Shashank Gupta

CONTACT

Email: shashank.nlp@gmail.com; shagup@microsoft.com; Phone: (+1) 217-904-6006

RESEARCH INTERESTS

Learning from Natural Language Instructions & Interactions; Efficient Fine-tuning; Language Modeling; Continual Learning; Explainable AI

EDUCATION

University of Illinois at Urbana Champaign *(Aug'15 - Dec'17)*
M.S., Computer Science
Thesis Adviser: Prof. Dan Roth

Birla Institute of Technology and Science, Pilani, India *(Aug'08 - June'12)*
B.E. (Hons.), Computer Science

INDUSTRIAL EXPERIENCE

- **Applied Scientist 2**: Microsoft AI, Redmond, WA *(May'20 - Present)*
  Themes: Multi-Task Learning; Mixture-of-Experts; Efficient Fine-tuning
- **Applied Scientist**: Microsoft AI, Redmond, WA *(Apr'18 - May'20)*
  Themes: Dialogue Systems; Model Compression; Domain Adaptation
- **Research Assistant**: Yahoo Labs, India *(June - Dec'12)*
  Themes: Display Ad-Platform; User-Response Prediction
- **Research Intern**: Yahoo R&D, India *(Jan - June'12)*
  Themes: Search Ad-Platform; User-Response Prediction

ACADEMIC RESEARCH EXPERIENCE

Research Assistant:

- **UIUC**: Cognitive Computation Group *(Aug'15 - Dec'17)*
  Themes: Zero-shot Text Classification; Text Generation; Structured Prediction
- **Max Planck Institute (MPI)**, Databases & Info. Sys. Group *(Aug'14 - April'15)*
  Themes: Entity Disambiguation; Automated KB Construction
- **IIT-Bombay**: InfoLab *(Jan'13 - June'14)*
  Themes: Entity Search & Disambiguation; Distributed Indexing

SELECTED PUBLICATIONS

5. Sparsely Activated Mixture-of-Experts are Robust Multi-Task Learners.
   S. Gupta, S. Mukherjee, K. Subudhi, E. Gonzalez, D. Jose, A. H. Awadallah, J. Gao
   Under submission, 2022 [pdf]

4. Knowledge Infused Decoding.
   R. Liu, G. Zheng, **S. Gupta**, R. Gaonkar, C. Gao, S. Vosoughi, M. Shokouhi, and A.H. Awadallah
   International Conference on Learning Representations (ICLR), 2022 [code][pdf]

3. Exploring Low-Cost Transformer Model Compression for Large-Scale Commercial Reply Suggestions.
   V. Shrivastava*, R. Gaonkar*, **S. Gupta***, A. Jha
   *Equal Contribution
   Microsoft Journal of Applied Research (MSJar), 2021 [pdf]

2. CogCompNLP: Your Swiss Army Knife for NLP
   D. Khashabi, M. Sammons, B. Zhou, T. Redman, C. Christodoulopoulos, V. Srikumar, N. Rizzolo, L. Ratinov, G. Luo, Q. Do, C. T. Tsai, S. Roy, S. Mayhew, Z. Feng, J. Wieting, X. Yu, Y. Song, **S. Gupta**, S. Upadhyay, N. Arivazhagan, Q. Ning, S. Ling, D. Roth
   International Conference on Language Resources and Evaluation (LREC), 2018 [code][pdf]

1. Web-scale Entity Annotation Using MapReduce
   S. Gupta, V. Chandramouli, S. Chakrabarti
   International Conference on High Performance Computing (HiPC), 2013 [project][pdf]

TEACHING EXPERIENCE

Teaching Assistant:

- **UIUC**: Machine Learning, CS446 *(Aug - Dec'16)*
- **IIT-Bombay**: Web Search and Mining, CS635 *(July - Nov'13)*
- **BITS-Pilani**: Operating Systems, CS C372 *(Aug - Dec'11)*

Technical Skills

Languages: *Proficient*: Python | *Intermediate*: Java | *Basic*: C++, HTML/CSS, JavaScript
Toolkits: PyTorch, Tensorflow, HuggingFace, AzureML, Hadoop, CogComp-NLP

Recent Projects

Sparse Multi-Task Learning using Mixture-of-Experts *(Sept'21 - Present)*
Mentors: Subho Mukherjee, and Ahmed Awadallah; Microsoft Research

(See Publication #5)
Introduced task-aware gating in Mixture-of-Experts architectures for Multi-task learning (MTL) that outperformed existing dense and sparse MTL models on three key dimensions: 1) transfer to low-resource tasks during MTL training. 2) sample-efficient generalization to unseen related tasks. 3) robustness to catastrophic forgetting on the addition of unrelated tasks. Scaling experiments demonstrated the efficacy of the approach regardless of the model scale and task diversity.

