IRSTI 28.23.11 **https://doi.org/10.58805/kazutb.v.1.22-249**

TEXT GENERATION MODELS FOR PARAPHRASE ON KAZAKH LANGUAGE

A.M. Kassenkhan*, N.K. Mukazhanov, S. Nuralykyzy, Z.B. Kalpeyeva,
Satbayev University, Almaty, Kazakhstan,
e-mail: a.kassenkhan@satbayev.university

This study delves into the relatively unexplored domain of natural language processing for the Kazakh language- a language with limited computational resources. The paper dissects the effectiveness of diffusion models and transformers in generating text, specifically paraphrases, which is a critical aspect of machine learning applications such as chatbots, virtual assistants, and automated translation services.

The researchers methodically adapt these advanced models to understand and generate Kazakh text, tackling the unique challenges posed by the language's complex morphology. The paper is comprehensive in its approach, covering everything from the initial adaptation of the models to the Kazakh language context, to the creation of specialized tokenizer tools, to the translation and preparation of datasets for effective training.

Through rigorous testing and performance analysis, the study identifies the strengths and weaknesses of each model type. This is critical as it informs the direction of future research and model development, with the goal of enhancing the fluency and accuracy of automated Kazakh text generation. The paper also discusses the broader impact of its findings, suggesting that the methodologies and insights gained could inform similar efforts in other low-resource languages, thereby contributing to the global field of NLP.

The research concludes with reflections on the implications of their findings for the ongoing development of machine learning technologies, asserting the potential of these technologies to accommodate the intricacies of any language, given the right approach and resources. This work not only advances the technical capabilities for Kazakh text generation but also serves as a testament to the potential of machine learning to bridge language gaps and foster greater digital inclusivity.

Keywords: Diffusion models; Transformer models; Machine learning; NLP; Paraphrase; Kazakh language;

МОДЕЛИ ГЕНЕРАЦИИ ТЕКСТА ДЛЯ ПЕРЕФРАЗА НА КАЗАХСКОМ ЯЗЫКЕ

А.М. Касенхан*, Н.К. Мукажанов, С. Нуралыкызы, Ж.Б. Кальпеева
Satbayev University, Алматы, Казахстан,
e-mail: a.kassenkhan@satbayev.university

Данное исследование углубляется в относительно неисследованную область обработки естественного языка казахского языка - языка с ограниченными вычислительными ресурсами. В статье анализируется эффективность моделей диффузии и преобразователей при создании текста, в частности перефразирования, что является важнейшим аспектом приложений машинного обучения, таких как чат-боты, виртуальные помощники и службы автоматического перевода.

Исследователи методично адаптируют эти передовые модели для понимания и создания казахского текста, решая уникальные проблемы, возникающие из-за сложной морфологии языка. Статья носит комплексный подход и охватывает все: от первоначальной адаптации моделей к контексту казахского языка до создания специализированных инструментов токенизатора, перевода и подготовки наборов данных для эффективного обучения.

Благодаря тщательному тестированию и анализу производительности исследование выявляет сильные и слабые стороны каждого типа модели. Это имеет решающее значение, поскольку определяет направление будущих исследований и разработки моделей с целью повышения беглости и точности автоматического создания казахского текста. В документе также обсуждается более широкое влияние своих выводов, предполагая, что полученные методологии и идеи могут послужить основой для аналогичных усилий на других языках с ограниченными ресурсами, тем самым способствуя глобальному развитию НЛП.

Исследование завершается размышлениями о последствиях их выводов для продолжающегося развития технологий машинного обучения, утверждая потенциал этих технологий для решения сложностей любого языка при правильном подходе и ресурсах. Эта работа не только расширяет технические возможности создания казахского текста, но и служит свидетельством потенциала машинного обучения для преодоления языковых разрывов и содействия большей цифровой инклюзивности.

Ключевые слова: диффузионные модели, Модели-трансформеры, Машинное обучение, НЛП, Парафраз, казахский язык.

ҚАЗАҚ ТІЛІНДЕГІ ПАРАФРАЗАЛАРҒА АРНАЛҒАН МӘТІН ЖАСАУ МҮЛДЕРІ

А.М. Қасенхан*, Н.К. Мукажанов, С. Нуралықызы, Ж.Б. Кальпеева

Satbayev University, Алматы

e-mail: a.kassenkhan@satbayev.university

Бұл зерттеу қазақ тілі үшін табиғи тілді өңдеудің салыстырмалы түрде зерттелмеген саласын - есептеу ресурстары шектеулі тілді зерттейді. Бұл мақалада чат-боттар, виртуалды көмекшілер және автоматтандырылған аударма қызметтері сияқты машиналық оқыту қолданбаларының маңызды аспектісі болып табылатын мәтінді, атап айтқанда парафразаларды құрудағы диффузиялық модельдер мен трансформаторлардың тиімділігі қарастырылады.

Зерттеушілер тілдің күрделі морфологиясы тудыратын ерекше қиындықтарды шеше отырып, қазақ тіліндегі мәтінді түсіну және жасау үшін осы жетілдірілген үлгілерді әдістемелік тұрғыдан бейімдейді. Бұл мақалада үлгілердің қазақ тіліндегі контекстке бастапқы бейімделуінен бастап, арнайы токенизатор құралдарын құруға, тиімді оқыту үшін деректер жинақтарын аударуға және дайындауға дейін барлығы қамтылған.

Қатаң тестілеу және өнімділікті талдау арқылы зерттеу әрбір үлгі түрінің күшті және әлсіз жақтарын анықтайды. Бұл өте маңызды, өйткені ол автоматтандырылған қазақша мәтінді құрудың еркіндігі мен дәлдігін арттыру мақсатында болашақ зерттеулер мен модельдерді әзірлеу бағытын көрсетеді. Сондай-ақ, мақалада оның нәтижелерінің кеңірек әсері талқыланып, алынған әдістемелер мен түсініктер басқа аз ресурс тілдеріндегі ұқсас күш-жігерді хабардар ете алады, осылайша NLP-нің жаһандық өрісіне үлес қосады.

Зерттеу дұрыс тәсіл мен ресурстарды ескере отырып, кез келген тілдің қыр-сырын орналастыру үшін осы технологиялардың әлеуетін растай отырып, машиналық оқыту технологияларының үздіксіз дамуына олардың нәтижелерінің салдары туралы ой-пікірлермен аяқталады. Бұл жұмыс қазақ тіліндегі мәтінді құрудың техникалық мүмкіндіктерін дамытып қана қоймайды, сонымен қатар, тіл олқылықтарын жою және цифрлық инклюзивтілікті арттыру үшін машиналық оқытудың әлеуетінің дәлелі болып табылады.

