Large Language Models for Summarizing Czech Historical Documents and Beyond

Václav Tran[1][a], Jakub Šmíd[1,2][b], Jiří Martínek[1,2][c], Ladislav Lenc[1,2][d] and Pavel Král[1,2][e]

[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*

nuva@students.zcu.cz, {jaksmid, llenc, jimar, pkral}@kiv.zcu.cz

Abstract: Text summarization is the task of shortening a larger body of text into a concise version while retaining its essential meaning and key information. While summarization has been significantly explored in English and other high-resource languages, Czech text summarization, particularly for historical documents, remains underexplored due to linguistic complexities and a scarcity of annotated datasets. Large language models such as Mistral and mT5 have demonstrated excellent results on many natural language processing tasks and languages. Therefore, we employ these models for Czech summarization, resulting in two key contributions: (1) achieving new state-of-the-art results on the modern Czech summarization dataset SumeCzech using these advanced models, and (2) introducing a novel dataset called Posel od Čerchova for summarization of historical Czech documents with baseline results. Together, these contributions provide a great potential for advancing Czech text summarization and open new avenues for research in Czech historical text processing.

1 INTRODUCTION

The rapid evolution of Natural Language Processing (NLP) techniques has elevated the performance of text summarization systems. While most advances focus on high-resource languages like English, the Czech language, particularly historical variations, remains underrepresented. Historical Czech documents pose unique challenges due to linguistic shifts, outdated vocabulary, and inconsistent syntax. These nuances create a significant gap in the development of automated summarization systems capable of handling this domain effectively.

Therefore, this paper addresses two interlinked challenges. First, it seeks to establish new state-of-the-art benchmarks on SumeCzech, the most comprehensive dataset for modern Czech text summarization using modern Large Language Models (LLMs),

namely Mistral (Jiang et al., 2023) and mT5 (Xue et al., 2021b). Second, recognizing the lack of resources tailored for historical Czech, we introduce a newly created dataset derived from the historical journal Posel od Čerchova. The dataset is specifically designed to facilitate summarization tasks in historical contexts, enabling future researchers to address the linguistic complexities inherent in this domain. This corpus is freely available for research purposes[1].

By combining model advancements and dataset innovation, this research aims to drive progress in the Czech summarization field and open venues for applications in cultural preservation, historical research, and digital humanities.

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 summa-

---

[a] https://orcid.org/0009-0003-0250-2821
[b] https://orcid.org/0000-0002-4492-5481
[c] https://orcid.org/0000-0003-2981-1723
[d] https://orcid.org/0000-0002-1066-7269
[e] https://orcid.org/0000-0002-3096-675X

---

[1] corpora.kiv.zcu.cz/posel_od_cerchova/

rization 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) (Christian et al., 2016), which scores sentence importance based on term frequency relative to rarity across a corpus. Similarly, TextRank (Mihalcea and Tarau, 2004) represents sentences as nodes in a graph and ranks them using the PageRank algorithm (Page et al., 1999).

Neural networks advanced both extractive and also abstractive summarization by modeling sequences with Recurrent Neural Networks (RNNs) (Elman, 1990). One extractive approach involves sequence-to-sequence architectures where LSTM models capture the contextual importance of each sentence within a document (Nallapati et al., 2017). Hierarchical attention networks combine sentence-level and word-level attention to better capture document structure and relevance for summarization (Yang et al., 2016). This approach has proven effective in summarizing longer and more complex documents. Hybrid approaches combining BERT embeddings (Devlin et al., 2019) with K-Means clustering (Lloyd, 1982) to identify key sentences (Miller, 2019) have shown excellent performance for abstractive summarization.

