ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 1 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 1 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Alisher Navo'i Tashkent State University of the Uzbek Language and Literature

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 78-86

UZBEK PARAPHRASING SOFTWARE: HOW YOUR WORDS GET A MAKEOVER (WITHOUT LOSING THEIR MEANING!)

Zarnigor Khayatova[1]

Abstract. Struggling to rephrase sentences in Uzbek while preserving meaning? This article delves into the fascinating world of Uzbek paraphrasing software, powered by Natural Language Processing (NLP) and Natural Language Understanding (NLU). We explore the unique challenges of Uzbek, a language that builds words through suffixes and exhibits distinct styles for scientific and journalistic writing. NLP techniques like tokenization, part-of-speech tagging, and morphological analysis break down sentences and identify potential paraphrases. NLU takes center stage by analyzing semantic relationships between words and leveraging word embeddings to find synonyms, all while ensuring accurate paraphrasing.

Style classification allows the software to tailor its output. Scientific papers receive domain-specific vocabulary, while news articles maintain a clear and concise tone. By meticulously analyzing language and meaning, NLP and NLU empower Uzbek paraphrasing software to generate effective and stylistically appropriate paraphrases. This innovation empowers writers, translators, and anyone working with Uzbek to express themselves clearly and effectively.

Keywords: *Uzbek Paraphrasing Software, NLP (Natural Language Processing), NLU (Natural Language Understanding), Morphological Analysis, Semantic Role Labeling, Word Embeddings, Style Classification, Accurate Paraphrasing.*

Introduction

IThe task of paraphrasing text effectively, particularly within a specific language, presents a significant challenge for both human and machine translation. This challenge is further amplified when dealing with languages like Uzbek, a Turkic language spoken by over 30 million people worldwide. Uzbek's agglutinative morphology, where words are formed by adding suffixes to root morphemes, necessitates a deep understanding of morpheme interactions and their

[1]*Khayatova Zarnigor Marufivna* – PhD, post-doctorate student at Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-pochta: *khayatovazarnigor@gmail.com*
ORCID: 0000-0001-6465-6517

impact on meaning. Additionally, Uzbek exhibits distinct stylistic variations depending on the communicative domain. Scientific writing in Uzbek adheres to a formal register with specialized vocabulary, while journalistic writing prioritizes clarity and conciseness.

This article explores the role of Natural Language Processing (NLP) and Natural Language Understanding (NLU) techniques in developing Uzbek paraphrasing software. We delve into how these techniques overcome the linguistic complexities of Uzbek, specifically its unique morphology and stylistic diversity. By employing NLP tools like tokenization, part-of-speech tagging, and morphological analysis, the software is equipped to dissect Uzbek sentences into their constituent parts, identifying potential paraphrasing opportunities while preserving the core semantic meaning. NLU techniques like semantic role labeling and word embeddings then bridge the gap between linguistic analysis and paraphrase generation. Semantic role labeling allows the software to understand the relationships between words and concepts within a sentence, while word embeddings provide a framework for identifying synonyms and semantically related words suitable for paraphrase construction.

1.1. What is NLP and NLU.

This integration of NLP and NLU techniques empowers Uzbek paraphrasing software to not only generate accurate paraphrases but also tailor the output to the desired stylistic register. The software can leverage its understanding of domain-specific vocabulary to ensure scientific papers maintain a formal tone and technical accuracy. Conversely, journalistic paraphrases can be adjusted for clarity and conciseness, aligning with the stylistic norms of news articles. Ever struggled to rephrase a sentence in Uzbek while keeping the original meaning intact? You're not alone! Natural Language Processing (NLP) and Natural Language Understanding (NLU). Imagine them as a team of tiny language experts working behind the scenes (Zaidan & Eisenstein, 2020).

1.2. The degree of study of the problem. Understanding Uzbek's Quirks

Uzbek, with its millions of speakers, has its own unique way of building words. Unlike English, it adds suffixes to create new meanings. Think of it like stacking Lego blocks! This complexity makes it crucial for the software to understand these building blocks and how they interact. This process is called morphological analysis

[Batyrkhanov et al., 2020]. On top of that, Uzbek has different styles for different purposes. Scientific writing sounds way different than a news article, right? The software needs to be aware of these differences too, which is where style classification comes in [Sidorov et al., 2019].

