UDC 577.29

TAILORING MT5 FOR THE UZBEK LANGUAGE: A COMPACT MODEL FOR NLP APPLICATIONS.

Adilova F.T.[1], *Davronov R.R.*[1], *Kushmuratov S.I.*[1]

[1] V.I. Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Tashkent, Uzbekistan
fatadilova@matinst.uz, rifqat.davronov@mathinst.uz, bekmezonali@gmail.com

Abstract. *Despite being spoken by nearly 50 million individuals, the Uzbek language remains underrepresented in Natural Language Processing (NLP). One primary reason is the limited availability of Uzbek linguistic resources. With the rising prominence of the Transformer architecture in NLP, it has overtaken earlier methods like convolutional and recurrent neural networks. The T5 (Text-to-Text Transfer Transformer) standardizes linguistic tasks in English by converting them into a text-to-text format. The mT5, its multilingual version, has shown promising outcomes in various NLP tasks spanning multiple languages. However, the considerable dimensions of the mT5 pose challenges for applications focused on a singular language. In our study, we fine-tuned the mT5 specifically for Uzbek, leading to a more compact T5 model. We compared this tailored model's efficiency with the mT5 on Automatic Text Summarization (ATS) and Named Entity Recognition (NER) tasks using identical protocols and datasets. Our adapted model surpassed the performance of the mT5, indicating the feasibility of developing a more compact pre-trained model with nearly half the size, without compromising results. This streamlined model also benefits from reduced memory usage, faster startup, and swifter processing times. For access to this model, please reach out.*

Keywords: *model compression, transformer, pre-trained model, automatic text summarization, named entity recognition.*

I. INTRODUCTION

Neural network models have significantly transformed the field of Natural Language Processing (NLP) and machine translation. Although Recurrent Neural Networks (RNNs) have made substantial contributions, they have certain inherent limitations. These challenges have pushed researchers towards Transformer models, particularly the "Uzbek Version of Multilingual T5 (Text-To-Text Transfer Transformer) Transformer", which offers remarkable performance across various language applications.

Historically, RNNs have been pivotal for tasks like language translation, sentiment analysis, and predicting time series. However, their effectiveness is hampered due to challenges like the vanishing and exploding gradients, making it difficult for them to learn long-term dependencies. Additionally, their sequential data processing nature doesn't exploit the full potential of modern parallel computing, leading to prolonged training durations.

On the other hand, Transformer models address many of these issues. Their bidirectional self-attention mechanisms can capture inter-relationships in input data irrespective of the distance between elements. Plus, their structure supports parallel processing, ensuring more efficient training. Google's T5 Transformer advances this by converting

all NLP tasks into a unified text-to-text approach, resulting in a highly adaptable model that delivers outstanding outcomes for numerous tasks.

The T5 model is available in variations like T5-Base, T5-Large, T5-3B, and T5-11B. Each differs in terms of size, computational needs, and NLP performance. Typically, larger models possess a deeper understanding of intricate language structures but require more computational resources.

The core focus of this paper is the creation and assessment of an Uzbek-focused Multilingual T5 Transformer. Considering the limited representation of the Uzbek language in current NLP studies, our research addresses a significant void, proving that the efficient performance of the T5 Transformer can be adapted successfully for lesser-studied languages.

In the following sections, we elaborate on the multilingual Text-to-Text Transfer Transformer (mT5) and outline the steps to produce several Uzbek-centric models from this multilingual base. We then measure the performance of these models against the mT5, focusing on tasks like Summarization and Named Entity Recognition (NER). Additionally, model efficiency comparisons encompass loading and inference durations as well as memory usage. The paper concludes with a detailed analysis and summary of our research findings.

II. RELATED WORK

Advancements in Natural Language Processing (NLP) and machine translation have been particularly notable with the emergence of transformer-based architectures, notably Google's Multilingual T5 (mT5) model [1]. This section provides a brief discussion on the progress in this domain, emphasizing the derivation of language-specific models from broader multilingual counterparts, and the unique challenges faced by languages, for instance, Uzbek.

