UDC: 004.65, 8, 81’2, 81’28

MARKOV MODEL FOR TEXT GENERATION IN UZBEK: PRACTICAL APPLICATION AND RESULTS

Sharipov Maksud Siddiqovich,
Candidate of Technical Sciences, Docent
maqsbek72@gmail.com
Urgench State University

Kurbonova Ruzikajon Ulug‘bek qizi,
ruzikajonqurbanova0809@gmail.com
PhD student, Urgench State University

Annotatsiya: Ushbu maqolada o‘zbek tilida matn generatsiyasi uchun Markov modeli qo‘llanilishi o‘rganilgan. Tadqiqotda Markov modelining xususiy holati trigram modeli asosida matn yaratish jarayoni va natijalari tahlil qilingan. O‘zbek tilida hali bu borada yetarlicha ishlar amalga oshirilmagan. Eksperimentlar natijasida Markov modelining matn generatsiyasida mantiqiy bog‘liqlikni saqlashda cheklovlari borligi aniqlangan. Modelning aniqlik darajasi 0.73% bo‘lib, generatsiya qilingan matnlarning grammatik sifatini oshirish uchun N-gram uzunligini oshirish yoki chuqur o‘rganish (LSTM, Transformer) modellari bilan taqqoslash zarurligi qayd etilgan. Maqola Markov modelining o‘zbek tilida matn yaratishda imkoniyatlari va cheklovlarini tahlil qiladi.

Аннотация. В данной статье исследуется применение марковской модели для генерации текста на узбекском языке. В рамках исследования был проанализирован процесс создания текста на основе частного случая марковской модели — триграммной модели, а также её результаты. В узбекском языке пока еще недостаточно работы в этом направлении. Эксперименты показали, что марковская модель имеет ограничения в сохранении логической связности при генерации текста. Точность модели составила 0.73%, что указывает на необходимость увеличения длины N-граммы или сравнения с моделями глубокого обучения (LSTM, Transformer) для улучшения грамматического качества сгенерированных текстов. Статья анализирует возможности и ограничения марковской модели в контексте генерации текста на узбекском языке.

Abstract. This article explores the application of the Markov model for text generation in the Uzbek language. The study analyzes the process and outcomes of text generation using a specific case of the Markov model – the trigram model. Not enough work has been done in this regard in the Uzbek language yet. Experimental results revealed limitations of the Markov model in preserving logical coherence during text generation. The model’s accuracy rate was 0.73%, highlighting the need

to either increase the N-gram length or compare it with deep learning architectures (e.g., LSTM, Transformer) to improve the grammatical quality of generated texts. The paper critically examines the potential and limitations of Markov models for text generation in Uzbek.

Kalit so‘zlar: *Markov modeli, matn generatsiyasi, o‘zbek tili, tabiiy tilni qayta ishlash (NLP), N-gram modeli, trigram, perplekslik, aniqlik darajasi.*

Introduction

Today, the widespread use of information systems and the internet across various fields has led to the generation of massive volumes of data. Consequently, the need for automated processing and searching of multilingual data is growing rapidly. Text generation, a critical branch of natural language processing (NLP), involves creating logical and coherent text based on provided data. In recent years, advancements in deep learning technologies have revolutionized text generation, enabling the production of human-like texts with unprecedented accuracy.

Natural language processing (NLP) is one of the key directions in the field of text generation. While numerous studies have been conducted earlier for languages such as English, French, and others, there is still insufficient scientific research on text generation for the Uzbek language. The Markov model is one of the simplest and most effective approaches used in natural language processing, enabling the generation of new text based on a probabilistic chain derived from a given input. The aim of this article is to explore the potential of Markov models for text generation in Uzbek, evaluate the results, and analyze its advantages and limitations.

Literature Review

In many studies involving morphologically rich languages, deep learning models such as LSTM and Transformers have outperformed traditional n-gram approaches in terms of semantic coherence and grammatical accuracy. For example, Transformer-based architectures like BERT and GPT have shown state-of-the-art performance in text generation tasks.

