Towards compact and efficient Slovak summarization models

Sebastián Petrík and Giang Nguyen
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 84 216 Bratislava, Slovakia
xpetriks1@stuba.sk, giang.nguyen@stuba.sk

Abstract

Language models, especially LLMs, often face significant limitations due to their high resource demands. While various model compression methods have emerged, their application to smaller models in multilingual and low-resource settings remains understudied. Our work evaluates selected decoder and embedding pruning methods on T5-based models for abstractive summarization in English and Slovak using a parallel dataset. The results reveal differences in model performance degradation and expand the limited Slovak summarization resources and models.

1 Introduction

One of the most prominent limitations of language models (especially LLMs) is their high memory and computational resource demand, which is especially limiting in low-resource environments. Although various model compression methods have emerged, intending to make models more effective, inclusive, and less resource-demanding, there is not much attention paid to smaller models and low-resource languages, such as the Slovak language.

In this work, we focus on **abstractive summarization** of short news articles. One of the challenges of this task in Slovak is the limited dataset options, especially on a larger scale (Ondrejova and Suppa, 2024). Furthermore, Slovak is limited in terms of pre-trained language models, and leveraging pre-trained multilingual models can address this limitation. Our motivation is to address these limitations and investigate the application of decoder and embedding pruning methods with the goal of producing small and efficient Slovak summarization models.

In section 2 we briefly mention the related work. We provide description of our data and methods in section 3 and follow with our experiments and evaluation in section 4. Finally, we conclude the work in section 5 and describe its limitations.

2 Related work

In deep learning, using large numbers of parameters often leads to success in many tasks. However, not all parameters contribute equally, and models often become overparameterized (Han et al., 2015). This leads to higher computational, memory, and power requirements, especially when it comes to large language models (LLMs) (Deng et al., 2020).

Pruning is a model compression method that aims to reduce the size and complexity of a deep learning model by removing redundant components/parameters, which have the lowest contribution to the model performance (Li et al., 2017; Zhu et al., 2023). The pruning methods are commonly divided into unstructured methods (sparse models) and structured methods (targeting entire structural components). Many pruning techniques exist, from simple structured layer pruning to more sophisticated methods, such as magnitude pruning (assigning importance to weights based on their magnitude) (Han et al., 2015), SparseGPT (Frantar and Alistarh, 2023) and hybrid methods such as LoRAPrune (Zhang et al., 2023). However, one of the limitations of more complex methods is model compabililty.

In this work, we focus only on simpler approaches, such as structured pruning of layers, an approach similar to "shrink and fine-tune" (Shleifer and Rush, 2020) and NASH pruning (Ko et al., 2023), and an approach based on embeddings pruning, similar to the Vocabulary Trimmer (Ushio et al., 2023) approach and TextPruner (Yang et al., 2022). While some of these approaches were evaluated on multilingual level, Slovak was not included in their evaluation.

3 Materials and methods

3.1 Data

We use the Gigaword dataset (Graff et al., 2003), specifically, the version intended for abstractive

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 58–68

summarization (Rush et al., 2015). The dataset contains 4 million pairs of short news articles and summaries. We apply additional text cleaning to the dataset and also use the second half of the original validation dataset as a test set to achieve more precise evaluation (3,783,821 samples for train set, 94,405 for validation set and 94,406 for test set). We label the final processed dataset as **Gigatrue** (GT). Furthermore, we introduce a Slovak translation of GT by using machine translation with the Seamless M4T-v2 model (Seamless Communication, 2023), which took approximately 100 hours on an Nvidia A4000 GPU.

3.2 Decoder pruning

Since the decoder of a transformer model is responsible for generation, it directly affects inference speed. By removing decoder layers, the model parameters are reduced and inference speed is increased (Ko et al., 2023). In this work, we prune a series of the middle decoder layers in the following configurations: D-5, D-4-6, and D-3-7, where the numbers indicate the range of layer indices pruned (the models used in our experiments have 8 layers). The approach is similar to decoder pruning in NASH (Ko et al., 2023), where they create a shallow decoder by pruning decoder layers uniformly, while also pruning encoder layers. We use a pre-training pruning approach, where we first remove layers from the base pre-trained model, and then we perform fine-tuning on the prepared variant of the GT dataset.

3.3 Embeddings pruning

The second approach focuses on post-training pruning of the embeddings and vocabulary. The motivation is that not all tokens and embeddings available in the model might be necessary for good summarization performance in a specific domain and specific language, which is especially true for multilingual models which contain tokens from various languages. Furthermore, in the case of the mT5 model, the embeddings take up 86% of model parameters (Ushio et al., 2023) and are shared for both the encoder and the decoder.