Google's mT5 model is an evolved multilingual version of the T5 model [2] and has been trained on a multilingual internet text corpus. Demonstrating prowess in numerous NLP tasks, mT5, as a seq2seq model, has been particularly effective in translation. Yet, its utility for underrepresented languages like Uzbek remains under investigation.

Earlier research has established the feasibility of isolating single language models from their multilingual versions. Pires, T. et al. [3] were at the forefront of this initiative, revealing that monolingual models encapsulated within multilingual BERT (mBERT) could be harnessed for specific linguistic tasks. However, extrapolating this principle to transformer architectures like T5 or mT5 hasn't been extensively probed.

Historically, research on language-specific models has mainly centered on well-resourced languages. Lewis, M., Liu, Y. et al. [4] presented BART, a denoising autoencoder adept at pretraining sequence-to-sequence architectures, marking notable strides in several applications. Yet, its adaptation for languages with fewer resources hasn't been a primary concern.

Work related to the Uzbek language and NLP is relatively sparse, owing to its status as a low-resource language. F.Adilova and R. Davronov, 2021 developed model UzRoBerta, demonstrating the challenges and potential in this field.

Dedicated deep learning models tailored for the Uzbek language remain an intriguing avenue for research. The multilingual Transformer model delineated in [5] refines the mBERT vocabulary, truncating a significant portion of

parameters, predominantly in the embedding layer. This strategy curtails up to 49% of total parameters without undermining the average accuracy. An advantage of this method is bypassing the need to retrain the model from scratch, contrasting the technique in [6]. This compression concept, targeting frequent tokens while eliminating the remainder, was similarly utilized by [7] to distill the Russian language from mT5, though it awaits peer review. Our study adopts this strategy to craft a streamlined model specifically tailored for the Uzbek language.

III. METHODOLOGY

Multilingual T5, or mT5, is a version of T5 that's been pre-trained using the mC4 (Multilingual Colossal Clean Crawled Corpus) dataset. This dataset incorporates a vast variety of languages, including but not limited to 101 of them, Uzbek being one. Owing to its diverse language coverage, the model is understandably quite large. This model, made available by Google AI's team and hosted on the Hugging Face repository, comes in different versions. The mT5-base, for instance, is 2.33 GB, while the smallest model, the mT5-small, is 1.2 GB in size.

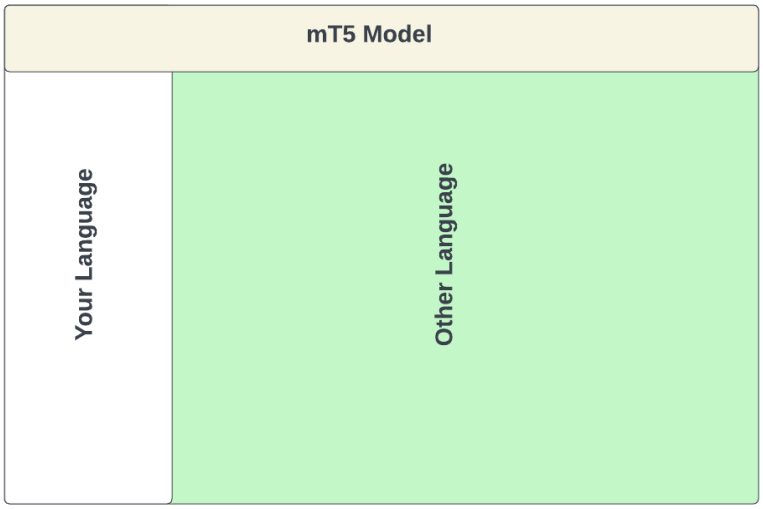

Fig.1. Flow of extraction.

To create the Uzbek language variant of mT5, we derive the Uzbek language from the mT5-base, choosing specific Uzbek vocabulary. Following this, we update the embedding layer and tokenizer to yield an appropriate model. Fig. 1 illustrates the comprehensive process of these stages.

3.1. Selecting Vocabularies

The Uzbek language corpus is processed utilizing the original tokenizer. Given that Uzbek texts frequently incorporate English and Russian, a limited assortment of English and Russian tokens are retained in the model. To determine the prevalence of various tokens, we sourced the corpus of Uzbek, Russian, and English phrases from the Leipzig corpora collection [8].

