Sean Bethard

Summary

I specialize in developing software for natural language processing, combining a strong foundation in linguistics and computer science with expertise in machine learning techniques and frameworks. I've been developing speech and text processing applications for twelve years, frequently on product teams. I studied linguistics in the linguistics department at UMass and computational linguistics in the computer science department at Brandeis. I've worked with the intelligence community (IARPA, IBM) and have delivered solutions in the public sector, healthcare, cybersecurity, aftermarket auto repair and other industries. I can help you leverage NLP and ML technologies to powerful effect.

Professional Experience

Curiouser AI, Sausalito, California

Machine Learning Engineer, NLP Lead **August 2023 – October 2023**

Delivered the backend of a generative AI service for creating marketing content and go-to-market strategies. Verified the endpoints with mock integrations and integrated them into Amplify. Established a containerization and deployment strategy as well as workflows for collaborative prompt testing, logging model runs and monitoring token consumption.

`GPTs, OpenAI API, LangChain, chat completions, conversation buffers, sequential chains, Amplify, API Gateway, GraphQL, Lambda, Vue, Weights & Biases`

CVS Health, Woonsocket, Rhode Island

AI Software Engineer[†] **March 2023 – July 2023**

Lead the migration of an Azure system with over 30 Microsoft contributors onto CVS infrastructure. Dodged bullets from both directions on the Microsoft-CVS bridge. Synced with Microsoft developers to understand what their services did, how to run and evaluate them. Migrated the services onto CVS tenants. Demonstrated how to run and evaluate the services internally. Demonstrated an end-to-end evaluation of the system to CVS stakeholders with focus on recall metrics, including techniques for how to improve the accuracy of the system by extending the cognitive search queries.

`Azure Cognitive Search, Text Analytics for Health, Form Recognizer, jq`

RAIN Agency, New York, New York

Senior Application Developer **August 2022 – January 2023**

Had a key role in the launch of `Ortho`, a voice application for aftermarket auto repair, at AAPEX 2023. Owned the speech recognition and NLU models. Identified error types and estimated level of effort for fixes and new features. Moved entities into entity lookups and mapped taxonomy entries to them unambiguously. Kept the models current with incoming user data without introducing new errors. Consistently suggested and delivered working solutions for new scenarios and integrated them without introducing new errors. Refactored ASR and NLU training data before product launch, improving accuracy, decreasing the number of intents and simplifying the integration.

`Amplify, ASR, conversational AI, Duckling, intent detection, MOTOR API, NER, NLU, Rasa, Speechly Annotation Language, Speechly CLI, text classification, Quasar`

Soffos AI, Limassol, Cyprus

Machine Learning Engineer, NLP Lead **December 2021 – May 2022**

Delivered a fast topic modeling service for suggesting document tags. Delivered a service for creating open-domain question-answer pairs. Delivered a service for creating open-domain multiple-choice questions.

`FastAPI, generative AI, few-shot, Gensim, GPT-J, non-negative matrix factorization, nvidia-smi, Prolog, PyTorch, T5`

Redflag AI, Berkeley, California

Machine Learning Engineer, NLP Lead **July 2020 – July 2021**

Improved the accuracy of a production model (LSTM) on all target labels by adding a model layer with part-of-speech information and adjusting the tokenization. Prepared resources for fine-tuning BERT's masked language modeling objective. Fine-tuned several BERTs and deployed one of them on an EC2 instance with an efficient inference pipeline. Predicted against hundreds of millions of sentences, typically in batches of around twenty million, until there were one million predictions for each label and used these

[†] This was a contract.

predictions to train lighter models (CNNs, RNNs) for use in production.

`AWS, BERT, Common Crawl, ConvNet, Dask, DistilBERT, DKPro, EC2, ELMo, GloVe, GRU, Leipzig Corpora Collection, LSTM, nvidia-smi, TensorFlow 1, TensorFlow 2, TensorFlow Hub, Python, PyTorch, RoBERTa, spaCy, screen`

Insight Engines, San Francisco, California

Founding NLP Engineer[‡] **September 2016 – October 2017**

Had a key role in raising a $15.8 million series A round (and a $12.5 million series AA round!) for `Cyber Security Investigator (CSI)` on an engineering team of four. Owned the semantic parser in `CSI`, a NLIDB for translating natural language expressions into valid Splunk queries. Contributed foundational improvements to the system and overall user experience, including recommending follow-up questions. Introduced syntactic information from a dependency parser to `CSI` in order to identify the types of expressions conjoined with *and* and *or* in order to disambiguate them and resolve them in the query language. Improved the accuracy of negation detection in `CSI`, further improving the quality of the resulting search queries and the overall experience of analysts using the application. Designed search schema to enable follow-up questions in
tt CSI in the context of kill chain workflows.