The approach uses a calibration dataset (we use the train set of the target dataset), which is tokenized, resulting in a distribution of used tokens. A token threshold is then set, splitting the distribution into two parts: 1. *before threshold* - all tokens and embeddings here are preserved; 2. *after threshold* - tokens (and embeddings) with zero occurrences in the calibration dataset are removed.

Figure 1 shows a comparison of the distributions on the English dataset variant with a 50k threshold. The motivation for using a threshold is an assumption that more important tokens are present at the start of the distribution, even if they are not present in the calibration dataset.

The embeddings are pruned by removing the rows in the embedding matrix for the removed tokens, essentially "shrinking" the matrix above the threshold. Finally, the tokenizer (and the SentencePiece model) used needs to be altered by removing pruned tokens and remapping the indices that have been altered. While we implemented this approach only for the T5 model family, it can be easily modified for other models.

3.4 Evaluation

For evaluation, we leverage common summarization metrics, including **BLEU** (Papineni et al., 2002), variants of **ROUGE** (Lin, 2004), and embedding-based **BERTScore** (Zhang et al., 2020). We chose **ROUGE-L** as our primary comparison metric. During experiments, we take note of various measurements, including: inference/training time, inference speed (θ [tok/s]), peak GPU memory consumption (M [GB]), parameters ($|E|$), and compression ratio c (Equation 1).

$$c = \frac{|E_c|}{|E|} \tag{1}$$

We also introduce the Baseline Degradation Estimate (BDE), a simple metric that estimates what percentage of model parameters can be removed until summarization performance drops to baseline levels (described in Equation 2 and Equation 3, where c^* is compression ratio, m is target metric and B is the baseline). BDE uses an exponential model to determine the relationship between compression ratio and performance degradation.

$$BDE_{m,B}(E) = 1 - c^* \text{ where } c^* \in [0, c_{\max}] \tag{2}$$

$$m(E_{c^*}) = m(B) \tag{3}$$

4 Experiments

We evaluate the selected approaches on the following encoder-decoder models:

- **mT5 small** (mt5) - small version (300 M parameters) of the multilingual T5 model, pretrained on multiple languages including English and Slovak (Xue et al., 2021).

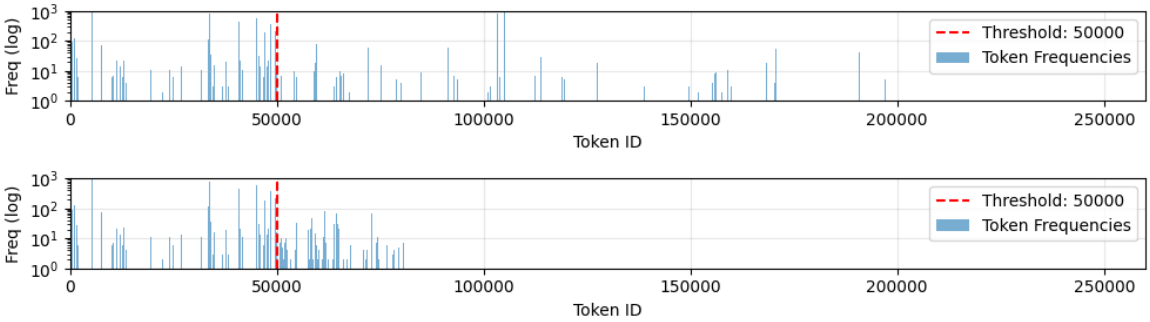

Figure 1: Token distribution of tokenized test set, before and after embedding pruning with 50k threshold.

- **Flan T5 small** (ft5) - a 76 M parameter English-only model, already fine-tuned on a mixture of tasks (Chung et al., 2022).

For **baseline** (**B**), we implement a simple algorithm, which summarizes input text by simply taking a fraction of the words from the beginning. In our approach, we use a 100/30 ratio, which means that a 100-word article will result in a summary by taking the first 30 words of the article. This ratio is based on the summarization ratio of words in the training set samples.

Table 2 compares the base models in terms of parameters, while Table 1 provides comparison of model inference speed, memory usage and summarization performance between base pre-trained models ("(b)" suffix), models fine-tuned on GT dataset and Slovak GT version ("-sk" suffix).

During evaluation, the inference has been performed on an Nvidia A4000 GPU with a batch size of 128.

E	$\|E\|$	|Enc|	|Dec|	M_f[GB]	BS	t_f[h]
mt5	300M	146M	153M	14.9	128	10.5
ft5	76M	35M	41M	14.2	256	6.3

Table 1: Model parameters, fine-tuning memory (M_f), time (t_f) and batch size.