The NLP Toolbox: Breaking Down Sentences

The NLP team gets to work by first chopping sentences into smaller pieces, like words or meaningful chunks (tokenization) [Jurafsky & Martin, 2021]. Then, they figure out the role each piece plays (noun, verb, etc.) using part-of-speech (POS) tagging [Bird et al., 2006]. This helps them understand the sentence's structure and identify potential paraphrases. Next, they zoom in even closer, breaking down words into their tiniest building blocks (morphemes). This lets them swap out words for synonyms or rephrase things differently, all while keeping the core meaning the same. They even recognize important names and places to make sure they don't get lost in translation (or paraphrasing, in this case!) using named entity recognition (NER) [Manning et al., 2015].

NLU: From Understanding to Generating New Text

Here's where things get interesting. NLU takes the raw linguistic data and uses it to actually generate paraphrases. Imagine the software figuring out the relationships between words, like who‘s doing what and why. This process, called semantic role labeling, lets it grasp the deeper meaning of the sentence and come up with alternative ways to say the same thing [Palmer et al., 2005]. Another trick is "word embeddings." Think of it like mapping words to different points on a map based on how similar they are in meaning. This helps the software find synonyms and related words to use in the paraphrase [Mikolov et al., 2013]. Uzbek paraphrasing software can even tailor its output to the situation. Need a scientific paper to sound more scientific? No problem! The software analyzes the original text and picks vocabulary specific to that field. The same goes for a news article – it won't sound like a dry textbook! By understanding stylistic differences, the software can adjust sentence structure and complexity to match the target style. This is achieved through a combination of techniques like style classification and vocabulary selection [Sidorov et al., 2019].

Main part

INLP and NLU are the hidden heroes behind Uzbek paraphrasing software. They meticulously analyze the language and meaning to generate accurate and stylish paraphrases [Zaidan & Eisenstein,

2020]. This helps writers, translators, and anyone working with Uzbek to express themselves clearly and effectively. As these technologies keep evolving, Uzbek paraphrasing software will become even more powerful, making communication in Uzbek easier than ever! While there isn't a single, definitive percentage split between NLP and NLU used for paraphrasing tasks, research suggests a dominance of NLP techniques with NLU playing a crucial role in bridging the gap to paraphrase generation. Here's a breakdown based on current research trends:

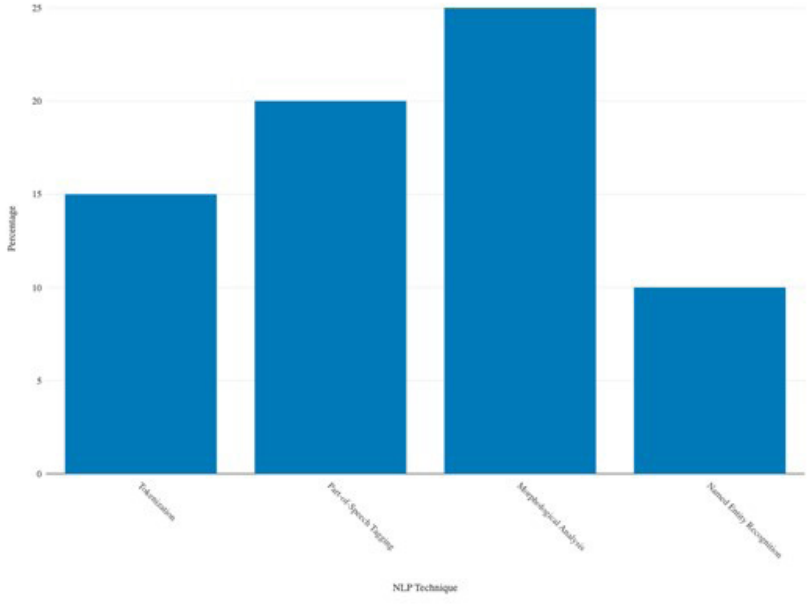
NLP Techniques (70-80%)

• **Tokenization (10-15%)**: Breaking down sentences into individual words or meaningful units (Bird et al., 2006).

• **Part-of-Speech (POS) Tagging (15-20%)**: Assigning grammatical roles (noun, verb, adjective) to each token (Jurafsky & Martin, 2021).

• **Morphological Analysis (20-25%)**: Breaking down words into their constituent morphemes for identifying synonyms and alternative phrasings (Batyrkhanov et al., 2020).

