大型语言模型用于总结捷克历史文档及其他

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

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

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been significantly explored in English and other high-resource languages, Czech text summarization, particularly for historical documents, remains underexplored due to linguistic complexities and a scarcity of annotated datasets.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

自然语言处理（NLP）技术的快速发展提升了文本摘要系统的性能。虽然大多数进展集中在英语等高资源语言上，但捷克语，尤其是其历史变体，仍然缺乏研究。由于语言变化、过时词汇和不一致的句法，历史捷克文档提出了独特的挑战。这些复杂性在开发能够有效处理该领域的自动摘要系统中造成了显著的空白。

因此，本文解决了两个相互关联的挑战。首先，它旨在针对 SumeCzech 建立新的最先进的基准，这是用于现代捷克文本摘要的最全面数据集，使用现代大型语言模型（LLMs），即 Mistral（Jiang et al., 2023）和 mT5（Xue et al., 2021b）。其次，鉴于缺乏为历史捷克语量身定制的资源，我们引入了一个由历史期刊 Posel od Č erchova 衍生的新创建数据集。该数据集专门设计用于促进历史背景下的摘要任务，使未来的研究人员能够应对该领域固有的语言复杂性。此语料库可免费用于研究目的[1]。

通过结合模型的进步和数据集的创新，本研究旨在推动捷克语总结领域的进展，并为文化保护、历史研究和数字人文学科的应用开辟途径。

2 RELATED WORK

文本摘要方法可以分为抽象和提取两类。提取式摘要从源文档中选择最具代表性的句子，而抽象式摘要则生成由新创句子组成的摘要。早期的摘要方法是提取式的，依赖于统计和基于图的方法，如 TF-IDF（词频-逆文档频率）（Christian et al., 2016），通过相对于语料库中稀有度的术语频率来评分句子重要性。类似地，TextRank（Mihalcea and Tarau, 2004）将句子表示为图中的节点，并使用 PageRank 算法（Page et al., 1999）对它们

[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/

Accepted at ICAART 2025. Official version: doi.org/10.5220/0013374100003890

进行排名。

神经网络通过使用循环神经网络(RNNs) (Elman, 1990) 来模拟序列，推进了抽取式和生成式摘要技术。其中一种抽取式方法涉及序列到序列的架构，在这种架构中，LSTM 模型捕捉文档中每个句子的上下文重要性 (Nallapati et al., 2017)。层次注意力网络结合句子级和词语级注意力，以更好地捕捉文档结构和摘要的相关性 (Yang et al., 2016)。这种方法在总结较长和较复杂的文档时被证明是有效的。结合 BERT 嵌入 (Devlin et al., 2019) 和 K-Means 聚类 (Lloyd, 1982) 的混合方法，用于识别关键句子 (Miller, 2019)，在生成式摘要中表现出色。基于 Transformer 的序列到序列模型的发展 (Vaswani et al., 2017) 彻底改变了生成式摘要。最近的模型如 T5 (Raffel et al., 2020a) 采用文本到文本的框架，由于在 C4 数据集上的预训练，在包括摘要在内的各种任务中表现优异。PEGASUS (Zhang et al., 2019) 通过在预训练期间引入缺口句子生成，以掩盖关键句子，在 12 个数据集上取得了强劲的表现。同样地，BART (Lewis et al., 2019) 使用去噪目标来实现稳健的文本摘要生成。多语言模型如 mT5 (Xue et al., 2021b) 和 mBART (Liu et al., 2020) 通过使用诸如 mC4 (Xue et al., 2021a) 和多语言公共爬网 [2] 的数据集，将这些能力扩展到包括捷克语在内的多种语言。

然而，这些模型在没有微调的情况下往往在非英语语料库上表现不佳。

3 DATASETS

以下部分简要回顾了主要现存的摘要数据集。此外，在本节的末尾还将详细介绍所创建的 未知原理公式 Č 的应用 语料库。

3.1 英文数据集

CNN/**每日邮报** (Hermann et al., 2015) 数据集包含超过 300,000 篇英文新闻文章，每篇文章配有作者写的重点摘要。它已被广泛用于摘要和问答任务，并经过多个版本演变以适应特定的自然语言处理任务。

