Postgraduate Seminar Series

Venue: Graduate Seminar Room
Date & Time: September 21, 2024 at 3:00 PM

Speaker Information

Sanzana Karim Lora (Std No. 0422052057) is a full-time M.Sc. student in the department of CSE, BUET. She completed her undergraduate studies from CSE, Ahsanullah University of Science and Technology (AUST) in 2021. Her research interests lie in natural language processing, specifically for low-resource languages with a focus on overcoming data scarcity challenges. She is currently doing her postgraduate thesis under the supervision of Prof. Dr. Rifat Shahriyar. She will be speaking about her ongoing research in this talk.

Data-Scarce Solution for Cross-Lingual Summarization without Direct Equivalents

Cross-lingual summarization (CLS) is a branch of Natural Language Processing that generates correct and coherent summaries in a target language from articles written for a different source language. The task provides unique challenges since models must be able to properly translate and summarize. Traditional pipeline approaches are computationally intensive and frequently produce unsatisfactory results. To deal with these difficulties, subsequent studies have used sequence-to-sequence and Transformer-based models to conduct CLS inside a single model framework. Despite improvements, the area still needs data-efficient solutions and effective training methodologies. Our ongoing research presents a novel strategy for CLS that incorporates contrastive learning characteristics, which have demonstrated promise in various text generation situations. In this work, we propose a novel approach for CLS leveraging the power of contrastive learning. We generate candidate summaries in different languages based on the given source document and contrast these summaries with reference summaries concerning the given documents. After that, we train the model with a contrastive ranking loss. Then, we rigorously evaluate the proposed approach against current methodologies, confirming its efficacy in creating high-quality summaries. Furthermore, we evaluate the model's robustness with a cross-lingual dataset, showing its reliability across languages, specifically for low-resource languages. Moreover, we compare our proposed method with the state-of-the-art Large Language Models (LLMs) like Gemini, GPT 3.5, and GPT 4 and find that our model performs better for low-resource languages' CLS. We believe our findings will pave the way for future research on more efficient and accurate cross-lingual summarizing techniques.