Advances in sequence-to-sequence Transformer-based models (Vaswani et al., 2017) have revolutionized abstractive summarization. Recent models like T5 (Raffel et al., 2020a) adopt a text-to-text framework and excel in various tasks, including summarization, due to pre-training on the C4 dataset. PEGASUS (Zhang et al., 2019) introduces gap sentences generation for masking key sentences during pre-training, achieving strong performance on 12 datasets. Similarly, BART (Lewis et al., 2019) uses denoising objectives for robust text summary generation. Multilingual models such as mT5 (Xue et al., 2021b) and mBART (Liu et al., 2020) extend these capabilities to multiple languages, including Czech, through datasets like mC4 (Xue et al., 2021a) and multilingual Common Crawl[2].

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.

3.1 English Datasets

CNN/Daily Mail (Hermann et al., 2015) dataset consists of over 300,000 English news articles, each paired with highlights written by the article authors. It has been widely used in summarization and question-answering tasks, evolving through several versions tailored for specific NLP tasks.

XSum (Narayan et al., 2018) contains 226,000 single-sentence summaries paired with BBC articles covering diverse domains such as news, sports, and science. Its focus on single-sentence summarization makes it less biased toward extractive methods.

Arxiv Dataset (Cohan et al., 2018) includes 215,000 pairs of scientific papers and their abstracts sourced from arXiv. It has been cleaned and formatted to ensure standardization, with sections like figures and tables removed.

BOOKSUM (Kryscinski et al., 2022) 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 (Hasan et al., 2021) 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 (Scialom et al., 2020) 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.

The above-mentioned datasets are for English summarization, and some are multilingual; however, Czech resources remain very limited.

3.3 SumeCzech

SumeCzech large-scale dataset (Straka et al., 2018) is a notable exception to the scarcity of Czech-specific resources. This dataset was created at the Institute

---

[2] http://commoncrawl.org/

of Formal and Applied Linguistics at Charles University and is tailored for summarization tasks in the Czech language. It comprises one million Czech news articles. These articles are sourced from five major Czech news sites: České Noviny, Deník, iDNES, Lidovky, and Novinky.cz. Each document is structured in JSONLines format, with fields for the URL, headline, abstract, text, subdomain, section, and publication date. The preprocessing includes language recognition, duplicate removal, and filtering out entries with empty or excessively short headlines, abstracts, or texts.

This dataset supports multiple summarization tasks, such as headline generation and multi-sentence abstract generation. The training, development, and testing splits are in roughly 86.5/4.5/4.5 ratio. The average word count is 409 for full texts and 38 for abstracts.

Nevertheless, this dataset caters exclusively to modern Czech and fails to address the needs of historical text processing.

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[3].

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 construct 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 (OpenAI, 2024) and Claude 3 Opus (Anthropic, 2024) (Opus) (specifically the `claude-3-opus-20240229` version) for initial text summary creation. These models were selected 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-level summaries were created; the first one was on the page level, and the second one summarizes a whole article that is usually composed of several pages. We thus summarized 432 pages, effectively resulting in the creation of 100 issue summaries. The subset containing page summaries is hereafter referred to as *POC-P*, while the issue 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 accessible for research purposes[4].

---

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

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

4 MODELS

The experiments employ two advanced Transformer-based models, Multilingual Text-to-Text Transfer Transformer (mT5) (Xue et al., 2021b) and Mistral 7B (Jiang et al., 2023).

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 (Xue et al., 2021a), which includes Czech, and effectively handles a wide range of languages. The model is based on Transformer encoder-decoder architecture and uses a Sentence-Piece tokenizer (Kudo and Richardson, 2018) to process complex language structures, including Czech morphology. Pre-trained using a span corruption objective (Raffel et al., 2020b), 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) (Ainslie et al., 2023) and Sliding Window Attention (SWA) (Beltagy et al., 2020). 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 (Gholami et al., 2021) and Low-Rank Adaptation (LoRA) (Hu et al., 2021) for efficient fine-tuning on limited hardware, enabling it to handle longer inputs effectively.