Based on the token count within the Uzbek corpus, we discovered that a mere 17.7% of the model's vocabulary was utilized. English accounted for 18.9%, while Russian comprised 14.5%. Notably, there exists a similarity of over 35% between the Uzbek and English lexicons, 14% between the Uzbek and Russian lexicons, and more than 15% between the Russian and English lexicons. This may be due to the inclusion of English words

or words in Latin script within Uzbek texts. Furthermore, the top 20,000 tokens represent more than 99.4% of both the Uzbek and Russian vocabularies. A similar statistic is observed for English.

This study determines the number of tokens to be 57.5K based on vocabulary filters. These tokens are composed of 1K top tokens from the original tokenizer, 9089 English tokens, which alone account for over 95% of the English corpus, and an additional 100 special tokens utilized by T5. For the Uzbek language, the remaining tokens have not yet been allocated among the three token groups, resulting in a total of 57.5K tokens. This represents 23% of the 250K tokens in the multilingual version. Other tokens have been removed. Of the 57.5K tokens, 9061 are allocated for the Uzbek language. This occurrence is due to overlapping tokens.

3.2. Model Update

In a recent update to our neural network, the input and output parameters were replaced with predefined ones, resulting in a reduction of the model's size by 57% - a decrease from 2330 MB to 1100 MB. The revamped model now houses 244 million parameters, equivalent to 49 percent of the parameters in the multilingual model.

In parallel, the tokenizer also underwent an update. The Protobuf representation was adopted in line with the Sentencepiece tokenizer employed by T5. Following these modifications, the restructured model has been uploaded to the Transformers Hub, enhancing its accessibility for the Natural Language Processing (NLP) communities, particularly those focusing on the Uzbek language.

It should be noted that the updated model shares similarities with the pre-trained mT5 model. Consequently, fine-tuning is required for its application in various NLP tasks, as the mT5 model is exclusively trained on unsupervised tasks for predicting missing words.

A pre-trained model refers to a model that has previously been trained on a vast dataset and possesses the capacity to be utilized across diverse natural language tasks without necessitating additional training. The use of pre-trained models turns out to be useful, since it eliminates the training of a model from scratch, thereby significantly reducing time and reducing computational costs.

Fine-tuning, a technique commonly employed within the realm of Natural Language Processing (NLP), serves to adapt a pre-trained model to a specific task. Within the fine-tuning process, the pre-trained model undergoes further training on a novel dataset with the aim of enhancing its performance in relation to a particular task. Obviously, this approach is considered more effective compared to starting model training from scratch.

IV. RESULTS AND DISCUSSION

The model we propose, uzT5, is a derivative of mT5, and similarly to its progenitor, it necessitates fine-tuning for specific natural language processing (NLP) tasks. As a scaled-down version of the mT5 model, uzT5's performance may potentially fall short when compared to the original mT5. This section aims to establish a comparative analysis between uzT5 and mT5 to determine their respective performances when fine-tuned for NLP tasks.

For the purpose of this comparison, we conducted fine-tuning on a number of NLP tasks for both models, namely, Summarization and Named Entity Recognition (NER). To ensure a fair and robust comparison, identical methods, datasets, and configurations were

employed for each model. We elected to use only ten epochs for the fine-tuning process across all tasks. This decision was not driven by a pursuit of optimal performance, but rather to compare the performance of the two models under the same conditions.

The fine-tuning process and subsequent comparisons were performed within the NVIDIA DGX Station using the Hugging Face's Transformers Trainer class [9] in conjunction with PyTorch [10].

4.1. Automatic Text Summarization

Automatic Text Summarization (ATS) is the procedure of condensing a longer text document into a more concise version by utilizing Natural Language Processing (NLP) techniques. It aims to emphasize the most significant details within the text, following specific criteria.

We adopted the approach introduced by [11] to develop a model based on both uzT5 and mT5 for the automatic text summarization task. Our dataset was sourced from the kun.uz news website. The training set included 4000 contexts from Uzbek articles, each paired with a corresponding summary. Similarly, the development set contained 1000 contexts with their respective summaries. We fine-tuned the models for the summarization tasks using this training set and a specific set of hyperparameters: a learning rate of 1e-4 and 10 training epochs.