Түйін сөздер: Диффузиялық модельдер, трансформатор үлгілері, машиналық оқыту, NLP, парафраза, қазақ тілі.

Introduction. Traditionally, machine learning frameworks for natural language tasks have been geared towards languages with extensive resources, such as English. Nonetheless, the necessity to tailor these sophisticated frameworks to languages with fewer resources is becoming increasingly crucial to broaden the reach of NLP technologies. While there have been initiatives to modify generative pre-trained models for languages with limited resources, the research on tailoring these models for the Kazakh language, specifically for text generation, remains scant. The application of a unified model to evaluate machine translation quality across numerous language pairs has showcased the scalability of this approach [1]. Still, its practical effectiveness and flexibility for adapting to new linguistic contexts, especially for a language with as few resources as Kazakh, are aspects that continue to pose significant inquiries.

The given study applies diffusion models and transformers to the task of text generation and paraphrasing in Kazakh, contrasting with another research [2] that uses ontologies and parsing trees for sentence classification. Both are centered on enhancing NLP resources for the Kazakh language. However, the distinction lies in the approaches: the

given research is geared towards generating new text, while the comparative study focuses on analyzing and understanding existing sentences.

This paper aims to bridge this gap by investigating the adaptation of two state-of-the-art generative models - Diffusion models and Transformers - for text generation in Kazakh. Diffusion models like Denoising Diffusion Probabilistic Models have shown promising results in high-quality and diverse text generation for English [2]. The study makes a valuable contribution to the field of natural language processing for the Kazakh and Turkish languages, offering tools for the determination of grammatical categories. Its strengths lie in the use of machine learning algorithms and extensive datasets, which are balanced by the complexities of language processing and potential limitations in the applicability of the algorithms [3]. Similarly, pretrained transformers fine-tuned on downstream tasks have dominated leaderboards across various NLP benchmarks [4]. Although there has been some work on paraphrasing datasets [5]. The research focuses on a fundamental aspect of sample-based machine translation: identifying the degree of similarity between sentences. This involves aligning an input sentence with a corresponding example from a database, selecting segments of this sentence, and subsequently adjusting or paraphrasing them to produce the intended translation [6]. The articles under review present the development of new linguistic and algorithmic solutions for information retrieval technologies within search systems, taking into account syntax and elements of semantics, including for Turkic texts [7]. The document provides a detailed description of the method for summarizing Kazakh texts [8], these studies do not solve the problem we address. Additionally, there are efforts to define semantically close words in the Kazakh language [9], and some initial work on Kazakh text generation using generative pre-trained transformers, thr research involves an empirical evaluation of text generation models for the Kazakh language, characterized by its limited resources and complex morphology [10]. The research examined the grammatical peculiarities of the Kazakh language [11]. However, none of these works comprehensively address the challenges of text generation in Kazakh, a low-resource, morphologically rich Turkic language.

We detail customizing techniques like developing specialized tokenizers and translating datasets to align these models with Kazakh's linguistic intricacies. Through meticulous experimentation, we compare the advantages and limitations of diffusion models versus transformers for sustained evolution of Kazakh text generation. Our primary contribution is pioneering the alignment of cutting-edge generative architectures with the specific needs of the Kazakh language. Findings will serve as benchmarks for future NLP advancements in similar low-resource contexts. Broader impacts include progressing text generation capabilities for understudied languages and highlighting considerations for multilingual model development.

Materials and methods. ***Analytical Techniques for Paraphrase Processing***. Given the absence of existing models and prior work on the topic of text paraphrasing, we chose to embark on this endeavor as young researchers in the field [2].

Our aim was to explore technologies and assess the feasibility of implementation, with the goal of attempting to create a model capable of addressing the challenges posed by text paraphrasing. This decision was motivated by the recognition of a research gap and the potential to contribute novel insights and advancements in the domain of text paraphrasing models. Our research pioneers the alignment of cutting-edge technologies with Kazakh's specific needs, contributing to machine learning discourse in less commonly studied languages. The initiative to delve into unexplored territories underscores the importance of pushing the boundaries of research and innovation to pave the way for advancements in natural language processing and text generation [12], a technology for constructing and visualizing the semantic image of a full-text document using ontology [13].

The utilization of diffusional models and transformers is pivotal for assessing the efficacy of methods in interacting with the most advantageous approach. Our selected models, Shark-NLP/DiffuSeq and chatgpt_paraphraser_on_T5_base, extend beyond mere paraphrasing, demonstrating utility in text summarization and more precise language translation to various languages, particularly leveraging T5[t5]. The multifaceted benefits of employing paraphrasing encompass enhancing understanding, clarity, simplification, incorporation of different styles, adaptation to diverse audiences, language improvement, summarization, avoidance of redundancy, and seamless integration of quotations. The exploration of these potentials provides compelling and rational motivations for their application.

DIFFUSEQ, tailored for SEQ2SEQ tasks, harnesses diffusion to enhance generation quality and diversity. Featuring a minimum Bayes risk decoding algorithm, it surpasses its counterparts in text generation quality

and diversity. Theoretical connections to AR and NAR models establish DIFFUSEQ as a robust extension of iterative-NAR models. Empirical results underscore its effectiveness, marking a significant stride in SEQ2SEQ learning. Notably, our findings indicate consistent performance improvements with larger models, aligning with the trend of more accessible and potent hardware. However, there are contexts, such as client-side inference or federated learning, where smaller models prove advantageous. Transfer learning emerges as a valuable tool for achieving optimal performance in low-resource tasks, advocating for research into methods delivering robust performance with cost-effective models [14].

The subsequent examination on English datasets raised a critical question regarding efficiency. Shark-NLP/DiffuSeq, QPP with a mere 144,000 text rows for paraphrasing [14], and chatgpt_paraphraser_on_T5_base, boasting over 6 million rows for the same task [15], exhibited proficiency. Importantly, diffusion models displayed optimization, requiring less data to function effectively. Transitioning to the evaluation on a Kazakh dataset, both models yielded unsatisfactory results. Tokenization issues specific to the Kazakh language led to a loss of efficiency and the semantic meaning of source sentences. Despite efforts to adjust settings, only 30% success was achieved, preserving the base semantic meaning in an average of 3.4 sentences out of 10. Paraphrasing efficiency remained unchanged, with no replacements in sentences of 5 words and minimal replacements in longer sentences. Diverse settings yielded paraphrased sentences, but the base meaning of the source sentence was invariably lost.