Today, numerous researchers worldwide are conducting studies on automated text generation using natural language processing (NLP) tools. For instance, A.M. Kassenkhan proposed adapting two modern generative models – Diffusion models and Transformers – for text generation in the Kazakh language, demonstrating the effectiveness of Transformer models in generating paraphrased text [1]. Meanwhile, Junyi Li and Tianyi Tang presented a comprehensive analysis of using Pre-trained Language Models (PLMs) for text generation. They highlighted three key aspects of applying PLMs:

1. Encoding input expressions into forms that preserve semantic meaning and integrating them into PLMs;

2. Designing PLMs that function effectively as generative models;
3. Efficiently fine-tuning PLMs based on reference texts to ensure unique characteristics in generated outputs.

They also identified major challenges and solutions related to these aspects [2].

In text generation research, statistical and neural approaches dominate. Statistical methods, such as N-gram models, generate text based on sequential probabilities. While these methods are efficient due to their simplicity, their limited ability to handle long-range context makes achieving accurate results challenging [3], [4]For example, N-gram language models – a straightforward and intuitive approach – have been applied to text generation in Uzbek, though with notable limitations [3].

Neural approaches, on the other hand, leverage deep learning architectures [5], [6]. Transformer models like BERT and GPT, trained on large datasets, produce high-quality text by capturing contextual information. These models excel at learning semantic and syntactic features of language, which is critical for managing complex grammatical structures in Uzbek [6].

Research on the Uzbek language remains in its early stages[7]. Current efforts focus on tasks such as corpus preparation and text generation using N-gram models. Preliminary experiments adapting BERT-like models for Uzbek have shown promising results [4], [8].

Research Methodology

In this study, we analyzed the process of text generation in the Uzbek language using a Markov model.

A Markov model assumes that the next event in a sequence depends only on the current state. In text generation, this means that a word’s occurrence depends on the previous N words.

Markov models are often implemented using N-gram models. Examples include:

- Bigram (n=2): "men bugun" → "kitob"
- Trigram (n=3): "men bugun kitob" → "o‘qidim"

To build a functional Markov model, a large Uzbek text corpus is required.

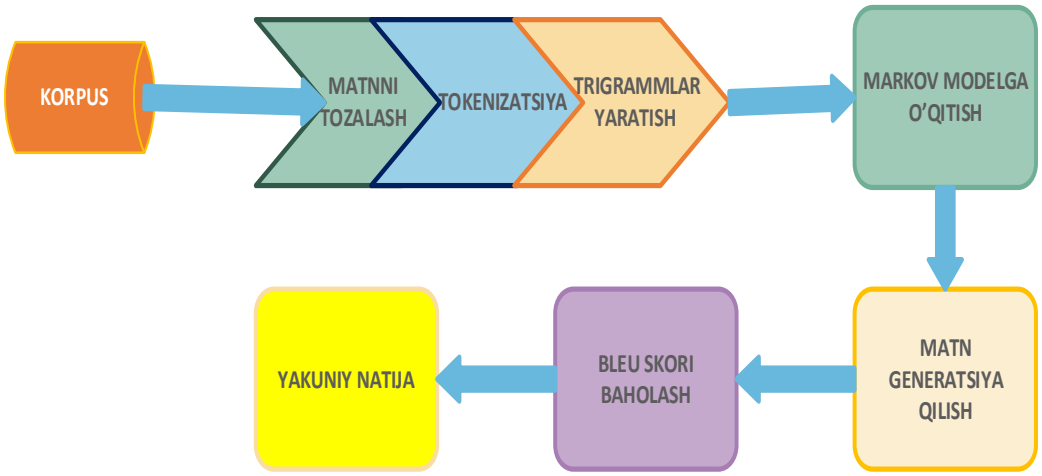

Figure 1: Stages of initial processing of language corpus texts

The following methodological approach was employed:

1. Data Collection and Preparation

- Corpus Selection: The "odilyakubov.txt" text corpus was chosen for experiments.
- Text Cleaning:
  - Removed extra spaces, special characters, and standardized uppercase/lowercase distinctions.
- Tokenization: The text was tokenized (split into words).

2. Model Training

- Model Selection: A trigram model ($n=3$) was implemented.
- Training Data: The entire corpus was used as the training set.
- Training Process: The Markov model was trained on this dataset to predict subsequent words based on probabilistic transitions.

