Abstractive Summarization for the Ukrainian Language: Multi-Task Learning with Hromadske.ua News Dataset

Svitlana Galeshchuk
Arval BNP Paribas / Rueil-Malmaison, France
West Ukrainian National University / Ternopil, Ukraine
svitlana.galeshchuk@gmail.com

Abstract

Despite recent NLP developments, abstractive summarization remains a challenging task, especially in the case of low-resource languages like Ukrainian. The paper aims at improving the quality of summaries produced by mT5 for news in Ukrainian by fine-tuning the model with a mixture of summarization and text similarity tasks using summary-article and title-article training pairs, respectively. The proposed training set-up with small, base, and large mT5 models produce higher quality résumé. Besides, we present a new Ukrainian dataset for the abstractive summarization task that consists of circa 36.5K articles collected from Hromadske.ua until June 2021.

1 Introduction

Reading a large number of documents is a time-consuming and frequently tedious process that requires a substantial investment of human resources. That is why creating pithy abstracts for financial articles, social media news, or even bug reports originated many real-life use cases for automatic summarization.

In the meantime, the rapid development of AI methodology and the latest NLP progress with large Transformer language models pushed the boundaries of text generation. Producing a résumé for a document constitutes one of the applications for text generation that keeps attracting more attention of the academic community (see Section 2.1) and practitioners.

Abstractive summarization is a generative task that foresees automatic creation of document summary by synthesizing an input while preserving its gist. Observed limitations of language models (see Section 2.1) frequently challenge this definition. Recent papers discuss the problem of information distortion when it comes to solutions for the English language; however, for low-resource languages like Ukrainian the differences between real and expected results might be even more significant. This paper tries to improve the ability of language models to capture a gist of text in order to generate summaries of better quality for news articles in Ukrainian by finetuning the multilingual T5 Transformer on the corpora that exploits training data for both summarization and text similarity tasks simultaneously and thus guiding the model to the essence of each article. The second objective is to construct and introduce the dataset of Ukrainian news that can be further exploited for abstractive summarization.

The next section presents problems of abstractive summarization and discusses mT5 architecture and training. Section 3 focuses on training data, methodology and evaluation strategy. Section 4 concludes with results and discussion.

2 Overview of automatic summarization

2.1 Challenges of Abstractive Summarization

The recent growth of transfer learning solutions with Transformer-like decoder architectures contributed to development of fine-tuned models apt for abstractive summarization (such as BART, T5, GPT). However, current research identifies significant issues which make automatic summarization a challenging task (see the papers that conduct in-depth research on the topic: Erdem et al., 2022; Ji et al., 2023). We highlight the following problems:

• How to evaluate a summary? We address the issues of summary evaluation in Section 3.3.

• Summaries suffer from hallucinations, i.e., information leaked to the output from the outside of source text. However, Cao et al., 2022 find that much hallucinated content is mainly consistent with world knowledge.

• Summaries do not convey a gist of text, which is especially noticeable in multi-document summarization. Our study concentrates on "helping" the mT5 model to pay attention to an essential message

Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 49–53

expressed in the article.

We can find a plethora of models pre-trained and fine-tuned on English corpora. However, language resources for Ukrainian are still limited, which penalizes models' performance and limits the number of available monolingual solutions. Among language models suitable for summarization BART, PEGASUS, T5 and GPT/GPT-2/GPT-3 are the most well regarded pre-trained solutions as they include a decoder part in their architecture. We use the multilingual T5 model in our experimentation (see Section 2.3).

2.2 Related Works

Training resources for summarization in Ukrainian are limited. XL-SUM (Hasan et al., 2021) multilingual dataset stands for a silver standard as it comprises more than 58K of BBC news articles in Ukrainian. While this number is higher than for Arabic or Chinese, the performance of the model trained with XL-SUM is still better for the latter languages. No human evaluation was conducted for the Ukrainian language as the authors focus mainly on top 10 spoken languages. In spite of the need for further investigation of the Ukrainian corpora quality, we consider this dataset as a benchmark for comparison and evaluation in our study.

MassiveSumm (Varab and Schluter, 2021) is another multilingual dataset that contains 594,415 news articles in Ukrainian. The data is collected from the sources that follow OpenGraph standard (see Grusky et al., 2018). While the corpus is large, there is no profound analysis of its quality presented. The reported summarization results are less convincing than with XL-SUM for the same languages.