Further investigations revealed that temperature, repetitions penalty, and certain parameters had limited impact due to tokenization constraints. Length of sentence, num_beams, num_beams_group, and return_sequence was found to influence output. Amidst challenges, our focus shifted to identifying and rectifying these problems.

Dataset preparation and preprocessing constituted the initial step, unveiling a scarcity of high-quality Kazakh datasets [1]. Consequently, meticulous translation and linguistic pre-processing of English datasets were undertaken, ensuring adaptability for Kazakh paraphrasing. Despite challenges associated with a narrowly focused dataset sourced from social media platforms, rife with exotic jargon and shortened sentences, we persevered. After training a stable diffusion model selected for adaptation, unsatisfactory results ensued, as paraphrased sentences lost their basic meanings, with word replacements proving unnecessary and incorrect.

In the rapidly evolving landscape of machine learning models, the transformer architecture has emerged as a frontrunner in numerous natural language processing tasks. It has shown a particularly impressive capacity for paraphrasing in English. Given its proven competence, we deemed it prudent to adapt the transformer model, which has demonstrated remarkable results in English paraphrasing, for the Kazakh language.

Model Source. The transformer model we employed originates from the Hugging Face repository, which offers a plethora of pre-trained models. Our chosen model, specifically trained for English paraphrasing, provides a solid foundation upon which we sought to build our Kazakh paraphrasing model.

Google Colab, or Colaboratory, constitutes a free service provided by Google designed for machine learning training and data analysis research. It offers a cloud-based code execution environment built on Jupyter Notebook and provides resources for utilizing Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs). Several key features and capabilities of Google Colab are highlighted below:

Free Access to GPU and TPU: Colab grants free access to Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), rendering it convenient for high-performance machine learning model training.

Integration with Google Drive: Seamless import and export of data to and from Google Drive simplify data storage and sharing.

Support for Various Libraries: Colab comes pre-installed with a plethora of libraries such as TensorFlow, PyTorch, Keras, OpenCV, and others, streamlining the development and training of machine learning models.

Data Handling and Visualization: Colab facilitates easy uploading, analysis, and visualization of data using Python libraries like Pandas, Matplotlib, and Seaborn.

Collaborative Access and Real-time Collaboration: Users can invite others for real-time collaboration on projects, making Colab a convenient tool for teamwork.

The utilization of this platform enabled us to assess the efficacy of DiffuSeq in practical applications due to its minimal data requirements. With an available dataset suitable for paraphrasing and a structure similar to QPP, minor adjustments sufficed to align it with

our objectives. Following these configurations, the process progressed to tokenization, utilizing the same tokenizer employed by DiffuSeq. Creating a tokenized word list proved swift owing to the modest dataset size. A Kazakh tokenizer was introduced into the model, and subsequent training on a Kazakh dataset, taking approximately 4 hours, revealed an unexpected challenge. The QPP tokenizer proved incompatible as the pre-trained model persisted in generating English words in the output. Occasionally, the output featured combined sentences with words from both languages. This underscored the need for a language-specific tokenizer, prompting a reevaluation of our approach.

The generated output exhibited inaccuracies not solely due to the inadequately chosen tokenizer but also owing to linguistic peculiarities, including the following factors:

Morphological Characteristics: Complex Word Structure. The Kazakh language is characterized by a rich morphology, encompassing affixes (prefixes, suffixes), and flexional changes. This complexity in word structure, where a single Kazakh word may consist of multiple morphemes, poses challenges for tokenization, diverging from the simpler structure of English words.

Declension and Conjugation: The Kazakh language employs a system of declension and conjugation, complicating the extraction of root forms and their corresponding forms in a sentence. These changes often carry significant semantic load, and mishandling them can distort the meaning of the sentence.

Word Formation: Challenges in Defining Word Boundaries. Certain Kazakh words can be intricate, containing multiple lexemes. Defining boundaries between individual words becomes problematic, especially for tokenizers trained on languages with simpler word structures.

Lexical Characteristics: Synonymy and Homonymy: Kazakh may feature words with rich semantic nuances, synonymy, or homonymy, complicating the task of accurately selecting adequate replacements during paraphrasing. Understanding context becomes critically important for proper interpretation.

Ethnic and Cultural Specificities: The Kazakh language may include terms and expressions unique to the culture and history of Kazakhstan. Some of these expressions may lack direct analogs in other languages, adding complexity to the tasks of translation and paraphrasing.

Semantic Changes Based on Word Usage in Different Contexts: Polysemy. Certain Kazakh words may have multiple meanings depending on context. This complexity introduces challenges in determining the best translation or paraphrasing, as the choice must be conditioned not only by the word itself but also by the context of its usage.

Grammatical Characteristics: Word Order. Kazakh has its own word order, which may differ from English. For example, in Kazakh, word order can be free, and it is crucial to preserve syntactic structure during paraphrasing. Tokenizers and models must account for these nuances.

Lack of Definite Article: The Kazakh language does not use a definite article, which can also influence the tokenization and translation processes.

Diversity of Lexical Styles: Official and Informal Styles. Like many languages, Kazakh can exhibit different levels of formality and style depending on the context. Acknowledging this diversity is crucial during paraphrasing.

In light of these linguistic intricacies, our initial assumption that a generic tokenizer would suffice proved untenable. The need for a language-specific tokenizer became evident to address the linguistic idiosyncrasies inherent in the Kazakh language. This realization prompted a reevaluation of our approach, emphasizing the importance of linguistic considerations in the development of machine learning models for text generation in Kazakh.

After numerous attempts to rectify the identified issue proved unsuccessful, a decision was made to explore whether a transformer could address the task at hand. Among the models deemed suitable and of interest, the ”chat_gpt_paraphraser_based_on_T5” model emerged. Its notable feature lay in its capacity to handle translation tasks into French, Romanian, and German, in addition to possessing modules for detecting semantic meaning, akin to DiffuSeq. For the sake of time and computational efficiency, we selected only the module designed for sentence paraphrasing.

Upon executing the model on an attached dataset comprising 400,000 sentences, the model’s performance deviated from expectations [16] and the growing challenges in text processing that arise from the increasing volume of information on the Internet [17]. The discrepancy stemmed from the fact that the original model operated on a modified algorithm for generating paraphrased sentences, courtesy of the ChatGPT-2 algorithm. Due to this algorithm, the dataset structure transitioned from one text

corresponding to one array of 5 paraphrased versions of that text to one text aligning with an array of 15 paraphrased versions. The final dataset comprised a total of 6 million sentences.

