developed statistical models and supporting infrastructure for adaptive learning platform. The adaptive learning platform recommended the next module for a student to work on in an online course.

- Worked on infrastructure for online updating and serving hundreds of millions of model parameters using Cassandra, Kafka, ZooKeeper, and Amazon CloudFormation.

- Implemented a Gibbs-sampled item response theory model in Python.

Sulia **New York, NY**
Lead Software Engineer *Oct 2009–Oct 2011*

- Developed a distributed pipeline for crawling users, lists, and tweets with Twitter's API. Collected and kept current a database of the 20 million most active tweeters and two million lists.

- Created a search index of users and lists with a public API using Solr, which was used in lieu of Twitter's own search by such clients as FlipBoard, TweetDeck, UberSocial and Mashable. Sulia's API served 13 million requests per day.

- Bootstrapped a classifier for Twitter Lists by hand-seeding categories and using a weighted KNN model.

- Trained language detection for tweets using a character n-gram model.

- Implemented a recommendation system based on a user's Twitter or Facebook connections. Used a mixture model with smoothed MLE to suggest topics to users as they sign in.

○ **Networked News**   **New York, NY**
Cofounder/CTO   *Apr 2008–Oct 2009*
- Developed a classifier that tagged news articles with related Wikipedia pages. Used Hadoop on EC2 to index Wikipedia.

- Built an RSS news reader using Django that allowed filtering by these tagged topics.

Teaching Experience

○ **The University of Washington**   **Seattle, WA**
Teaching Assistant for Noah A. Smith   *Winter 2017*
Introduction to Natural Language Processing
39 undergraduate students

○ **Carnegie Mellon University**   **Pittsburgh, PA**
Teaching Assistant for Alon Lavie, Chris Dyer, and Robert Frederking   *Fall 2014*
Algorithms for Natural Language Processing
50 graduate students

Peer-reviewed Publications

(https://scholar.google.com/citations?user=g6MislAAAAAJ)

Conference Long Papers

Peng, Hao, Roy Schwartz, **Sam Thomson**, and Noah A. Smith (2018). "Rational Recurrences." In: *Proc. of EMNLP. (to appear)*.

Swayamdipta, Swabha, **Sam Thomson**, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith (2018). "Syntactic Scaffolds for Semantic Structures." In: *Proc. of EMNLP. (to appear)*.

Peng, Hao, **Sam Thomson**, and Noah A. Smith (2018). "Backpropagating through Structured Argmax using a SPIGOT." In: *Proc. of ACL*. [Best Long Paper Honorable Mention].

Schwartz[*], Roy, **Sam Thomson**[*], and Noah A. Smith (2018). "SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines." In: *Proc. of ACL. ([*]Equal contribution)*.

Peng, Hao, **Sam Thomson**, Swabha Swayamdipta, and Noah A. Smith (2018). "Learning Joint Semantic Parsers from Disjoint Data." In: *Proc. of NAACL*.

Zellers, Rowan, Mark Yatskar, **Sam Thomson**, and Yejin Choi (2018). "Neural Motifs: Scene Graph Parsing with Global Context." In: *Proc. of CVPR*.

Peng, Hao, **Sam Thomson**, and Noah A. Smith (2017). "Deep Multitask Learning for Semantic Dependency Parsing." In: *Proc. of ACL*.

Liu, Fei, Jeffrey Flanigan, **Sam Thomson**, Norman Sadeh, and Noah A. Smith (2015). "Toward Abstractive Summarization Using Semantic Representations." In: *Proc. of NAACL*.

Flanigan, Jeffrey, **Sam Thomson**, Jaime Carbonell, Chris Dyer, and Noah A. Smith (2014). "A Discriminative Graph-Based Parser for the Abstract Meaning Representation." In: *Proc. of ACL*. [Best Long Paper Honorable Mention, 100+ citations].

Conference Short Papers

Kshirsagar, Meghana, **Sam Thomson**, Nathan Schneider, Jaime Carbonell, Noah A. Smith, and Chris Dyer (2015). "Frame-Semantic Role Labeling with Heterogeneous Annotations." In: *Proc. of ACL*.

Thomson, **Sam**, Brendan O'Connor, Jeffrey Flanigan, David Bamman, Jesse Dodge, Swabha Swayamdipta, Nathan Schneider, Chris Dyer, and Noah A. Smith (2014). "CMU: Arc-Factored, Discriminative Semantic Dependency Parsing." In: *Proc. of SemEval*.

Workshop Papers

Flanigan, Jeffrey, **Sam Thomson**, David Bamman, Jesse Dodge, Manaal Faruqui, Brendan O'Connor, Nathan Schneider, Swabha Swayamdipta, Chris Dyer, and Noah A. Smith (2014). "Graph-based Algorithms for Semantic Parsing." In: *ACL 2014 Workshop on Semantic Parsing*.

Professional Activities

- Program comittee member, ACL 2018, NAACL 2018, ACL 2017, EMNLP 2017, CoNLL 2017, ACL 2016, EMNLP 2016, EMNLP 2015, and other various conferences and workshops.

- Open source:

  - Soft Patterns (1st/2 contributors): Text classifier using neural WFSAs.

  - SEMAFOR (1st/4 contributors): Frame-semantic parser using log-linear models.

  - Chu-Liu-Edmonds (1st/1 contributor): Efficient reference implementation of CLE.

  - open-SESAME (2nd/2 contributors): Semantic role labeler using a softmax-margin SegRNN.

  - JAMR (2nd/3 contributors): AMR parser using the MSCG algorithm and Lagrangean relaxation.

  - NeurboParser (2nd/3 contributors): Multitask semantic parser using joint MAP inference in a neural factor graph.

  - Various smaller contributions: XGBoost (docs); Morpha (stemming bugfixes); Spire (GCD for polynomials); Purescript Lists, Maps (stack safety); Purescript Flare (UI for Lists); etc.