Concerning attempts to build automatic summarization model, most of research until recently focused on extractive summarization (see Shakhovska and Cherna, 2019). Abstractive summarization is mainly represented by finetuned multilingual models with XL-SUM[1] or extracted Ukrainian model from multilingual version[2]. Comparing to these works we present a sequence-to-sequence language model trained with a mixture of tasks for the newly developed dataset of Ukrainian news.

[1] see "mT5multilingualXLSum" at https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum

[2] see "ukmt5base" at https://huggingface.co/kravchenko/uk-mt5base

2.3 Multilingual-T5 and Multitask Training

Text-to-Text Transfer Transformer, or simply T5, is a Transformer model with encoder-decoder architecture well suited for sequence-to-sequence tasks. The encoder comprises blocks with two main elements: a self-attention layer and a feed-forward network. The decoder has a slightly adjusted structure with standard attention added after each autoregressive self-attention layer.

No monolingual T5 model exists for Ukrainian. Hence multilingual version called mT5 is used. Similar to its original version, mT5 has been pretrained on a large dataset cleaned with a set of heuristic rules (i.e., removal of all texts with less than three lines and 200 characters). The corpora cover more than 100 languages, and the Ukrainian part accounts for 1.51% with 41B tokens and 39M pages (see Xue et al., 2020).

We choose mT5 model as its training foresees transforming multi-task learning into finetuning on a mixture of datasets with the same text-to-text objective (see Raffel et al., 2020). "Prefix", i.e. some context provided to the model that is later used when making predictions, is added to the input text and helps model separate tasks. Thus, after pretraining, the model is further finetuned on a mixture of tasks in a sequence-to-sequence manner: the output is always a string, even in the case of quantitative values. This unified text-to-text approach in multi-task learning is a key element in our study as we mix Hromadske.ua data with summaries as target together with the same Hromadske.ua' articles and titles with "similar" as expected output.

3 Experimental setup

3.1 Methodology

The pre-trained mT5 checkpoint serves our experimentation as a baseline model. We considerate two downstream tasks for further training:

• *Summarization* with a respectful prefix that defines the task for the model. Here we use an article as input and a summary as a target.

• *Similarity* that learns the similarity between a text and its title. Here we use the same set of articles (sentence 1) together with the articles' titles (sentence 2) as an input and a string "similar" as a target.

This setup builds on the original idea of training T5 with a mixture of several tasks with the same text-to-text objective. Raffel et al., 2020 use independent datasets for each of the task. In contrast,

we train the model with the same collection of texts adjusted for both tasks. Here, mT5 can see an article twice but with different target. This approach helps the model catch the gist of text usually reflected in its title and produce a more meaningful, topic-focused summary.

We concatenated adjusted versions of the dataset creating the mixed tasks for multi-task learning (see Figure 1). We refer to it as an extended dataset. Because of task mixing, the T5 approach does not require changes in model design for classification output on similarity as it is usually designed in multi-output settings (i.e., in Nan et al., 2021 supplementary classification head in the decoder of BART helped identify summary-worthy named entities to tackle hallucination problem).

Different checkpoints of mT5 are released: mT5-small, base, large and XL. Moreover, recently mT5 model fine-tuned on XL-SUM dataset made another step towards better summarization (see Hasan et al., 2021). We thus use mT5-small and mT5-large pretrained original checkpoints together with mT5-base model finetuned on a XL-SUM single task to assess empirically the presented hypothesis.

3.2 Training Dataset

In our experiment setup, we use the dataset of 36,488 articles collected and filtered from the website Hromadske.ua from September 2018 to June 2021. The total corpus shared by ***UberText***[3] initiative contains approximately 130K articles from Hromadske.ua, but this value decreased drastically after we applied several layers of heuristic rules.

Although the total Ubertext collection contains other news sources (Ukrainska Pravda, Unian, etc.), the task of automatic distinguishing of a summary from an article is quasi-impossible for most of them. In contrast, many articles from Hromadske.ua have consistent editorial structure:

Part 1: summary + Part 2: source of the news + Part 3: the article itself.

