FRI Natural language processing course 2021
Project report

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

Automatic generation of Slovenian traffic news for RTV Slovenija

Katarina Gojković, Alen Kurtagić, and Žan Terplan

Abstract

/

Keywords
Traffic report, LLM

Advisors: Slavko Žitnik

Introduction

Large Language Models (LLMs) have revolutionized natural language processing (NLP) by achieving state-of-the-art performance across a wide range of tasks, including text generation, translation, and summarization. Models like GPT-3 [1] and BERT [2] have transformed the field by demonstrating advanced capabilities in processing and generating human-like text. However, despite their success, LLMs face challenges in less spoken languages with low resources.

Related Work

LLMs have been increasingly used in automated news generation domains, especially in fields like finance, sports, and weather reporting. Despite this, traffic news generation is still a fairly unexplored area. However, some studies have explored this topic, for example, Wan et al. [3] leveraged social media data to complement traditional traffic reporting systems. Ouyang et al. [4] used AI agents to process multi-scale traffic data, conduct analysis and generate suggestions.

Besides traffic news generation, data-to-text generation in general has seen significant advancements with the application of neural network architectures. Several recent studies have explored different methodologies for improving the quality and accuracy of generated text from structured data.

One key advancement in data-to-text generation is the incorporation of entity modeling. Puduppully et al. (2019) [5] proposed an entity-centric neural architecture that dynamically updates entity-specific representations and employs hierarchical attention for generating coherent and factually accurate text. Their approach outperformed existing models on both the benchmark ROTOWIRE dataset and a newly introduced baseball dataset, demonstrating improvements in content selection and text fluency.

Another significant development in the field has been the comparison between pipeline and end-to-end architectures. Castro Ferreira et al. (2019) [6] systematically compared neural pipeline models with end-to-end approaches for generating text from RDF triples. Their findings indicated that pipeline models, which introduce explicit intermediate steps, produce more fluent and accurate text than end-to-end models. Additionally, the pipeline approach was found to generalize better to unseen inputs, reinforcing the notion that structured intermediate representations play a crucial role in high-quality text generation.

Curriculum learning has also been explored to enhance data-to-text generation, particularly in cross-lingual and noisy data settings. Hari et al. (2024) [7] introduced an approach that applies curriculum learning principles by ordering training samples based on an alignment score and using an annealing schedule. Their method resulted in significant improvements in BLEU scores and faithfulness metrics across multiple Indian languages and English.

A recent study by Zhang et al. (2024) [8] proposed optimized LLMs tailored for medical record generation, incorporating data augmentation techniques and domain-specific customization. Their models significantly improved faithfulness scores compared to state-of-the-art general models, demonstrating the importance of high-quality annotated data in generating accurate medical records.

Overall, these studies highlight various approaches to improving data-to-text generation, from entity modeling and pipeline architectures to curriculum learning and domain-specific optimizations.

Table 1. Candidate models

Metric	Architecture	Size	Slovene	Instruct-tuned
GaMS-1B-Chat [9]	Decoder	Medium	Yes	Yes
Mixtral-8x7B-Instruct	Decoder	XL	Yes	Yes
mt0-large	Encoder-Decoder	Medium	Yes	Yes
aya-101	Decoder	XL	Yes	Yes
Llama-3.1-8B	Decoder	XL	Partial	Partial
gemma-2b-it	Decoder	Medium	Partial	Yes
t5-sl-small	Encoder-Decoder	Small	Yes	No

Proposed Approach

Model Selection Criteria

Selecting an appropriate pre-trained language model is a critical step in designing any language generation system. Especially for specialized data-to-text tasks such as generating Slovenian traffic news reports. LLMs differ significantly in terms of architecture, size, multilingual support, and their ability to follow natural language instructions, all of which must be considered.

Firstly, since our input consists of structured data, we consider the encoder-decoder architecture to be the most suitable for this task. Unlike classical text generation, data-to-text generation benefits from models that can explicitly encode the input structure before producing the output. In encoder-decoder models, the encoder processes the input data into a meaningful representation, which the decoder then uses to generate a coherent output. This separation often leads to better consistency and can help reduce hallucinations. That said, we do not rule out the use of decoder-only architectures, especially given their recent success in following instructions.

