Tahmid Hasan
Email: tahmidhasan@cse.buet.ac.bd
Website: https://tahmid04.github.io/ [Google Scholar] [GitHub]

Research Interests

- Low-Resource, Multilingual, and Cross-Lingual Natural Language Processing
- Data- and Compute-Efficient Deep Learning for Natural Language Processing

Education

- **Bangladesh University of Engineering and Technology (BUET)** Dhaka, Bangladesh
  M.Sc. in Computer Science and Engineering; CGPA: 3.90/4.0 July 2019 - Present
- **Bangladesh University of Engineering and Technology (BUET)** Dhaka, Bangladesh
  B.Sc. in Computer Science and Engineering; CGPA: 3.98/4.0 February 2015 - April 2019

Research Experience

1. **Adapting XL-Sum for Cross-Lingual Summarization**: The target language of a multilingual model on cross-lingual summarization is limited to only the language it is fine-tuned on, and we have observed that fine-tuning with multiple languages without cross-lingual supervision cannot help control the language of the generated summaries. In this work, we generate summaries in any target language for a given article by fine-tuning multilingual models with explicit (albeit limited) cross-lingual signals. Aligning identical articles across languages via cross-lingual retrieval on the XL-Sum dataset, coupled with a multi-stage sampling technique, we perform large-scale cross-lingual summarization for 45 languages.
   Supervisors: *Prof. Rifat Shahriyar* and *Dr. Yuan-Fang Li* Status: Ongoing

2. **Paraphrase Generation via Knowledge Distillation from NMT Models**: Instead of doing round-trip translation to generate synthetic paraphrase pairs, in this work, we directly distill the paraphrasing knowledge of translation models into a paraphrase model. Using a forward and a backward NMT model as teachers, we distill the cross-attention and output distributions into a student paraphrase model. In contrast to traditional KD, here we have two teachers instead of one and the task of the student model is different from the teachers'.
   Supervisors: *Prof. Rifat Shahriyar* and *Dr. Wasi Uddin Ahmad* Status: Ongoing

3. **XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages**: We present *XL-Sum*, a comprehensive and diverse dataset comprising 1 million professionally annotated article-summary pairs in 44 languages from BBC News, extracted using a set of carefully designed heuristics. We show higher than 11 ROUGE-2 scores on 10 languages we benchmark on, with some of them exceeding 15, as obtained by multilingual training. We release the dataset and evaluation scripts for future research on multilingual summarization.
   Supervisors: *Prof. Rifat Shahriyar* and *Dr. Yuan-Fang Li* Status: Published in *Findings of ACL, 2021.*

4. **BanglaBERT: Limitations of Embedding Barrier for Low-Resource Language Understanding**: In this work, we build *BanglaBERT* – a Transformer-based Bangla NLU model pre-trained on 18.6 GB data crawled from top Bangla sites, and through comprehensive experiments identify a major shortcoming of multilingual models, which we name the 'Embedding Barrier', that hurts performance for low-resource languages that don't share writing scripts with any high resource one.
   Supervisors: *Prof. Rifat Shahriyar* Status: Submitted to *ACL Rolling Review*

5. **Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation**: In this work, we build a customized sentence segmenter for Bengali and propose two novel methods for parallel corpus creation on low-resource setups: *aligner ensembling* and *batch filtering*. Our proposed methods improve sentence alignment F-1 score by 3.38% and translation BLEU score by 2.5 points. We release the data and code for future research on low-resource machine translation.
Supervisors: *Prof. Rifat Shahriyar* and *Prof. M. Sohel Rahman* Status: Published in *EMNLP, 2020.*

6. **CoDesc: A Large Code–Description Parallel Dataset**: In this study, we present CoDesc – a large parallel dataset composed of 4.2 million Java methods and natural language descriptions. With extensive analysis, we identify and remove prevailing noise patterns from the dataset. We demonstrate the proficiency of CoDesc in two complementary tasks for code-description pairs: code summarization and code search.
Supervisors: *Prof. Rifat Shahriyar* and *Dr. Wasi Uddin Ahmad* Status: Published in *Findings of ACL, 2021.*

7. **BERT2Code: Can Pretrained Language Models be Leveraged for Code Search?**: We leverage the efficacy of pretrained word and code embeddings using a simple, lightweight neural network for semantic code search. We show that our model learns the inherent relationship between the embedding spaces and further probe into the scope of improvement by empirically analyzing the embeddings. We show that the quality of the code embeddings is the bottleneck for our model's performance, and discuss future directions in this area.
Supervisor: *Prof. Rifat Shahriyar* Status: Completed

