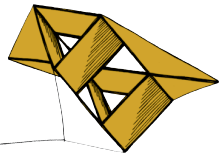

Automatic Creation of Summaries of Historical Documents

Václav Tran[1]

1 Introduction

In the domain of automatic text summarization, neural networks show promising performances. This work probes into the task of automatic summarization of Czech historical documents, a largely unexplored niche area with a scant amount of datasets available. To evaluate and improve the performance of our methods, we created our own dataset constructed from a corpus of historical documents. Then we fine-tuned and utilized Transformer-based models Mistral 7B and mT5. We also implemented and evaluated a method, which we refer to as Translation-Summarization-Translation, where we utilize state-of-the-art machine translation and English summarization methods to generate Czech summaries. The performance of these methods set a new baseline for the task of summarizing Czech historical documents.

2 Methods

We fine-tuned a 512-million parameter variant of mT5 from Xue et al. (2021) and the Mistral 7B model from Jiang et al. (2023) on SumeCzech dataset from Straka et al. (2018), which contains collection of articles from Czech news sites alongside their summaries. Additionally, we curated our own dataset, which we abbreviated as POC, comprising of historical documents from *Posel od Čerchova* journals and their respective summaries. Mistral 7B was additionally fine-tuned on the POC dataset. By using various training optimization methods, we managed to fit the entire process of fine-tuning Mistral 7B on a single NVIDIA A40 45 GB GPU.

For Translation-Summarization-Translation (TST) method, we used a machine translation model ALMA from Xu et al. (2024) and instruct fine-tuned Mistral 7B for text summarization. TST translates the given Czech text to English, summarizes it using the preferred English summarization model and translates the English summary back to Czech.

3 Results

Mistral 7B additionally fine-tuned on POC is abbreviated as M7B-POC. The mT5 model fine-tuned on SumeCzech dataset is abbreviated as mT5-SC. Using POC dataset, we evaluated these three methods on POC-I and POC-P using $\text{ROUGE}_{\text{RAW}}$ metric by Straka et al. (2018), where M7B-POC achieved the highest overall performance. POC-I is a subset of POC that contains summaries of issues and POC-P is a subset of POC which contains summaries of pages of individual issues. However, our limited observations of the generated summaries suggest that higher performance on the chosen evaluation metric does not necessarily indicate superior summarization quality, particularly with regard to factuality.

[1] student of the bachelor degree program Applied Sciences, field of study Computer Science, e-mail: nuva@students.zcu.cz

Table 1: Results of implemented methods on POC-P. M7B-POC was evaluated only on summaries it has not been trained on.

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**
TST	17.2	**25.1**	**19.9**	2.5	**3.8**	2.9	11.3	**16.4**	13.0
mT5-SC	20.2	8.2	11.1	1.4	0.5	0.7	14.9	6.1	8.2

Table 2: Results of implemented methods on POC-I. M7B-POC was evaluated only on summaries it has not been trained on.

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**
TST	14.0	**24.8**	17.5	1.7	**3.1**	2.1	9.1	**16.3**	11.4
mT5-SC	18.2	5.9	8.6	1.0	0.3	0.4	**14.0**	4.5	6.5

Acknowledgement

I would like to thank my thesis advisor Doc. Ing. Pavel Král, Ph.D. for their constant support and help. I would also like to thank everyone, who supported me throughout the creation of this Bachelor thesis, including Zach for checking my work for grammatical errors. 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.

References

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

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou (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. https://doi.org/10.18653/v1/2021.naacl-main.41

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Singh Chaplot, D., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., and Kim, Y. J. (2024). Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. arXiv preprint arXiv:2401.08417.