Model	$\theta[t/s]$	$M[GB]$	R-L	BLEU	BERT
B	-	-	0.227	0.039	0.872
B-sk	-	-	0.206	0.045	0.712
ft5(b)	93.2	0.89	0.287	0.056	0.884
ft5	87.6	0.89	0.384	0.157	0.908
mt5(b)	43.6	2.03	0.003	0.000	0.797
mt5	66.1	2.03	0.363	0.142	0.904
mt5-sk	68.5	2.03	0.278	0.092	0.764

Table 2: Base models performance.

4.1 Decoder layer pruning

English

The decoder pruning approach is first evaluated on English, comparing the Flan T5 and mT5 models (Table 3). Figure 2 describes the degradation of model performance on ROUGE-L with further removal of parameters, while Figure 3 compares the models in terms of inference speed.

The results show differences in model degradation. Flan T5, despite being smaller, degraded more slowly, reaching up to 38% decoder compression before reaching baseline, while the mT5 model allowed pruning of only 10% of decoder parameters. We also observed that the Flan T5 model with ~20% of decoder parameters pruned was able to achieve the same performance as an unpruned mT5 model. The approach resulted in a significant increase in inference speed and reduced memory usage for both models; however, the quality degradation was significant, which was also verified qualitatively.

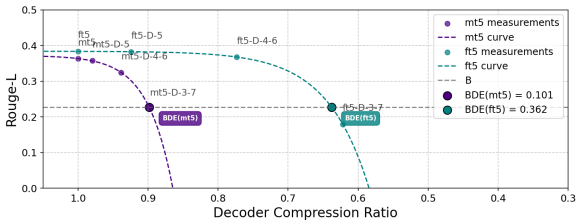

Figure 2: Decoder pruning - English mt5 vs ft5 on ROUGE-L.

Slovak vs English

The approach is then also applied to mT5 variants trained on the Slovak and English datasets (Table 4). Figure 4 describes the degradation of the models. The approach produced similar degradation curves, the Slovak model reached its baseline after removing only 3% less parameters than the

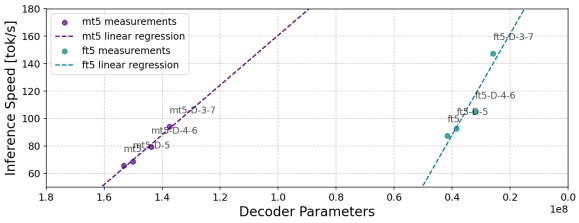

Figure 3: Decoder pruning - English mt5 vs ft5 inference speed.

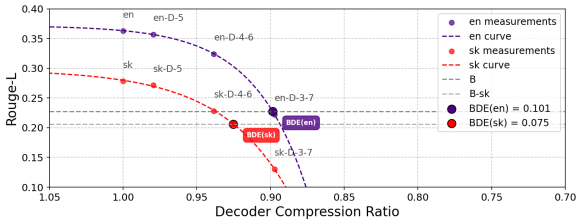

Figure 4: Degradation of Rouge-L on mT5 decoder pruning - SK vs EN.

Model	c_{Dec}	$\theta[t/s]$	$M[GB]$	R-L	BLEU	BERT
B	-	-	-	0.227	0.039	0.872
mt5	1.00	65.9	2.03	0.363	0.142	0.904
mt5-D-5	0.98	68.6	1.95	0.357	0.138	0.903
mt5-D-4-6	0.94	79.3	1.80	0.324	0.114	0.894
mt5-D-3-7	0.90	94.1	1.65	0.222	0.060	0.871
ft5	1.00	87.6	0.89	0.384	0.157	0.908
ft5-D-5	0.92	92.8	0.81	0.382	0.154	0.908
ft5-D-4-6	0.77	104.9	0.66	0.368	0.138	0.905
ft5-D-3-7	0.62	147.4	0.59	0.179	0.039	0.863

Table 3: Decoder pruning - mT5 vs Flan T5 degradation.

English model. We conclude that the difference between the languages using this approach is only minimal.

Model	c_{Dec}	$\theta[t/s]$	$M[GB]$	R-L	BLEU	BERT
B-sk	-	-	-	0.206	0.045	0.712
B	-	-	-	0.227	0.039	0.872
en	1.00	65.9	2.03	0.363	0.142	0.904
en-D-5	0.98	68.6	1.95	0.357	0.138	0.903
en-D-4-6	0.94	79.3	1.80	0.324	0.114	0.894
en-D-3-7	0.90	94.1	1.65	0.222	0.060	0.871
sk	1.00	68.6	2.03	0.278	0.092	0.764
sk-D-5	0.98	71.9	1.95	0.272	0.088	0.762
sk-D-4-6	0.94	81.6	1.80	0.228	0.066	0.742
sk-D-3-7	0.90	93.2	1.65	0.130	0.023	0.683

Table 4: Degradation of mT5 on decoder pruning - SK vs EN dataset.

