Large Language Models for the Summarization of Czech Documents: From History to the Present

Václav Tran[1][0009−0003−0250−2821], Jakub Šmíd[1,2][0000−0002−4492−5481], Ladislav Lenc[1,2][0000−0002−1066−7269], Jean-Pierre Salmon[3][0009−0002−0049−7726] and Pavel Král[1,2][0000−0002−3096−675X]

[1] Department of Computer Science and Engineering, University of West Bohemia in Pilsen, Univerzitní, Pilsen, Czech Republic
[2] NTIS - New Technologies for the Information Society, University of West Bohemia in Pilsen, Univerzitní, Pilsen, Czech Republic
[3] LaBRI, UMR 5800, F-33400, Univ. de Bordeaux
https://nlp.kiv.zcu.cz/
{jaksmid,llenc,pkral}@kiv.zcu.cz,jean-pierre.salmon@u-bordeaux.fr

Abstract. Text summarization is the task of automatically condensing longer texts into shorter, coherent summaries while preserving the original meaning and key information. Although this task has been extensively studied in English and other high-resource languages, Czech summarization, particularly in the context of historical documents, remains underexplored. This is largely due to the inherent linguistic complexity of Czech and the lack of high-quality annotated datasets.

In this work, we address this gap by leveraging the capabilities of Large Language Models (LLMs), specifically Mistral and mT5, which have demonstrated strong performance across a wide range of natural language processing tasks and multilingual settings. In addition, we also propose a translation-based approach that first translates Czech texts into English, summarizes them using an English-language model, and then translates the summaries back into Czech. Our study makes the following main contributions: We demonstrate that LLMs achieve new state-of-the-art results on the SumeCzech dataset, a benchmark for modern Czech text summarization, showing the effectiveness of multilingual LLMs even for morphologically rich, medium-resource languages like Czech. We introduce a new dataset, Posel od Čerchova, designed for the summarization of historical Czech texts. This dataset is derived from digitized 19th-century publications and annotated for abstractive summarization. We provide initial baselines using modern LLMs to facilitate further research in this underrepresented area.

By combining cutting-edge models with both modern and historical Czech datasets, our work lays the foundation for further progress in Czech summarization and contributes valuable resources for future research in Czech historical document processing and low-resource summarization more broadly.

Keywords: Czech Text Summarization · Deep Neural Networks · Mistral · mT5 · Posel od Čerchova · SumeCzech · Transformer Models

1 Introduction

The rapid advancement of Natural Language Processing (NLP) techniques has substantially improved the capabilities of automated text summarization systems. However, the majority of these developments have been concentrated on high-resource languages such as English, leaving other languages, especially Czech, comparatively underexplored. This disparity becomes even more apparent in the case of historical Czech, which poses unique challenges due to linguistic shifts, archaic vocabulary, and inconsistent syntactic structures.

Effectively summarizing historical Czech texts is not only a technical challenge, but also a task with significant practical value. Many such documents contain culturally and historically important information, often hidden in long and complex narratives. Automated summarization of these texts can greatly support digital humanities, historical research, archival work, and the accessibility of heritage content to broader audiences. By enabling faster navigation and comprehension of large volumes of historical material, summarization tools can accelerate academic workflows and promote the reuse of digitized content in education, journalism, and public history initiatives.

These factors highlight a critical need for the development of robust summarization systems tailored specifically to the domain of historical Czech documents.

To address the current limitations in Czech text summarization, this paper tackles two interrelated challenges. First, we aim to establish new state-of-the-art benchmarks on SumeCzech, the most comprehensive dataset available for modern Czech summarization. We leverage the capabilities of advanced Large Language Models (LLMs), specifically Mistral [14] and mT5 [40], to evaluate their effectiveness in handling the linguistic richness and morphological complexity of contemporary Czech. Moreover, we propose a translation-based approach that first translates Czech texts into English, summarizes them using an English-language model, and then translates the summaries back into Czech.

Then, in light of the evident lack of resources tailored to historical Czech, we introduce a novel dataset constructed from the 19th-century journal Posel od Čerchova. This corpus is carefully curated to support summarization tasks in historical domains, where the language exhibits substantial diachronic variation, including archaic vocabulary and evolving syntactic patterns. The dataset is intended to provide a solid foundation for future research into the summarization of historical Czech texts, where domain-specific challenges differ significantly from those in modern language processing. The full corpus is freely available for research purposes[4], promoting openness and reusability.

