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

Automatic generation of Slovenian traffic news for RTV Slovenija

Matej Brodnik, Žan Mencigar, Primož Mihelak

Abstract

With the recent developments of artificial intelligence and large language models (LLMs), many simple tasks can be streamlined and automatized. We will attempt to implement a robust solution that will enable automatic generation of traffic news in Slovene, given key traffic related events as input. We will employ prompt engineering and fine-tune the algorithm using Slovene LLMs to generate short traffic news automatically. We will then analyze and present the results and discuss the algorithm's performance.

Keywords

GaMS, LLM, Prompt engineering, Text generation, Traffic news

Advisors: Slavko Žitnik

Introduction

Traffic news are an important part of RTV Slovenija's radio program, where presenters read out traffic updates every 30 minutes. Currently, students manually check the promet.si portal, select relevant traffic events, and write short reports based on internal guidelines. This process takes time and could be made more efficient.

The goal of this project is to automate the generation of Slovene traffic news using GaMS 9B large language model (LLM) [1]. GaMS is based on Google's Gemma 2 family and was originally trained designed as 1B model, but was extended for 2B, 9B and 27B parameters. It was specifically trained for Slovene language, using a corpora of primarily Slovene, but also Croatian, Serbian and Bosnian documents consisting of 28 billion tokens [2].

The idea is to first explore how well we can solve the task using prompt engineering, and then improve the system by fine-tuning an LLM with provided data. We will evaluate the model using criteria such as how well it identifies important events, uses correct road names, follows text-length limits, and respects the style of RTV Slovenija's reports, while also measuring performance improvements between different methods, such as prompt engineering and fine-tuning. The goal is to create a system that can automatically generate traffic news similar to what is currently done manually.

Data

The traffic news data was scraped from promet.si over years 2022-2024 and is saved in Excel format, making up over 170k rows of data in total. Each row contains a timestamp of acquisition and 3 separate segments (A, B, C) that contain data based on their importance, with A containing urgent news, B containing more general but still important news and C containing less significant announcements that may remain unchanged for several months. This data is also split across 10 thematic categories (ie. accidents, weather, obstacles, ...). We also have access to existing RTV Slo bulletins from rtvslo.si containing human-written traffic news in regular 30 minute intervals in the same time period as the scraped data, serving as a reference to what the final output should look like. Finally, we also have instructions detailing the news structure, road naming conventions, lecturer's notes, example formulations and general guidelines regarding the formatting of the news.

Related work

Since the emergence of LLMs, many similar projects have been conceived. Yuan et al. [3] describes the use of their LLM named HEAL, which was purpose-built for medical conversations and was measured on automated scribing.

A related approach to automated report generation described by Guan et al. [4], is the SRAG (Summary Report Auto-Generation) framework which uses large language models to transform query data into high-quality summary reports through hierarchical corpus retrieval across general fields.

Similarly, Su et al. [5] introduce a model named TableGPT2, which is pre-trained and fine-tuned. Its extensive training enables it to excel in table-centric tasks while maintaining strong general language and coding abilities.

Zhao et al. [6] investigate the table-to-text generation capabilities of different large language models in different real-world information seeking scenarios.

Lastly, Tian et al. [7] propose ReSpark, which is a method based on LLMs to generate new reports using previous similar-topic data reports as references.

Methods

Prompt engineering

Constructing an effective prompt is essential for producing meaningful and accurate outputs from our language model. Our prompt includes:

- **General Instructions** to establish the tone and purpose of the task,
- **Specific Guidelines** for structuring and formatting traffic reports in line with RTV Slovenija's editorial standards,
- A **Template** to ensure consistency and clarity,
- And a real **Example** based on previously published traffic reports.

Additionally, we provide the model with **processed input data** extracted from our structured Excel database. The data for each prompt is collected from all events that happened in the last 45 min.