XSum (Narayan et al., 2018) 包含 226,000 个单句摘要，这些摘要与涵盖多个领域的 BBC 文章配对，如新闻、体育和科学。其对单句摘要的关注使得它对抽取式方法的偏向性更小。

Arxiv **数据集** (Cohan et al., 2018) 包含了来自 arXiv 的 215,000 对科学论文及其摘要。数据已被清理和格式化以确保标准化，其中如图表之类的部分已被移除。

书籍摘要 (Kryscinski et al., 2022) 是一个专门用于总结长文本的数据集，如小说、戏剧和故事，并在段落、章节和整本书的层面提供总结。文本和摘要均来自古登堡计划和其他网络档案，支持抽取式和生成式摘要。

3.2 多语言数据集

XLSum (Hasan et al., 2021) 提供了超过一百万篇文章摘要对，覆盖了 44 种语言，从孟加拉语和斯瓦希里语这样的低资源语言到英语和俄语这样的高资源语言。这一数据集从各种 BBC 网站中提取，是多语言摘要研究的重要资源。

MLSUM (Scialom et al., 2020) 包含 150 万篇文章-摘要对，共有五种语言：德语、俄语、法语、西班牙语和土耳其语。该数据集通过存档知名报纸的新闻文章创建，包括《世界报》和《国家报》，重点在于确保广泛的主题覆盖。

上述数据集用于英文摘要，其中一些是多语言的；然而，捷克语资源仍然非常有限。

3.3 署名捷克

SumeCzech 大规模数据集 (Straka et al., 2018) 是针对捷克语资源匮乏的一个显著例外。该数据集由查尔斯大学的形式与应用语言学研究所创建，专为捷克语的摘要任务而定制。数据集包含了一百万篇捷克新闻文章。这些文章来自五个主要的捷克新闻网站：České Noviny, Deník, iDNES, Lidovky 和 Novinky.cz。每个文档都是以 JSONLines 格式构建的，包含 URL, 标题, 摘要, 正文, 子域, 版块和发布日期等字段。预处理过程包括语言识别、去除重复项，以及过滤掉那些标题、摘要或正文为空或过短的条目。

这个数据集支持多种摘要任务，比如标题生成和多句子摘要生成。训练、开发和测试集的比例大约为 86.5/4.5/4.5。全文的平均字数为 409，摘要的平均字数为 38。

然而，该数据集仅适用于现代捷克语，并未解决历史文本处理的需求。

3.4 Č 的消息

为了构建数据集，我们使用了来自历史期刊 Posel od Č erchova (POC) 的数据，该期刊在档案门户 Porta fontium [3] 上是可获取的。数据集的构建涉及解决为所提供的文本创建摘要的挑战，这些文本是用历史捷克语编写的，极少数情况下甚至是德语。文本还涵盖了各种不同的主题，从围绕多玛兹利采(捷克共和国的一个历史悠久的小镇)的本地新闻、意见文章、各种本地广告到国内和全球政治和连载小说。此外，构建足够大规模的数

[2] http://commoncrawl.org/

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

据集以确保评估的准确性和可靠性也很重要。这些方面增加了摘要任务的复杂性。

为了克服上述问题，我们采用了最先进的（SOTA）LLM 模型 GPT-4 （OpenAI, 2024）和 Claude 3 Opus （Anthropic, 2024）（Opus）（特别是 claude-3-opus-20240229 版本）用于初始文本摘要的创建。选择这些模型是基于它们在许多 NLP 任务中的 SOTA 性能以及在一些初步总结实验中的出色表现。

在生成摘要时，确保简洁性是至关重要的。由于大多数实现的方法都是在 SumeCzech 数据集上进行了微调，我们旨在通过以新闻风格创建摘要来保持一致性，从而反映数据集的特征。为此，用于生成摘要的提示包括明确的说明，如下所示：

- 创建以下文本的摘要，采用新闻记者的风格。句子数量 <= 5 ；（EN：创建以下文本的摘要，采用新闻记者的风格。句子数量 <= 5 ）

在摘要任务中，我们观察到虽然两个模型都能生成质量非常好的摘要，但 Opus 倾向于创造更简洁且风格适当的摘要，与新闻记者的格式非常契合。然而，有些情况下，Opus 生成的摘要过于集中于单一主题。