4.2 Embeddings pruning

Figure 5 and Table 5 compare the summarization performance of the embedding-pruned mT5 model variants on Slovak and English test sets at different thresholds, while Figure 6 compares the model variants in terms of inference speed. Our results show that, using this method and this specific dataset, the threshold does not play a significant role, and the method is able to remove up to 69% of mT5 parameters with minimal impact on summarization performance in English. However, we also notice that the Slovak model reaches only up to 62% of pruned parameters, indicating that there are slightly more tokens needed to represent the Slovak text. The inference speed is also increased (Figure 6).

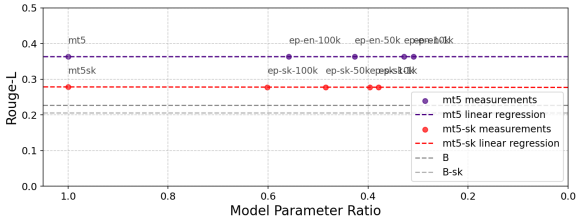

Figure 5: Degradation of mT5 on ROUGE-L with different embedding pruning thresholds (EN + SK).

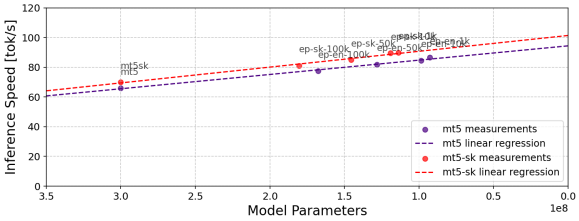

Figure 6: Inference speed of mT5 with different embedding pruning thresholds (EN + SK).

Model	c	$\theta[t/s]$	$M[GB]$	R-L	BLEU	BERT
B	-	-	-	0.227	0.039	0.872
B-sk	-	-	-	0.206	0.045	0.712
en	1.00	65.9	2.03	0.363	0.142	0.904
en-100k	0.56	77.4	1.83	0.363	0.142	0.904
en-50k	0.43	81.9	1.63	0.363	0.142	0.904
en-10k	0.33	84.6	1.47	0.363	0.142	0.904
en-1k	0.31	86.5	1.44	0.363	0.142	0.904
sk	1.00	69.9	2.03	0.278	0.092	0.847
sk-100k	0.60	81.1	1.89	0.278	0.094	0.765
sk-50k	0.48	85.0	1.71	0.277	0.094	0.765
sk-10k	0.40	89.4	1.58	0.277	0.094	0.765
sk-1k	0.38	89.9	1.55	0.277	0.094	0.765

Table 5: Embedding pruning of mT5 at different thresholds (SK + EN).

4.3 Comparison against LLM

In order to provide a more fair comparison, we also evaluate our models and dataset with OpenAI GPT 3.5 using a few-shot summarization prompt (in order to output similar sentence length) on 5,000 samples of test set. After evaluating the GPT summaries, the metrics indicate that our dataset test set summaries slightly differ from the GPT outputs (Table 6). This might be affected by the different text output style of GPT (thus, affecting the sim-

pler metrics), however, after empirical evaluation we also concluded that the GPT summaries have slightly higher quality than the provided dataset summaries. We also evaluate model variants using GPT as a reference (Table 7), which show that the fine-tuned models produce summaries that are more similar to GPT than the test set. However, in both cases, we notice similar semantic similarity score between the dataset and GPT summaries (BERTScore). These observations reveal the limitations of our dataset and evaluation metrics, however, we concluded that the model degradation observations are not significantly affected.

Model	EN			SK		
	R-L	BLEU	BERT	R-L	BLEU	BERT
GPT	0.345	0.085	0.905	0.264	0.061	0.755
mt5	0.357	0.121	0.904	0.294	0.096	0.767
mt5(D)	0.211	0.043	0.869	0.137	0.018	0.679
mt5(E)	0.357	0.121	0.904	0.292	0.091	0.767
ft5	0.374	0.135	0.908	-	-	-
ft5(D)	0.166	0.028	0.862	-	-	-

Table 6: Comparison of GPT and model variants - GT test set as reference.

Model	EN			SK		
	R-L	BLEU	BERT	R-L	BLEU	BERT
testset	0.345	0.086	0.905	0.264	0.060	0.755
mt5	0.430	0.124	0.915	0.396	0.123	0.798
mt5(D)	0.245	0.044	0.871	0.166	0.018	0.677
mt5(E)	0.431	0.124	0.915	0.386	0.112	0.794
ft5	0.446	0.135	0.918	-	-	-
ft5(D)	0.178	0.021	0.861	-	-	-

Table 7: Comparison of dataset and model variants - GPT 3.5 as reference.