Knowledge Infused Decoding *(June - Sept'21)*
Mentor: Ahmed Awadallah; Microsoft Research

(See Publication #4)
Introduced a novel decoding algorithm (KID) for generative LMs that dynamically retrieves and infuses external knowledge into each step of the LM decoding. KID outperformed task-optimized state-of-the-art models and existing knowledge-infusion techniques on six diverse knowledge-intensive NLG tasks.

Efficient model compression for Commercial Suggested Replies *(Aug - Dec'20)*
Mentor: Milad Shokouhi; Microsoft AI

(See Publication #3)
Explored low-cost model compression techniques for PLMs to successfully reduce the training and inference times of a commercial email reply suggestion system by 42% and 35% respectively. Studied the impact of the dataset size and PLM quality (random init; domain adaptation) on the efficacy of compression techniques and obtained some key recommendations for industrial applications.

Automated Suggested Replies *(July'18 - June'21)*
Mentor: Milad Shokouhi; Microsoft AI [Web]

Developed and productionized a PLM-based automated reply suggestion feature for millions of Outlook and Teams users. The project involved modeling it as an information retrieval task with large-scale training of a transformer-based bi-encoder model, developing pipelines for generating candidate responses, identifying and addressing gender bias and response diversity issues, and carrying out principled offline and A/B online experiments to measure the impact on user engagement.

Selected Previous Projects

Zero-shot Text Classification *(Aug'16 - Aug'17)*
Mentor: Prof. Dan Roth; UIUC [Technical Report]

The goal was to develop a zero-shot topic classification methodology that classifies documents into topics by requiring only a semantic description of the topic. The key idea was to embed documents & topics using some world knowledge (e.g., Wikipedia) and then compute the similarity between the representations for classification. Developed novel topic-informed dense word and entity representations using Wikipedia by augmenting the word2vec loss to address the limitations of state-of-the-art sparse word representations (ESA; explicit-semantic-analysis).

Conditional Text Generation *(Jan - May'17)*
Mentor: Prof. Svetlana Lazebnik; UIUC

The goal of this course project was to compare Conditional GANs and VAEs for sentiment-conditioned review generation. Used Policy-Gradient and Gumbel-Softmax with Curriculum Learning to stabilize the GAN training. Human evaluations showed VAEs to be superior for conditional text generation.

Joint NER, Relation Extraction and CoReference Resolution *(Jan - May'16)*
Mentor: Prof. Dan Roth; UIUC [code]

The goal of this course project was to jointly model NER, Relation Extraction, and Coreference Resolution. Simple coupling of individual classifiers without constraints showed poor performance. Developed a framework for joint inference with constraints using Constrained-Conditional Models.

Scalable Entity Disambiguation and Search *(Jan’13 - June’14)*
Mentor: Prof. Soumen Chakrabarti, IIT-Bombay [Web]

(See Publication #1)
Designed a scalable entity disambiguation and search system using MapReduce. Developed a load-balancing protocol that led to a 5.4x speedup compared to a vanilla MapReduce baseline.

User Response Prediction for Non-Guaranteed Display Ad Delivery *(June - Dec’12)*
Mentor: Prof. Sanjay Chawla, Prof. Shivaram Kalyanakrishnan, Yahoo Labs

Improved the accuracy of the user-click prediction model by mining new features. Analyzed Petabytes of data for feature signal & coverage. Used that analysis to find a training data partitioning strategy that showed promise when different models were trained on those different partitions.

Automated Campaign Optimization for Search Advertising *(Jan - June’12)*
Guide: Ajay Sharma, Director, UDA, Yahoo R&D

Protoyped a tool that automated the account optimization for advertisers. Developed models for predicting #impressions, #clicks, #conversions, and handled sparsity issues by using community detection algorithms to cluster competitors together. Ultimately, given a budget, the tool used resource allocation algorithms to select appropriate bid amounts for various targeting combinations.

Relevant Coursework

Machine Learning, NLP, Structured Learning, Recent Trends in Deep Learning, Graphical Models, Web Search & Mining, Organization of Web Information, Advanced Data Mining

References

Milad Shokouhi, *Partner Applied Scientist, Microsoft AI* | milads@microsoft.com *[Ex-manager]*
Ahmed Awadallah, *S. Principal Research Manager, MSR* | hassanam@microsoft.com *[Collaborator]*
Dan Roth, *Distinguished Professor, UPenn* | danroth@seas.upenn.edu *[M.S. adviser]*
Soumen Chakrabarti, *Professor, IIT-Bombay* | soumen@cse.iitb.ac.in *[R.A. adviser]*
Denilson Barbosa, *Associate Prof., Univ. of Alberta* | denilson@ualberta.ca *[R.A. adviser]*
Shivaram Kalyanakrishnan, *Associate Prof., IIT-Bombay* | shivaram@cse.iitb.ac.in *[R.A. adviser]*