Sam Thomson — Curriculum Vitae

✉ sam@samthomson.com

About

- Natural language processing (NLP) and machine learning (ML) researcher with expertise in semantic parsing, NL2code, large language models (LLMs), graph neural networks (GNNs), task-oriented dialogue, and voice assistants
- Seeking an industry research position in an ambitious and highly collaborative group.

Experience

- **Microsoft** **Berkeley, CA / Northampton, MA (remote)**
  Senior Researcher, Semantic Machines *Mar 2019–present*
  - Deep learning for NL2Code: designed and conducted experiments aimed at increasing accuracy, robustness, and scope of the NL2Code component in a production task-oriented dialogue system. Iterated on model architectures, training/optimization, feature sources, and output representations, increasing accuracy from 68.2% to 80.7% (on our public dataset). Engaged with the academic community by publishing in top-tier conferences and journals, releasing open-source code, data, benchmarks, and leaderboards.
  - LLM usage: helped capitalize on Microsoft's early access to GPT inference.
    - Constrained decoding for NL2Code: Extended Earley's algorithm to handle left-to-right streaming decoding, constrained by context-sensitive grammars (for SQL, Python, Scala, Lisp, etc.). Used context-sensitivity to handle variable assignment and reference, type inference (including typeclasses), SQL column references, etc. Designed synchronous grammars to transpile to/from more predictable English-like DSLs (patent #20220327288, "Semantic Parsing of Utterance Using Contractive Paraphrasing," 2022). Ideas and techniques are being ported to production (for tool usage in Copilot).
    - Medium-shot learning: experimented with a form of retrieval-augmented generation (RAG) which finds the most helpful examples that fit into a prompt; this technique helps teams transition from an initial few-shot prototyping stage into a more robust, accurate, and deployable medium-shot stage (roughly 500-example training set).
  - Mentoring: managed two summer PhD interns, leading to two conference publications (one Outstanding Paper award) and one patent (#20230367602, "Speculative Execution of Dataflow Program Nodes," 2023). Explored two applications of real-time parsing of streaming automatic speech recognition (ASR): speculative tool usage, and mixed dictation and commanding for text-editing by voice.
- **Carnegie Mellon University** **Pittsburgh, PA**
  Graduate Research Assistant, Advisor: Noah A. Smith *Aug 2012–Mar 2019*

- Visiting Ph.D. Student at University of Washington, Seattle, Sep 2015–Mar 2019
- 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.
- Supervised undergraduate researchers, summer 2014 and fall 2017.

○ **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.

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

Education

○ **Carnegie Mellon University, School of Computer Science** **Pittsburgh, PA**
Ph.D. in Language and Information Technology *Aug 2014–Mar 2019*
Dissertation: 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*

Publications

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

Journal Articles

Andreas, Jacob, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, **Sam Thomson**, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov (2020). "Task-Oriented Dialogue as Dataflow Synthesis." In: *TACL* 8.

Conference Long Papers

Roy, Subhro, **Sam Thomson**, Tongfei Chen, Richard Shin, Adam Pauls, Jason Eisner, and Benjamin Van Durme (2023). "BenchCLAMP: A Benchmark for Evaluating Language Models on Syntactic and Semantic Parsing." In: *Proc. of NeurIPS (Datasets and Benchmarks)*.

Li, Belinda Z., Jason Eisner, Adam Pauls, and **Sam Thomson** (2023). "Toward Interactive Dictation." In: *Proc. of ACL*.

Stengel-Eskin, Elias, Emmanouil Antonios Platanios, Adam Pauls, **Sam Thomson**, Hao Fang, Benjamin Van Durme, Jason Eisner, and Yu Su (2022). "When More Data Hurts: A Troubling Quirk in Developing Broad-Coverage Natural Language Understanding Systems." In: *Proc. of EMNLP*.

Belyy, Anton, Chieh-yang Huang, Jacob Andreas, Emmanouil Antonios Platanios, **Sam Thomson**, Richard Shin, Subhro Roy, Aleksandr Nisnevich, Charles Chen, and Benjamin Van Durme (2022). "Guided K-best Selection for Semantic Parsing Annotation." In: *Proc. of ACL (Demo)*.

Zhou, Jiawei, Jason Eisner, Michael Newman, Emmanouil Antonios Platanios, and **Sam Thomson** (2022). "Online Semantic Parsing for Latency Reduction in Task-Oriented Dialogue." In: *Proc. of ACL*. [Outstanding Paper Award].

Shin, Richard, Christopher Lin, **Sam Thomson**, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme (2021). "Constrained Language Models Yield Few-Shot Semantic Parsers." In: *Proc. of EMNLP*. [100+ citations].

Platanios, Emmanouil Antonios, Adam Pauls, Subhro Roy, Yuchen Zhang, Alexander Kyte, Alan Guo, **Sam Thomson**, Jayant Krishnamurthy, Jason Wolfe, Jacob Andreas, and Dan Klein (2021). "Value-Agnostic Conversational Semantic Parsing." In: *Proc. of ACL*.

Yin, Pengcheng, Hao Fang, Graham Neubig, Adam Pauls, Emmanouil Antonios Platanios, Yu Su, **Sam Thomson**, and Jacob Andreas (2021). "Compositional Generalization for Neural Semantic Parsing via Span-level Supervised Attention." In: *Proc. of NAACL*.

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

Swayamdipta, Swabha, **Sam Thomson**, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith (2018). "Syntactic Scaffolds for Semantic Structures." In: *Proc. of EMNLP*. [100+ citations].

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*. [900+ citations].

Peng, Hao, **Sam Thomson**, and Noah A. Smith (2017). "Deep Multitask Learning for Semantic Dependency Parsing." In: *Proc. of ACL*. [100+ citations].

Liu, Fei, Jeffrey Flanigan, **Sam Thomson**, Norman Sadeh, and Noah A. Smith (2015). "Toward Abstractive Summarization Using Semantic Representations." In: *Proc. of NAACL*. [300+ citations].

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, 300+ 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 Rolling Review (ARR) 2021–2024; ACL 2016–2018; EMNLP 2015–2017; NAACL 2018–2019; Repl4NLP 2017–2018; CoNLL 2017, and other various conferences and workshops.

- Open source contributor:
  - Online Semantic Parsing: Graph-based prefix2code semantic parser, with speculative execution and evaluation harness
  - Semantic Parsing with Constrained Language Models: LLM benchmark for parsing by Earley-constrained decoding
  - Task-Oriented Dialogue as Dataflow Synthesis: Large dialogue dataset (100k+ examples) with baseline models
  - Soft Patterns: Text classifier using neural WFSAs
  - SEMAFOR: Frame-semantic parser using log-linear models
  - Chu-Liu-Edmonds: Efficient reference implementation of CLE
  - open-SESAME: Semantic role labeler using a softmax-margin SegRNN
  - JAMR: AMR parser using the MSCG algorithm and Lagrangean relaxation
  - NeurboParser: Multitask semantic parser using joint MAP inference in a neural factor graph
  - Smaller contributions: XGBoost (docs); Morpha (stemming bugfixes); Spire (GCD for polynomials); Purescript Lists and Maps (stack safety); Purescript Flare (UI for Lists); etc.