

Ehsan Shareghi

✉ ehsan.shareghi@gmail.com • 🌐 eehsan.github.io

Fields of Interest: Natural Language Processing, Machine Learning

I lead a small team of grad and undergrad students working on predictive models for language (text and speech). During my PhD and early parts of Postdoc, I worked on Bayesian nonparametric models and Bayesian learning of neural models and their application in language modelling and parsing. Since 2019 I have been working on various tasks and settings where training data is not sufficient for generalization at test or deployment due to several reasons; insufficient training data, knowledge-intensive nature of the task, absence of symbolic constraints. While in principle the solutions we investigate are agnostic to the underlying models, we try to leverage pretrained large models as much as possible. Hence, in recent years we have been also investigating various shortcomings of such models. Some recent areas of research:

- Deep generative models (e.g., generation, semi-supervised learning, sparsity and disentanglement)
- Self-supervised learning (e.g., applications in text, speech, and graph)
- Utilising graphs for knowledge intensive tasks (e.g., graph-to-text generation, applications in biomedical domain)
- Probing and augmenting pretrained large models (e.g., representational and computational efficiency, reliability)
- Bayesian learning and inference (e.g., applications in training with small data)

Employment

- **Department of Data Science and Artificial Intelligence, Monash University** **Melbourne, AUS**
Tenured Assistant Professor in Machine Learning *Jan 2021–Present*
- **Language Technology Lab, University of Cambridge** **Cambridge, UK**
Affiliated Lecturer *Oct 2020–Present*
- **Department of Electronic and Electrical Engineering, University College London** **London, UK**
Tenured Lecturer in Machine Learning *April 2020–Dec 2020*
- **Language Technology Lab, University of Cambridge** **Cambridge, UK**
Postdoctoral Researcher (Advisor: Anna Korhonen) *Jan 2018–April 2020*
I worked on training deep neural models in resource poor settings by compensating for the lack of labelled data via (i) developing Bayesian learning and efficient approximate inference schemes, and (ii) building on the generative capacity of deep generative models for semi-supervised learning to leverage unlabeled data, and (iii) sharing training across tasks and languages.

Education

- **Faculty of Information Technology, Monash University** **Melbourne, AUS**
PhD in Computer Science *Oct 2013–Oct 2017*
Thesis: Scalable Non-Markovian Sequential Modelling for Natural Language Processing
Advisors: Gholamreza Haffari, Trevor Cohn (University of Melbourne), Ann Nicholson
Internal Examiners: Wray Buntine, Ingrid Zukerman
External Examiners: Brian Roark (Google), Mark Dras (Macquarie University)
- **Department of Computer Science, Concordia University** **Montreal, CA**
M.Sc. in Computer Science *Jan 2011–April 2013*
Thesis: Feature Combination for Measuring Sentence Similarity
Advisor: Sabine Bergler
Examiners: Adam Krzyzak (Concordia University), Leila Kosseim (Concordia University)
- **Department of Computer Science, Shahid Beheshti University (National University of Iran)** **Tehran, IR**
B.Sc. in Computer Science *Sept 2004–2009*
Thesis: Text Summarization with Harmony Search Algorithm-Based Sentence Extraction
Advisor: Leila Sharif