4.4 Translation quality

Finally, we also evaluate the quality of the Slovak translation using the OpenAI GPT-4.1 model on the first 1,000 samples of the test set. Each text (articles and summaries) is evaluated using a prompt that assigns a score (0 to 5) on multiple criteria: 1. accuracy (A) - how well does the translation capture the original meaning, 2. grammar errors (G), 3. wrong word choice or meaning (W), 4. missing crucial words (M), 5. unnecessary words (U), 6. incorrect word order (O), 7. stylistic issues (S), 8. cultural sensitivity (C) - such as interpreting idioms correctly.

Figure 7 describes the distribution of scores with different criteria, indicating that the translation is not perfect and contains slight semantic and linguistic errors.

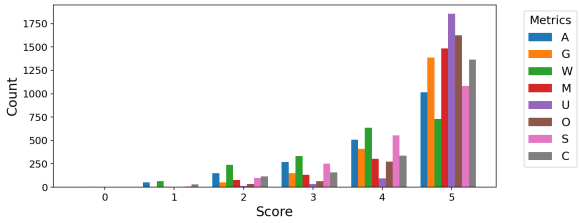

Figure 7: GPT-4.1 translation scores distribution.

4.5 Metric sensitivity

As can be seen in the results, the metrics show a level of divergence in their sensitivity when it comes to both the pruning approaches and other evaluations. These differences come from the nature of the metrics themselves. The ROUGE-L (LCS based) and BLEU (n-gram based) are highly sensitive to changes in vocabulary and word order in the summaries. A significant difference in grammar or vocabulary therefore results in stronger degradation of the metrics. On the other hand, the BERTScore metric leverages contextual embeddings for a semantic comparison, and is less sensitive to such changes. A significant degradation of the BERTScore metric therefore indicates that there might be a significant loss of meaning.

5 Conclusion

In our work, we presented a case study of simple pruning methods on both the decoder and the embeddings of the T5 model family. After producing a synthetic translation to the Slovak language, we provide a parallel English-Slovak variant of the Gigaword dataset for summarization. Using the decoder layer pruning approach, we were able to prune significantly more parameters from the Flan T5 model than from the multilingual mT5 alternative. When comparing decoder layer pruning of mT5 in English and Slovak, the degradation of the Slovak variant was only slightly faster (with decreasing parameters). In the case of embedding (vocabulary) pruning, we were able to reach up to 69% parameter reduction in English with minimal degradation, however, the Slovak variant achieved 7% lower maximum reduction, which we believe is due to more tokens required to represent the Slovak text. Both methods resulted in memory usage reduction and increased inference speed, and can be further combined. The methods are not difficult to implement and are applicable beyond the scope of this work, with models outside the T5 family and other non-English languages.

Limitations

Although the work shows positive results, it is limited in various aspects. In the case of the dataset used, the usage of machine translation for producing a parallel dataset can result in text of lower quality, including some level of grammatical and syntax errors, or alteration of the meaning of the original text altogether, which has been confirmed by evaluation using the GPT 4.1 model.

The next point is that the Gigaword dataset has a very short context length, and the behavior of applied methods on longer sequences is yet to be explored. Another limitation is that although the simplicity of the decoder layer pruning method can be seen as an advantage when considering its application flexibility, other pruning methods could be explored in this context, such as targeting smaller model components, attention heads, or using a more sophisticated framework. The decoder pruning method also affects the architecture on a high level, and is less sensitive to the choice of language.

While the baseline used provides a minimal method for summarization, it shares characteristics with extractive approaches as it preserves the original sentence structure without modifications, only selecting relevant segments. Although this simplicity serves as a fundamental starting point, the absence of other basic models trained on this specific dataset limits our comparison options. The development of more sophisticated baseline methods remains subject to future work.

Finally, the evaluation could benefit from better data quality, such as dataset enhancements through human and/or LLM evaluation for both summarization and translation to Slovak. However, this was not possible due to our time and resource limitations. Nevertheless, our work serves as a solid foundation for further improvements.

Ethical considerations

Machine-translated text from the Seamless M4T-v2 model by Meta AI is used for fine-tuning of the models, which implies that any bias and ethical issues that are caused by the translation model can be present in the Slovak variant of our dataset, and therefore in the Slovak models. Additionally, both translation and summarization using artificial intelligence can result in incorrect preservation of cultural nuances of the languages in the produced text. We conclude that for research purposes, this is acceptable; however, when considering using the dataset in a real-world scenario, further analysis of the translation model and the source dataset is needed.

