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

Automatic generation of Slovenian traffic news for RTV Slovenija

Petra Kuralt, Erik Mažgon, Nina Sangawa Hmeljak

Abstract

The abstract goes here. The abstract goes here.

Keywords

Traffic News Generation, Large Language Models (LLM), RTV Slovenija, Prompt Engineering

Advisors: Slavko Žitnik

Introduction

Traffic news plays a crucial role in informing the public about road conditions, accidents, and other transportation-related updates. At RTV Slovenija, the current approach involves students manually reviewing and compiling traffic news every 30 minutes. This process is labor-intensive and time-consuming. The objective of this project is to leverage a Large Language Model (LLM) to automate the generation of Slovenian traffic news, making the process more efficient and accurate. We will employ fine-tuning and prompt engineering techniques to ensure the model aligns with the structured guidelines used by radio presenters. Additionally, we will establish evaluation criteria to assess the generated outputs based on correctness, relevance, and readability.

Related works

The application of Natural Language Processing (NLP) and Large Language Models (LLMs) in automated news generation has been extensively explored across various domains. In our research we focused particularly traffic reporting. Recent advancements focus on enhancing the accuracy, readability, and contextual relevance of automatically generated reports through prompt engineering, fact-enhanced generation, and fine-tuning techniques.

One of the most relevant works in this domain is the BBC's automated news system, which leverages structured journalism principles combined with template-based NLP. The BBC used Arria Studio, an AI-driven tool, to allow journalists to design customizable templates that automatically generate news stories from structured data. Their semi-automated local content (SALCO) system integrates predefined story structures with NLP-based text generation, ensuring that output aligns with journalistic standards while allowing human oversight in edge cases. [1]

A key study in traffic news automation is an NLG-based tweet generation system for real-time traffic updates. This system processes structured incident data and uses a semantic alignment model to map data fields to natural language representations. It then applies concept-to-text generation models to produce short, human-like traffic incident reports in a location-aware manner. By incorporating prompt engineering techniques, the system ensures concise and contextually relevant outputs for different user preferences. [2]

Further advancements in data-to-text generation are demonstrated in GameRecapper, a system for generating sports summaries. While primarily designed for football match reporting, it employs template-based sentence structuring and domain-adaptive text generation, showcasing how predefined prompt patterns and structured data inputs can produce coherent and varied text outputs. This approach is particularly relevant for automating structured traffic reports, where predefined patterns can enhance clarity and fluency. [3]

In the domain of LLMs and fact-enhanced generation,

FACTGEN introduces a hybrid approach that integrates external knowledge retrieval with LLM-driven text generation. This model applies fact-enrichment techniques to ensure that generated news remains factually consistent and contextually rich. FACTGEN's methodology aligns closely with LLM fine-tuning for domain-specific traffic news, where additional datasets can improve accuracy and relevance. [4]

Recent advancements in NLP-driven traffic reporting have explored the integration of social media mining with fine-tuned LLMs to enhance real-time traffic alerts. One such approach employs a BERT model to classify and extract relevant traffic incident details from social media posts. By leveraging question-answering techniques, this system identifies key attributes such as location, time, and severity, transforming them into structured alerts for navigation assistants. This method highlights the potential of LLM-based prompt engineering to improve the responsiveness and accuracy of automated traffic reporting. [5]

In the field of AI-generated radio news, recent research has tested ChatGPT's ability to produce concise, broadcast-ready news reports. By iteratively refining prompts, researchers assessed the model's adherence to journalistic standards such as clarity, brevity, and objectivity. While the AI successfully adapted to structured news writing, it faced challenges with grammatical nuances, numerical expressions, and language-specific syntax. This study underscores the role of prompt engineering in optimizing LLM outputs for domain-specific news automation, particularly in multilingual contexts such as Slovenian and Slovak traffic reports. [6]

Collectively, these studies demonstrate the evolving role of NLP, LLM fine-tuning, and prompt engineering in automated traffic news generation. They highlight the importance of structured data processing, real-time adaptation, and factual accuracy, key considerations for deploying LLM-based Slovenian traffic news automation.

Methods

The project involves two key datasets: Podatki - Prometno-Porocilo_2022_2023_2024.xlsx, which contains structured traffic data, and Podatki - rtvslo.si, which includes archived traffic news manually written by students. Additionally, we have an instructional document, PROMET.docx, which outlines the formatting and structuring of traffic reports. Since the raw Excel data consists of various columns such as event categories, timestamps, operator details, and HTML-formatted text fields, preprocessing is necessary to ensure the data is usable. This preprocessing involves extracting relevant traffic event types (e.g., accidents, roadworks, congestion, weather conditions, and obstacles), cleaning and parsing HTML text fields, standardizing road names and directions based on PROMET.docx, and filtering out redundant or irrelevant entries. These steps are essential for improving the accuracy and consistency of the generated traffic reports.

As additional input we will experiment how additional external data such as weather may improve our output. We will use open source data from the Slovenian government that are on OPSI (such as weather information from ARSO or the past data about roadworks and car accidents). In our initial approach, we will focus on prompt engineering to generate traffic news on Llama 3.1 LLM. Llama is a free open-source LLM that does not officially support Slovene, but it can still generate decent-quality Slovenian text, making it a good option for our project. Alternatively we will use DeepSeek, another open-source LLM that supports Slovene language. We will experiment with different prompt techniques such as instruction tuned, self-consistency, prompt chaining, structuring the prompt with tags, adopting a persona, etc. ensuring generated text adheres to RTV Slovenija's guidelines from the instructional document.

Evaluation

To evaluate the generated traffic reports, we will compare the model generated reports with human-written reports from Podatki - rtvslo.si to assess accuracy and relevance. This will involve tokenization of reference and generated text so that we can evaluate it with precision, recall, and F1-score to measure the identification of key traffic events. We could also use BLEU that measures how closely the reference and generated text match by evaluating the overlap of words and phrases. Or use ROUGE scores that assess textual similarity with reference reports. The system can also be evaluated based on its ability to correctly identify and prioritize important events, ensuring that high-priority incidents such as wrong-way drivers, road closures, and severe congestion are emphasized. To achieve this, we will create a priority weight system based on the hierarchy of events from PROMET.docx. Lastly we can conduct human evaluation to rate reports based on clarity, correctness, and informativeness.

Results

TODO

Discussion

TODO

Acknowledgments

TODO

References

[1] Samuel Danzon-Chambaud. Automated news at the bbc. *BBC Research and Development white paper, 2021c*, 2021.

[2] Khoa Tran and Fred Popowich. Automatic tweet generation from traffic incident data. In *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, pages 59–66, 2016.

[3] João Pinto Barbosa Machado Aires. Automatic generation of sports news. Master's thesis, Universidade do Porto (Portugal), 2016.

[4] Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. Fact-enhanced synthetic news generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13825–13833, 2021.

[5] Xiangpeng Wan, Michael C Lucic, Hakim Ghazzai, and Yehia Massoud. Empowering real-time traffic reporting systems with nlp-processed social media data. *IEEE Open Journal of Intelligent Transportation Systems*, 1:159–175, 2020.

[6] Lucia Furtáková-L'ubica Janáčková. News at the speed of ai: Automating journalism through text generator. *MARKETING IDENTITY*, page 166, 2024.