Table 1. Generated summarize from kun.uz news.

Example 1	
Context	Yevropa Markaziy banki (YeMB) yuqori inflatsiyaga qarshi kurashda bazaviy foiz stavkalarini yana 0,25 foizga oshirdi. Tijorat banklari YeMBdan qarz olishi mumkin bo‘lgan asosiy stavka 4 foizga, depozitlar bo‘yicha bazaviy stavka esa 3,5 foizgacha oshirildi, deb xabar berdi regulyator 15 iyun, payshanba kuni. https://kun.uz/uz/news/2023/06/16/yevropa-markaziy-banki-bazaviy-foiz-stavkalarini-oshirdi
Target	Yevropa regulyatori inflatsiyaga qarshi kurash maqsadida bazaviy stavkalarni yana 0,25 foizga oshirdi.
mT5	O‘sishi mumkin bo‘lgan asosiy stavka 3,5 foizgacha oshiriladi; yalpi ichki mahsulotning narxi 3 foizga oshdi. Bu haqda ushbu bosqich yakuniga ko‘ra, xalqaro iqtisodiyot va kambag‘allikni qisqartirish maqsadida yuridik shaxslar ishtirokida e’lon qilingan.
uzT5	O‘sish bo‘yicha prognozlar joriy yilning mos davrida amalga oshiriladi. Bu haqda YeMB ma’lum qildi. "Yevro iqtisodiyotiga nisbatan inflatsiya darajasi ikkinchi bosqichda e’lon qilindi", - deyiladi xabarda. Qayd etilishicha, ushbu ko‘rsatkichning asosiy qismi – neft va gaz narxi – 1 foiz; yevro – 2 foizga yetib boradi.
Example 2	
Context	Prezidentlikka nomzod Shavkat Mirziyoyev o‘zining ilk saylovoldi tashviqotini Nukus shahrida, qoraqalpog‘istonlik saylovchilar qarshisida o‘tkazdi. Prezident o‘z nutqida qator yangiliklarni e’lon qildi. https://kun.uz/uz/news/2023/06/12/ochiq-budjetda-golib-bolish-uchun-15-ming-ovoz-yetarli-boladi
Target	Aholining mol-mulk va yer soliqlaridan qarzdorliklari bo‘yicha 500 milliard so‘m penya kechib yuboriladi,

mT5	Xalqaro moliya vazirligi matbuot xizmati xabariga ko‘ra, loyiha davomida iqtisodiyot va kambag‘allikni qisqartirish bo‘yicha qator yangiliklarni e’lon qildi. Bu haqda respublika prezidenti Jo Baydening ma’lum qilishicha, “Ochiq budjet” loyihasining g‘olib borilishi uchun 1,5 ming ovoz yetarli. Shuningdek, davlat rahbari Shavkat Mirziyoyev ushbu chora-tadbirlar amalga oshiriladi.
uzT5	Xalqaro moliya resurslarini joriy etish, unga ajratiladigan mablag‘larni 3 karra oshiradi. Loyiha amalga oshirilishi uchun 1,5 mingdan ortiq ovoz yetkazib beriladi. “Ochiq budjet” loyihasi g‘olib bo‘lishi mumkin.

To benchmark the two refined models, we assessed them using the development set. The models' performance was quantified using the evaluation package proposed by [12], which computes both BLEU (Bilingual Evaluation Understudy) [13] and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [14] metrics. BLEU assesses precision by determining the frequency of words in machine-produced summaries that match those in human reference summaries. Conversely, ROUGE evaluates recall by counting the instances of words in human-crafted summaries appearing in machine-made summaries.

Table 2. Performance comparison on summarize generation

Model	Size (GB)	Time (sec)	Rouge1	Rouge2	RougeL	BLUE
mT5	2,2	18	0,167	0,069	0,129	0,0275
uzT5	1	7,5	0,171	0,072	0,131	0,0295

Table 2 presents a detailed comparison between two models - mT5 and uzT5 - in terms of performance on the task of summary generation. The metrics include size (GB), time (sec), Rouge1, Rouge2, Rouge L, and BLUE score.