Statement on the use of AI assistants

We hereby declare that AI assistants based on LLMs (Claude AI and GitHub Copilot) have been used in: 1. grammatical corrections of the text in this paper and translation, 2. generating high-level ideas when approaching the problem in the early stages, 3. limited assistance during implementation (such as generating code for tables and visualisations). Furthermore, the Grammarly software was used for grammatical corrections.

Acknowledgments

This work is funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project Artificial Intelligence for Legal Professions (AILE) No. 09I05-03-V02-00038.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4):485–532.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. *Preprint*, arXiv:2301.00774.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. *arXiv*.

Jongwoo Ko, Seungjoon Park, Yujin Kim, Sumyeong Ahn, Du-Seong Chang, Euijai Ahn, and Se-Young Yun. 2023. Nash: A simple unified framework of structured pruning for accelerating encoder-decoder language models. *Preprint*, arXiv:2310.10054.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. 2017. Pruning filters for efficient convnets. *Preprint*, arXiv:1608.08710.

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

Viktoria Ondrejova and Marek Suppa. 2024. Slovak-Sum: A large scale Slovak summarization dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14916–14922, Torino, Italia. ELRA and ICCL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Yu-An Chung Mariano Coria Meglioli David Dale Ning Dong Mark Duppenthaler Paul-Ambroise Duquenne Brian Ellis Hady Elsahar Justin Haaheim John Hoffman Min-Jae Hwang Hirofumi Inaguma Christopher Klaiber Ilia Kulikov Pengwei Li Daniel Licht Jean Maillard Ruslan Mavlyutov Alice Rakotoarison Kaushik Ram Sadagopan Abinesh Ramakrishnan Tuan Tran Guillaume Wenzek Yilin Yang Ethan Ye Ivan Evtimov Pierre Fernandez Cynthia Gao Prangthip Hansanti Elahe Kalbassi Amanda Kallet Artyom Kozhevnikov Gabriel Mejia Robin San Roman Christophe Touret Corinne Wong Carleigh Wood Bokai Yu Pierre Andrews Can Balioglu Peng-Jen Chen Marta R. Costa-jussà Maha Elbayad Hongyu Gong Francisco Guzmán Kevin Heffernan Somya Jain Justine Kao Ann Lee Xutai Ma Alex Mourachko Benjamin Peloquin Juan Pino Sravya Popuri Christophe Ropers Safiyyah Saleem Holger Schwenk Anna Sun Paden Tomasello Changhan Wang Jeff Wang Skyler Wang Mary Williamson Seamless Communication, Loïc Barrault. 2023. Seamless: Multilingual expressive and streaming speech translation.

Sam Shleifer and Alexander M. Rush. 2020. Pre-trained summarization distillation.

Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. 2023. An efficient multilingual language model compression through vocabulary trimming. *Preprint*, arXiv:2305.15020.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Ziqing Yang, Yiming Cui, and Zhigang Chen. 2022. TextPruner: A model pruning toolkit for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 35–43, Dublin, Ireland. Association for Computational Linguistics.

Mingyang Zhang, Hao Chen, Chunhua Shen, Zhen Yang, Linlin Ou, Xinyi Yu, and Bohan Zhuang. 2023. Loraprune: Pruning meets low-rank parameter-efficient fine-tuning. *Preprint*, arXiv:2305.18403.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. *Preprint*, arXiv:2308.07633.

A Appendix

In the appendix, we include supplementary materials related to our experiments, with the goal of providing additional information and examples.

A.1 LLM prompts

In this section, we provide examples of prompts used for summarization and translation evaluation.

<SYSTEM>

I will show you some examples of article summaries. Learn from these examples to provide similarly concise short single-sentence summaries.

Article: India raised doubts on Thursday that a summit of seven South Asian nations could be held in Pakistan in January, saying there was no reason to meet unless progress has been made in the past year.

Summary: India raises doubts about next regional summit in Pakistan.

Article: Indonesia's top tourism officials on Thursday pleaded with international travelers to come back to its resort island of Bali, where a bomb attack Saturday on a nightclub killed nearly 314 people, many of them young tourists.

Summary: Indonesian officials ask travelers to help heal Bali's tourism industry.

Article: Indonesian police on Thursday were focusing their investigation into the Bali nightclub bombing on a group of eight suspects,

officials said, while the President won crucial parliamentary backing for an emergency anti terrorism decree.

Summary: Security Minister suspects foreign terrorist involvement in Bali bombing.

Article: Swiss pharmaceuticals giant Novartis on Thursday reported a fall in third quarter sales but said its profit had increased 2 percent on the same period last year.

Summary: Novartis reports third quarter sales fall profits up 2 percent.

