Dheeru Dua

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

Education

PhD in Natural Language Processing — Sept 2017 - Present
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

Bachelors in Computer Science and Engineering — Aug 2007 - June 2011
Indira Gandhi Institute of Technology, India
GPA (Core): 3.94/4.00

Research Publications and Patents

- Successive Prompting for Decomposing Complex Questions. D Dua, S Gupta, S Singh, M Gardner. EMNLP 2022.
- Tricks for Training Sparse Translation Models. D Dua, S Bhosale, V Goswami, J Cross, M Lewis, A Fan. NAACL 2022.
- 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.
- 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. Q Ning, H Wu, P. Dasigi, D. Dua, M. Gardner, R. Logan IV, A. Marasovic, Z. Nie, 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. ACL 2020. A Gottumukkala, D Dua, S Singh, M Gardner, ACL 2020.
- 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
- 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. J. Kang, R. Logan IV, Z. Chu, Y. Chen, D. Dua, K. Gimpel, S. Singh, N. Balasubramanian, NAACL 2019.
- Generating natural adversarial examples. Zhao Z, Dua D, Singh S, ICLR 2018.
- Generative adversarial network based modeling of text for natural language processing. D. Dua, C. N. Santos, B. Zhou, US Patent.

Industry Experience

Google Research — June 2022 - Dec 2022

- Non conservative domain generalization is not a well studied topic in open book question answering. We study domain generalization and propose solutions for zero-shot and few-shot domain adaptation from general purpose domain like Wikipedia to five extremely diverse domains: Biomedical, Legal, Stackoverflow, Reddit and News.
- Propose a method to estimate if zero-shot adaptation would be effective on the target domain with the help of a small target evaluation set.
- Preprint available on arXiv: To Adapt or to Annotate - Challenges and Interventions for Domain Adaptation in Open-Domain Question Answering.

Facebook AI Research — June 2021 - Sept 2021

- Mixture-of-Experts based multilingual translation models overfit to low resource and underfit to high resource languages.
- Explored tricks to overcome these shortcoming in sparse translation models.

Amazon Web Services, AI Labs — June 2020 - Sept 2020

- Passage retrievers take shortcuts while selecting passages for Multi-hop Question Answering
- Proposed a generative model which selects context pairs based on the likelihood of generating the gold question.

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

- Developed a framework in Lua Torch containing common data processing and model architecture modules that allowed for faster experimentation and release of neural network based models.
- Proposed a method for generating text with generative adversarial networks.

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

Skills and Achievements

- Talks:
  - Oral Spotlight talk at EMNLP 2021 for “Learning with Instance Bundles for Reading Comprehension”.
  - Oral Spotlight talk at EMNLP 2022 for “Successive Prompting for Decomposing Complex Question”.
  - Invited talk at summer seminar series in University of Texas at Dallas.
- Funded by fellowship from Hasso-Plattner Institute, Germany.
- Served as reviewer for ACL 2018, NAACL 2019, EMNLP 2019, NeurIPS 2019, EMNLP 2020 and NeurIPS 2020, NeurIPS 2021, IJCAI 2022, ACL 2022.
- 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: Pytorch (including Huggingface, Fairseq) and JAX.

Other Academic Projects

- Selected as one of the top 10 teams for Alexa Prize 2019 socialbot development challenge.
  - 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 base to ask commonsense questions around user hobbies like reading, swimming etc.

- Won second position in Event Detection and Co-reference task in TAC KBP. Developed an Event Mention Detection system using Conditional Random Fields with k-best label sequences in an online-passive aggressive manner.
- Won second position in NTCIR competition for World History QA task.
  - Extracted event frames from unstructured data and ordered them in a temporal sequence using Markov Logic networks.
  - Improved passage retrieval using FrameNet and Wikipedia linking.