Note that this paper presents an extended version of [33], including additional experiments, a more detailed methodological description, and a broader discussion of the results and their implications.

By combining methodological advances in LLMs with the creation of a new domain-specific dataset, our work contributes not only to the technical advance-

[4] https://corpora.kiv.zcu.cz/posel_od_cerchova/

ment of Czech summarization systems but also to broader applications in cultural heritage preservation, historical scholarship, and the digital humanities. We believe these contributions will help bridge the gap between modern NLP methods and the demands of underrepresented, historically valuable language resources.

2 Related Work

Text summarization methods can be categorized into abstractive and extractive ones. Extractive summarization selects the most representative sentences from the source document, while abstractive summarization generates summaries composed of newly created sentences.

Early summarization methods were extractive ones and relied on statistical and graph-based methods like TF-IDF (Term Frequency-Inverse Document Frequency) [4], which scores sentence importance based on term frequency relative to rarity across a corpus. Similarly, TextRank [23] represents sentences as nodes in a graph and ranks them using the PageRank algorithm [28].

Neural networks advanced both extractive and also abstractive summarization by modeling sequences with Recurrent Neural Networks (RNNs) [8]. One extractive approach involves sequence-to-sequence architectures where LSTM models capture the contextual importance of each sentence within a document [25]. Hierarchical attention networks combine sentence-level and word-level attention to better capture document structure and relevance for summarization [41]. This approach has proven effective in summarizing longer and more complex documents. Hybrid approaches combining BERT embeddings [7] with K-Means clustering [21] to identify key sentences [24] have shown excellent performance for abstractive summarization.

Advances in sequence-to-sequence Transformer-based models [35] have revolutionized abstractive summarization. Recent models like T5 [29] adopt a text-to-text framework and excel in various tasks, including summarization, due to pre-training on the C4 dataset. PEGASUS [42] introduces gap sentences generation for masking key sentences during pre-training, achieving strong performance on 12 datasets. Similarly, BART [18] uses denoising objectives for robust text summary generation. Multilingual models such as mT5 [40] and mBART [20] extend these capabilities to multiple languages, including Czech, through datasets like mC4 [39] and multilingual Common Crawl[5].

However, these models often underperform on non-English corpora without fine-tuning.

3 Datasets

The following section provides a brief review of the primary existing summarization datasets. Moreover, the created Posel od Čerchova corpus will also be detailed at the end of this section.

[5] http://commoncrawl.org/

3.1 English Datasets

CNN/Daily Mail dataset [12] comprises more than 300,000 English-language news articles, each accompanied by highlights written by the article authors. This extensive dataset has become a cornerstone in the field of natural language processing, particularly in research related to abstractive summarization and question-answering systems. Over time, it has evolved through multiple versions, with adaptations designed to address specific NLP tasks and to improve the quality and diversity of the training data.

XSum dataset [26] consists of 226,000 BBC articles paired with concise, single-sentence summaries. These articles span a wide range of domains, including news, sports, and science, providing a diverse and comprehensive dataset for summarization tasks. By emphasizing one-sentence summaries, XSum is specifically designed to challenge and promote the development of abstractive summarization models, which generate concise, human-like summaries rather than relying heavily on extractive methods.

ArXiv Dataset [5] comprises 215,000 pairs of scientific papers and their corresponding abstracts, all sourced from the arXiv preprint repository. This dataset has undergone extensive cleaning and standardization to facilitate research and analysis, which includes the removal of non-essential sections such as figures, tables, and references to focus solely on the textual content.

BOOKSUM [16] is a dataset tailored for summarizing long texts like novels, plays, and stories, with summaries provided at paragraph, chapter, and book levels. Texts and summaries were sourced from Project Gutenberg and other web archives, supporting both extractive and abstractive summarization.

3.2 Multilingual Datasets

XLSum [11] provides over one million article-summary pairs across 44 languages, ranging from low-resource languages like Bengali and Swahili to high-resource languages such as English and Russian. Extracted from various BBC sites, this dataset is a valuable resource for multilingual summarization research.