Model Size: The mT5 model, with a size of 2.2 GB, is significantly larger than the uzT5, which stands at 1 GB. Therefore, uzT5 has the advantage of being a lighter model, potentially making it more accessible and manageable for users with limited resources.

Processing Time: The mT5 model takes 18 seconds to process a summary generation task, more than double the time taken by the uzT5, which completes the task in 7.5 seconds. The uzT5 is thus substantially faster, demonstrating superior efficiency.

Rouge Scores: The Rouge1, Rouge2, and Rouge L scores are commonly used in NLP for assessing the quality of automatic summarization. These metrics show that the uzT5 slightly outperforms the mT5. Specifically, the uzT5 scores are 0.171, 0.072, and 0.131, respectively, compared to mT5 scores of 0.167, 0.069, and 0.129. Even though the difference is marginal, it suggests that uzT5 generates summaries that are marginally more similar to the reference summaries than those produced by the mT5.

BLUE Score: Similarly, the uzT5 surpasses the mT5 in terms of the BLUE score, which evaluates the quality of machine-generated text. The uzT5 achieves a score of 0.0295, slightly higher than the mT5's 0.0275. This indicates that uzT5 is likely to generate summaries that are marginally more coherent and semantically relevant than mT5's.

In summary, although both models perform similarly, uzT5 shows a marginal

advantage in terms of size, processing time, and the quality of generated summaries. It's worth noting that these differences are small and specific to this task of summary generation, so users may experience different results based on their specific needs and tasks.

4.2. Named Entity Recognition

Named Entity Recognition (NER) is a key component of Natural Language Processing (NLP) assigned to identify regions of text that contain references to entities. It is the process of identifying the informative part of data or applicable labels from unstructured data. In NER, data is gathered from unstructured data such as emails, blogs, newspapers, tweets, etc., to extract meaningful information.

Table 3. Named Entity Recognition from text.

1	
Context	Prezidentning tegishli farmoniga muvofiq 2023 yil 1 iyuldan boshlab O‘zbekiston Respublikasida «Jamoatchilik ekologiya nazoratchisi» tizimi joriy qilinadi.
Target	Prezidentning tegishli farmoniga muvofiq 2023 yil(DATE) 1 iyuldan(DATE) boshlab O‘zbekiston Respublikasida(LOC) «Jamoatchilik ekologiya nazoratchisi»(ORG) tizimi joriy qilinadi.
mT5	Prezidentning tegishli farmoniga muvofiq 2023 yil(DATE) 1 iyuldan boshlab O‘zbekiston Respublikasida «Jamoatchilik ekologiya nazoratchisi»(ORG) tizimi joriy qilinadi.
uzT5	Prezidentning(PER) tegishli farmoniga muvofiq 2023 yil(DATE) 1 iyuldan(DATE) boshlab O‘zbekiston Respublikasida(LOC) «Jamoatchilik ekologiya nazoratchisi»(ORG) tizimi joriy qilinadi.
2	
Context	Respublika Gidrometeorologiya markazi O‘zbekiston bo‘yicha 5 iyul, chorshanba kuni kuzatiladigan ob-havo ma’lumotini e’lon qildi.
Target	Respublika Gidrometeorologiya markazi(ORG) O‘zbekiston(LOC) bo‘yicha 5 iyul(DATE), chorshanba(DATE) kuni kuzatiladigan ob-havo ma’lumotini e’lon qildi.
mT5	Respublika Gidrometeorologiya markazi O‘zbekiston(LOC) bo‘yicha 5 iyul(DATE), chorshanba(DATE) kuni kuzatiladigan ob-havo ma’lumotini e’lon qildi.
uzT5	Respublika Gidrometeorologiya markazi(ORG) O‘zbekiston(LOC) bo‘yicha 5 iyul(DATE), chorshanba(DATE) kuni kuzatiladigan ob-havo ma’lumotini e’lon qildi.