We observed a lexical pattern that helps define Part 2 in many texts : it tends to start with "Pro tse" (English translation: "This is") + a source. For example, "Про це повідомляє Маріупольська міська рада." (Transliteration: "Pro tse povidomlyaye Mariupolska miska rada", English translation: "This is reported by the Mariupol City Council").

[3]see details at https://lang.org.ua/en/.

INPUT	TARGET
До «червоної» зони карантинних обмежень 22 березня потрапила Сумська область.Загалом жорсткі обмеження запровадили вже у 8 регіонах. Translated: On March 22, the Sumy region entered the "red" zone of quarantine restrictions. In general, strict restrictions have already been introduced in 8 regions.	підсумувати: Тепер жорсткий карантин за розподілом МОЗ діє у Житомирській, Закарпатській, Івано-Франківській, Київській (потрапила туди 21 березня), Чернівецькій, Сумській, Одеській областях та Києві. Також напередодні в Миколаєві місцева влада самостійно запровадила «червону» зону в місті через критичну ситуацію з госпіталізацією пацієнтів. У «помаранчевій» зоні за розподілом МОЗ опинилися 11 регіонів: Вінницька, Дніпропетровська, Донецька, Луганська, Львівська, Миколаївська, Полтавська, Тернопільська, Хмельницька, Черкаська, Чернігівська областіРешта областей залишаються в «жовтій» зоні карантину. З 24 лютого Україна повернулася до адаптивного карантину. Translated: summarize: Now strict quarantine according to the Ministry of Health is in effect in Zhytomyr, Zakarpattia, Ivano-Frankivsk, Kyiv (from March 21), Chernivtsi, Sumy, Odesa regions and Kyiv. Also, the day before in Mykolaiv, local authorities independently introduced a "red" zone in the city due to a critical situation with hospitalization of patients. According to the Ministry of Health, 11 regions were in the "orange" zone: Vinnytsia, Dnipropetrovsk, Donetsk, Luhansk, Lviv, Mykolaiv, Poltava, Ternopil, Khmelnytskyi, Cherkasy, Chernihiv oblasts. The rest of the oblasts remain in the "yellow" quarantine zone. Since February 24, Ukraine has returned to adaptive quarantine
речення1 (sentence1): Тепер жорсткий карантин за розподілом МОЗ діє у Житомирській, Закарпатській, Івано-Франківській, Київській (потрапила туди 21 березня), Чернівецькій, Сумській, Одеській областях та Києві. Також напередодні... повернулася до адаптивного карантину речення2 (sentence2): У Сумській області запровадили «червону» зону карантину. Загалом в Україні вже 8 таких регіонів.	подібні / similar

Figure 1: Example of input for multi-task training with mixture of datasets having the same text-to-text objective. English translation provided alongside. Note sentence 1 is truncated to save the page space.

Recall from Section 3.1 that training employs the dataset comprising the following components:

• input: article –> target: summary;

• input: sentence 1: same article, sentence 2: title –> target: “similar”

Figure 1 displays an example of input-target used for training accompanied with English translation for non-Ukrainian speakers.

Occasionally, a summary repeats a title. To avoid these issues, we adopted an n-gram approach to discard title-summary near-duplicates. We followed the guidance from the original T5 paper (Raffel et al., 2020) and lowercased texts before using them. In addition, we deleted the titles that contain digits as the set-up does not foresee an assessment of numerical values consistency. Topic analysis classifies the filtered articles into four main categories: politics, sport, culture and science with a majority of texts falling in the first category. Human evaluation of datasets is expensive and time-consuming. Hence, automatic approaches serve to understand better and clean the dataset. The following metrics measure the quality of the training input:

Abstractivity: a metric based on the matched text spans between a text and a summary (Grusky et al.,

Dataset	ABS	SBert	Rouge-L
Hromadske.ua	82.30	0.52	39.4
XL-SUM	75.70	0.63	35.8

Table 1: Evaluation of the presented dataset (Hromadske.ua) comparing to XL-SUM.

2018).

SBertScore similarity between a summary and a first sentence of an article to avoid duplication of content by simple paraphrasing, as the model may learn to pay attention only to the first sentence.

ROUGE-L: the score reflects the longest sequence of words shared. In this case the lower score is preferable.

