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.

Data preprocessing

The pipeline first loads the RTF files and extracts dates and times embedded in their content. It then selects the corresponding rows from the Excel file based on this timestamp information. From these selected rows, duplicate entries are removed, and the HTML content present in the columns is converted into plain text. The process further cleans the data by keeping only the relevant columns (namely A1, B1, C1, A2, B2, C2, ContentPomembnoSLO, ContentNesreceSLO, ContentZastojiSLO, ContentVremeSLO, ContentOvireSLO, ContentDeloNaCestiSLO, ContentOpozorilaSLO, ContentMednarodneInformacijeSLO, and ContentSplosnoSLO), which are then merged to form the input text for the model.

Prompt Engineering

After preprocessing the tabular data, we employed a simple prompt engineering approach to generate the traffic news. We started with a *few-shot* prompting [4] technique, where we provided the model with a few examples of preprocessed data and corresponding traffic news reports. This approach provides the model with contextual examples, helping it mimic the structure and tone of the target traffic news format.

Fine-tuning

As an extra step, we decided to fine-tune *cjvt/GaMS-9B-Instruct* on our input/ground truth pairs. Since the model has 9 billion parameters and demands substantial GPU memory, we employed QLoRA to enable efficient fine-tuning for the news generation task. This approach allowed us to update only a small number of low-rank adapter weights while keeping the quantized base model frozen. Training prompts were structured as alternating user and model turns using special tokens to clearly mark role boundaries, as shown below:

```
<start_of_turn>user
...pre-processed input data...
<end_of_turn>
<start_of_turn>model
...ground truth traffic news report...
<end_of_turn>
```

Evaluation metrics

The evaluation focuses on the BLEU score as a measure of translation quality, with the possibility of adding other metrics in the future.

BLEU Score: The BLEU (Bilingual Evaluation Understudy) score is a popular metric for evaluating the quality of machine-generated text, particularly for tasks like text summarization or machine translation. It is based on precision, specifically n-gram precision, where a higher BLEU score indicates that the machine-generated text closely matches the ground truth.

The BLEU score ranges from 0 to 1, where a higher value suggests better quality. However, BLEU has limitations, especially when dealing with languages that have more varied sentence structures or when evaluating more creative text generation tasks like traffic report creation. For example, it does not account for semantic meaning and may penalize diverse word choices or paraphrasing that is still correct. Additionally, BLEU heavily relies on n-gram matching, which may penalize acceptable lexical variations. For example, if the model uses different phrasing that still conveys the same meaning, it may receive a lower BLEU score, which does not necessarily reflect a poor-quality output.

4. Results and Evaluation

Computational Resources

Inference and training for all experiments was conducted on the ARNE HPC cluster. The cluster is equipped with NVIDIA's H100 80GB and V100 32GB GPUs, enabling efficient handling of LLMs.

Quantitative Results

V1

For each experimental setup, we ran 32 trials to ensure statistical reliability. During each trial, a randomly sampled set of input-output pairs was used, and the performance score was calculated for each pair.

The baseline results for few-shot prompting with GaMS-9B-Instruct [5] are listed in Table 1. We experimented with different configurations by varying the number of shots (input/output pairs) and the number of rows merged from the Excel file into a single input.

As expected, increasing the number of shots provided the model with more context, which typically improves its performance. However, increasing the number of shots also resulted in higher memory consumption, so we had to find a balance that remained within our system's limits.

Table 1. Average BLEU scores for different few-shot prompting setups using GaMS-9B-Instruct. Each configuration was evaluated over 32 trials. **Rows**: The number of Excel rows merged into a single input. **Shots**: The number of input/output pairs presented to the model for few-shot prompting.

Rows	Shots	Trials	BLEU
1	8	32	**0.1942**
3	2	32	**0.1576**

Interestingly, the average BLEU score decreased when we merged several Excel rows into the input. While we expected more data to improve the model's performance, the increase in data duplication led us to reduce the number of shots. With only 2 examples presented to the model, the model lacked sufficient context to recover all the relevant details from the ground truth, which likely led to missing parts in the generated reports.

V2

First, we fine-tuned *cjvt/GaMS-9B-Instruct*. We loaded the base model weights in 4-bit (NF4) precision with double quantization. LoRA adapters of rank 16 ($\alpha = 32$, dropout $= 0.1$) were applied to all query/value projection layers, with no bias terms and a causal-LM objective. Training used a JSON file of Input-GroundTruth pairs, each formatted as described in the Methods section, and truncated to 2048 tokens (including EOS), with 80% used for training and 20% held out for evaluation. Optimization employed an AdamW-style optimizer with peak learning rate 0.0002, linear warmup over 150 steps (3% of total 5000 steps), and gradient clipping at 0.3. Effective batch size was 1 (no accumulation), and mixed precision (FP16) was used.

