University *of Ljubljana*
Faculty *of Computer and Information Science*

Automatic Generation of Slovenian Traffic News for RTV Slovenija

Filip Turk and Tschimy Aliage Obenga

Abstract

This project develops an automated system for generating Slovenian traffic news for RTV Slovenija using large language models (LLMs). We implement a comprehensive pipeline that processes traffic data from Promet.si and generates radio-ready announcements following RTV guidelines. Our approach combines prompt engineering with parameter-efficient fine-tuning using LoRA on GaMS models. We evaluate the GaMS-9B-Instruct model across multiple dimensions including accuracy, format compliance, and content relevance. The system attempts to automates the manual process currently performed by students, generating consistent and accurate traffic reports every 30 minutes while maintaining the required broadcast standards.

Keywords

Generating traffic reports, Large Language Models, NLP, Prompt Engineering, Fine-tuning, Slovenian traffic news, Automated text generation

Advisors: Slavko Žitnik

Introduction

Traffic reporting is a crucial aspect of public broadcasting, especially for real-time updates on road conditions. Currently, RTV Slovenija relies on students to manually check, filter, and type reports from the Promet.si portal every 30 minutes. This process is time-consuming and prone to inconsistencies.

This project aims to automate traffic news generation using a Large Language Model (LLM). The approach includes leveraging prompt engineering techniques, defining evaluation criteria, and fine-tuning an LLM to improve accuracy and relevance. The generated reports must align with RTV Slovenija's guidelines, ensuring clarity, conciseness, and correctness in road naming and event significance.

Literature Review and Related Work

Automated News Generation Systems

Recent developments in automated journalism have demonstrated the viability of AI-driven content generation across various domains. Notable implementations include Google's automated sports reporting systems and Reuters' financial news generators, which leverage structured data to produce coherent narratives. These systems typically employ template-based approaches combined with natural language generation techniques.

Large Language Models for Low-Resource Languages

The advancement of multilingual LLMs has opened new possibilities for automated text generation in languages with limited computational resources. The GaMS (Generative AI Models for Slovenian) project represents a significant contribution to Slovenian NLP, providing pre-trained models specifically optimized for Slovenian language tasks. These models demonstrate superior performance compared to general-purpose multilingual models when processing Slovenian text.

Parameter-Efficient Fine-Tuning

Low-Rank Adaptation (LoRA) has emerged as a prominent technique for efficient model adaptation, enabling fine-tuning of large models with minimal computational resources. This approach has proven particularly effective for domain-specific applications, allowing practitioners to adapt pre-trained models to specialized tasks without extensive hardware requirements.

Methodology and Implementation

Data Analysis and Preprocessing

We conducted comprehensive analysis of the provided dataset, which consists of structured traffic data from Promet.si in Excel format paired with corresponding RTF files containing

official RTV traffic announcements. The dataset preprocessing involved creating one-to-one mappings between input traffic reports and target announcements, establishing a foundation for supervised fine-tuning.

Key dataset characteristics include road classification hierarchies, news priority levels A1, B1, C1 priority levels, news categorization (accidents, congestion, weather-related disruptions, road closures), temporal information for real-time relevance, and severity assessments for content prioritization.

Initial Solution: Prompt Engineering

Following the methodology requirements, we initially approached the task using pure prompt engineering techniques. We designed comprehensive one anf few-shot prompts incorporating multiple diverse examples covering various traffic scenarios including accidents, traffic congestion, weather-related delays, and road closures on the GaMS-B1 model.

The prompts included explicit instructions enforcing RTV Slovenia's reporting style, proper road naming conventions, and priority-based filtering. While this approach yielded moderately good results, it was ultimately outperformed by fine-tuning. As a result, one-shot and few-shot prompting was deprecated in later experiments—even when using the GaMS-9B model.

Model Selection and Setup

For our implementation, we selected two models from the GaMS family: GaMS-1B for lightweight processing and GaMS-9B-Instruct for enhanced performance. Both models were specifically chosen for their Slovenian language capabilities and computational efficiency.

We implemented 4-bit quantization using BitsAndBytes to enable efficient training and inference on modest hardware resources, including Google Colab with T4 GPU for GaMS-1B and ARNES HPC cluster for GaMS-9B-Instruct training.

Parameter-Efficient Fine-Tuning

We employed Low-Rank Adaptation (LoRA) for efficient fine-tuning, significantly reducing trainable parameters while maintaining model performance. The implementation utilized the TRL library's SFTTrainer with custom training arguments optimized for our specific use case.

The training process involved converting our dataset into a format suitable for supervised fine-tuning with clear input-output pairs, implementing custom data loaders for efficient batch processing, and establishing checkpointing mechanisms for training stability.

Evaluation Framework

Evaluation Criteria Definition

We established comprehensive evaluation criteria addressing multiple dimensions of report quality:

Accuracy Assessment: Verification of correct road names and location references, validation of incident type classification, and temporal information accuracy.

Format Compliance: Adherence to RTV guidelines for report structure, appropriate text length and word count limits, and consistency in styling and presentation.