MLSUM [30] consists of 1.5 million article-summary pairs in five languages: German, Russian, French, Spanish, and Turkish. The dataset was created by archiving news articles from well-known newspapers, including Le Monde and El Pais, with a focus on ensuring broad topic coverage.

While the above datasets represent significant progress in summarization research, especially for English and several other widely spoken languages, the availability of comparable resources for Czech remains extremely limited. In particular, Czech lacks large-scale, high-quality datasets for both modern and historical domains. This scarcity significantly hinders the development and evaluation of summarization models tailored to Czech, and motivates the need for

focused dataset creation and model adaptation in this underrepresented linguistic setting.

3.3 SumeCzech

SumeCzech is a large-scale dataset for Czech summarization [31] and represents a notable exception to the general scarcity of Czech-specific NLP resources. Created by the Institute of Formal and Applied Linguistics at Charles University, it is specifically designed to support summarization tasks in the Czech language.

The dataset consists of approximately one million news articles collected from five major Czech news outlets: České Noviny, Deník, iDNES, Lidovky, and Novinky.cz. Each entry is stored in JSONLines format and includes rich metadata such as the article URL, headline, abstract, full text, subdomain, section, and publication date. The creators applied preprocessing steps including language detection, duplicate removal, and filtering out records with empty or excessively short fields to ensure data quality.

SumeCzech is well-suited for multiple summarization tasks, including both headline generation and multi-sentence abstract generation, which makes it a versatile resource for training and evaluating models. The dataset is split into training, development, and test sets in roughly an 86.5/4.5/4.5 ratio. On average, full texts contain around 409 words, while the associated abstracts average 38 words, providing a reasonable length ratio for abstractive summarization.

Despite its scale and utility, SumeCzech is focused entirely on modern Czech language and content, limiting its applicability to historical text processing, which presents different linguistic characteristics such as archaic spelling, grammar, and vocabulary. This gap underscores the need for dedicated datasets targeting historical Czech, especially for tasks like summarization that rely on linguistic fluency and domain-specific adaptation.

3.4 Posel od Čerchova

To construct the dataset, we used data from the historical journal *Posel od Čerchova (POC)*, which is available on the archival portal Porta fontium[6].

The construction of the dataset involved addressing the challenge of creating summaries for the provided texts, which were composed in historical Czech and, in some rare cases, even German. The texts also covered a variety of different topics, from local news surrounding Domažlice (a historic town in the Czech Republic), opinion pieces, and various local advertisements to internal and worldwide politics and feuilletons. Furthermore, it was important to create a dataset of sufficient size to ensure the accuracy and reliability of the evaluation. These aspects added complexity to the summarization task.

To overcome the mentioned issues, we employed state-of-the-art (SOTA) LLMs GPT-4 [27] and Claude 3 Opus [2] (Opus) (specifically the `claude-3-opus-20240229` version) for initial text summary creation. These models were selected

[6] https://www.portafontium.eu

based on their SOTA performance in many NLP tasks and excellent performance in some preliminary summarization experiments.

While generating the summaries, it was essential to ensure conciseness. Since most of the implemented methods were fine-tuned on the SumeCzech dataset, we aimed to maintain consistency by creating summaries in a journalistic style, reflecting the dataset's characteristics. To achieve this, the prompts for generating the summaries included explicit instructions, as shown below:

- Vytvoř shrnutí následujícího textu ve stylu novináře. Počet vět <= 5; (EN: Create a summary of the following text in the style of a journalist. Number of sentences <= 5)

During the summarization task, we observed that while both models produced summaries of very good quality, Opus tended to create more succinct and stylistically appropriate ones, closely aligning with the news reporter format. However, there were instances where summaries generated by Opus exhibited an excessive focus on a single topic.

On the other hand, GPT-4 aimed to incorporate a greater level of detail within the five-sentence constraint but occasionally deviated from the specified stylistic prompt.

If the model-generated summary exhibited significant stylistic deviations or excessive focus on a single topic, we either modified or regenerated it until a correct version was achieved.