Selected Publications

- 2023 J. Zhao, G. Haffari, E. Shareghi, Generating Synthetic Speech from SpokenVocab for Speech Translation, Proceedings of the European Chapter of the Association for Computational Linguistics (**Findings of EACL**)
- 2022 J. Han, E. Shareghi, Self-supervised Graph Masking Pre-training for Graph-to-Text Generation, Proceedings of the Empirical Methods in Natural Language Processing (**EMNLP**)
- 2022 J. Zhao, H. Yang, G. Haffari, E. Shareghi, RedApt: Faster and Smaller Speech Translation without Quality Compromise, Proceedings of the Empirical Methods in Natural Language Processing (**Findings of EMNLP**)
- 2022 H. Yang, J. Zhao, G. Haffari, E. Shareghi, Self-supervised Rewiring of Pre-trained Speech Encoders: Towards Faster Fine-tuning with Less Labels in Speech Processing, Proceedings of the Empirical Methods in Natural Language Processing (**Findings of EMNLP**)
- 2022 J. Zhao, H. Yang, G. Haffari, E. Shareghi, M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation, Proceedings of the Conference of the International Speech Communication Association (**INTERSPEECH**)
- 2022 Y. Su, F. Liu, Z. Meng, T. Lan, L. Shu, E. Shareghi, N. Collier, TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning, Proceedings of the North American Chapter of Association for Computational Linguistics (**Findings of NAACL**)
- 2022 L. Zhang, W. Buntine, E. Shareghi, On the Effect of Isotropy on VAE Representations of Text, Proceedings of the Association for Computational Linguistics (**ACL**)
- 2022 Z. Meng, F. Liu, E. Shareghi, Y. Su, C. Collins, N. Collier, Rewire-and-Probe: A Contrastive Recipe for Probing Knowledge in Pretrained Biomedical Language Models, Proceedings of the Association for Computational Linguistics (**ACL**)
- 2022 D. K. Ryu, E. Shareghi, M. Fang, Y. Xu, S. Pan, G. Haffari, Fire Burns, Sword Cuts: Commonsense Inductive Bias for Exploration in Text-based Games, Proceedings of the Association for Computational Linguistics (**ACL**)
- 2021 Z. Meng, F. Liu, T. Clark, E. Shareghi, N. Collier, Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT, Proceedings of the Empirical Methods in Natural Language Processing (**EMNLP**)
- 2021 J. Zhao, P. Arthur, G. Haffari, T. Cohn and E. Shareghi, It Is Not As Good As You Think! Evaluating Simultaneous Machine Translation on Interpretation Data, Proceedings of the Empirical Methods in Natural Language Processing (**EMNLP**)
- 2021 L. Zhang, V. Prokhorov, E. Shareghi, Unsupervised Representation Disentanglement of Text: An Evaluation on Synthetic Datasets, Proceedings of the Workshop on Representation Learning for NLP (**Rep4NLP**), 2021
- 2021 V. Prokhorov, Y. Li, E. Shareghi, N. Collier, Learning Sparse Sentence Encoding without Supervision: An Exploration of Sparsity in Variational Autoencoders. Proceedings of the Workshop on Representation Learning for NLP (**Rep4NLP**), 2021
- 2021 T. Clark, C. Conforti, F. Liu, Z. Meng, E. Shareghi, N. Collier, Integrating Transformers and Knowledge Graphs for Twitter Stance Detection, Proceedings of the Workshop on Noisy User-generated Text (**W-NUT**), 2021
- 2021 M. Zhao, Y. Zhu, E. Shareghi, I. Vulić, R. Reichart, A. Korhonen, H. Schütze, A Closer Look at Few-Shot Crosslingual Transfer: The Choice of Shots Matters, Proceedings of the Association for Computational Linguistics (**ACL**)
- 2021 F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment Pretraining for Biomedical Entity Representations, Proceedings of the North American Chapter of the Association for Computational Linguistics (**NAACL**)
- 2021 Y. Zhu, E. Shareghi, Y. Li, R. Reichart, A. Korhonen, Combining Deep Generative Models and Multi-lingual Pretraining for Semi-supervised Document Classification, Proceedings of the European Chapter of the Association for Computational Linguistics (**EACL**)
- 2020 M. Basaldella, F. Liu, E. Shareghi, N. Collier, COMETA: A Corpus for Medical Entity Linking in the Social Media, Proceedings of the Empirical Methods in Natural Language Processing (**EMNLP**)
- 2019 V. Prokhorov, E. Shareghi, Y. Li, M.T. Pilehvar, N. Collier, On the Importance of the Kullback-Leibler Divergence Term in Variational Autoencoders for Text Generation, Proceedings of the Workshop on Neural Generation and Translation (**WNGT**), 2019
- 2019 E. Shareghi, Y. Li, Y. Zhu, R. Reichart, A. Korhonen, Bayesian Learning for Neural Dependency Parsing, Proceedings of the North American Chapter of the Association for Computational Linguistics (**NAACL**)
- 2019 E. Shareghi, D. Gerz, I. Vulić, A. Korhonen, A Bit of Progress and Stronger n -gram Language Modeling Baselines, Proceedings of the North American Chapter of the Association for Computational Linguistics (**NAACL**)
- 2017 E. Shareghi, G. Haffari, T. Cohn, Compressed Nonparametric (Bayesian) Language Modelling, Proceedings of the International Joint Conference on Artificial Intelligence (**IJCAI**)
- 2016 E. Shareghi, T. Cohn, G. Haffari, Richer Interpolative Smoothing Based on Modified Kneser-Ney Language Modeling, Proceedings of the Empirical Methods in Natural Language Processing (**EMNLP**)
- 2016 E. Shareghi, M. Petri, G. Haffari, T. Cohn, Fast, Small and Exact: Infinite-order Language Modelling with

- Compressed Suffix Trees, Transactions of the Association for Computational Linguistics (**TACL**)
- 2015 E. Shareghi, M. Petri, G. Haffari, T. Cohn, Compact, Efficient and Unlimited Capacity: Language Modeling with Compressed Suffix Trees, Proceedings of the Empirical Methods in Natural Language Processing (**EMNLP**)
- 2015 E. Shareghi, G. Haffari, T. Cohn, A. Nicholson, Structured Prediction of Sequences and Trees using Infinite Contexts, Proceedings of the European Conference on Machine Learning (**ECML**)