<USER>

Article: An American woman checked out of a hotel in central China with her 2 year old son Thursday night, saying she had resolved a custody standoff with her Chinese ex husband after they spent nine days sequestered in a suite conducting delicate negotiations.

Summary:

Example 1: Example summarization prompt for GPT 3.5 - English.

<SYSTEM>

I will show you some examples of article summaries in Slovak. Learn from these examples to provide similarly concise short single-sentence Slovak summaries.

Article (SK): India vo štvrtok vzbudzovala pochybnosti o tom, že by sa v januári v Pakistane mohol konat samit siedmich krajín Južnej Ázie, pričom uviedla, že nie je dôvod na stretnutie, pokiaľ sa v uplynulom roku nedosiahol pokrok.

Summary (SK): India vyvoláva pochybnosti o ďalšom regionálnom samite v Pakistane.

Article (SK): Indonézski úradníci pre cestovný ruch vyzvali cestujúcich z celého sveta, aby sa vrátili na ostrov Bali, kde v sobotu pri bombovom útoku na nočný klub zahynulo 314 ľudí, z ktorých mnohí boli mladí turisti.

Summary (SK): Indonézski úradníci žiadajú cestujúcich, aby pomohli uzdraviť cestovný ruch na Bali.

Article (SK): Indonézska polícia vo štvrtok zamerala svoje vyšetrovanie bombového útoku na nočný klub na Bali na skupinu ôsmich podozrivých, zatiaľ čo prezident získal kľúčovú podporu parlamentu pre núdzový dekrét proti terorizmu.

Summary (SK): Minister bezpečnosti podozrieva zahraničných teroristov z účasti na bombovom útoku na Bali.

Article (SK): Švajčiarsky farmaceutický gigant Novartis vo štvrtok oznámil pokles predaja v treťom štvrťroku, ale povedal, že jeho zisk sa v porovnaní s rovnakým obdobím minulého roka zvýšil o 2%.

Summary (SK): Novartis hlási, že predaj v treťom štvrťroku klesol a zisky vzrástli o 2%.

<USER>

Article (SK): Americká žena sa v štvrtok večer s dvojročným synom odhlásila z hotela v centrálnej Číne a povedala, že vyriešila spor o opatrovníctvo so svojím čínskym bývalým manželom po tom, čo strávili deväť dní v apartmáne, kde viedli delikátne rokovania.

Summary (SK):

Example 2: Example summarization prompt for GPT 3.5 - Slovak.

<SYSTEM>

You are a bilingual English-Slovak language expert.

<USER>

Evaluate the following translation from English to Slovak. Assign points based on these criteria:

A = Accuracy (0-5): 5 if the text captures the original meaning perfectly, 0 if the meaning is completely different.

G = Grammar errors / misspelled words (0-5): 5 if the text is grammatically correct, down to 0 if there are significant errors.

W = Wrong word choice/meaning: (0-5): 5 if all words are used correctly, down to 0 if there

are major misuses.

M = Missing words which should be present (0-5): 5 if no words are missing, down to 0 if many important words are absent.

U = Added unnecessary words (0-5): 5 if no extra words are present, down to 0 if many unnecessary words are included.

O = Incorrect word order (0-5): 5 if the word order is correct, down to 0 if the order is significantly incorrect.

S = Stylistic issues (0-5): 5 if the style is appropriate, down to 0 if there are major stylistic issues.

C = Cultural relevance (0-5): 5 if the translation is culturally appropriate, down to 0 if it is culturally very insensitive, such as literaly translating idioms or phrases that do not make sense in the target language.

First, briefly think about the different categories (letters), then provide a formatted JSON output (only categories and their points).

The texts:

Source (English): "American woman fighting for custody of son checks out of Chinese hotel says standoff with ex husband over."

Translation (Slovak): "Americká žena bojujúca o opatrovníctvo syna odchádza z čínskeho hotela a hovorí, že spor s bývalým manželom sa skončil."

Example 3: Example translation evaluation prompt for GPT 4.1.

A.2 Gigatrue samples

Following are some samples (labeled as X, Y and Z) from the Gigatrue test set, the summaries for these articles are then referenced in other examples, where we include only the summaries. We highlight any suspected mistakes or quality degradation in red color.

Article: Americká žena sa v štvrtok večer s dvojročným synom odhlásila z hotela v centrálnej Číne a povedala, že vyriešila spor o opatrovníctvo so svojím čínskym bývalým manželom po tom, čo strávili deväť dní v apartmáne, kde viedli delikátne rokovania.

Summary: Americká žena bojujúca o opatrovníctvo syna odchádza z čínskeho hotela a hovorí, že spor s bývalým manželom sa skončil.

Example 4: GT example X - EN.

