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

Research Interests

My research interests are in **Natural Language Processing** with a particular focus on **efficient data and compute utilization** under resource scarcity. My works have so far focused on **low-resource languages** and **multilingual/cross-lingual** language models. My long-term goal is to build practical and general-purpose NLP systems that can learn to communicate with speakers of all languages with limited supervision.

Education

- **Bangladesh University of Engineering and Technology (BUET)** — Dhaka, Bangladesh
  M.Sc. in Computer Science and Engineering; CGPA: 3.92/4.0 — June 2019 - May 2022
- **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

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. **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]

3. **BanglaBERT: Language Model Pretraining and Evaluation Benchmarks for Low-Resource Language Understanding Evaluation in Bangla**
   Abhik Bhattacharjee*, **Tahmid Hasan*** (*Equal contribution*), Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, M. Sohel Rahman, Anindya Iqbal, Rifat Shahriyar
   In *Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2022*. [PDF] [Code]

4. **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]

5. **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/Pre-print:

1. **BanglaNLG: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla**
   Abhik Bhattacharjee, **Tahmid Hasan**, Wasi Uddin Ahmad, Rifat Shahriyar
   ArXiv Pre-print, 2022. [PDF] [Code]

2. **CrossSum: Beyond English-Centric Cross-Lingual Abstractive Text Summarization for 1500+ Language Pairs**
   Tahmid Hasan, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, Rifat Shahriyar
   ArXiv Pre-print, 2021. [PDF] [Code]

3. **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]

Research Experience

1. **Adapting XL-Sum for Many-to-Many 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. We align identical articles across languages via cross-lingual retrieval on the XL-Sum dataset and curate a large-scale cross-lingual summarization dataset containing 1.65 million article-summary samples in over 1500 language pairs. To effectively train cross-lingual summarization models, we introduce a multistage data sampling algorithm and propose a metric for automatically evaluating summaries when references in the target language are unavailable.
   Supervisors: *Prof. Rifat Shahriyar*, *Dr. Yuan-Fang Li* and *Dr. Wasi Uddin Ahmad* Status: Ongoing

2. **Multilingual 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 multilingual machine translation models into a paraphrase generation model. Using a forward and a backward NMT model as teachers, we distill the cross-attention and output distributions into a student paraphrasing model. In contrast to traditional KD, here we have two teachers instead of one, and the student model's task 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 perform extensive evaluation to demonstrate the high-quality, conciseness and abstractiveness of XL-Sum. We show higher than 11 ROUGE-2 scores on ten languages tested, with some of them exceeding 15, as obtained by multilingual training. We release the dataset, evaluation scripts, and models 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 BERT-based Bangla NLU model pre-trained on 18.6 GB data we meticulously crawled from top Bangla sites. We establish the *'Bangla Language Understanding Evaluation' (BLUE)* benchmark and achieve strong baselines on all BLUE tasks with BanglaBERT. Through comprehensive experiments, we identify a significant shortcoming of multilingual models, which we name the 'Embedding Barrier,' that hurts performance for low-resource languages that do not share writing scripts with any high-resource language.
   Supervisors: *Prof. Rifat Shahriyar* and *Dr. Wasi Uddin Ahmad* Status: Submitted to *ARR 2021*

5. **Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation**: In this work, we identify that erroneous sentence segmentation

and presence of noise deteriorates the quality of sentence alignments for Bengali. Therefore, we build a customized sentence segmenter for Bengali and introduce two methods for sentence alignment from noisy comparable document corpora on low-resource setups: *aligner ensembling* and *batch filtering*. Our proposed methods improve 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 we 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. **Phylogenetic Tree Estimation from Quartets**: In this work, we propose a hybrid algorithm for accurate phylogenetic tree reconstruction from quartets combining distance- and frequency-based approaches. We prove the proposed algorithm to be statistically consistent and achieve good experimental results on simulated datasets.
Supervisors: *Prof. M. Sohel Rahman* and *Prof. Shamsuzzoha Bayzid* Status: Completed

9. **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 of long-lived TCP connections and in the process, improve keep-alive intervals of mobile devices. Our proposed methods outperform the native Android keep-alive algorithm and improve notification delivery success by 11%.
Supervisors: *Prof. M. Saifur Rahman* and *Prof. M. Sohel Rahman* Status: Published in *IEEE/ACM Transactions on Networking (Vol: 26-1, 2018)*.

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 April 2019 - Present
  Supervisor: *Prof. Rifat Shahriyar*

Teaching Experience (Selected)

- CSE 471 Machine Learning: Jul'21
- CSE 218 Numerical Methods: Jul'19, Jul'21

- CSE 305 Computer Architecture: Jan'22
- CSE 472 Machine Learning Sessional: Jan'20, Jul'21
- CSE 204 Data Structures & Algorithms I Sessional: Jul'19, Jul'21
- CSE 412 Simulation & Modeling Sessional: Jan'20
- CSE 308 Software Engineering Sessional: Jan'21
- CSE 216 Database Sessional: Jul'21
- CSE 306: Computer Architecture Sessional
- CSE 208 Data Structures & Algorithms II Sessional: Jan'20
- CSE 462 Algorithm Engineering Sessional: Jan'20
- CSE 408 Software Development: Jul'19, Jan'22
- CSE 108: Object Oriented Programming Language Sessional: Jan'21, Jan'22
- CSE 308: Software Engineering Sessional: Jan'21

Honors & Awards

- University Merit Scholarships in each semester for excellent postgraduate results: 2019 - 2022
- Dean's Award in each academic year for excellent undergraduate results: 2015 - 2019
- University Merit Scholarships in each semester for excellent undergraduate results: 2015 - 2019
- Bronze Medal, Asian Pacific Mathematics Olympiad: 2013
- National Champion, Bangladesh Mathematical Olympiad: 2010, 2011, 2013, 2014
- National Champion, Bangladesh Olympiad in Informatics: 2012, 2013

Technical Skills

- **Programming Languages**: Python, C/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, 2022)
- **Member**
  BUET CSE Academic Curriculum Modification Committee (2020 - 2022)

Reference

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

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

Wasi Uddin Ahmad
Applied Scientist
AWS AI
Email: wuahmad@amazon.com