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

Automatic generation of Slovenian traffic news for RTV Slovenija

Janez Kuhar, Veljko Dudić, 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.

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

Li et al. [1] proposed a decoder-only model for headline generation, incorporating multihead attention, sentiment features, and part-of-speech information. 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.

In a more recent work on the topic of text summarization [2], a ChatGPT-based algorithm for summarizing Russian texts was proposed. The suggested approach combined preprocessing (e.g., word segmentation, stop word removal) with BERT for contextual understanding and a pointer mechanism for out-of-vocabulary words.

Sha et al. [3] proposed an order-planning mechanism for text generation from tabular data, combining link-based attention with traditional content-based attention to model the sequence of information. Their approach was further enhanced by a copy mechanism to handle rare words, improving the model's ability to generate accurate and coherent summaries from structured data.

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. Methods

Project dataset

The primary dataset was provided to us by the TA. It consists of three important sets of files:

1. **A structured Excel** file containing historical traffic report information from *promet.si*.
2. **Word documents** with lexical notes and the prescribed structure of a traffic report.
3. **Ground truth data** in Rich Text Format (RTF) files, providing text in a ready-to-use format for radio hosts.

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.

[3] Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. Order-planning neural text generation from structured data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.