Article: An American woman checked out of a hotel in central China with her 2 year old son Thursday night, saying she had resolved a custody standoff with her Chinese ex husband after they spent nine days sequestered in a suite conducting delicate negotiations.

Summary: American woman fighting for custody of son checks out of Chinese hotel says standoff with ex husband over.

Example 5: GT example X - SK.

Article: Thousands of Norwegians joined a nationwide one hour strike on Thursday to protest the government's national budget proposal for next year, saying it threatens jobs and welfare benefits.

Summary: Thousands strike against Norwegian government's proposed budget.

Example 6: GT example Y - EN.

Article: Tisíce Nórov sa vo štvrtok pripojili k celonárodnému hodinovému štrajku, aby protestovali proti vládnemu návrhu národného rozpočtu na budúci rok, ktorý podľa nich ohrozuje pracovné miesta a sociálne dávky.

Summary: Tisíce ľudí štrajkujú proti navrhovanému rozpočtu nórskej vlády.

Example 7: GT example Y - SK.

Article: One person was killed and two injured in a helicopter crash in Russia's Yaroslavl region, officials said.

Summary: One person killed two injured in helicopter crash in Russia.

Example 8: GT example Z - EN.

Article: Jedna osoba zahynula a dvaja boli zranení pri havárii vrtuľníka v ruskej oblasti Jaroslavl, uviedli úradníci.

Summary: Jedna osoba zahynula a dvaja boli zranení pri havárii vrtuľníka v Rusku.

Example 9: GT example Z - SK.

A.3 GPT-generated summaries

The following section contains summaries of X, Y and Z example articles by the GPT 3.5 model.

Summary (EN): American woman resolves custody standoff with Chinese ex-husband in central China hotel.

Summary (SK): Americká žena sa s synom odhlásila z čínskeho hotela po vyriešení sporu o opatrovníctvo so svojím bývalým manželom.

Example 10: GPT 3.5 - X.

Summary (EN): Norwegians strike against government's budget proposal.

Summary (SK): Nóri protestujú proti vládnemu návrhu národného rozpočtu na budúci rok prostredníctvom celonárodného štrajku.

Example 11: GPT 3.5 - Y.

Summary (EN): Helicopter crash in Russia's Yaroslavl region leaves one dead, two injured.

Summary (SK): Jedna obet a dva zranení po havárii vrtuľníka v ruskej oblasti Jaroslavl.

Example 12: GPT 3.5 - Z.

A.4 Decoder pruning examples

In this section, we include a comparison of model outputs between base models and most pruned models using decoder layer pruning, showing the degradation in performance.

Base: American woman checks out of hotel in central China.

D-5: American woman checks out of hotel with son.

D-4-6: American woman says she resolves custody standoff with Chinese ex husband.

D-3-7: American woman arrested in China.

Example 13: Flan T5 - X (EN) - decoder pruning.

Base: Americká žena sa odhlásila z hotela v Číne.

D-5: Americká žena sa odhlásila z hotela v centrálnej Číne.

D-4-6: Americká žena sa s dvojročným synom odhlásila z hotela v centrálnej Číne.

D-3-7: Americká žena odhlásila s čínskym bývalého manželom.

Example 14: mT5 - X (SK) - decoder pruning.

Base: Norwegians strike to protest budget proposal.

D-5: Norwegians protest budget proposal.

D-4-6: Norwegians protest budget proposal.

D-3-7: Norwegians join in a week to protest.

Example 15: Flan T5 - Y (EN) - decoder pruning.

Base: Tisíce Nórov protestujú proti vládnemu návrhu rozpočtu na budúci rok.

D-5: Tisíce Nórov protestujú proti vládnemu návrhu rozpočtu.

D-4-6: "Tisíce Nórov sa pripojili k národnému štrajku, aby protestovali proti vládnemu návrhu."

D-3-7: "Veľké vláde sľubuje, že sľubuje s vládnou vládnou vládu."

Example 16: mT5 - Y (SK) - decoder pruning.

Base: One killed two injured in helicopter crash in Russia.

D-5: One killed two injured in helicopter crash in Russia.

D-4-6: One killed in helicopter crash in Russia.

D-3-7: Two killed in helicopter crash.

Example 17: Flan T5 - Z (EN) - decoder pruning.

Base: Jedna osoba zahynula pri havárii vrtuľníka v ruskej oblasti Jaroslavl.

D-5: Jedna osoba zahynulo pri havárii vrtuľníka v Rusku.

D-4-6: Čína zabití v havárii vrtuľníka v Rusku.

D-3-7: V oblasti zranených pri oblasti a zranených pri oblasti a zranených.

Example 18: mT5 - Z (SK) - decoder pruning.