另一方面，GPT-4 旨在在五句限制内融入更多的细节，但有时会偏离指定的风格提示。

如果模型生成的摘要在风格上有显著偏差或过度集中于某一主题，我们会修改或重新生成摘要，直到获得正确版本为止。

创建了两级摘要；第一级在页面层面，第二级则总结了一整篇文章，通常由若干页面组成。因此，我们总结了 432 页，有效地创建了 100 个期刊摘要。包含页面摘要的子集在下文中称为 POC-P ，而期刊摘要称为 现场初步验证-I 。请注意，所有创建的摘要均由两位捷克本地人士手动检查和校正。

数据集采用 .json 格式，并包含以下信息：

- 文本：从给定页面提取的文本，原始印刷内容的数字版本；
- 总结：页面的总结，不超过 5 句话；
- 年份：期刊的出版年份；
- 期刊：来源期刊的具体说明：日期、月份和期号包含在此标识符中；
- page_src：被转换为文本的源图像文件的名称；
- page_num: 页码。

该数据集旨在支持捷克历史背景下的摘要任务，为研究人员提供工具来应对此领域特有的语言挑战。语料库可免费用于研究用途 [4] 。

实验采用了两个先进的基于 Transformer 的模型，分别是多语言文本到文本转换 Transformer (mT5) (Xue et al., 2021b) 和 Mistral 7B (Jiang et al., 2023) 。

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

3.5 多语言文本到文本转移转换器

多语言文本到文本转换 Transformer（mT5）是 T5 模型的一个变种，专为多语言任务设计。该模型是在多语言 mC4 数据集 (Xue et al., 2021a) 上训练的，其中包括捷克语，并有效处理多种语言。该模型基于 Transformer 编码器-解码器架构，并使用 SentencePiece 分词器 (Kudo and Richardson, 2018) 处理复杂的语言结构，包括捷克语形态。通过 span corruption 目标 (Raffel et al., 2020b) 进行预训练，mT5 预测被遮盖的文本片段，使其能够学习语义和上下文关系。

mT5 模型有多种尺寸，涵盖了从小型的 3 亿参数到 XXL 型的 130 亿参数，因此可以适应不同的计算需求。在进一步的实验中，使用了包含 5.8 亿参数的 mT5 基础变体。

3.6 密西特拉尔语言模型

Mistral 语言模型（Mistral LM）是一种高效的大型语言模型，以其在各种自然语言处理任务中的强大性能而闻名。它被设计为结合高准确性和计算效率，在推理、文本生成、摘要和其他 NLP 应用中实现最先进的结果。Mistral 7B 具有 70 亿个参数，在计算效率和任务性能之间取得了平衡，在多个基准测试中超过了更大的模型，如 13B 或 34B。

该模型利用了先进的注意力机制，例如分组查询注意力（GQA） (Ainslie et al., 2023) 和滑动窗口注意力（SWA） (Beltagy et al., 2020) 。GQA 通过将注意力头分组以关注相同的输入数据来提高处理速度，而 SWA 通过限制标记注意力至附近的标记来降低计算成本。该模型支持量化 (Gholami et al., 2021) 和低秩适配（LoRA）(Hu et al., 2021) 等技术，以便在有限硬件上进行高效微调，使其能够有效处理更长的输入。

4 EXPERIMENTS

4.1 评估指标

使用了以下评价指标。

ROUGE（面向召回的概要评估代理）(Lin, 2004) 是一组用于评估摘要质量的指标，通过比较系统生成的摘要与参考文本之间的 n-gram 重叠来实现。主要的 ROUGE 指标包括 ROUGE-N（用于 n-gram 重叠）和 ROUGE-L（用于最长公共子序列）。

ROUGERAW (Straka and Straková, 2018) 是 ROUGE 的一种变体，它在预测文本和参考文本之间进行原始标记级重叠评估，不进行任何预处理，例如词干提取或词形还原。它测量标记的精确匹配，使其适用于精确标记对齐重要的任务。

