Dheeru Dua

ddua.github.io dheeru_dua@hotmail.com Dheeru Dua

EDUCATION

PhD in Natural Language Processing — Sept 2017 - Present
Information and Computer Science, University of California at Irvine
GPA (Core): 3.98/4.00

Masters in Intelligent Information Systems — Aug 2014 - Dec 2015
Language Technologies Institute, Carnegie Mellon University
GPA (Core): 3.91/4.00

Masters in Intelligent Information Systems — Aug 2014 - Dec 2015
Language Technologies Institute, Carnegie Mellon University
GPA (Core): 3.94/4.00

SELECT PUBLICATIONS

- Tricks for Training Sparse Translation Models. D Dua, S Bhosale, V Goswami, J Cross, M Lewis, A Fan. Work done during internship at FAIR (Under Review).
- Learning with Instance Bundles for Reading Comprehension. D Dua, P Dasigi, S Singh, M Gardner. EMNLP 2021
- Generative Context Pair Selection for Multi-hop Question Answering. D Dua, CN Santos, P Ng, B Athiwaratkun, B Xiang, M Gardner, S Singh. EMNLP 2021. Word done during internship at Amazon.
- Evaluating models' local decision boundaries via contrast sets. M Gardner, Y Artzi, V Basmov, J Berant, B Bogin, S Chen, P Dasigi, D Dua, Y Elazar, A Gottumukkala, N Gupta, H Hajishirzi, G Ilharco, D Khashabi, K Lin, J Liu, N F Liu, P Mulcaire, Q Ning, S Singh, N A Smith, S Subramanian, R Tsarfaty, E Wallace, A Zhang, B Zhou EMNLP 2020
- Easy, reproducible and quality-controlled data collection with CrowdAQ. EMNLP 2020
- Benefits of intermediate annotations in reading comprehension. D Dua, S Singh, M Gardner, ACL 2020.
- Dynamic Sampling Strategies for Multi-Task Reading Comprehension. A Gottumukkala, D Dua, S Singh, M Gardner, ACL 2020.
- DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. D Dua, Y Wang, P Dasigi, G Stanovsky, S Singh, M Gardner, NAACL 2019.
- PoMo: Generating entity-specific post-modifiers in context
- ORB: An Open Reading Benchmark for Comprehensive Multi-Dataset Evaluation of Reading Comprehension. D Dua, A Gottumukkala, A Talmor, S Singh, M Gardner, MRQA Workshop 2019
- Generating natural adversarial examples. Zhao Z, Dua D, Singh S. ICLR 2018

INDUSTRY EXPERIENCE

IBM Research – Statistical Language and Discovery Team — May 2016 - Aug 2017

- Applied Policy-gradient and Deep-Q Network based techniques for extractive document summarization.
- Developed a framework in Lua Torch containing common data processing and model architecture modules (similar to Allennlp) that allowed for faster experimentation and release of neural network based models.

Microsoft Corporation July 2011 – August 2014

- Knowledge Repository for Bing Search: Performed entity resolution and disambiguation to extract the right set of knowledge graph triples and surface information cards about the entity being searched on the search engine.
- SuperFresh Pipeline for Knowledge Repository: Developed an infrastructure for performing fast point updates in the knowledge graph, especially for popular events.
- Developed REST based apis for various utilities like image comparison, object detection to perform regression testing on the

ACADEMIC EXPERIENCE

- Participated in Alexa Prize 2019 socialbot development team.
  - Trained a question generation model to pre-populate a cluster index with new questions from news articles to create a list of interesting ice-breakers and improve engagement.
  - Used ATOMIC and ConceptNet knowledge to ask commonsense questions around user hobbies like reading, swimming etc.
- Participated in Event Detection and Co-reference task in TAC KBP 2015. Developed an Event Mention Detection system using Conditional Random Fields trained on k-best label sequences in an online-passive aggressive manner.
- Worked on a relation classification system that involved introducing features extracted from Path Ranking Algorithm (NELL) into a distantly supervised document level multi-label relation classifier (MultiR) which achieved better results over the stateof-the art at the time.
- Working on two components of a Question-Answering System for NTCIR QA Task on World History Questions.
  - Events Knowledge Base – Extraction of event frames from unstructured data and finding temporal sequence amongst them using Markov Logic networks. This is further used for automated event-ontology extraction.
  - Machine Reading – Finding more domain specific documents based on various facets of an entity extracted from FrameNet and Wikipedia to build a domain-specific corpus.
- Built a Search Engine over a semester with different retrieval and feedback models.
  - Implemented various retrieval mechanisms ranging from Boolean operators, BM25 model to Bayesian Indri Model.
  - Worked on query enrichment using pseudo-documents.
  - Trained a Learning to Rank Model incorporating features from all different retrieval models, pseudo-relevance feedback and other features like PageRank.
- Used techniques like dual decomposition and ADMM to impose linear constraints on structured predictions tasks at inference.

OTHER SKILL AND ACHIEVEMENTS

- Served as reviewer for ACL 2018, NAACL 2019, EMNLP 2019, NeurIPS 2019, EMNLP 2020 and NeurIPS 2020.
- Oragnized Socal NLP 2018 symposium at University of California Irvine (UCI)
- Co-maintainer of UCI machine learning repository.
- Received best poster award at Socal ML 2017 symposium.
- Experience in working with programming languages such as Python, C/C++, Java, HTML/CSS, Javascript
- Experience in working with deep learning frameworks such as Pytorch, Huggingface, Fairseq, Allennlp.