Additionally, to expedite the translation process, a restructuring of the dataset was required. Specifically, each dataset entry contained 15 identical texts in the left column and 15 distinct paraphrased versions in the right column. The outcomes were consistent, utilizing identical hyperparameters for text generation as with DiffuSeq, and the functionality of the hyperparameters proved identical.

Dataset Acquisition. Initially, we procured a dataset [18] encompassing 140,000 rows of source-target pairs, where the source represented the original text and the target its paraphrased counterpart.

Challenges with the Initial Dataset. While our initial dataset offered a good starting point, we observed certain limitations during the preliminary stages of training. The scope of the dataset, limited to 130k rows, might not have captured the rich linguistic diversity and intricacies of the Kazakh language. Additionally, considering the complexity and capabilities of the transformer architecture, there was a potential risk of the model overfitting to this smaller dataset. These concerns compelled us to reconsider our data strategy.

Preparation of an Expansive Dataset. In our pursuit to refine the potential capabilities of our model, we delved into more extensive datasets. Our exploration led us to the Hugging Face platform, where we identified a dataset [19] that comprised an impressive 6 million rows of source-target pairs. We believed that translating and adapting this voluminous dataset to Kazakh would offer our model a broader linguistic spectrum, potentially augmenting its paraphrasing proficiency. However, due to time constraints, we did not proceed with training on this dataset. We propose that future researchers consider leveraging this enriched dataset to possibly attain superior results.

This necessitated the translation of the dataset to avoid utilizing incorrect data and attempting to replicate minimal results akin to the original model. For this task, a semi-automatic translation method utilizing machine translation was devised. This process, coupled with token creation for the dataset and subsequent training, spanned an entire week.

Translation and Dataset Refinement. The predominance of English in our selected datasets necessitated their translation into Kazakh. With meticulous attention to detail, we ensured that the semantic essence remained intact, thus upholding the data's contextual fidelity. Once translated, the datasets were subjected to thorough preprocessing, involving the rectification of irregularities and removal of superfluous data, making them primed for potential future training exercises.

Tokenization. Challenges in Existing Tokenizers for the Kazakh Language [20]. Our journey began with an exhaustive search for pre-existing tokenizers tailored for the Kazakh language. While there are numerous tokenization tools available for major languages, we quickly discerned a gap in robust, efficient tokenizers specifically engineered for the linguistic nuances of Kazakh. Many existing tools either lacked the precision required for our paraphrasing task or were not compatible with the transformer model's requirements.

Development of a Custom Tokenizer. Recognizing this void, we embarked on the path of creating our bespoke tokenizer. Our tokenizer's design was influenced by the Sentencepiece [6] tokenizer's underlying principles, renowned for its proficiency in handling a wide array of languages. By building upon the foundational concepts of Sentencepiece and incorporating specific adaptations for Kazakh, we endeavored to engineer a tokenizer that was both efficient and linguistically adept.

Contribution to the Community. Understanding the broader implications of our work and in the spirit of promoting open research, we have made our custom Kazakh tokenizer available to the public. It has been uploaded to the Hugging Face model hub, a prominent platform known for its extensive collection of machine learning models and tools. By doing so, we aim to bridge the existing gap in Kazakh language processing tools and hope that our contribution will assist researchers, developers, and linguists in their respective endeavors.

For those interested in leveraging our tokenizer or furthering its development, it can be accessed on the Hugging Face platform under our repository.

Model Training. Initial Training. Our preliminary training was undertaken on Google Colab Pro, tapping into the computational prowess of a V100 GPU with 16 VRAM. For this phase, which used the 140k row dataset, our chosen parameters were:

Batch Size (bsz): 22

Epochs: 3

Training Duration: Approximately 2 hours

Figure 1 illustrates the training loss trajectory of the 'rut5-multitask' model over the course of training.

Notably, the plot exhibits an exponential decrease in loss values up to 12,000 steps, suggesting rapid and substantial learning by the model during the initial training phase. This promising trend indicates the effectiveness of the chosen parameters and highlights the potential of the 'rut5-multitask' model for linguistic tasks.

Figure 1- Training loss of the 'rut5-multitask' model

Experimentation with 'rut5-multitask'. Post our initial training endeavors, we chanced upon a multi-language model for paraphrasing named 'rut5-multitask'. There was a study which offered a solution to the problem of summarizing texts in the Kazakh language, considering the process as two tasks: extracting the most important sentences and simplifying them. The TF-IDF method was used for extraction, and Seq2Seq neural network technology was utilized for sentence simplification. Due to the lack of a Kazakh dataset for training, transfer learning with a model trained on Simple English from Wikipedia was proposed. The main scientific contribution of the work is the transfer learning technology for simplifying Kazakh sentences using English language data [8]. Another study we looked at focuses on determining semantic similarity of words to enhance information retrieval tasks in the Kazakh language. This research employs vector representations of words and methods to calculate semantic closeness, with the aid of Apache Spark for distributed computing. It also details the use of pre-trained sentence transformers to grasp sentence-level semantics and speed up searches using semantic indexes. The results indicate that the proposed methods can be an effective solution for identifying semantically similar words and texts in the Kazakh language [9]. With keen interest, we experimented with this model and observed encouraging results right from the initial training phase. This presented a promising alternative to our previous approach and underscored the potential of utilizing multilanguage models for specific linguistic tasks.

Expanded Training. With the acquisition of the larger 6 billion row dataset, there was an evident need for enhanced computational resources. Hence, we migrated our training regimen to a server fortified with dual GPUs, each boasting 24 VRAM. The parameters adjusted for this expansive dataset were:

Batch Size (bsz): 64

Epochs: 2 (Projected)

Training Duration: Estimated around 30 hours (Note: Training is ongoing, and this section will be updated post completion)

Model Foundation. It's imperative to clarify that we did not train our model from scratch. Initially, our training process was anchored on the "chatgpt_paraphraser_on_T5_base" [21] model from

Hugging Face. This strategy allowed us to leverage the knowledge encapsulated in this pre-trained model, transitioning from a model fine-tuning approach rather than an exhaustive end-to-end training.

Training Observations. Throughout the training journey, we were astute in our monitoring of the model's loss trajectory. A discernible plateau in the loss served as an indicator that the model was possibly nearing its optimal state and might not gain significantly from protracted training. Recognizing this, we made the judicious decision to terminate the training. Such attentive oversight not only ensures judicious resource deployment but also mitigates the risk of model overfitting.

Consideration of Alternative Models. Diffusion Paraphrase Model for the Kazakh Language. During our literature review and exploration of potential models, we stumbled upon a promising diffusion-based model for paraphrasing [14], as detailed in a recently published paper. Accompanying the paper was an open-source Git repository, which provided a comprehensive walkthrough of their methodology and implementation.