`semantic parsing, Splunk SPL, NLIDB, Whoosh, FuzzyWuzzy, spaCy, pytest.`

IBM, Arlington, Virginia

NLP Engineer **July 2015 – September 2016**

Authored a *wh-tracer* for walking slot grammar parses. Used the *wh-traces* to measure the syntactic diversity of question-answer pairs. Contributed source to `Watson Discovery` that improved the quality of its responses to questions containing ordinal numbers. Extended the tooling for evaluating `Watson Discovery` and assessing the effectiveness of domain adaptation. Supported projects with the Australian border patrol, Miami-Dade County, Apple and the 2020 US census. Supported on-site projects in Ireland and at the Department of Economic Development in Dubai.

`BeakerX, Hadoop, Java, Python, Shiny, UIMA, Watson Discovery Advisor, Watson Engagement Advisor, Watson Explorer`

Brigham and Women's Hospital, Wellesley, Massachusetts

Research Assistant **January 2015 – July 2015**

Improved negation detection in `MTERMS`, a system for processing unstructured data in clinical documents.
Implemented a bottom-up chart parser with well-formed substring table for identifying multiword expressions in clinical documents..

`lexical databases, OpenNLP, cTAKES, UIMA, Java, Jython, Scala, Python, MetaMap, MTERMS, SNOMED CT, SPECIALIST Lexicon, UMLS`

Brandeis Lab for Linguistics and Computation, Waltham, Massachusetts

Research Assistant **Summer 2013 – Spring 2014**

Prepared linguistic corpora in ISO Space working group. Adjudicated SpaceBank annotations with MAE.
Used resource to create a model capable of parsing spatial language in text, such as motion predicates and spatial prepositions in descriptions of travel. Showed that the model could reliably interpret motion and resolve ambiguity, such as by distinguishing the motion sense of *run* from its figurative sense.
Supported IARPA research on discovering emergent technologies in patents and scientific literature.
Recovered empty categories from treebank parses to explore the phenomenon of zero anaphora in Chinese.
Proctored the open round of the North American Computational Linguistics Olympiad (NACLO) and graded submissions.

`Python, Java, CoreNLP, CMU-C LMTK, GATE, MALLET, SRILM, empty categories, syntactic parsing, epistemic logic, modal logic, MAE, linguistic corpora, WordNet, Switchboard Dialog Act Corpus, speech recognition, SemEval, the semantics of motion`

Vioby, Boston, Massachusetts

Natural Language Engineer **Spring 2013**

Implemented a similarity function for matching search queries to product descriptions.

`split, Perl, PyDev, Beautiful Soup, NLTK, scikit-learn`

[‡] Last on-site position.

Lexalytics, Amherst, Massachusetts

Intern **Fall 2012**

Prepared a linguistic resource for a text processing task.

Education

2014 M.A. Computational Linguistics, *unfinished*, Brandeis University.
2013 M.A. Linguistics, *unfinished*, University of Massachusetts Amherst.
2011 B.A. Linguistics, Japanese Language & Literature, *cum laude*, University of Massachusetts Amherst.
2011 B.S. Biology, *cum laude*, University of Massachusetts Amherst.

Workshop Proceedings

2016. Charley Beller, Graham Katz, Allen Ginsberg, Chris Phipps, Sean Bethard, Paul Chase, Elinna Shek, and Kristen Summers. Watson Discovery Advisor: Question-answering in an industrial setting. *Proceedings of the Workshop on Human-Computer Question Answering (NAACL).*

2015. James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. SemEval-2015 Task 8: SpaceEval. *Proceedings of the 9th International Workshop on Semantic Evaluation (ACL).*

Patents

U.S. Patent 10,133,724 Sean Bethard, Graham Katz, Chris Phipps. Syntactic classification of natural language sentences with respect to a targeted element. November 20th, 2018. Provisional filed August 22nd, 2016. Armonk, New York. International Business Machines Corporation. Armonk, New York.

U.S. Patent 10,394,950 Sean Bethard, Graham Katz, Chris Phipps. Generation of a grammatically diverse test set for deep question answering systems. August 27th, 2019. Provisional filed August 22nd, 2016. International Business Machines Corporation. Armonk, New York.

U.S. Patent 10,956,463 Charley Beller, Sean Bethard, Will Dubyak, Alex Tonetti, Sean Thatcher, Julie Yu. System and method for generating improved search queries from natural language questions. March 23rd 2021. Provisional filed January 18th, 2019. International Business Machines Corporation. Armonk, New York.

Skills

Linux, Perl, Prolog, Python, Java, Scala, React, TypeScript
NLP, IR, ML, LLMs, vector space models, generative models, few shot, fine-tuning, NLU
Naïve Bayes, HMMs, PGMs, CRFs, ConvNets, ResNets, GANs, RNNs, seq2seq, T5, BERTs, GPTs
NLTK, scikit-learn, UIMA, DKPro, Gensim, spaCy, PyTorch, TensorFlow 1, TensorFlow 2, FAISS
Word embeddings, word2vec, GloVe, BPE, TikTokens, search, document search
Data sourcing, linguistic corpora, model development and optimization, deployments
Multi-tenant architectures, AWS, EC2, ECS, ECR, Bedrock, SageMaker, Lambda, Amplify