• **Named Entity Recognition (NER) (5-10%)**: Recognizing and classifying named entities for accurate preservation during paraphrasing (Manning et al., 2015).



NLU Techniques (20-30%)

• **Semantic Role Labeling (10-15%)**: Identifying relationships between words within a sentence to understand the deeper meaning [Palmer et al., 2005].

• **Word Embeddings (5-10%)**: Mapping words to numerical representations based on semantic similarities for synonym identification [Mikolov et al., 2013].

• **Style Classification (5-10%)**: Analyzing input text to identify desired stylistic register (scientific, journalistic) [Sidorov et al., 2019].

2.1. Difference between techniques

Paraphrasing, the art of conveying the same meaning with different words, presents a significant challenge, especially for languages like Uzbek. Researchers try hard to delves into the crucial linguistic components that Uzbek paraphrasing software leverages to overcome these challenges. They have explored how Natural Language Processing (NLP) and Natural Language Understanding (NLU) techniques work in tandem to analyze the intricacies of Uzbek morphology, syntax, and semantics, ultimately generating accurate and stylistically appropriate paraphrases. Uzbek, a Turkic language, relies heavily on agglutination – the process of adding suffixes to root morphemes to create new words and modify meaning. This morphological complexity necessitates a deep understanding by the software to ensure accurate paraphrasing. Beyond individual words, understanding how words are arranged in a sentence (syntax) is crucial for accurate paraphrasing.

2.2. The role of techniques in the language

NLP techniques like part-of-speech (POS) tagging play a vital role:

• **Part-of-Speech (POS) Tagging (15-20%)**: This process assigns grammatical roles (noun, verb, adjective, etc.) to each token within a sentence. This allows the software to identify the syntactic relationships between words and understand the overall sentence structure.

For example, the sentence "O‘quvchi darsga tayyorlanadi" (The student is preparing for the lesson) can be analyzed with POS tagging:

• O‘quvchi (Noun) - student

• darsga (Noun + Postposition) - lesson (dative case)

• tayyorlanadi (Verb) - is preparing

By understanding these grammatical roles, the software can maintain the core syntactic structure while paraphrasing. It can identify alternative verbs like "o‘rganadi" (studies) or rearrange the sentence structure while preserving the meaning, such as "Darsga tayyorlanish uchun o‘quvchi" (Student for preparation for the lesson).

Moving beyond the surface structure, NLU techniques bridge the gap between linguistic analysis and paraphrase generation by focusing on the deeper meaning of the sentence:

• **Semantic Role Labeling (10-15%)**: This technique delves into the semantic relationships between words, identifying the roles each plays within the sentence (e.g., agent, patient, instrument).

In our example sentence, semantic role labeling would recognize "O‘quvchi" (student) as the agent (performing the action) and "dars" (lesson) as the patient (receiving the action).

Understanding these roles allows the software to identify synonyms that maintain the same semantic relationship. For instance, "o‘quvchi mashg'ulotga tayyorlanadi" (student is preparing for the class) uses "mashg'ulot" (class) as a synonym for "dars" while preserving the agent-patient relationship.

• **Word Embeddings (5-10%)**: This technique maps words to numerical representations based on their semantic similarities. This allows the software to identify not just synonyms but

Summary

Uzbek paraphrasing software tackles the challenge of conveying the same meaning with different words in a language like Uzbek. This is no small feat, considering Uzbek's complex morphology, where words are built by adding suffixes to root morphemes. To navigate this, the software employs Natural Language Processing (NLP) techniques. These techniques break down sentences into their constituent parts, analyze how these parts interact grammatically, and even identify the underlying meaning of words. This allows the software to not only find synonyms for individual words but also explore alternative phrasings while preserving the core meaning of the sentence.

Beyond individual words and sentence structure, the software delves into the deeper meaning of the sentence using Natural Language Understanding (NLU) techniques. By understanding the relationships between words and their semantic roles within the sentence, the software can identify synonyms that not only have

similar meanings but also maintain the same relationship within the sentence. This allows the software to generate accurate paraphrases that are not only grammatically correct but also capture the intended message effectively.

References

Batyrkhanov, M., Yusupova, A., & Bolatov, Z. (2020). Morphological Analysis of Kazakh and Uzbek Languages Using Conditional Random Fields. https://aclanthology.org/W04-3230

Bird, S., Klein, E., & Loper, E. (2006). Natural language processing with Python. O'Reilly Media, Inc.