8. **Using Adaptive Heartbeat Rate on Long-Lived TCP Connections**: We propose a set of iterative probing techniques, namely binary, exponential, and composite search, that detect the middle-box binding timeout with varying degree of accuracy, and in the process, keep improving the keep-alive interval used by the client device. Our proposed methods improve over Google Firebase's keep-alive algorithm by 6%.
Supervisors: *Prof. M. Saifur Rahman* and *Prof. M. Sohel Rahman* Status: Published in *IEEE/ACM Transactions on Networking (Vol: 26-1, 2018).*

Publications

1. **XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages**
Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, Rifat Shahriyar
In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* [PDF] [Code]

2. **CoDesc: A Large Code–Description Parallel Dataset**
Masum Hasan, Tanveer Muttaqueen, Abdullah Al Ishtiaq, Kazi Sajeed Mehrab, Md. Mahim Anjum Haque, **Tahmid Hasan**, Wasi Ahmad, Anindya Iqbal, Rifat Shahriyar
In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* [PDF] [Code]

3. **Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation**
Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, Rifat Shahriyar
In *Proceedings of the Empirical Methods in Natural Language Processing, EMNLP 2020.* [PDF] [Code]

4. **Using Adaptive Heartbeat Rate on Long-Lived TCP Connections**
M. Saifur Rahman, Md. Yusuf Sarwar Uddin, **Tahmid Hasan**, M. Sohel Rahman, M. Kaykobad
In *IEEE/ACM Transactions on Networking (Volume: 26, Issue: 1, Feb. 2018).* [PDF] [Code]

Under Review:

1. **BanglaBERT: Combating Embedding Barrier in Multilingual Models for Low-Resource Language Understanding**
   Abhik Bhattacharjee*, **Tahmid Hasan*** (*Equal contribution*), Kazi Samin, Md Saiful Islam, M. Sohel Rahman, Anindya Iqbal, Rifat Shahriyar
   ArXiv Pre-print, 2021. [PDF] [Code]

2. **BERT2Code: Can Pretrained Language Models be Leveraged for Code Search?**
   Abdullah Al Ishtiaq, Masum Hasan, Md. Mahim Anjum Haque, Kazi Sajeed Mehrab, Tanveer Muttaqueen, **Tahmid Hasan**, Anindya Iqbal, Rifat Shahriyar
   ArXiv Pre-print, 2021. [PDF]

Professional Experience

- **Bangladesh University of Engineering and Technology (BUET)** — Dhaka, Bangladesh
  Lecturer, Department of CSE, BUET — October 2019 - Present
- **Bangladesh University of Engineering and Technology (BUET)** — Dhaka, Bangladesh
  Graduate Research Assistant, Department of CSE, BUET — July 2019 - Present
  Supervisor: *Prof. Rifat Shahriyar*

Teaching Experience (Selected)

- Jan 2021 CSE 308: Software Engineering Sessional
- Jan 2020 CSE 208: Data Structures & Algorithms II Sessional
- Jan 2020 CSE 412: Simulation & Modeling Sessional
- Jan 2020 CSE 462: Algorithm Engineering Sessional
- Jan 2020 CSE 472: Machine Learning Sessional
- Jul 2019 CSE 204: Data Structures & Algorithms I Sessional
- Jul 2019 CSE 218: Numerical Methods
- Jul 2019 CSE 408: Software Development

Honors & Awards

- (2019 - Present) University Merit Scholarships in each semester for excellent postgraduate results
- (2015 - 2019) Dean's Award in each academic year for excellent undergraduate results
- (2015 - 2019) University Merit Scholarships in each semester for excellent undergraduate results

Technical Skills

- **Programming Languages**: Python, C++, Java
- **Frameworks**: PyTorch, Keras, TensorFlow

Selected Courses

- Artificial Intelligence
- Advanced Artificial Intelligence
- Machine Learning
- Data Mining
- Pattern Recognition
- Distributed Computing Systems

Services

- **Coach**
  BUET International Collegiate Programming Contest Teams (2020, 2021)
- **Member**
  BUET CSE Academic Curriculum Modification Committee (2020 - 2021)

Reference

Rifat Shahriyar
Professor
Department of CSE, BUET
Email: rifat@cse.buet.ac.bd

Yuan-Fang Li
Senior Lecturer
Deaprtment of Data Science & AI, Monash University
Email: yuanfang.li@monash.edu

Wasi Uddin Ahmad
Applied Scientist
Amazon Inc.
Email: wasiahmad@ucla.edu