Two levels of summaries were created: the first at the page level, and the second summarizing an entire article, which typically consists of several pages. In total, we summarized 432 pages, which effectively resulted in the creation of 100 article-level summaries. The subset containing the page-level summaries is hereafter referred to as *POC-P*, while the article-level summaries are referred to as *POC-I*. Note that all created summaries were checked and corrected manually by two native Czech speakers.

The dataset is in the **.json** format and contains the following information:

- **text:** Text extracted from the given page, a digital rendition of the original printed content;
- **summary:** Summary of the page, which is no more than 5 sentences long;
- **year:** Publication year of the journal;
- **journal:** Specification of the source journal: the day, month, and the number of the issue is contained within this identifier;
- **page_src:** Name of the source image file converted into the text;
- **page_num:** Page number.

This dataset is designed to support summarization tasks within Czech historical contexts, providing researchers with the tools to tackle the linguistic challenges unique to this domain. The corpus is freely available for research purposes[7].

[7] https://corpora.kiv.zcu.cz/posel_od_cerchova/

4 Methods

The experiments employ two advanced Transformer-based models, Multilingual Text-to-Text Transfer Transformer (mT5) [40] and Mistral 7B [14].

Moreover, we propose an alternative approach to abstractive summarization that translates the Czech text into English, summarizes it using an English model, and then translates it back, hereafter referred to as the *Translation-Summarization-Translation (TST)* method.

4.1 Multilingual Text-to-Text Transfer Transformer

The Multilingual Text-to-Text Transfer Transformer (mT5) is a variant of the T5 model designed for multilingual tasks. This model is trained on the multilingual mC4 dataset [39], which includes Czech, and effectively handles a wide range of languages. The model is based on Transformer encoder-decoder architecture and uses a SentencePiece tokenizer [17] to process complex language structures, including Czech morphology. Pre-trained using a span corruption objective [29], mT5 predicts masked spans of text, enabling it to learn semantic and contextual relationships.

The mT5 model is available in various sizes, from small with 300 million parameters to XXL with 13 billion parameters, and is therefore adapted to different computational needs. The base variant of the mT5, which contains 580 million parameters, is used for further experiments.

4.2 Mistral Language Model

The Mistral Language Model (Mistral LM) is a highly efficient large language model known for its robust performance across diverse natural language processing tasks. It is designed to combine high accuracy with computational efficiency, achieving state-of-the-art results in reasoning, text generation, summarization, and other NLP applications. Mistral 7B, with its 7 billion parameters, strikes a balance between computational efficiency and task performance, surpassing larger models like 13B or 34B in several benchmarks.

This model utilizes advanced attention mechanisms like Grouped-Query Attention (GQA) [1] and Sliding Window Attention (SWA) [3]. GQA enhances processing speed by grouping attention heads to focus on the same input data, while SWA reduces computational costs by limiting token attention to nearby tokens. The model supports techniques such as quantization [9] and Low-Rank Adaptation (LoRA) [13] for efficient fine-tuning on limited hardware, enabling it to handle longer inputs effectively.

4.3 Translation-Summarization-Translation (TST) Method

As already stated, this approach involves translating the Czech text into English, summarizing it using an English-language model, and then translating the

summary back into Czech. For translation, we use ALMA-R model, introduced by Xu et al. [38][8] which exceeds the performance of GPT-4 [27] and WMT[9], winners on various translation benchmarks.

The model was trained using Contrastive Preference Optimization (CPO), a novel training method that was proposed in the paper above. Traditional methods such as Supervised Fine-Tuning (SFT) [10] train models to mimic reference translations, where the resulting model performance relies on dataset quality. Xu et al. demonstrate that translations by advanced models can be superior to reference translations. Its training objective is designed not to minimize the error between the model output and a single reference translation, but rather to increase the likelihood of generating a preferred translation and decrease the likelihood of generating a dispreferred one. This involves generating a triplet of translations for a given source sentence: one from a reference (human-generated), one from GPT-4, and one from an ALMA [37] model (prior to CPO application). Each translation in the triplet is then scored using reference-free translation quality evaluation models and the translations are ranked based on their quality. The highest-scoring translation is labeled as the preferred translation.

5 Experiments

5.1 Evaluation Metrics

The following evaluation metrics are used.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [19] is a set of metrics used to evaluate the quality of summaries by comparing n-gram overlaps between a system-generated summary and reference texts. Key ROUGE metrics include ROUGE-N (for n-gram overlap) and ROUGE-L (for the longest common subsequence).