However, while this diffusion model posed as an intriguing candidate, its computational demands significantly exceeded the resources at our disposal. The complexity of the diffusion model, combined with the intricacies of the Kazakh language, would necessitate substantial computational power to train effectively and efficiently. Given our constraints as students with limited access to high-end computational infrastructure, we had to make a pragmatic decision. Thus, while the diffusion model remained an enticing avenue for potential exploration, our current resources dictated that we prioritize the more feasible transformer architecture, which also had a proven track record in paraphrasing tasks.

It's worth noting that while our focus shifted to the transformer model, the diffusion model's potential merits in the context of the Kazakh language remain an area of interest. Future endeavors, especially with enhanced computational capabilities, could revolve around revisiting this model to ascertain its effectiveness in paraphrasing the Kazakh language.t.

Evaluation metrics:

To quantitatively evaluate the performance of our model, we leveraged three distinct similarity metrics: Cosine Similarity, Jaccard Similarity, and FuzzyWuzzy Similarity. Each metric was chosen to provide a multifaceted understanding of how the paraphrased content aligns with the original.

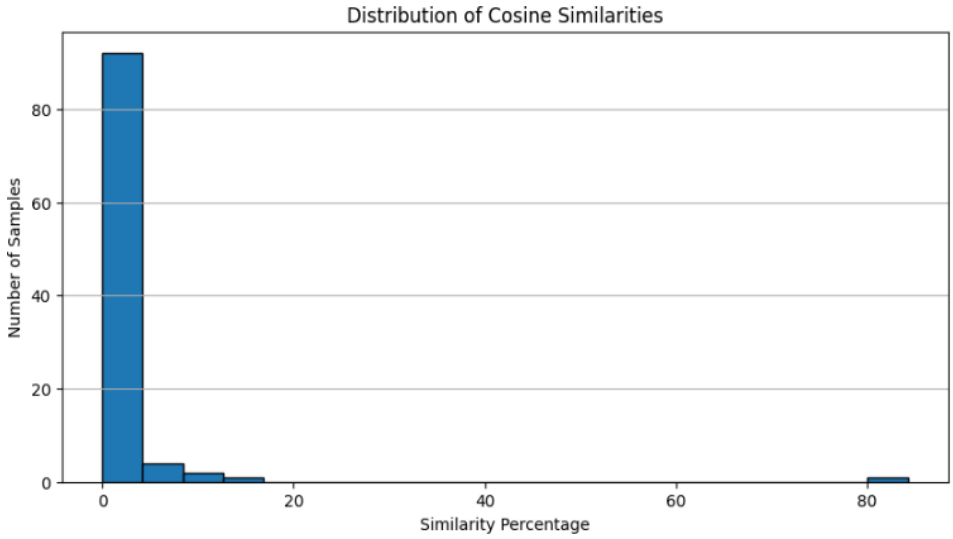

Figure 2 - Cosine Similarity between source sentences and model-generated responses from our test set of 100 examples

Cosine Similarity. Cosine similarity measures the cosine of the angle between two vectors in a multidimensional space, making it suitable for textual comparison when texts are represented as frequency vectors.

Text Vectorization: Utilizing TF-IDF, we transformed each text into a vector representation,

highlighting the importance of each word in relation to the overall corpus.

Similarity Calculation: We then calculated the cosine of the angle between the vectors of the original and paraphrased texts.

Figure 2 depicts the cosine similarity between source sentences and model-generated responses from our test set of 100 examples. The figure illustrates the computed cosine similarities between the vectors of original sentences and their paraphrased counterparts, providing insight into the model's performance in generating responses that align with the semantics of the source sentences.

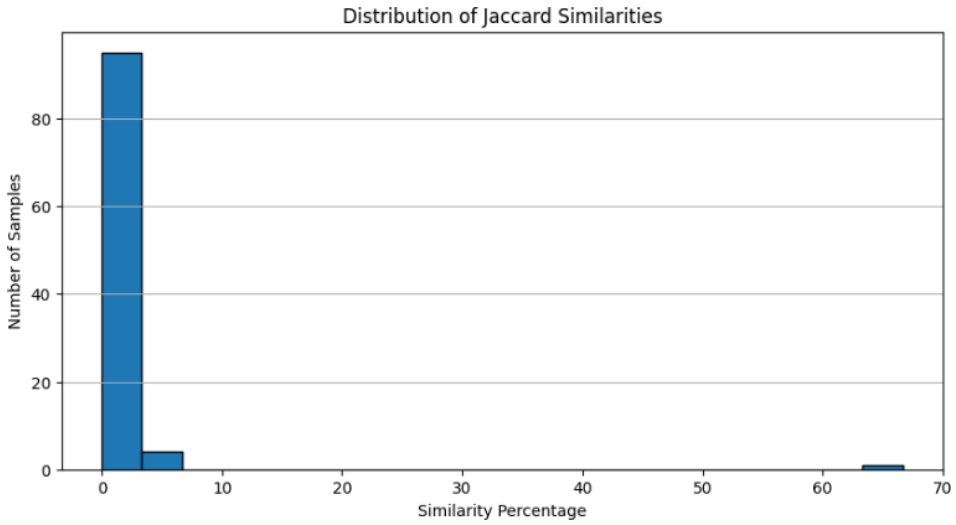

Figure 3 - Jaccard Similarity between source sentences and model-generated responses from our test set of 100 examples

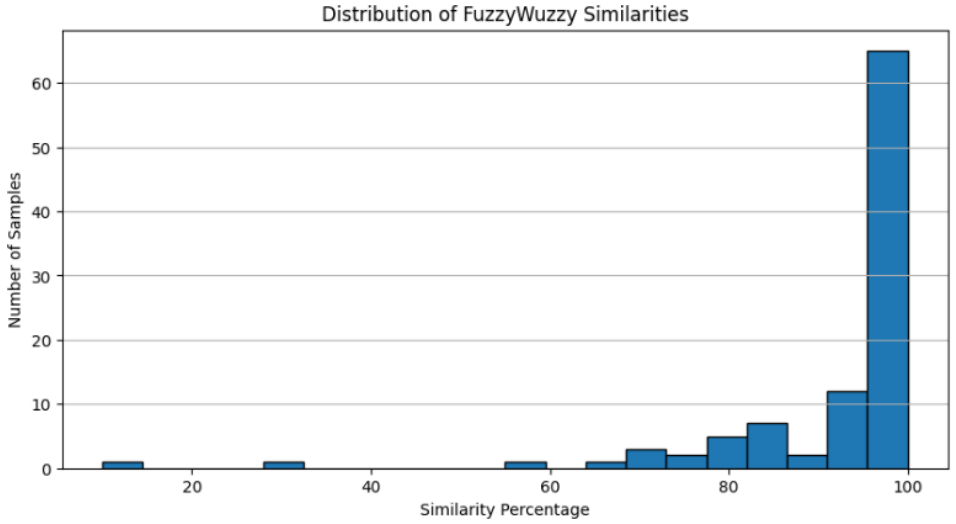

Figure 4 - Similarity between source sentences and model-generated responses based on FuzzyWuzzy Ratio from our test set of 100 examples

