Abhik Bhattacharjee

38/9/A, Mirpur-11, Dhaka, Bangladesh
+880-1954689352

abhik@ra.buet.ac.bd
https://abhik.vercel.app/
Google Scholar | GitHub

Research Interests

- Data and Computing-Efficient Deep Learning
- Controllable Natural Language and Code Generation
- Low-Resource, Multilingual, and Cross-Lingual Natural Language Processing

Education

- **Bangladesh University of Engineering and Technology (BUET)** — Dhaka, Bangladesh
  B.Sc. in Computer Science and Engineering — February 2016 - February 2021
  - *CGPA: 3.69 on a scale of 4.00*
  - *Major GPA: 3.81 on a scale of 4.00*
  - *Position: Ranked* 26^{th} *in a class of 143 students*

Publications

(* indicates equal contribution)

1. **CrossSum: Beyond English-Centric Cross-Lingual Summarization for 1,500+ Language Pairs**
   Abhik Bhattacharjee*, Tahmid Hasan*, Wasi U. Ahmad, Yuan-Fang Li, Yong-Bin Kang, Rifat Shahriyar
   Proceedings of 61st Annual Meeeting of the Association for Computational Linguistics: ACL 2023. [PDF] [Code]

2. **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
   Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. [PDF] [Code]

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

4. **BanglaNLG: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla**
   Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Rifat Shahriyar
   Findings of the Association for Computational Linguistics: EACL 2023. [PDF] [Code]

5. **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
   Proceedings of the Empirical Methods in Natural Language Processing, EMNLP 2020. [PDF] [Code]

6. **GEMv2: Multilingual NLG Benchmarking in a Single Line of Code**
   Sebastian Gehrmann, **Abhik Bhattacharjee**, ⋯
   Proceedings of the Empirical Methods in Natural Language Processing, EMNLP 2022. [PDF] [Code]

7. **BanglaParaphrase: A High-Quality Bangla Paraphrase Dataset**
   Ajwad Akil*, Najrin Sultana*, **Abhik Bhattacharjee**, Rifat Shahriyar
   Proceedings of the Asia-Pacific Chapter of the Association for Computational Linguistics: AACL 2022. [PDF] [Code]

Selected Research Projects

1. **Improving Document-Level Event Argument Extraction with Coreference resolution**

   Supervisor: *Prof. Rifat Shahriyar (BUET)*
   Status: Ongoing

   In this work, we observe the efficacy of explicit coreference resolution in the conditional generation framework for document-level event argument extraction. Inspired from this, we employ a contrastive learning loss formulation among the entity mentions from the same coreference cluster to improve the end-to-end argument identification performance.

2. **Multilingual Paraphrase Generation via Knowledge Distillation from NMT Models**

   Supervisors: *Prof. Rifat Shahriyar (BUET)* and *Dr. Wasi Uddin Ahmad (AWS AI)*
   Status: Ongoing

   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.

3. **Test case aware Program Synthesis with Deep Reinforcement Learning**

   Supervisor: *Abdus Salam Azad (UC Berkeley)*
   Status: Ongoing

   We develop a test case generation framework from programming problem descriptions using CodeT5. Conditioned on these synthetic test cases, we design a critic network that provides dense feedback signals on the functional correctness of a generated program to guide an actor Language Model generation. We further propose using a weighted ensemble of these signals based on their relative importance to refine the generated code in multiple rounds.

4. **CrossSum: Beyond English-Centric Cross-Lingual Summarization for 1,500+ Language Pairs**

   Supervisor: *Prof. Rifat Shahriyar (BUET)* and *Dr. Wasi Uddin Ahmad (AWS AI)*
   Status: Published in *ACL, 2023*

   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 1.7 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.

5. **XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages**

   Supervisors: *Prof. Rifat Shahriyar (BUET)* and *Dr. Yuan-Fang Li (Monash Uni.)*
   Status: Published in *Findings of ACL, 2021.*

   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.

6. **BanglaBERT: Language Model Pretraining and Evaluation Benchmarks for Low-Resource Language Understanding Evaluation in Bangla**

   Supervisor: *Prof. Rifat Shahriyar (BUET)* and *Dr. Wasi Uddin Ahmad (AWS AI)*
   Status: Published in *Findings of NAACL 2022*

   In this work, we present *BanglaBERT* – a BERT-based Bangla NLU model pre-trained on 27.5 GB data we meticulously crawled from 110 top Bangla sites. We establish the *'Bangla Language Understanding Benchmark' (BLUB)*, introducing two new downstream datasets. BanglaBERT achieves state-of-the-art results on BLUB, outperforming larger monolingual and multilingual models. We additionally build *BanglishBERT* – a model jointly trained on Bangla and English data to facilitate strong zero-shot cross-lingual transfer performance.

7. **Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation**

   Supervisors: *Prof. Rifat Shahriyar (BUET)* and *Prof. M. Sohel Rahman (BUET)*
   Status: Published in *EMNLP, 2020.*

   We identify that erroneous sentence segmentation and presence of noise deteriorates the quality of sentence alignments for Bengali. To alleviate this issue, 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.

Professional Experience

- **Bangladesh University of Engineering and Technology (BUET)** — Dhaka, Bangladesh
  Graduate Research Assistant, Department of CSE, BUET — March 2021 - Present
  Supervisor: *Prof. Rifat Shahriyar*
- **Samsung R&D Institute** — Dhaka, Bangladesh
  Adjunct Research Assistant — August 2023 - December 2023
- **Intelligent Machines Limited** — Dhaka, Bangladesh
  AI Research Collaborator — April 2021 - September 2021
- **Bangladesh University of Engineering and Technology (BUET)** — Dhaka, Bangladesh
  Undergraduate Research Assistant, Department of CSE, BUET — February 2019 - February 2021
  Supervisor: *Prof. Rifat Shahriyar*

Honors & Awards

- **Dean's List Award**: Bangladesh University of Engineering and Technology — 2018 - 2019
- **ICT Innovation Fund**: Government of Bangladesh — 2020
- **University Merit Scholarship**: Bangladesh University of Engineering and Technology — 2018 - 2019
- **Board Merit Scholarship**: Government of Bangladesh — 2016 - 2021

Professional Services

- **Reviewer**: NAACL 2024
- **Program Committee**: BNLP Workshop 2023, GEMv2 Workshop 2022

Technical Skills

- **Programming Languages**: Python, Rust, C/C++, Java, TypeScript, MATLAB, SQL, Bash, LaTeX
- **Libraries/Frameworks**: PyTorch, Keras, TensorFlow, HF Transformers, NLTK, Scikit-learn, Flask, Next.js
- **Tools/Platforms**: Docker, Git, Kubernetes, Amazon Web Services, Google Cloud Platform

Selected Courses

- Artificial Intelligence
- Machine Learning
- Computer Networks
- Probability & Statistics
- Advanced Algorithms
- Operating Systems

Reference

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

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

Tahmid Hasan
Assistant Professor
Deparment of CSE, BUET
Email: tahmidhasan@cse.buet.ac.bd