ROUGERAW [32] is a variant of ROUGE that evaluates raw token-level overlaps between predicted and reference texts without any preprocessing like stemming or lemmatization. It measures exact matches of tokens, making it suitable for tasks where precise token alignment is important.

5.2 Experimental Setup

We used AdamW optimizer [22] with a learning rate set to 0.001 as suggested by authors of mT5 [40] for the training of this model. For Mistral 7B, we utilized QLoRA [6], a method that integrates a 4-bit quantized model with a small, newly introduced set of learnable parameters. During fine-tuning, only these additional parameters are updated while the original model remains frozen, thereby substantially reducing memory requirements.

We used the models from the HuggingFace Transformers library [36]. For training both models, we used a single NVIDIA A40 GPU with 45 GB VRAM.

[8] https://github.com/fe1ixxu/ALMA
[9] https://machinetranslate.org/wmt

5.3 Model Variants

We use the following model variants in our experiments:

- M7B-SC: The Mistral 7B model fine-tuned on the SumeCzech dataset;
- M7B-POC: The Mistral 7B model further fine-tuned on the POC dataset;
- mT5-SC: The mT5 model fine-tuned on the SumeCzech dataset;
- TST: used ALMA-R 13B model for translation, English summarization was done using the 4-bit pre-quantized, fine-tuned variant of Mistral 7B[10].

5.4 Results on the SumeCzech Dataset

This experiment compares the results of the proposed mT5-SC and M7B-SC models with related work on the SumeCzech dataset, see Table 1.

The first comparative method, HT2A-S [15], is based on the mBART model, which is further fine-tuned on the SumeCzech dataset. The other methods provided by the authors of the SumeCzech dataset [31] are as follows: First, Random, Textrank and Tensor2Tensor [34].

Table 1 demonstrates that the proposed M7B-SC method is very efficient, outperforming all other baselines and achieving new state-of-the-art results on this dataset. Furthermore, the second proposed approach, mT5-SC, also performs remarkably well, consistently obtaining the second-best results.

Table 1. Results of various methods on SumeCzech dataset with precision (P), recall (R), and F1-score (F) [33].

Method	$ROUGE_{raw}$-1			$ROUGE_{raw}$-2			$ROUGE_{raw}$-L		
	P	R	F	P	R	F	P	R	F
M7B-SC	**24.4**	**19.7**	**21.2**	**6.5**	**5.3**	**5.7**	**17.8**	**14.5**	**15.5**
mT5-SC	22.0	17.9	19.2	5.3	4.3	4.6	16.1	13.2	14.1
HT2A-S [15]	22.9	16.0	18.2	5.7	4.0	4.6	16.9	11.9	13.5
First [31]	13.1	17.9	14.4	0.1	9.8	0.2	1.1	8.8	0.9
Random [31]	11.7	15.5	12.7	0.1	2.0	0.1	0.7	10.3	0.8
Textrank [31]	11.1	20.8	13.8	0.1	6.0	0.3	0.7	13.4	0.8
Tensor2Tensor [31]	13.2	10.5	11.3	0.1	2.0	0.1	0.2	8.1	0.8

5.5 Results on Posel od Čerchova Dataset

This section evaluates the proposed methods on the Posel od Čerchova dataset. Table 2 shows the results on the *POC-P* subset containing summaries for every page (106 pages), while Table 3 depicts the results on the *POC-I* subset, which is composed of the summaries of every article (25 issues).

[10] https://huggingface.co/unsloth/mistral-7b-instruct-v0.2-bnb-4bit

These tables show that, as in the previous case, M7T-POC model gives significantly better results than the mT5-SC model, and it is with a very high margin. Additionally, the TST approach slightly outperformed the M7T-POC model in some cases.

Table 2. Results of implemented methods on the *POC-P* subset from Posel od Čerchova dataset with precision (P), recall (R), and F1-score (F).