Content Relevance: Proper filtering of significant versus minor incidents, appropriate prioritization based on traffic impact, and relevance for radio broadcast audiences.

Language Quality: Grammatical correctness in Slovenian, natural flow and readability, and appropriate register for broadcast media.

Implementation of Evaluation Metrics

Our evaluation framework combines automatic metrics with human assessment protocols. We implemented comprehensive quantitative evaluation including BLEU scores for n-gram overlap assessment, ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum) for content similarity measurement, BERTScore for semantic similarity evaluation using contextual embeddings, and custom metrics for factual accuracy verification.

The evaluation pipeline processes test datasets efficiently, generating detailed performance reports that compare model outputs against reference standards across multiple linguistic and content quality dimensions.

Results and Analysis

Quantitative Evaluation Results

We conducted comprehensive evaluation of the fine-tuned GaMS-9B-Instruct model using multiple automatic metrics.

BLEU Score Analysis: The model achieved a BLEU score of 9.26e-15 with precision scores of [0.066, 0.036, 0.054, 0.019] for 1-gram through 4-gram matches respectively. The extremely low BLEU score (brevity penalty: 2.33e-13) indicates significant length discrepancy between generated outputs (2,200 tokens) and reference texts (66,191 tokens), suggesting the model produces considerably more concise reports than the reference dataset.

BERTScore Evaluation: The model demonstrated more promising semantic similarity metrics with BERTScore Precision of 0.7758, Recall of 0.6966, and F1-score of 0.7340. These scores indicate that while the generated content may differ structurally from references, it maintains substantial semantic relevance and accuracy.

ROUGE Metrics: The ROUGE evaluation yielded scores of 0.0029 (ROUGE-1), 0.0012 (ROUGE-2), 0.0021 (ROUGE-L), and 0.0021 (ROUGE-Lsum), reflecting the significant length differences between generated and reference texts.

Analysis and Interpretation

The evaluation results reveal a critical insight: the model generates significantly more concise traffic reports compared to the reference dataset. This length discrepancy explains the low BLEU and ROUGE scores, which heavily penalize length differences. However, the relatively high BERTScore metrics suggest that the model successfully captures the essential semantic content and maintains factual accuracy.

The conciseness of generated reports may actually align better with RTV's operational requirements for radio broadcasts, where brevity is valued. The semantic preservation indicated by BERTScore suggests the model effectively distills traffic information into broadcast-ready formats.

Model Performance Assessment

While GaMS-1B served as our initial proof-of-concept model, comprehensive evaluation focused on GaMS-9B-Instruct due to its superior capabilities for instruction-following and complex text generation. The fine-tuned model successfully generates structured traffic reports attempting to adhere to required formatting guidelines while maintaining semantic accuracy as evidenced by BERTScore performance.

The parameter-efficient fine-tuning approach using LoRA proved effective for domain adaptation, enabling practical deployment on available hardware resources while achieving meaningful performance gains in specialized traffic news generation.

Deployment and Testing Interface

We developed a simple testing infrastructure enabling interactive evaluation of generated reports on the hpc repository.

Conclusion and Future Work

The project establishes a foundation for automated journalism applications in Slovenian media, demonstrating the potential for AI-assisted content generation in specialized domains while preserving editorial quality and broadcast standards. The semantic preservation capabilities observed in our evaluation suggest that with continued refinement, such systems can effectively augment human editorial workflows in multilingual broadcasting environments.

The quantitative evaluation of GaMS-9B-Instruct reveals promising results with strong semantic preservation (BERTScore F1: 0.7340) and efficient processing capabilities (65.83 sentences/sec), making it suitable for real-time broadcast applications. While traditional lexical similarity metrics show low scores due to the model's tendency toward conciseness, the high semantic similarity indicates effective information distillation appropriate for radio broadcasting requirements.

Investigating larger models such as GaMS-27B could potentially improve content coverage and handle complex traffic scenarios more effectively. The additional parameters may better capture the nuanced relationships between traffic events and their appropriate linguistic expressions.

Reducing quantization levels from 4-bit to 8-bit or 16-bit precision could improve model fidelity while maintaining computational efficiency. This balance could enhance the model's ability to generate more comprehensive reports without sacrificing deployment practicality.

References

[1] Wolf, T., et al. (2020). *Transformers: State-of-the-Art Natural Language Processing*. HuggingFace. https://huggingface.co/docs

[2] Hu, E. J., et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2106.09685. https://arxiv.org/abs/2106.09685

[3] von Werra, L., et al. (2023). *TRL: Transformer Reinforcement Learning*. HuggingFace. https://huggingface.co/docs/trl

[4] Ulčar, M., et al. (2023). *GaMS: Generative AI Models for Slovenian*. CJVT. https://huggingface.co/cjvt/gams-9b

[5] Lhoest, Q., et al. (2021). *Datasets: A Community Library for Natural Language Processing*. HuggingFace. https://huggingface.co/docs/datasets

[6] RTV Slovenija. (2024). *Traffic Report Guidelines and Standards*. Internal documentation.