3. Text Generation

- Input: Starting words were provided as input.
- Output: The model generated new text sequentially from the starting words.
- Evaluation: Generated texts were assessed for grammatical correctness and logical coherence.

Key Findings

- Capabilities: The Markov model demonstrated basic feasibility for generating Uzbek text.
- Limitations:
  - Struggled to preserve long-range logical coherence (e.g., topic consistency over long sentences).

- Limited grammatical accuracy due to the model’s inability to handle complex Uzbek morphological rules.

Importance of research

- Provides a baseline for comparing traditional (Markov) and advanced (e.g., Transformer, LSTM) models in low-resource languages like Uzbek.
- Highlights the need for hybrid approaches (e.g., combining N-gram models with neural architectures) to address grammatical and contextual challenges.

Preparing the Text Corpus

For experimentation, the "odilyakubov.txt" dataset was used. The text underwent:

- Cleaning: Removed unnecessary characters and spaces.
- Normalization: Converted all text to lowercase.
- Tokenization: Split into individual words.

BLEU (Bilingual Evaluation Understudy) score

How does this method work?

- BLEU is an evaluation method used for machine translation (MT) or text generation.
- It measures how similar the text generated by the model is to the ideal text.
- The BLEU score ranges from 0 to 1, with 1 → perfect match.

$$BLUE = BP \times \exp\left(\sum_{n=1}^{N} w_n log p_n\right)$$

- BP (Brevity Penalty) – A penalty based on the length of the text.
- pnp_npn – N-gram matching level.

Disadvantages:

- Does not understand semantic meaning, only compares based on n-grams.
- Due to the large number of synonyms in the Uzbek language, sometimes incorrect evaluations are possible.

Matnni boshlash uchun 3 ta so‘z kiriting: men kitobni o‘qishim

Yangi generatsiya qilingan matn:

men kitobni o‘qishim kerakligini esladim . . . . . .

✅ *Modelning BLEU skori bo‘yicha aniqligi: 0.73*

Analysis and Results

Model Performance

1. Text Quality: The Markov model successfully generates Uzbek text while maintaining basic syntactic structures. However, the generated text sometimes lacks semantic coherence, leading to inconsistencies in meaning.

2. Dependence on Local Context: The model makes accurate predictions within a short context but fails to capture long-range dependencies, resulting in disjointed and fragmented sentences over extended texts.

3. Repetitive Patterns: Due to the probabilistic nature of the model, certain words and phrases tend to be repeated, reducing the variability and creativity of the generated text.

Common Issues Observed

1. Logical Inconsistencies: The generated text does not always maintain a coherent structure, leading to sudden topic shifts.

2. Grammatical Errors: The model struggles with Uzbek morphology, particularly case endings and verb conjugations.

3. Word Prediction Limitations: Since the Markov model relies on fixed-length sequences, it cannot infer contextual meanings beyond its immediate N-gram scope.

Model Evaluation and Suggested Improvements

1. Increasing N-gram Size: Expanding n=4 or n=5 could help capture more context and improve text fluency.

2. Hybrid Approaches: Integrating deep learning models (LSTM, Transformer) could enhance grammatical accuracy and logical consistency.

3. Data Augmentation: A larger and more diverse corpus could improve the model’s ability to generate varied and contextually appropriate text.

4. Contextual Awareness Mechanisms: Enhancing the model with memory-based techniques or semantic embeddings could mitigate logical inconsistencies.

One possible way to address the model’s inability to capture long-range dependencies and morphological agreements is to integrate deep learning models, particularly sequence-aware architectures like LSTM and Transformer networks. These models can learn hierarchical representations and capture semantic nuances more effectively, especially in morphologically complex languages like Uzbek.

Conclusions

This study demonstrates that the Markov model offers a simple yet foundational approach for generating text in the Uzbek language. Its effectiveness lies in its ability to model short-range dependencies and maintain local syntactic structure using

probabilistic transitions between N-grams, particularly trigrams in our implementation. However, the results also reveal notable limitations that prevent the model from producing high-quality, coherent, and grammatically sound long-form text.