Method	$ROUGE_{raw}$-1			$ROUGE_{raw}$-2			$ROUGE_{raw}$-L		
	P	R	F	P	R	F	P	R	F
M7B-POC	**23.5**	17.4	19.6	**4.8**	3.5	**4.0**	**16.6**	12.2	**13.8**
mT5-SC	20.2	8.2	11.1	1.4	0.5	0.7	14.9	6.1	8.2
TST	17.2	**25.1**	**19.9**	2.5	**3.8**	2.9	11.3	**16.4**	13.0

Table 3. Results of implemented methods on *POC-I* subset from Posel od Čerchova dataset with precision (P), recall (R), and F1-score (F).

Method	$ROUGE_{raw}$-1			$ROUGE_{raw}$-2			$ROUGE_{raw}$-L		
	P	R	F	P	R	F	P	R	F
M7B-POC	**19.3**	17.6	**18.0**	**3.2**	2.8	**2.9**	13.7	12.4	**12.8**
mT5-SC	18.2	5.9	8.6	1.0	0.3	0.4	**14.0**	4.5	6.5
TST	14.0	**24.8**	17.5	1.7	**3.1**	2.1	9.1	**16.3**	11.4

6 Discussion

We provide two examples of generated summaries. The first one, depicted in Table 4, presents page-level summaries from the 38th issue of Posel od Čerchova (1882). The second example, shown in Table 5, contains issue-level summaries from the 52nd issue of the same publication and year.

The summaries in Table 4 generally convey similar information as the reference summary to a certain extent. However, the mT5-SC summary is just a composition of the first two sentences of the original text, which is not shown here due to its length. Both the M7B-POC and TST summaries contain grammatical errors. While the M7B-POC summary is slightly more coherent and better structured, it also includes more factual inaccuracies, as evident in its concluding sentence, where it mistakenly mentions the city of Prague as one of the donors and attributes the Weihburggasse street to the wrong village.

The following example in Table 5 shows comparable results. The mT5-SC summary was irrelevant. The M7B-POC summary has slightly better coherence

and grammar in comparison to TST summary, however, it misinforms by not stating that the death of "sultánka" is a fictional event.

In our limited observations of these generated summaries and many other generated summaries, M7B-POC tended to create slightly more coherent summaries with better grammar, but TST produced less factual errors than M7B-POC.

Table 4. Summaries of 'Posel od Čerchova,' page 5 of the 38th issue, 1882

Reference
V Petrohradě došlo k otravě osmi běloušů, kteří měli táhnout korunovační vůz cara Alexandra III. při jeho korunovaci v Moskvě. Koně byli nalezeni mrtví ve stáji carského paláce. Obce Milaveč a Radonice v okresu Domažlice přispěly na znovuzřízení Národního divadla, čímž mohou jít příkladem ostatnímu venkovu. Článek dále uvádí tržní ceny různých produktů v Domažlicích a Praze.
M7B-POC
V Petrohradě byli otráveni osm carští bělouši, kteří měli táhnouti korunovační vozku pro cara Alexandra III. v Moskvě. V Domažlicích byly zaznamenány tržní ceny potravin a dalších zboží, včetně cen za talíře a sklenice. V Praze a Weihburggasse 26 v Milaveči a Radonicích I. byly oznámeny příspěvky na obnovu Národního divadla.
mT5-SC
Carští koně otráveni. Oněch osm běloušů, kteří měli korunní vůz při korunovací carovi Alexandra III. v Moskvě táhnouti, nalezli služebníci ve stáji paláce carského v Petrohradě zdatná ležeti Dla -x-1 byly snad jediné obce v okresu Domažlice.
TST
V Petrohradě byly v koníčkách palácových stájí nalezeny osm bílých koní, kteří zemřeli na otravu, určených pro korunovaci cara Alexandra III. V českém okrese Domažlice se loni dvě obce podílely na obnově Národního divadla a další mohou následovat jejich příklad. V následujícím textu jsou uvedeny ceny různých položek v Praze: Neffův obchod na Staroměstském náměstí za 39 J. Milaveč a Radonice I. v ulici Weihburggasse 26 vybraly za měsíc 10 zl. díky úsilí několika odvážných občanů.

7 Conclusions

This paper investigates the use of state-of-the-art large language models—specifically Mistral 7B and mT5—for Czech text summarization, addressing both contemporary and historical domains. In addition to direct summarization, we propose a novel translation-based pipeline, which first translates Czech input texts into English, performs summarization using an English-language model, and subsequently translates the summaries back into Czech. This method leverages the strength of English-language models in low-resource scenarios.