Evaluation

For evaluation, we used Recall-Oriented Understudy for Gisting Evaluation (ROUGE), a set of metrics traditionally used for evaluating how similar a generated summary is to a human-written one. Its results may be flawed however, as synonyms are not detected, while the occurrence of shared common words potentially artificially inflates the metrics. The recorded metrics are:

- ROUGE-N: overlap of N-grams (unigrams, bigrams)
- ROUGE-L: longest common subsequence (LCS)
- ROUGE-Lsum: average LCS of each sentence

For each of them, recall, precision and F1 can be calculated. Recall for ROUGE-N measures how much of the reference matches generated text:

$$R = \frac{\text{Overlapping N-grams}}{\text{Total N-grams in reference}} \tag{1}$$

Precision for ROUGE-N measures how much of the generated text is relevant:

$$P = \frac{\text{Overlapping N-grams}}{\text{Total N-grams in generated text}} \tag{2}$$

F1 score balances the two scores, giving a more general purpose approximation of how well the compared texts match:

$$P = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

For ROUGE-L, recall computes the proportion of LCS length to reference length, while precision computes the proportion of LCS length to generated text length.

Results

The model achieves its highest scores in ROUGE-1 and ROUGE-Lsum, indicating it captures relevant unigrams and preserves a coherent sentence structure. ROUGE-2 scores are noticeably lower, indicating that the model struggles more with producing natural-sounding word combinations and fluent phrasing.

Metric	Precision	Recall	F1-Score
ROUGE1	0.4305	0.5052	0.4268
ROUGE2	0.2804	0.3113	0.2674
ROUGEL	0.3538	0.4141	0.3467
ROUGELsum	0.4036	0.4638	0.3956

Table 1. Results of ROUGE metrics over 10 examples

A manual review of the outputs reveals that the model often closely reproduces the input data with minimal rephrasing or abstraction. While the formatting of the generated text typically aligns with the structure defined in the prompt, the content sometimes reflects direct copying rather than paraphrasing. In a few instances, the model appears to incorporate specific details from the example included in the prompt, rather than relying solely on the provided input data.

These findings suggest that while the model can produce syntactically correct and well-structured traffic reports, its generative behavior leans more toward extractive summarization. This limits its ability to generalize and compose reports.

Discussion

There are still many ways to improve upon the current solution: the model can be fine-tuned through different prompt structure and instructions, finding optimal sets of parameters and streamlining the input data. While we have begun with automatic methods of analysis such as ROUGE, such a task also requires thorough manual evaluation: are the generated news sensible and decipherable from a human standpoint? Do they capture all relevant information and do not add unnecessary or misleading information?

We can also test our solution using other LLMs. Despite GaMS's training on Slovene data, there are other LLMs that have stronger reasoning skills and may be able to follow the logic of our prompts more accurately. We could then compare the performances between general purpose multilingual LLMs and focused LLMs designed for a specific language.

Acknowledgments

References

[1] Huggingface. Gams 9b, 2025.

[2] Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik Šikonja. Generative model for less-resourced language with 1 billion parameters. *Zenodo*, 2024.

[3] Dong Yuan, Eti Rastogi, Gautam Naik, Sree Prasanna Rajagopal, Sagar Goyal, Fen Zhao, Bharath Chintagunta, and Jeffrey Ward. A continued pretrained llm approach for automatic medical note generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 565–571, 2024.

[4] Xinxin Guan, Zekai Ye, Bin Cao, and Jing Fan. Summary report auto-generation based on hierarchical corpus using large language model. *SSRN*, 2023. Available at SSRN: https://ssrn.com/abstract=5095836 or http://dx.doi.org/10.2139/ssrn.5095836.

[5] Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, et al. Tablegpt2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*, 2024.

[6] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. Investigating table-to-text generation capabilities of llms in real-world information seeking scenarios. *arXiv preprint arXiv:2305.14987*, 2023.

[7] Yuan Tian, Chuhan Zhang, Xiaotong Wang, Sitong Pan, Weiwei Cui, Haidong Zhang, Dazhen Deng, and Yingcai Wu. Respark: Leveraging previous data reports as references to generate new reports with llms. *arXiv preprint arXiv:2502.02329*, 2025.