5 EXPERIMENTS

5.1 Evaluation Metrics

The following evaluation metrics are used.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) 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 (Straka and Straková, 2018) 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 Set-up

We used AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate set to 0.001 as suggested by authors of mT5 (Xue et al., 2021b) for the training of this model. For Mistral 7B, we utilized QLoRA (Dettmers et al., 2024), 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 employ the models from the HuggingFace Transformers library (Wolf et al., 2020). For training both models, we used a single NVIDIA A40 GPU with 45 GB VRAM.

5.3 Model Variants

We use three variants of the models 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.

Table 1: Results of various methods on SumeCzech 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-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 (Krotil, 2022)	22.9	16.0	18.2	5.7	4.0	4.6	16.9	11.9	13.5
First (Straka et al., 2018)	13.1	17.9	14.4	0.1	9.8	0.2	1.1	8.8	0.9
Random (Straka et al., 2018)	11.7	15.5	12.7	0.1	2.0	0.1	0.7	10.3	0.8
Textrank (Straka et al., 2018)	11.1	20.8	13.8	0.1	6.0	0.3	0.7	13.4	0.8
Tensor2Tensor (Straka et al., 2018)	13.2	10.5	11.3	0.1	2.0	0.1	0.2	8.1	0.8

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

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

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 (Krotil, 2022), 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 (Straka et al., 2018) are as follows: First, Random, Textrank and Tensor2Tensor (Vaswani et al., 2018).

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.

5.4.1 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).

These tables show clearly 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.

6 CONCLUSIONS

This paper explored the application of state-of-the-art large language models, specifically Mistral 7B and mT5, for summarization of Czech texts, addressing both modern and historical contexts. Our experiments demonstrated that the proposed M7B-SC model establishes a new benchmark for the SumeCzech dataset, achieving state-of-the-art performance, while the mT5-SC model also performed strongly, consistently ranking second.

Furthermore, we introduced a novel dataset, Posel od Čerchova, dedicated for the summarization of historical Czech documents. By leveraging this dataset, we provided baseline results and highlighted the unique challenges posed by historical Czech texts.

These contributions not only advance the field of Czech text summarization but also pave the way for future research in processing historical documents, offering significant opportunities in cultural preservation and digital humanities. Future work could focus on further enhancing summarization quality, exploring hybrid modeling approaches, and extending the dataset for multilingual and cross-temporal studies.

ACKNOWLEDGEMENTS

REFERENCES

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints.

Anthropic (2024). The Claude 3 Model Family: Opus, Sonnet, Haiku.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer.

Christian, H., Agus, M., and Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7:285.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2024). Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. (2021). A survey of quantization methods for efficient neural network inference.

Hasan, T. et al. (2021). Xlsum: A multilingual dataset for summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2133–2149.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1693–1701.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

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., and Sayed, W. E. (2023). Mistral 7b.

Krotil, M. (2022). Text summarization methods in czech. Bachelor's thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Cybernetics.

Kryscinski, W., Rajani, N., Agarwal, D., Xiong, C., and Radev, D. (2022). BOOKSUM: A collection of datasets for long-form narrative summarization. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, E. and Lu, W., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into text. In Lin, D. and Wu, D., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Miller, D. (2019). Leveraging bert for extractive text summarization on lectures.

Nallapati, R., Zhai, F., and Zhou, B. (2017). 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)*, pages 3075–3081.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Extreme summarization (xsum). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 931–936.

OpenAI (2024). Gpt-4 technical report.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking : Bringing order to the web. In *The Web Conference*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020a). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020b). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Scialom, T. et al. (2020). Mlsum: Multilingual summarization dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2146–2161.

Straka, M., Mediankin, N., Kocmi, T., Žabokrtský, Z., Hudeček, V., and Hajič, J. (2018). SumeCzech: Large Czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Straka, M. and Straková, J. (2018). Rougeraw: Language-agnostic evaluation for summarization. *Proceedings of the International Conference on Computational Linguistics*.

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., and Uszkoreit, J. (2018). Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021a). mC4: A massively multilingual cleaned crawl corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7517–7532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021b). 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., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). 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*, pages 1480–1489.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2019). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.