Table 4. Comparative evaluation of models mT5 and uzT5

	mT5			**uzT5**		
Class	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
ORG	0,65	0,59	0,62	0,45	0,39	0,42
PER	0,78	0,46	0,58	0,85	0,74	0,79
LOC	0,61	0,53	0,57	0,71	0,77	0,74
DATE	0,86	0,41	0,56	0,78	0,61	0,68
TIME	0,92	0,46	0,61	0,94	0,66	0,78
NUMBER	0,41	0,26	0,32	0,54	0,54	0,54

Table 4 presents a comparative evaluation of two models, mT5 and uzT5, over several metrics, namely precision, recall, and f1-score across various categories like ORG, PER, LOC, DATE, TIME, and NUMBER.

Starting with the ORG category, the mT5 model demonstrates a precision of 0.65, a recall of 0.59, and an f1-score of 0.62. In comparison, the en_uzT5 model shows lower results in this category with a precision of 0.45, a recall of 0.39, and an f1-score of 0.42. Therefore, the mT5 model outperforms the en_uzT5 in the ORG category.

In the PER category, the mT5 model shows a precision of 0.78, a recall of 0.46, and an f1-score of 0.58. On the other hand, the uzT5 model performs significantly better with a precision of 0.85, a recall of 0.74, and an f1-score of 0.79. This implies that the uzT5 model is superior to the mT5 model in the PER category.

Moving onto the LOC category, the mT5 model possesses a precision of 0.61, a recall of 0.53, and an f1-score of 0.57. Contrastingly, the uzT5 model shows an improved performance with a precision of 0.71, a recall of 0.77, and an f1-score of 0.74, making the uzT5 model better in the LOC category than the mT5 model.

For the DATE category, the mT5 model exhibits a precision of 0.86, a recall of 0.41, and an f1-score of 0.56. Conversely, the uzT5 model displays a slightly lesser precision of 0.78 but an increased recall of 0.61 and an f1-score of 0.68, indicating a better performance by the uzT5 model in the DATE category.

In the TIME category, the mT5 model has a precision of 0.92, a recall of 0.46, and an f1-score of 0.61. Comparatively, the uzT5 model maintains a similar precision of 0.94 but a higher recall of 0.66 and an f1-score of 0.78, suggesting that the uzT5 model is more effective in the TIME category.

Lastly, in the NUMBER category, the mT5 model showcases a precision of 0.41, a recall of 0.26, and an f1-score of 0.32. The uzT5 model, in this case, outperforms with a precision of 0.54, an equal recall and f1-score of 0.54, demonstrating that the uzT5 model is superior in the NUMBER category.

Consequently, the uzT5 model surpasses the mT5 model in all categories except ORG, making it a more effective model overall for named entity recognition tasks.

V. CONCLUSION

A key benefit of employing a multilingual Transformer is its capability to process multiple languages simultaneously, negating the need for a unique model per language. This characteristic proves advantageous in scenarios demanding multi-lingual text processing. However, when a single language task is to be performed, the multilingual model falls short in efficiency. The extensive size of this model can impact memory distribution and processing speed, and memory constraints must be considered when deploying Transformers on public cloud platforms.

This article introduces the method to generate specific single-language models from multilingual Transformer models. With this approach, we can craft models tailored to individual languages, significantly smaller than the multilingual parent model. We validated this approach using the mT5-base Transformer model, extracting Uzbek and a minor part of English due to its frequent occurrence in Uzbek texts. The resultant model, reduced by half in size compared to the original multilingual model, showcases the potential of our method in creating more

streamlined models specific to certain languages.

We fine-tuned the derived smaller model and the mT5 models on two distinct natural language processing tasks - Automatic Text Summarization and Named Entity Recognition, to examine their relative performances. Interestingly, the trimmed-down model's performance was on par with the mT5 model. In addition, the smaller model consumed less memory, had quicker loading times, and faster inference times, despite no alterations to the model architecture. These findings highlight the promise held by this approach to produce more efficient and compact single-language models from multilingual ones.

REFERENCES

[1] *Xue, L., Constant, N., Roberts, A., and Raffel, C.* (2020). "mT5: A massively multilingual pre-trained text-to-text transformer." arXiv preprint arXiv:2010.11934.

[2] *Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J.* (2019). "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683.