Supervisions

- Ongoing Jiuzhou Han [PhD,2023 -], Monash University
- Topic: Graph-based Generation
- Ongoing Chang Shu [PhD,2023 -], University of Cambridge (Co-Advise with Prof. Collier)
- Topic: Augmenting Large Language Models
- Ongoing Yinhong Liu [PhD,2022 -], University of Cambridge (Co-Advise with Prof. Collier)
- Topic: Guided Decoding
- Ongoing Jinming Zhao [PhD,2021 -], Monash University
- Topic: Speech Translation
- Ongoing Dongwon Ryu [PhD,2021 -], Monash University
- Topic: Reinforcement Learning and Interactive Games
- Alumni Xiny Li [Honours,2022], Monash University
- Thesis: In-Context Knowledge Elicitation from Large Language Models
- Alumni Hao Yang [Msc,2022], Monash University
- Thesis: Self-supervised Rewiring for Faster Fine-tuning with Less Labels in Speech Understanding Tasks
- Alumni Yi Zhu [PhD,2019-2022], University of Cambridge (Co-Advise with Prof. Korhonen)
- Thesis: Multi-Lingual Learning with Few Labels
- Alumni Jirarote Jirasirikul [MSc,2021], Monash University
- Thesis: Fine-tuning Pretrained Language Models for Biomedical Tasks
- Alumni Victor Prokhorov [PhD,2019-2021], University of Cambridge (Co-Advise with Prof. Collier)
- Thesis: Injecting Inductive Biases into Distributed Representations of Text
- Alumni Aiswarya Kaladharan [MSc,2020], UCL
- Thesis: Privacy-Preserving Neural Representation for Audio Speech
- Alumni Lan Zhang [MSc,2020], UCL
- Thesis: Towards Understanding the Latent Space in Variational Autoencoders
- Alumni Lei Yang [MSc,2020], UCL
- Thesis: Deep Neural Models for Generating Recipes

Grants

- Tracking media and parliamentary narratives around cycles of disadvantage in Australia (AUD 500k), Chief Investigator, Paul Ramsay Foundation (2022-2023)
- Big Idea Fund (AUD 50k), Chief Investigator, Monash University (2021)
- Research Gift (USD 5k), Adobe Research (2021)
- Early Career Grant (AUD 15k), Monash University (2021)
- Research Gift (USD 5k), Adobe Research (2019)

Teaching

- Natural Language Processing (FIT5217), Monash University, 2022, 2023
- Data Analysis for Semi-structured Data (FIT5212), Monash University, 2022, 2023
- Applied Machine Learning II (ELEC0135), Co-Lecturer, UCL, 2020
- Quantitative Methods in Analyzing Linguistic Data (QMALD), University of Cambridge, 2020
- Computational Linguistics (Li18), Guest Lecturer, University of Cambridge, 2018, 2019

Talks

- **On the Limits of Large Neural Language Models** **Melbourne, AUS**
- *Venue: The 1st Monash Interdisciplinary Workshop on Language.* *November 2022*

- **Utilising the Unlabelled Text via Deep Generative Models** **Melbourne, AUS**
March 2021
Venue: Monash ML Seminars.
- **Language Modelling : Something Old, Something New** **London, UK**
February 2020
Venue: UCL Engineering Dept.
- **Modeling Solutions for Low-resource Parsing** **Cambridge, UK**
October 2018
Venue: Apple Siri Lab
- **Scalable Non-Markovian Language Modeling** **Cambridge, UK**
April 2018
Venue: DTAL, University of Cambridge
- **Scalable and Unbounded Smoothing for Language Modeling** **Melbourne, AUS**
December 2016
Venue: Australasian Language Technology Association (ALTA)
- **Unbounded Smoothing for NLP (Trees and Sequences)** **Melbourne, AUS**
September 2016
Venue: IBM Research Australia

Internships

- **IBM Research** **Melbourne, AUS**
2016–2016
Part-Time Research Intern
- **IBM Research** **Melbourne, AUS**
2015–2015
Full-Time Research Intern

PhD Theses Examinations

- *Multilingual Representations and Models for Improved Low-Resource Language Processing*, Masoud Jalili Sabet (LMU, Advisor: Hinrich Schütze), 2022

Technical and Personal skills

- **Programming Languages:** Python, C++ (and some Shell)
- **Natural Language Processing:** GATE API, CoreNLP, NLTK, SpaCy
- **Machine Learning/Statistics tools:** PyTorch, DyNet, R, Scikit, Weka API
- **Languages:** Persian (Native), English (Fluent), Turkish-Azeri (Fluent)

Services

- **Organization:** Area Chair of Multilinguality Track in EACL2023, General Chair of The 1st Monash Interdisciplinary Workshop on Language 2022, General Chair of The 3rd DeepLo2022 Workshop (co-located with NAACL22), Assistant for General Chairs of ACL2019
- **Invited to Review for:** TACL, ACL, EMNLP, NAACL, ARR Rolling Review, AAAI, IJCAI, CoNLL, COLING
- **Theater:** Dramaturge in Monash University Student Theater (2014)