Jaccard Similarity . Jaccard Similarity calculates the ratio of intersection over union for two sets, offering insights into their resemblance. Word Splitting: Texts are partitioned into individual word sets. Similarity Calculation: Employing the Jaccard formula, we quantified the overlap between the word sets of the original and paraphrased texts. Figure 3 presents the Jaccard Similarity, a measure derived from the ratio of intersection over union for two sets, revealing their likeness. Our approach involves breaking down texts into individual word sets, and using the Jaccard formula, we assess the overlap between the word sets of both the original and paraphrased texts. This method sheds light on the semantic alignment between source and generated texts.

FuzzyWuzzy Similarity . Grounded on the Levenshtein distance, FuzzyWuzzy gauges the edits required to transform one string into another, reflecting their similarity. Distance Calculation: We determined the edit distance between the original and paraphrased texts. Normalization: The resultant distance was normalized to yield a percentage score, indicative of textual similarity.

Discussion and Results. ***TA. Model Performance.*** Model Implied Hyper-parameters for Paraphrasing. For our paraphrasing model, several hyperparameters were meticulously tuned to optimize the quality of the output. Herein, we elucidate the significance of each:

- num_beams (5) - Beam search is an algorithmic optimization used during inference to explore multiple possible output sequences simultaneously and select the best one. num_beams specifies the number of beams or paths that the algorithm will explore concurrently. Increasing the number of beams typically enhances the quality of model outputs at the expense of computational time;

- num_beam_groups (5) - This parameter divides the beams into groups and ensures that each group is diverse from the others. It helps in diversifying the generated outputs. A higher number of beam groups can lead to more diverse paraphrasing outputs;

- num_return_sequences (5) - Specifies the number of sequences or outputs the model should return for every input. Helpful in generating multiple paraphrase options for a given input, providing a varied set of alternatives;

- repetition_penalty (10.0) - Penalizes words that are repeated in the generated output. A higher penalty value discourages the model from producing repetitive words or phrases in the paraphrased text;

- diversity_penalty (3.0) - Encourages the production of diverse tokens by adding a penalty for generating similar tokens within the beam groups. Ensures that the returned paraphrases within the beam groups are different from one another;

- no_repeat_ngram_size (2) - This parameter ensures that no n-grams (in this case, 2-grams or bigrams) are repeated in the generated output. Prevents over-repetition of certain word combinations, enhancing the fluency and diversity of the output;

- temperature (0.7) - A scaling factor applied to the logits before softmax during inference. It adjusts the distribution from which tokens are sampled. Higher values increase randomness in token selection, whereas lower values make the model's decisions more deterministic;

- max_length (128) - The maximum number of tokens in the output sequence. Constrains the length of the generated paraphrase, ensuring it doesn't become excessively verbose;

B. Model Trained on rut5-multitask

In our quest to further refine our paraphrasing capabilities, we turned our attention to the 'rut5-multitask' model. Upon training, we subjected the model to evaluation using diverse Kazakh sentences to ascertain its ability to generate appropriate paraphrases.

For a clearer understanding, we present a comparative analysis between the original sentence, the expected (or reference) paraphrased answer, and the output generated by our model. A representative snapshot of our findings is illustrated in Table 1.

It's evident from the results presented in Table 1 that while our model has shown competency in generating paraphrased content, there remains room for improvement. Instances where the model output aligns closely with the original sentence highlight the potential challenges and limitations of the paraphrasing process.

We aim to delve deeper into these results, analyzing instances of accurate paraphrasing against instances of repetition or minor deviations. By dissecting these results, we hope to shed light on areas of potential enhancement and avenues for future research.

Amidst the diverse range of model variations subjected to training on all four tokenizers, it is noteworthy that the obtained results demonstrated a remarkable consistency under identical hyperparameter configurations. Notwithstanding, a notable exception emerged with the humarin [5] tokenizer. This particular tokenizer, uniquely configuring both models for

English text generation, underscored the nuanced impact that tokenization choices can exert on model behavior and output.

Table 1- Comparative analysis of original sentences, reference paraphrases, and model outputs

Source	Target	Answer
Консервативтік коммунистер блогы 239 орынға ие болды; Ұлттық кеңеске айналған Демократиялық блокта 125 депутат болды.	Консервативтік коммунистік топта 239 орын болды, ал бұрын Демократиялық блок деп аталатын Ұлттық кеңесте 125 өкіл болды.	Консервативтік коммунисте Маг блогы 239 орынға ие Үндістанның олды; Ұлттық сарбаздар еңеске аладыайнал Сицилияан Демокра 2011 шы 125 депута әлемде Тапсырыс о әдістерід команда.
Бұл оқиға грузин саясатын түбегейлі өзгертіп, көпшілікті Кеңес өкіметі жалғасқаннан гөрі тәуелсіздік артық деп тұжырымдауға итермеледі.	Осы оқиғаға байланысты Грузияның саясаты радикалдануға ұшырады, бұл көптеген адамдардың тәуелсіздік Кеңес өкіметі кезінде қалудан гөрі жақсы нұсқа екеніне сенді.	Бұл оқиға грузин әлемде түбегейлі өзганртіп, көпшілсkңізбен Кеңес аборигендеркімет Қайсы жалғдіңсқанна ала ретіндеөрі аладытәуелсіздік тұжырымдауғ жасай.
14 сәуірде Горбачев Джумбер Патиашвилиді Грузия Коммунистік партиясының бірінші хатшысы қызметінен алып, оның орнына Грузияның бұрынғы КГБ басшысы Гиви Гумбаридзені тағайындады.	Горбачев орнына келді. Джумбер Патиашвили Грузия Коммунистік партиясының бірінші хатшысы ретінде Грузияның бұрынғы КГБ басшысы Гиви Гумбаридземен бірге 14 сәуірде.	Сәуірде Горбачев Джумбер Патиашвилиді биіктік рузия Коммунистік парт қайсысыясының ханіңшыс- қызмет Қайсын алып, он артық қайған сәттіна саласындаршыз бөртпеян.