Since the model will not be deployed on embedded or resource-constrained devices, there is no strict requirement for it to be extremely lightweight. At the same time, the task does not demand highly creative or open-ended text generation, which is where larger models typically excel. Furthermore, the inference time requirements are relatively relaxed, as traffic news reports are generated on a fixed schedule (e.g., every 30 minutes) and not in real-time. Therefore, a mid-sized model that balances performance, generation quality, and reasonable inference time is likely to be the most appropriate choice for this application.

One of the most important factors for this task is the model's ability to generate text in Slovenian, a low-resource language with limited available training data. Due to this limitation, it is common practice to first pretrain multilingual models on high-resource languages such as English, and then fine-tune them on smaller languages. While the exact mechanisms behind this transfer are complex, the shared linguistic patterns among languages can enable the model to generalize better. In the case of Slovenian, additional training data from closely related South Slavic languages (such as Croatian, Bosnian, and Serbian) are often used to improve performance, as these languages share similar vocabulary [10].

Another important consideration is the choice between base models and instruction-tuned models. Base models, such as the original versions of GPT, T5, or BLOOM are trained to predict the next word in a sequence without any explicit understanding of task-specific instructions. As a result, they function more like autocomplete systems and are not suitable for tasks such as answering questions and summarizing. In contrast, instruction-tuned models, such as ChatGPT, FLAN-T5, or BLOOMZ have been further fine-tuned on datasets containing natural language instructions paired with expected outputs. This enables them to follow prompts like ”summarize this,” ”translate this,” or, in our case, ”generate a report from this data.” Since our task depends on instruction-driven behavior and will involve additional fine-tuning on report generation instructions, we focus our selection to mostly instruction-tuned models.

Table 1 lists the models selected as initial candidates for experimentation and potential fine-tuning, based on the outlined criteria.

Prompt Engineering

Before applying fine-tuning techniques, we will first explore whether Slovenian traffic report generation can be effectively solved using prompt engineering. We will employ role-based prompting, where we explicitly assign the model the role of a traffic reporter, and few-shot prompting, where we provide examples to guide the output. The structured prompt follows:

> *You are a traffic reporter. Generate a concise report based on the following data.*
>
> *Example 1: (Structured Data) → (Formatted Report)*
> *Example 2: (Structured Data) → (Formatted Report)*
> *Now generate a report for: (Structured Data)*

Additionally, we will explore step-by-step prompting, where the task is explicitly broken down into sequential steps, and chain-of-thought prompting, where the model is encouraged to reason through its output before generating a response.

Dataset

The dataset for fine-tuning our selected model is provided by RTV Slovenija and consists of structured traffic data paired with human-written reports. It includes a table with approximately 50 thousand entries, each describing real-time traffic incidents through 27 attributes such as affected roads, incident type, severity, and timestamps. Additionally, it contains around 28 thousand manually composed textual reports, which summarize traffic conditions in a concise and informative manner for radio broadcasting.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Xiangpeng Wan, Michael 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, 01 2020.

[4] Jinhui Ouyang, Yijie Zhu, Xiang Yuan, and Di Wu. Trafficgpt: Towards multi-scale traffic analysis and generation with spatial-temporal agent framework, 2024.

[5] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with entity modeling, 2019.

[6] Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures, 2019.

[7] Kancharla Aditya Hari, Manish Gupta, and Vasudeva Varma. Curriculum learning for cross-lingual data-to-text generation with noisy data, 2024.

[8] Xuanyi Zhang, Genghong Zhao, Yi Ren, Weiguang Wang, Wei Cai, Yan Zhao, Xia Zhang, and Jiren Liu. Data augmented large language models for medical record generation. *Applied Intelligence*, 55(2):1–22, 2025.

[9] Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, Iztok Lebar Bajec, Timotej Petrič, and Marko Robnik-Šikonja. GaMS-1B-Chat: Generative model for slovene. https://huggingface.co/cjvt/GaMS-1B-Chat, 2024. University of Ljubljana and XLAB. Instruction-tuned version of GaMS developed within the PoVeJMo research program.

[10] Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. Generative model for less-resourced language with 1 billion parameters. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia, 2024. Znanstvena založba Filozofske fakultete, Univerza v Ljubljani.