Table 5. Summaries of 'Posel od Čerchova,' 52nd issue, 1882

Reference
Domažlický týdeník "Posel od Čerchova" vstupuje do 11. ročníku svého vydávání s čtvrtletním předplatným variujícím dle způsobu doručení. V aktuálním vydání týdeník pokrývá široké spektrum témat, od politických událostí v Bosně, Tyrolsku, až po tragické incidenty v Srbsku a soudní případy v Paříži. Součástí obsahu je také pokračování povídky "Ze života sultánky", odhalení pomníku knížete Michajla v Bělehradě, a záměr francouzského ministra války rozšířit armádu. Zajímavostí jsou novinky na trhu, jako Hellerovy hrací stroje, a opatření pražské městské rady proti novoroční žebrotnické praxi mezi městskými zaměstnanci.
M7B-POC
Posel od Čerchova informuje o politických událostech, jako jsou předlohy o výpomoci v Tyrolsku a Korutanech, rakousko-německém napětí a ruské vojenské aktivitě. Dále se diskutuje o stavebních pracích v Tyrolsku a Korutanech, financovaných zemskou zárukou. Zmíněna je nová předloha o založení samostatného církevního systému v Sarajevu, umožňující větší autonomii muslimské komunity. Dále je zmíněna popisná smrt sultánky ze Cařihradu, jejího vyhnanství po milostném vztahu s bratrem, a slavnostní odhalení pomníku knížeti Miloše Obrenoviče v Bělehradě. Dále jsou zmíněny kulturní události, jako je nová výstava uměleckých prací v Plzni, a společenské události, jako je nová výzva do sněmu v Bělehradě, a trestní události, jako jsou vraždy a zatčení v několika městech.
mT5-SC
Vánoční svátky jsou v plném proudu a s nimi i přípravy na vánoční svátky. Přinášíme vám plné znění projevu prezidenta republiky Václava Klause, který přednesl ve Vladislavském sále Pražského hradu.
TST
ČERCHOVSKÝ TÝDENÍK (POŠLA OD ČERCHOVA), noviny z Domažlic, pokrývají politická, obecná a komunitní témata v daném regionu. Ceny předplatného se liší. V Bosně byla založena náboženská organizace pro bosenskou muslimskou církev, zatímco ve městě Krivošiji byly z důvodu počasí pozastaveny stavební práce. V Srbsku zůstává Královský sněm krále Milana prázdný a situace je po volbách, jejichž cílem bylo nahradit radikální poslance, klidná. V Egyptě jsou odsouzení vůdci arabského povstání veřejně zesměšňováni a poté vysídleni, v České republice došlo k fatálnímu střetu mezi sousedy. Probíhají různé akce, například oslavy jmenování pana Justina V. Práška na funkci ředitele, jmenování paní Karol Fremuthové do funkce a hudební zkoušky pro taneční zábavy 26. prosince. Dále se zde koná loterie ve prospěch organizace „Matica“ a je zveřejněna kniha „Humory“. Společnost Singer nabízí originální šicí stroje jako ideální vánoční dárek.

Our experimental results demonstrate that the proposed Mistral-based model (M7B-SC) sets a new performance benchmark on the SumeCzech dataset, achieving state-of-the-art results. The mT5-based model (mT5-SC) also delivered consistently strong performance, ranking second across most metrics. Moreover, the translation-based approach (TST) proved to be highly competitive, further validating its potential as a viable strategy for Czech summarization tasks.

To support historical text summarization, we introduced Posel od Čerchova, a newly curated dataset of historical Czech documents. Alongside this resource, we provide baseline results and discuss the linguistic and structural challenges that distinguish historical Czech from modern usage.

These contributions not only advance the field of Czech Together, these contributions significantly advance Czech-language summarization research. They also establish a foundation for broader investigations in the processing of historical texts, with practical applications in cultural heritage preservation and digital humanities. Future work may explore improving model robustness, developing hybrid architectures, and expanding the dataset to enable multilingual and diachronic studies.