The apex of our outcomes materialized through the utilization of the final tokenizer, as explicated in the narrative of this study. The intrinsic congruence observed in the results across models sharing identical datasets further emphasizes the reproducibility and stability of our findings. Such uniformity not only substantiates the reliability of our approach but also suggests the intrinsic characteristics of the chosen tokenizer and dataset interplay significantly in shaping model performance.

In the broader context of machine learning model training, our exploration of machine translation techniques revealed nuanced considerations. The endeavor to transpose an English dataset into Kazakh, while not deemed an optimally effective strategy for model training, was recognized as a pragmatic recourse in the face of a paucity of substantial, high-quality Kazakh datasets. It is crucial to underscore that the appropriateness of such translation methodologies is contingent upon the nature of the target domain. In scenarios characterized by the mundane and commonplace, the use of translation services like Yandex/Google Translate introduces minimal distortions. However, within specialized domains such as medicine, scientific discourse, and politics, the imperative for reliance solely on manually translated datasets becomes apparent, acknowledging the intricacies and specificities inherent to these domains.

Computational resources and the diversity of existing models emerge not as deterministic factors but rather as moderating influences. Their impact is contingent upon the proficiency and acumen of the model operator. In the absence of a nuanced understanding of the intricacies of less efficient models, misguided impressions and fruitless endeavors may ensue. Practical considerations, encompassing the available time and the uninterrupted continuity of model training operations, should be weighed judiciously. This deliberation, naturally, presupposes a commensurate level of resources akin to those at the disposal of undergraduate students.

The central tenet of our endeavor was to illustrate that models of this nature necessitate adaptation not only to widely spoken languages but can, with concerted efforts, be extended to less popular languages. This endeavor is particularly noteworthy given the context of our study, executed by a group of final-year bachelor students. It underscores the potential democratization of advanced natural language processing (NLP) techniques, emphasizing the importance of accessible avenues for linguistic diversity.

The successful adaptation of such models to less common languages, as demonstrated in our study,

speaks to the viability of these endeavors even within resource-constrained settings. While our study serves as a testament to the feasibility of adaptation by bachelor students, it is crucial to recognize that the process can be further optimized and expanded with the integration of cutting-edge technologies, enhanced computational resources, and the expertise of more qualified and certified practitioners in the field of NLP.

Leveraging specialized technologies and substantial computational power, along with the guidance of seasoned NLP experts or accomplished professionals in the medical domain, this process can be comprehensively realized. The infusion of advanced tools and the expertise of certified NLP professionals can elevate the adaptation of such models to lesser-known languages, ensuring not only linguistic accuracy but also semantic fidelity. This collaborative approach, merging technological capabilities with expert insights, promises a more nuanced and sophisticated integration of advanced language models, thereby expanding their applicability to a broader linguistic spectrum.

In the culmination of our scientific pursuits, the attainment of our primary objective- the construction of a model proficient in sentence paraphrasing- is a testament to the systematic exploration of various models and datasets. Situating our research within the broader narrative of advancing machine learning applications in linguistic tasks, particularly for languages grappling with limited available resources, positions our findings as a meaningful contribution to the evolving landscape of natural language processing.

Conclusion. In conclusion, our research endeavours centred on the development and adaptation of machine learning models for paraphrasing in the Kazakh language. Through a meticulous exploration of various models and datasets, we achieved our primary objective of constructing a model proficient in sentence paraphrasing. Our findings underscore the adaptability of such models to less popular languages, exemplified here by Kazakh, and the potential democratization of advanced natural language processing techniques.

This study, conducted by a group of final-year bachelor students, demonstrates the feasibility of undertaking complex NLP tasks with limited resources. The success achieved in the adaptation of models to the Kazakh language highlights the potential impact of similar endeavors in linguistically diverse contexts.

Moreover, the study emphasizes that while the adaptation of models by bachelor students is feasible, further optimization and scalability can be achieved with the integration of advanced technologies, increased computational resources, and the guidance of more experienced NLP experts. This collaborative approach holds promise for enhancing linguistic accuracy and semantic fidelity in the adaptation of machine learning models to lesser-known languages.

Our study has delved into the realm of machine learning for paraphrasing in the Kazakh language, employing two semi-automatic translation algorithms- QPP and chat_gpt_paraphraser_T5. Both models demonstrated commendable performance on the paraphrase task, paving the way for efficient adaptation to various languages beyond English.

The creation of two translated datasets, encompassing 140k sentences and an extensive 6 billion sentences, underpins the adaptability of our approach to datasets of varying scales. This versatility is instrumental in accommodating the linguistic nuances of different languages, a crucial factor in the effectiveness of paraphrasing models.

Our study encapsulates a comprehensive exploration of machine learning for paraphrasing, emphasizing adaptability, linguistic diversity, and methodological rigor. By bridging the gap between English-centric models and lesser-known languages, we contribute to the ongoing evolution of natural language processing technologies, fostering inclusivity and accessibility across diverse linguistic landscapes.

As we transition beyond the English-centric paradigm, the inclusion of two robust paraphrasing models and the translation of datasets from English to Kazakh mark significant strides toward linguistic inclusivity. The adaptation of models and datasets for lesser-known languages aligns with the broader narrative of democratizing advanced natural language processing technologies.

To gauge the unique generation capability of our models, we employed four evaluation metrics, providing a multifaceted assessment of paraphrased outputs. This methodological rigor enhances the reliability of our findings and contributes to the robustness of our study.

Looking forward, others future research could delve into refining the adaptation process, exploring more extensive datasets, and investigating the specific challenges posed by languages with intricate linguistic structures. Additionally, the integration of domain-specific expertise, such as medical or scientific knowledge, could further enhance the performance of paraphrasing models in specialized contexts.

In essence, our study contributes to the broader discourse on the applicability of machine learning models in linguistic tasks for languages with limited resources. It opens avenues for further research, encouraging the exploration of innovative solutions and collaborative efforts in advancing the accessibility and effectiveness of natural language processing technologies across diverse linguistic landscapes.

References

1.Thompson, B., & Post, M. (2020). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).- pp. 90-121. https://doi.org/10.18653/v1/2020.emnlp-main.8

2.Yelibayeva, G., Sharipbay, A., Mukanova, A., & Razakhova, B. (2020, September). Applied ontology for the automatic classification of simple sentences of the Kazakh language. In 2020 5th International Conference on Computer Science and Engineering (UBMK) .- pp. 13-18.- IEEE.

