Shashank Gupta

CONTACT

Email: shashank91.bits@gmail.com; shagup@microsoft.com; Phone: (+1) 217-904-6006
Homepage: https://shatu.github.io/

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 *(May'20 - Present)*
  Themes: Multi-Task Learning; Mixture-of-Experts; PLM Domain Adaptation; Few-shot Learning
- **Applied Scientist**: Microsoft AI *(Apr'18 - May'20)*
  Themes: Dialogue Systems; Model Compression; Responsible AI; Text Generation

RESEARCH EXPERIENCE

Research Assistant:

- **UIUC**: Cognitive Computation Group *(Aug'15 - Dec'17)*
  Themes: Unsupervised Text Classification; Text Generation; Structured Learning
- **Max Planck Institute (MPI)**, Databases & Info. Sys. Group *(Aug'14 - April'15)*
  Themes: Named Entity Disambiguation; Automated KB Construction
- **IIT-Bombay**: InfoLab *(Jan'13 - June'14)*
  Themes: Entity Search & Disambiguation; Distributed Training and Indexing
- **Yahoo Labs**: Ad-Predict Team *(June - Dec'12)*
  Themes: Display Ad-Platform; User-Response Prediction
- **Yahoo R&D**: User Data & Analytics Team *(Jan - June'12)*
  Themes: Search Ad-Platform; User-Response Prediction; Automated Account Optim.

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
   Currently under review, 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 [pdf]

3. Exploring Low-Cost Transformer Model Compression for Large-Scale Commercial Reply Suggestions.
   V. Shrivastava*, R. Gaonkar*, **S. Gupta***, A. Jha
   (Preprint), 2021 [arxiv]

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) [project][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)*

RESEARCH INTERESTS

NLP: Pre-trained Language Models; Text Generation; Dialogue Systems; Commonsense Reasoning; Multi-modal Learning
Machine Learning: Mixture-of-Experts; Multi-Task Learning; Structured Learning; Distillation

Technical Skills

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

Recent Projects

Sparse Multi-Task Learning using Mixture-of-Experts *(Sept'21 - Present)*
Mentors: Subho Mukherjee, and Ahmed Awadallah
(See Publication #5)
Introduced task-aware gating for Multi-task learning (MTL) using Mixture-of-Experts (MoE) architecture that outperformed existing dense and sparse MTL models on 3 dimensions: 1) transfer to low-resource tasks during MTL training 2) sample efficient generalization to unseen related tasks 3) robustness to catastrophic forgetting on 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)*
Collaborators: Ruibo Liu, Guoqing Zheng, and Ahmed Awadallah
(See Publication #4)
Introduced a novel decoding algorithm (KID) for generative LMs, which dynamically 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
(See Publication #3)
Explored several low-cost model compression and domain adaptation techniques for PLMs, and successfully reduced the training and inference times of a commercial email reply suggestion system by 42% and 35% respectively.

Automated Suggested Replies *(July'18 - May'21)*
Mentor: Milad Shokouhi Web
Developed and shipped a PLM-based email reply suggestion feature for millions of Outlook and Teams users. Project involved creating pipelines for response candidate generation, large-scale training of a transformer-based matching model on millions of examples, and addressing gender-bias and response diversity issues.

Selected Previous Projects

Zero-shot Text Classification *(Aug'15 - Dec'17)*
Mentor: Prof. Dan Roth, UIUC Technical Report
Goal was to do zero-shot text classification of documents into user-specified topics. The key idea was to embed documents & topic name using some world knowledge, and then computing similarity between the representations for unsupervised text classification. Developed topic-sensitive word and entity embeddings using Wikipedia by augmenting the Word2Vec loss, and used their composition to create document representations.

Joint NER, Relation Extraction and CoReference Resolution *(Jan - May'16)*
Mentor: Prof. Dan Roth, UIUC Web | Github
Aim was to jointly model NER, Relation Extraction and CoRef using explicit constraints. Simple coupling of classifiers without constraints showed poor performance. Developed a framework for joint training with Constrained-Conditional Models, using Illinois-SL and CogComp-NLP.

Agile NERD for KB-Lifecycle *(Aug'14 - April'15)*
Mentor: Prof. Gerhard Weikum, Prof. Denilson Barbosa, MPI Web
Identified the problem of separating mentions of emerging entities from mentions worthy of abstention as the key hurdle in achieving real-time KBs and iterative entity annotation on corpus. Used the disagreement between an ensemble of annotators to signal abstention on a given mention.

Scalable Entity Disambiguation and Search *(Jan'13 - June'14)*
Mentor: Prof. Soumen Chakrabarti, IIT-Bombay Web | Publication | CSAW
(See Publication #1)
Designed a scalable entity disambiguation and indexing framework by developing custom-key partitioning strategies to mitigate the load-skew problem of a simple MapReduce implementation. Further improved the accuracy of the entity disambiguation system by extracting more training data from Wikipedia and engineering features.

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

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 Web

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]*
Subho Mukherjee, *Senior Researcher, MSR* | subhabrata.mukherjee@microsoft.com *[Collaborator]*
Abhishek Jha, *ML Engineering Manager, Stripe* | abhija@stripe.com *[Ex-manager]*
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]*