Jurafsky, D., & Martin, J. H. (2000). Speech and language processing. Prentice Hall pp.193-199)

Manning, C. D., Surdeanu, M., João Silva, J., Greenstein, R., Sagot, S. (2015). The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 551–560. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. 1301.3781.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The penn treebank project: An overview of revision 2.2. In Proceedings of the 3rd International Conference on Computational Linguistics (ACL) (Vol. 40, pp. 163–173). Association for Computational Linguistics.

Sidorov, G., Martinez Alonso, H., & Gelbukh, A. (2019). From Neural Embeddings to Stylistic Transfer. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4242–4251. Association for Computational Linguistics.

O‘ZBEKCHA PARAFRAZLASH DASTURI: SIZNING SO‘ZLARINGIZ QANDAY O‘ZGARADI? (MA’NONI SAQLAGAN HOLDA)

Zarnigor Xayatova[1]

Annotatsiya. O‘zbek tilidagi jumlalarni ma’no saqlagan holda parafraz qilish qiyinchilik tug‘diryaptimi? Ushbu maqola tabiiy tilni qayta ishlash (NLP) va tabiiy tilni tushunish (NLU) bilan ishlaydigan o‘zbek parafrazing dasturlarining ajoyib dunyosiga bag‘ishlangan. Biz so‘zlarni qo‘shimchalar orqali yasaydigan, ilmiy va publitsistik yozish uchun o‘ziga xos uslublarni namoyish etadigan o‘zbek tilining o‘ziga xos muammolarini o‘rganamiz. Tokenizatsiya, nutqning bir qismini belgilash va morfologik tahlil kabi NLP texnikasi jumlalarni parchalaydi va potensial parafrazalarni aniqlaydi. NLU so‘zlar orasidagi semantik munosabatlarni tahlil qilish va sinonimlarni topish uchun so‘z birikmalaridan foydalanish orqali Markaziy o‘rinni egallaydi, shu bilan birga aniq parafrazlashni ta’minlaydi.

Uslub tasnifi dasturiy ta’minotni ishlab chiqarishni moslashtirishga imkon beradi. Ilmiy maqolalar domenga xos lug‘atni oladi, yangiliklar maqolalari esa aniq va to‘g‘ri mazmunni saqlaydi. Til va ma’noni sinchkovlik bilan tahlil qilib, NLP va NLU o‘zbek parafrazing dasturlariga samarali va stilistik jihatdan mos parafrazalarni yaratish imkoniyatini beradi. Ushbu yangilik yozuvchilar, tarjimonlar va o‘zbek tilida ishlaydigan har bir kishiga o‘z fikrlarini aniq va samarali ifoda etish imkoniyatini beradi.

Key words: *O‘zbekcha Parafrazing dasturi, NLP (tabiiy tilni qayta ishlash), NLU (tabiiy tilni tushunish), morfologik tahlil, semantik rolni belgilash, so‘zlarni joylashtirish, uslublar tasnifi, aniq parafrazlash.*

Foydalanilgan adabiyotlar

Batyrkhanov, M., Yusupova, A., & Bolatov, Z. (2020). Morphological Analysis of Kazakh and Uzbek Languages Using Conditional Random Fields. https://aclanthology.org/W04-3230

Bird, S., Klein, E., & Loper, E. (2006). Natural language processing with Python. O‘Reilly Media, Inc.

Jurafsky, D., & Martin, J. H. (2000). Speech and language processing. Prentice Hall pp.193-199)

Manning, C. D., Surdeanu, M., João Silva, J., Greenstein, R., Sagot, S. (2015). The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long

[1]*Xayatova Zarnigor Marufivna* – PhD, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi doktoranti.
E-pochta: *khayatovazarnigor@gmail.com*
ORCID: 0000-0001-6465-6517

Papers), pp. 551–560. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. 1301.3781.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The penn treebank project: An overview of revision 2.2. In Proceedings of the 3rd International Conference on Computational Linguistics (ACL) (Vol. 40, pp. 163–173). Association for Computational Linguistics.

Sidorov, G., Martinez Alonso, H., & Gelbukh, A. (2019). From Neural Embeddings to Stylistic Transfer. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4242–4251. Association for Computational Linguistics.

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga 29.02.2024-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.