3.Yerimbetova, A., Tussupova, M., Sambetbayeva, M., Turdalyuly, M., & Sakenov, B. Grammatical categories determination for Turkish and Kazakh languages based on machine learning algorithms and fulfilling dictionaries of link grammar parser.- Eastern-European Journal of Enterprise Technologies.- 2021.-Vol. 5(2). - pp. 113.

4.Lu, X.-Q., Ren, F., Huang, Z.-D., & Yao, T.-S. Sentence similarity model and the most similar sentence search algorithm. Dongbei Daxue Xuebao.-Journal of Northeastern University.-2003.- Vol. 24(6).- pp. 531-534. - Retrieved from

https://www.researchgate.net/publication/289662239_Sentence_similarity_model_and_the_most_similar

_sentence_search_algorithm

5.Leippold, M. Sentiment spin: Attacking financial sentiment with GPT-3. Finance Research Letters. - 2023.- Vol.55.- Article 103957. https://doi.org/10.1016/j.frl.2023.103957

6.Kamanur, U., Sharipbay, A., Altenbek, G., Bekmanova, G., & Zhetkenbay, L. (2016, October). Investigation and use of methods for defining the extends of similarity of Kazakh language sentences. In Proceedings of the China National Conference on Chinese Computational Linguistics International Symposium on Natural Language Processing Based on Naturally Annotated Big Data/ - 2016, October.- pp.14 - 17.https://doi.org/10.1007/978-3-319-47674-2_14

7.Yerimbetova, A. S., Sagnayeva, S. K., Murzin, F. A., & Tussupov, J. A. Creation of tools and algorithms for assessing the relevance of documents. -2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). https://doi.org/10.1109/rpc.2018.8482202

8.Zhabayev, T., & Tukeyev, U. Development of technology for summarization of Kazakh text. International Journal of Advanced Computer Science and Applications.- 2021.-Vol. 12(9). -pp.11-116.

9.Ayazbayev, D., Bogdanchikov, A., Orynbekova, K., & Varlamis, I. Defining semantically close words of Kazakh language with distributed system Apache Spark. Big Data and Cognitive Computing.- 2023.-Vol. 7 (4). - pp. 2 - 13.

10.Tolegen, G., Toleu, A., Mussabayev, R., Zhumazhanov, B., & Ziyatbekova, G. Generative Pre-Trained Transformer for Kazakh text generation tasks. In 2023 19th International Asian School-Seminar on Optimization Problems of Complex Systems (OPCS).- 2023, August - pp. 1-5. IEEE.

11.Zura, D., & Doyle, W. J. A grammar of Kazakh. Durham: Duke University, Duke Center for Slavic.- Eurasian, and East European Studies.- 2018. - 69 p.

12.Kasekeyeva, A. B., Batura, T. V., Efimova, L. V., Murzin, F. A., Tussupov, J. A., Yerimbetova, A. S., & Doshtayev, K. Zh. (2020). Link grammar and formal analysis of paraphrased sentences in a natural language. Journal of Theoretical and Applied Information Technology.- 2020.- Vol. 98(10). - pp. 1724-1736.

http://www.jatit.org/volumes/Vol98No10/10Vol98No10.pdf

13.Sizykh, O. V., Zhelobtsova, S. F., Barashkova, N. N., Burtseva, M. A., & Zhelobtsov, F. F. Problems and literary characters in the world prose of the 20-21st century: I. S. Shmelev, D. Setterfield, Su Tong. Indian Journal of Science and Technology.-2016.- Vol. 9(20).- pp.1-7.

14.Gong, S., Li, M., Feng, J., Wu, Z., & Kong, L. (2023). DiffuSeq-v2: Bridging discrete and continuous text spaces for accelerated Seq2Seq diffusion models.- Singapore.-2023.- pp.9868-9875.- arXiv preprint arXiv:2310.05793.

15.Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. - mT5: A massively multilingual pre-trained text-to-text transformer.-2021.- pp.483-498. 10.18653/v1/2021.naacl-main.41

16.cointegrated. (2021). rut5-base-multitask [Model repository]. Hugging Face Model Hub. Retrieved [15.11.2023], from https://huggingface.co/cointegrated/rut5-base-multitask

17.Batura, T. V., Murzin, F. A., Semich, D. F., Sagnayeva, S. K., Tazhibayeva, S. Z., Bakiyev, M. N. Using the link grammar parser in the study of Turkic languages. Eurasian Journal of Mathematical and Computer Applications.- 2016.-Vol.4(2). - pp. 14-22.

https://doi.org/10.32523/2306-6172-2016-4-2-14-22

18. CCRs. (n.d.). small-chatgpt-paraphrases-kz [Dataset]. Retrieved from https://huggingface.co/datasets/CCRs/small-chatgpt-paraphrases-kz

19.Kudo, T., & Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations Brussels, Belgium: Association for Computational Linguistics.- 2018.- pp. 66-71

20.CCRss. (n.d.). tokenizer_t5_kz [Software]. Hugging Face. Получено с https://huggingface.co/CCRss/tokenizer_t5_kz

21.Pan, X., Chen, Z., & Komachi, M. Query generation using GPT-3 for CLIP-based word sense disambiguation for image retrieval. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (StarSEM 2023), co-located with ACL. - Toronto, Canada.- 2023.- pp. 417-422. https://aclanthology.org/2023.starsem-1.36.pdf

Information about the authors

Kasenkhan A.M. - doctor Ph.D.,Satbayev University, Almaty, Kazakhstan,e-mail: a.kassenkhan@satbayev.university;

Mukazhanov N. K. - doctor Ph.D., associate professor Satbayev University, Almaty, Kazakhstan, e-mail:

n.mukazhanov@satbayev.university;

Nuralykyzy S. - senior lecturer, Satbayev University, master of Technical Sciences, Almaty, Kazakhstan, e-mail:

s.nuralykyzy@satbayev.university;

Kalpeeva Zh. B. - doctor Ph.D., associate professor, Satbayev University, Almaty, Kazakhstan, e -mail:

z.kalpeyeva@satbayev.university.

Сведения об авторах

Қасенхан А.М - доктор Ph.D, Satbayev University, Алматы, Казахстан,

e-mail: a.kassenkhan@satbayev.university;

Мукажанов Н. К. - доктор Ph.D., асс. профессор Satbayev University, Алматы, Казахстан,

e -mail: n.mukazhanov@satbayev.university;

Нуралыкызы С.- Satbayev University, магистр технических наук, Алматы, Казахстан,

e-mail: s.nuralykyzy@satbayev.university;

Кальпеева Ж. Б. - доктор Ph.D., ассоциированный профессор, Satbayev University, Алматы, Казахстан,

e-mail: z.kalpeyeva@satbayev.university