[3] *Pires, T., Schlinger, E., and Garrette, D.* (2019). "How multilingual is multilingual BERT?." arXiv preprint arXiv:1906.01502.

[4] *Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L.* (2020). "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461.

[5] *Abdaoui. C. Pradel. and G. Sigel,* "Load Wliat You Need: Smaller Versions of Multilingual BERT," arXiv preprint arXiv:2010.05609, 2020.

[6] *S. Mehta, R. Koncel-Kedziorski. M. Rastegari, and H. Hajishirzi,* "Define: Deep factorized input token embeddings for neural sequence modeling," arXivpreprint arXiv: 1911.12385, 2019.

[7] *D. Dale.* "How to adapt a multilingual T5 model for a single language." https://towardsdatascience.com/how-to-adapt-a-iiiultilingual-t5-model-for-a-siiigle-language-b9f94f3d9c90.

[8] *D. Goldhahn. T. Eckart, and U. Quasthoff.* "Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages." in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012. pp. 759-765.

[9] *T. Wolf et al.,* "Huggingface's transformers: State-of-the-art natural language processing," arXivpreprint arXiv: 1910.03771, 2019.

[10] *A. Paszke et al.,* "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, 2019.

[11] https://github.com/abhimishra91/transformers-tutorials/blob/master/transformers_summarization_wandb.ipynb

[12] *S. Sharma, L. E. Asri, H. Schulz, and J. Zumer.* "Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation." arXiv preprint arXiv:1706.09799, 2017.

[13] *K. Papineni, S. Roukos. T. Ward, and W.-.T. Zhu*, "Bleu: a method for automatic evaluation of machine translation." in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311-318.

[14] *C.-Y. Lin.* "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004. pp. 74-81.

Поступила в редакцию 15.07.2023

Citation: *Adilova F.T., Davronov R.R., Kushmuratov S.I.* 2023. Tailoring mT5 for the Uzbek Language: A Compact Model for NLP Applications. *International Journal of Theoretical and Applied Issues of Digital Technologies*. 3(5): 7-16.

АДАПТАЦИЯ MT5 ДЛЯ УЗБЕКСКОГО ЯЗЫКА: КОМПАКТНАЯ МОДЕЛЬ ДЛЯ ПРИЛОЖЕНИЙ NLP.

Адилова Ф.Т.[1], *Давронов Р.Р.*[1], *Кушмуратов С.И.*[1]

[1] Академия наук Республики Узбекистан Математический институт имени В.И. Романовского, Ташкент, Узбекистан

fatadilova@matinst.uz, rifqat.davronov@mathinst.uz, bekmezonali@gmail.com

Аннотация. *Несмотря на то, что на нем говорят почти 50 миллионов человек, узбекский язык по-прежнему недостаточно представлен в системе обработки естественного языка (NLP). Одной из основных причин является ограниченная доступность узбекских лингвистических ресурсов. С ростом популярности архитектуры Transformer в NLP она обогнала более ранние методы, такие как сверточные и рекуррентные нейронные сети. T5 (преобразователь преобразования текста в текст) стандартизирует лингвистические задачи на английском языке, преобразуя их в формат преобразования текста в текст. mT5, его многоязычная версия, показала многообещающие результаты в различных задачах NLP, охватывающих несколько языков. Однако значительные размеры mT5 создают проблемы для приложений, ориентированных на один язык. В нашем исследовании мы доработали mT5 специально для узбекского языка, в результате чего модель T5 стала более компактной. Мы сравнили эффективность этой адаптированной модели с mT5 в задачах автоматического суммирования текста (ATS) и распознавания именованных сущностей (NER) с использованием идентичных протоколов и наборов данных. Наша адаптированная модель превзошла производительность mT5, что указывает на возможность разработки более компактной предварительно обученной модели почти вдвое меньшего размера без ущерба для результатов. Эта оптимизированная модель также выигрывает от меньшего использования памяти, более быстрого запуска и сокращения времени обработки. Чтобы получить доступ к этой модели, пожалуйста, свяжитесь с нами.*

Ключевые слова: *сжатие модели, преобразователь, предварительно обученная модель, автоматическое суммирование текста, распознавание именованных объектов.*