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

Automatic generation of Slovenian traffic news for RTV Slovenija

John Doe, Jane Doe, and Marko Rozman

Abstract

In this report we present the related works, initial ideas and proposed dataset for our project.

Advisors: Slavko Žitnik

Introduction

The project focuses on replacing the manual production of traffic reports with an automated system. Traffic data provided in tabular form and traffic data from the promet.si website is used to generate traffic reports according to provided specifications.

0.1 Project Context and Motivation

- Timely Information: Reliable traffic updates are critical for public safety and efficient transportation management.
- Manual Process Limitations: Currently, students manually verify and type these reports every 30 minutes, leading to potential delays and inaccuracies.
- Automated Solution: By utilizing LLMs and advanced prompt engineering (inspired by works on news headline generation), we aim to generate clear, concise, and contextually accurate traffic news.

1. Related Work

Recent studies in neural summarization and headline generation provide valuable insights for our project:

- HG-News: News Headline Generation Based on a Generative Pre-Training Model: This paper presents a decoder-only architecture that incorporates pointer mechanisms and n-gram language features. Its focus on generating succinct news headlines while accurately handling out-of-vocabulary terms (such as specific road names) can be directly applicable to our task. [1]
- Algorithm for Automatic Abstract Generation of Russian Text under ChatGPT System: This work leverages the ChatGPT system to preprocess and generate concise summaries of Russian texts. Although in Russian, Its detailed data preprocessing pipeline (including tokenization, removal of stop words, and punctuation handling) and the integration of pointer mechanisms to capture key information offer a robust framework that can be maybe adapted to Slovene to ensure the fluency and accuracy of traffic news. [2]

2. Initial ideas

- Initial Prompt Engineering: Experiments will start by generating news texts directly from the structured traffic data using prompts.
- Selection of important data. Our observations show that the final news report is mostly made from only few important data points from the excel file. We intend to filter the data to remove duplicate enteries and automaticly select the important data points relavant for our traffic news generation.
- Enhanced Generation through Fine-Tuning: The project will then incorporate parameter-efficient fine-tuning (e.g., LoRA) and retrieval techniques to further refine the generated content, ensuring correct road naming and event descriptions.
- Evaluation: A robust evaluation framework will be established using both automatic metrics (such as ROUGE, precision, recall, and F1) and human judgment to verify that the generated texts meet RTV Slovenija's standards.
- Pointer Mechanisms: Integrate pointer mechanisms to ensure that domain-specific terminology (e.g., road names and traffic event descriptors) is accurately reproduced, minimizing the risk of omitting critical details.

- Advanced Preprocessing: Apply rigorous preprocessing techniques—such as tokenization, normalization, and filtering—to both the traffic data and guideline documents. This will help standardize the input and ensure adherence to established formats.
- N-gram Language Features: Incorporate n-gram features to improve language fluency and ensure that generated sentences are coherent and stylistically consistent with existing RTV Slovenija news.

3. Project dataset

The primary dataset is the Excel file `Podatki - PrometnoPorocilo_2022_2023_2024.xlsx`, which contains detailed real-time and historical traffic information from the *promet.si* portal. The documents `PROMET.docx` and `PROMET osnove.docx` provide the rules for road naming, event hierarchy, and news formulation. We will also use Python notebooks such as `Prompting_and_lora_fine-tuning.ipynb`, and `Retrieval_augmented_generation.ipynb` for insights into prompt engineering and fine-tuning, which will guide our approach. We intend to normalize the data by enforcing standard naming conventions, filtering out irrelevant details, and tagging urgent events (e.g., *nujna prometna informacija*).

References

[1] Ping Li, Jiong Yu, Jiaying Chen, and Binglei Guo. HG-News: News Headline Generation Based on a Generative Pre-Training Model. 9:110039–110046.

[2] Lina Hou. Algorithm for Automatic Abstract Generation of Russian Text Under ChatGpt System. In *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–5.