RaMToE-FINE: Randomized Multilingual Training with Ordered Evaluation via Fine-Tuning and Incremental Data Expansion

Pradeep Sai Teja Sanka[1], Raja Pavan Vignesh Kajjayam[1], Sai Karthik Nallamothu[1], Yeshwanthvenkatakumar Vidhuvaluru[1], Tapas Kumar Mishra[1].

[1]*Department of Computer Science and Engineering, SRM University – AP, Andhra Pradesh, India*

Email: pradeepsaiteja_sanka@srmap.edu.in, rajapavanvignesh_k@srmap.edu.in, saikarthik_nallamothu@srmap.edu.in, yeswanthvenkatakumar_v@srmap.edu.in, tapaskumar.m@srmap.edu.in

Abstract: Multilingual textual summarization is a demanding task due to the semantic, syntactic and structural variance across many languages. In this research, we experiment with mT5 (Multilingual Text-to-Text Transfer Transformer) to do abstractive summarization across seven different major languages (Telugu, Urdu, Marathi, Hindi, Tamil, Bengali, English) spoken in India to analyze how a mixed multilingual data volume influences model performance as the dataset increases. The experiment steps involve incrementally increasing the number of data samples per language, starting with 10 instances per language and progressively upgrading up to 1000 samples and beyond. This controlled dataset expansion allows us to methodically check the impact of training data volume on the quality of output summaries. Model performance is evaluated using metrics such as ROUGE and BLEU scores. Generally, mT5 showcases good generalization capabilities, its effectiveness in many low resource languages is significantly influenced by the size of data available. To further improve summarization quality for, we use transfer learning techniques and fine-tuning methods. This research experiment provides promising understanding of adaptability and scalability of large multilingual transformer-based models, giving a path for optimizing summarization tasks in a wide range of linguistic works. Our key findings contribute to booming the domain of multilingual NLP, particularly in inventing efficient summarization methods for inadequate languages.

Keywords: Multilingual Summarization; mT5; Abstractive Summarization; Low-Resource Languages; Dataset Size; Transfer Learning; ROUGE; BLEU; NLP.