Next we compared inference on finetuned model with the baseline *cjvt/GaMS-9B-Instruct*. The results are shown in Figure 1.

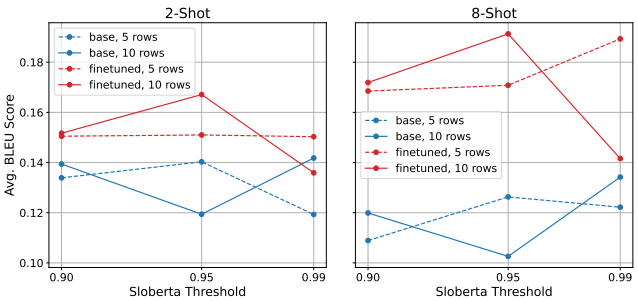

Figure 1. Inference results comparing fine-tuned and baseline models. X-axis: Sloberta cosine similarity thresholds (lower = shorter input). Red: fine-tuned model, Blue: baseline. Left: 2-shot, Right: 8-shot prompting. Fine-tuned model shows higher BLEU scores.

4.1 Qualitative Example / Error Analysis

V1

To complement the quantitative BLEU scores reported in Table 1, we asked the GPT-o3 assistant to act as an external evaluator: it was shown one randomly selected test case from the same evaluation set and requested to assign an adequacy score on the 1–10 scale described earlier.

Ground-truth bulletin

Podatki o prometu. Zaradi del so zastoji na regionalni cesti Vrhnika–Ljubljana pri avtocestnem priključku Brezovica v obe smeri. Na primorski avtocesti proti Kopru je zaradi razlitega goriva vozišče spolzko med Vrhniko in Logatcem. Na vipavski hitri cesti je zaradi pokvarjenega tovornjaka med predorom Podnanos in priključkom Vipava še vedno zaprt vozni pas proti Novi Gorici. Na mejnem prehodu Obrežje vozniki tovornih vozil na vstop v državo čakajo več kot 2 uri.

Model-generated bulletin

Podatki o prometu. Na ljubljanski obvoznici je zastoj proti Primorski, ki sega do razcepa Kozarje. Čas potovanja se tam podaljša za približno četrt ure. Na cesti Maribor – Hoče je zastoj dolg 2 kilometra. Na regionalni cesti Vrhnika – Brezovica – Ljubljana je zastoj pred delovno zaporo pri Brezovici proti Podpeči. Na cesti Podpeč – Brezovica je zastoj od Podpeči proti Brezovici. Priporočamo obvoz po avtocesti med Vrhniko in Ljubljano, vendar le za vozila opremljena z vinjeto ali napravo DarsGo. Na dolenjski avtocesti je med predoroma Mali vrh in Debeli hrib oviran promet zaradi okvare vozila. Na mejnem prehodu Obrežje vozniki tovornih vozil na vstop v državo čakajo do 2 uri.

Table 2. Illustrative example of expected versus generated traffic bulletin used in the OpenAI o3 adequacy evaluation.

o3 adequacy score. Using the 1–10 scale described in our annotation guidelines, this example received an **adequacy score of 4**. The score reflects the following observations:

- **Coverage:** Only one of the four critical events in the ground truth is reproduced; two are omitted entirely (slick lane Vrhnika–Logatec, lane closure Podnanos–Vipava) and one is partially correct (queues Vrhnika–Brezovica but only one direction).
- **Hallucinations:** Five incidents not present in the ground-truth bulletin are introduced (e.g. Ljubljana ring, Maribor–Hoče, Podpeč queues).
- **Detail accuracy:** Waiting time at Obrežje is understated ("do 2 uri" vs. "več kot 2 uri"), softening the severity.

Take-aways. This qualitative evidence clarifies why the BLEU score (≈ 0.19) remains modest despite grammatical fluency: the model tends to *omit* salient facts and *invent* unrelated incidents, both of which automatic n-gram metrics penalise weakly.

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.

[4] Prompting Guide. Few-shot prompting, 2024. Accessed: 2025-04-30.

[5] CJVT. Gams-9b-instruct. https://huggingface.co/cjvt/GaMS-9B-Instruct, 2025. Accessed: 2025-04-30.