Acknowledgments. This work has been supported by the Grant No. SGS-2025-022 – New Data Processing Methods in Current Areas of Computer Science. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., Sanghai, S.: Gqa: Training generalized multi-query transformer models from multi-head checkpoints (2023)
2. Anthropic: The Claude 3 Model Family: Opus, Sonnet, Haiku (3 2024), https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
3. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer (2020)
4. Christian, H., Agus, M., Suhartono, D.: Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). ComTech: Computer, Mathematics and Engineering Applications **7**, 285 (12 2016). https://doi.org/10.21512/comtech.v7i4.3746
5. Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N.: A discourse-aware attention model for abstractive summarization of long documents. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 615–621. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2097, https://aclanthology.org/N18-2097
6. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: efficient finetuning of quantized llms. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2024)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
8. Elman, J.L.: Finding structure in time. Cognitive Science **14**(2), 179–211 (1990)

9. Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W., Keutzer, K.: A survey of quantization methods for efficient neural network inference (2021)
10. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
11. Hasan, T., et al.: Xlsum: A multilingual dataset for summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 2133–2149 (2021)
12. Hermann, K.M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 1693–1701 (2015), https://arxiv.org/abs/1506.03340
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
14. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
15. Krotil, M.: Text Summarization Methods in Czech. Bachelor's thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Cybernetics (5 2022)
16. Kryscinski, W., Rajani, N., Agarwal, D., Xiong, C., Radev, D.: BOOKSUM: A collection of datasets for long-form narrative summarization. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 6536–6558. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.findings-emnlp.488, https://aclanthology.org/2022.findings-emnlp.488
17. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Blanco, E., Lu, W. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012, https://aclanthology.org/D18-2012
18. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019)
19. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
20. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020). https://doi.org/10.1162/tacl_a_00343, https://aclanthology.org/2020.tacl-1.47
21. Lloyd, S.: Least squares quantization in pcm. IEEE Transactions on Information Theory **28**(2), 129–137 (1982). https://doi.org/10.1109/TIT.1982.1056489
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Mihalcea, R., Tarau, P.: TextRank: Bringing order into text. In: Lin, D., Wu, D. (eds.) Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 404–411. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-3252

24. Miller, D.: Leveraging bert for extractive text summarization on lectures (2019)
25. Nallapati, R., Zhai, F., Zhou, B.: Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI). pp. 3075–3081 (2017)
26. Narayan, S., Cohen, S.B., Lapata, M.: Extreme summarization (xsum). In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 931–936 (2018)
27. OpenAI: Gpt-4 technical report (2024)
28. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking : Bringing order to the web. In: The Web Conference (1999), https://api.semanticscholar.org/CorpusID:1508503
29. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
30. Scialom, T., et al.: Mlsum: Multilingual summarization dataset. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 2146–2161 (2020)
31. Straka, M., Mediankin, N., Kocmi, T., Žabokrtský, Z., Hudeček, V., Hajič, J.: SumeCzech: Large Czech news-based summarization dataset. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://www.aclweb.org/anthology/L18-1551
32. Straka, M., Straková, J.: Rougeraw: Language-agnostic evaluation for summarization. Proceedings of the International Conference on Computational Linguistics (2018), https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2615
33. Tran, V., Šíd, J., Martínek, J., Lenc, L., Král, P.: Large language models for summarizing czech historical documents and beyond. In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART. pp. 798–804. INSTICC, SciTePress (2025). https://doi.org/10.5220/0013374100003890
34. Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A.N., Gouws, S., Jones, L., Kaiser, L., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., Uszkoreit, J.: Tensor2tensor for neural machine translation. CoRR **abs/1803.07416** (2018), http://arxiv.org/abs/1803.07416
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
36. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6
37. Xu, H., Kim, Y.J., Sharaf, A., Awadalla, H.H.: A paradigm shift in machine translation: Boosting translation performance of large language models (2024)
38. Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Durme, B.V., Murray, K., Kim, Y.J.: Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation (2024)

39. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mC4: A massively multilingual cleaned crawl corpus. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7517–7532. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.593, https://aclanthology.org/2021.emnlp-main.593
40. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A massively multilingual pre-trained text-to-text transformer. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 483–498. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.41, https://aclanthology.org/2021.naacl-main.41
41. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
42. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization (2019)