Not many datasets are available to train summarization model in Ukrainian. We find XL-SUM (Hasan et al., 2021) the most advanced and reliable benchmark for an intrinsic comparison with our dataset. Table 1 reports the comparison.

The evaluation proves a reasonable abstractiveness of the Hromadske.ua dataset, which is higher than XL-SUM. The Rouge-L score is also higher in our case, reflecting better originality of the benchmark summaries yet.

3.3 Metrics and evaluation

The benchmark metric for abstractive summarization tasks adopted by the research community is the ROUGE score. The metric compares a generated summary against a reference. We employ three sub-categories of the ROUGE score:

- ROUGE-1: unigram overlap
- ROUGE-2: bigram overlap
- ROUGE-L: Longest Common Subsequence

The evaluation strategy foresees a split of the available dataset into the training-validation-test set with the ratio 80:10:10. The validation and test comprise only summary-article pairs, as we do not tend to assess similarity task. Thus, the reported results include only summaries of previously non-seen articles ignoring the evaluation of titles' similarity.

4 Main Findings

4.1 Results

This section reports the results of training the following mT5 checkpoints:

1. mT5-small with 300M parameters pretrained ("mT5-small")

Checkpoint	Baseline	One task	Two tasks
mT5-small	Not tested	9.50/2.12/9.43	13.26/2.71/13.40
mT5-SUM	11.72/3.41/11.74	19.69/5.52/19.48	21.46/6.12/21.55
mT5-large	1.52/1.01/1.63	19.55/4.89/19.77	22.09/7.04/22.12

Table 2: ROUGE-1/2/L scores on test set

2. mT5-base with 580M fine-tuned only with XL-Sum dataset ("mT5-SUM")

3. mT5-large pretrained with 1.2B parameters ("mT5-large")

Each training includes tokenization with vocabulary given with mT5 checkpoint. The input is truncated to 1024 tokens with a maximum output length equal to 128. The constant learning rate of 0.001 mimics the original setup. No dropout is applied. The models have been trained with circa 10000 steps (compared to XL-SUM with 37000 steps).

Table 2 concludes the empirical findings on test split by comparing with the baseline (column 1), training with only articles-summary pairs (column 2), and training with article-summary and article-title similarity test (column 3). mT5 large one task[4] and mT5 large two tasks[5] model may be tested at HuggingFace hub with proposed text examples.

All setups show better performance of the models with two-task learning rather than fine-tuning on a sole summarization downstream objective. The values are usually more important with Rouge-1/2 scores than Rouge-L. The output for mT5-SUM baseline is lower than in the original paper. However, Hasan et al., 2021 adjust the Rouge score for the languages. It may explain the reported difference.

4.2 Discussion

The improvement of generative models' ability to produce better quality summaries and an introduction of the Ukrainian news dataset constitute two main objectives and contributions of the paper. An adjusted multi-task learning setup for mT5 models is employed to achieve the first goal. The heuristics and evaluation behind the Hromadske.ua dataset satisfy the second objective. Concerning further research, we plan to use BertScore (Zhang et al., 2019) to better assess a model's ability to grasp the gist of an article with contextual similarity. The proposed approach may especially benefit multi-text

[4] see mT5-large-one-task model at https://huggingface.co/SGaleshchuk/t5-large-ua-news

[5] see https://huggingface.co/SGaleshchuk/mT5-sum-news-ua

summarization. Testing with available multi-article datasets in English together with a construction of such source in Ukrainian create a basis for further research. Moreover, the presented training setup may be fully reproducible for other low-resource languages.

Limitations

This Section highlights the following limitations of the presented setup:

• Although we received satisfactory scores with the extended dataset of Hromadske.ua, more computational resources could allow longer training and thus better assessment of the model performance.

• Rouge score may penalize abstractiveness of generated summaries. Metrics that assess factuality could evaluate better the model results.

• Expert evaluation of the dataset's sample reveals summaries that sound like introduction rather than abstract of article.

References

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nataliya Shakhovska and Taras Cherna. 2019. The method of automatic summarization from different sources. *arXiv preprint arXiv:1905.02623*.

Daniel Varab and Natalie Schluter. 2021. Massivesumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.