4.2 设置

我们使用 AdamW 优化器 (Loshchilov and Hutter, 2017)，根据 mT5 (Xue et al., 2021b) 作者的建议，将学习率设置为 0.001 来训练该模型。对于 Mistral 7B，我们采用了 QLoRA (Dettmers et al., 2024)，这是一种将 4 位量化模型与一小组新引入的可学习参数相结合的方法。在微调过程中，只更新这些额外参数，而原始模型保持不变，从而大大减少了内存需求。我们使用了来自 HuggingFace Transformers 库 (Wolf et al., 2020) 的模型。为了训练这两个模型，我们使用了一张具有 45 GB 显存的 NVIDIA A40 GPU。

我们在实验中使用了三种模型的变体：

- M7B-SC：在 SumeCzech 数据集上微调的 Mistral 7B 模型；
- M7B-POC：对 POC 数据集进一步微调的 Mistral 7B 模型；
- mT5-SC：在 SumeCzech 数据集上微调的 mT5 模型。

、。

4.3 在 SumeCzech 数据集上的结果

本实验将提出的 mT5-SC 和 M7B-SC 模型的结果与 SumeCzech 数据集上的相关工作进行比较，见表 1 。

第一个比较方法，HT2A-S (Krotil, 2022)，基于 mBART 模型，进一步在 SumeCzech 数据集上进行了微调。SumeCzech 数据集的作者提供的其他方法 (Straka et al., 2018) 如下：首先是 Random、Textrank 和 Tensor2Tensor (Vaswani et al., 2018) 。

表 1 显示，所提出的 M7B-SC 方法非常高效，性能优于所有其他基线，并在此数据集上取得了新的最先进结果。此外，第二个提出的方法 mT5-SC 也表现出色，始终获得第二好的结果。

4.3.1 在 Č erchova 数据集上的 Posel 结果

本节评估了提议的方法在 Posel od Č erchova 数据集上的表现。表 2 显示了包含每一页（106 页）摘要的 POC-P 子集的结果，而表 3 则展示了由每篇文章（25 期）摘要构成的 POC-I 子集的结果。

这些表格清楚地显示，与前一种情况一样，M7T-POC 模型的结果明显优于 mT5-SC 模型，而且差距非常大。

5 CONCLUSIONS

本文探讨了最先进的大型语言模型，特别是 Mistral 7B 和 mT5，在捷克语文本摘要中的应用，涵盖了现代和历史背景。我们的实验表明，所提出的 M7B-SC 模型为 SumeCzech 数据集建立了新的基准，达到了最先进的性能，而 mT5-SC 模型也表现强劲，始终排名第二。

此外，我们引入了一个新的数据集 Posel od Č erchova，专门用于历史捷克文档的摘要。通过利用这个数据集，我们提供了基准结果，并突出了历史捷克文本所带来的独特挑战。

这些贡献不仅推进了捷克文本摘要领域的发展，也为处理历史文献的未来研究铺平了道路，在文化保护和数字人文学科中提供了重要的机会。未来的工作可以集中于进一步提高摘要质量，探索混合建模方法，并扩展数据集以进行多语言和跨时间的研究。

6

ACKNOWLEDGEMENTS 本研究得到了项目 R & D 的部分支持，该项目为皮尔森大都市区 (DigiTech) 高级数字化技术项目 No. CZ.02.01.01/00/23_021/0008436，以及资助号 SGS-2022-016 高级数据处理和分析方法。计算资源由 e-INFRA CZ 项目 (ID:90254) 提供，该项目由捷克共和国教育部、青年和体育部支持。

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 Inter-

Table 1: 在 SumeCzech 数据集上使用各种方法所得结果，其中包含准确率（P）、召回率（R）和 F1 分数（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: 在 Posel od Č erchova 数据集的 POC-P 子集上实现的方法的结果，包括精度（P）、召回率（R）和 F1 得分（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: 在 Posel od Č erchova 数据集的 POC-I 子集上实现的方法的结果、精准度（P）、召回率（R）和 F1 分数（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

national 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 Computa-

tional 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.