While the generated sentences exhibit basic fluency and partial contextual relevance, the model fails to maintain semantic consistency and long-range coherence, especially in more complex or abstract expressions. Furthermore, it struggles with morphological nuances of the Uzbek language, such as suffix handling, verb conjugation, and case agreement—factors that are critical for natural-sounding Uzbek text[9].

To address these challenges, the following strategies are recommended:

- Expanding the N-gram window size (e.g., to 4-grams or 5-grams) can improve context awareness by including more prior words in each prediction.
- Integrating deep learning models such as LSTM and Transformer architectures would allow the model to capture deeper contextual relationships and better handle the complex grammatical structure of the language.[10], [11]
- Training on larger, domain-diverse corpora can enhance the model’s ability to produce varied, topic-relevant, and semantically rich text outputs.

In addition, the study underscores the importance of incorporating evaluation metrics like BLEU to quantify output quality and guide model development.

Future Work

Future research should explore hybrid modeling approaches, combining the strengths of traditional statistical methods (like Markov chains) with the capabilities of neural language models. Such approaches can potentially offer greater accuracy, fluency, and contextual richness, particularly for low-resource languages like Uzbek where annotated data is limited. Moreover, expanding this work to domain-specific generation tasks (e.g., education, medicine, legal documents) can provide more practical applications and further refine model adaptability.

References.

1. A. M. Kassenkhan, N. K. Mukazhanov, S. Nuralykyzy, and Z. B. Kalpeyeva, “Text generation models for paraphrase on kazakh language,” *КазУТБ*, vol. 1, no. 22, Mar. 2024, doi: 10.58805/kazutb.v.1.22-249.

2. S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, “Semantic Similarity Metrics for Evaluating Source Code Summarization,” in *IEEE International Conference on Program Comprehension*, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/nnnnnnn.nnnnnnn.

3. E.B. Boltayevich. N-gramm til modellari vositasida o‘zbek tilida matn generatsiya qilish. // “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” Xalqaro ilmiy-amaliy konferensiya 2022. [Online]. Available: http://compling.navoiy-uni.uz/

4. B Elov *et al.*, O‘zbek tili korpusi matnlarini qayta ishlash usullari. //Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali. Volume 1, ISSUE 3, October 2023.

5. N. Fatima, A. S. Imran, Z. Kastrati, S. M. Daudpota, and A. Soomro, “A Systematic Literature Review on Text Generation Using Deep Neural Network Models,” *IEEE Access*, vol. 10, pp. 53490–53503, 2022, doi: 10.1109/ACCESS.2022.3174108.

6. T. Iqbal and S. Qureshi, “The survey: Text generation models in deep learning,” Jun. 01, 2022, *King Saud bin Abdulaziz University*. doi: 10.1016/j.jksuci.2020.04.001.

7. K. Makhija, T. N. Ho, and E. S. Chng, “Transfer learning for punctuation prediction,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, Institute of Electrical and Electronics Engineers Inc., Nov. 2019, pp. 268–273. doi: 10.1109/APSIPAASC47483.2019.9023200.

8. M. S. Sharipov, H. S. Adinaev, and E. R. Kuriyozov, “Rule-Based Punctuation Algorithm for the Uzbek Language,” in *International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, EDM*, 2024, pp. 2410 – 2414. doi: 10.1109/EDM61683.2024.10615061.

9. S. Maksud, K. Elmurod, Y. Ollabergan, and S. Ogabek, “UzbekVerbDetection: Rule-based Detection of Verbs in Uzbek Texts,” in *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, 2024, pp. 17343 – 17347. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195974039&partnerID=40&md5=ff3003ea2644833f3dc45277429459f7

10. Z. Y. Peng and P. C. Guo, “A Data Organization Method for LSTM and Transformer When Predicting Chinese Banking Stock Prices,” *Discrete Dyn Nat Soc*, vol. 2022, 2022, doi: 10.1155/2022/7119678.

11. N. R. Bhowmik, M. Arifuzzaman, and M. R. H. Mondal, “Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms,” *Array*, vol. 13, 2022, doi: 10.1016/j.array.2021.100123.