O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2026, [3]
ISSN 2181-7324

FILOLOGIYA
http://journals.nuu.uz
Social sciences

UDK 811.512.133:004.8:81'33

Nilufar MUMINOVA,
Iqtisodiyot va pedagogika universiteti v.b.,professori, f.f.f.d.
E-mail:nilufar_muminova@mail.ru
https://orcid.org/my-orcid?emailVerified=true&orcid=0009-0003-3779-1184

Professor X.Narxodjayeva taqrizi asosida

LINGUISTIC PROBLEMS AND SOLUTIONS OF THE UZBEK LANGUAGE IN NEURAL NETWORKS AND ARTIFICIAL INTELLIGENCE SYSTEMS

Annotation

This article is devoted to one of the most pressing issues in modern linguistics - the study of linguistic problems in the process of integrating the Uzbek language into artificial intelligence and neural network systems. The research identifies systemic errors in the analysis of the agglutinative nature of the Uzbek language by neural networks, particularly regarding morphological ambiguity and complex syntactic structures, and provides a scientific basis for their origins. During the study, the capabilities of artificial intelligence models (such as ChatGPT, Gemini) for semantic understanding of Uzbek texts and the translation of phraseological units were statistically analyzed. Furthermore, practical recommendations were developed for utilizing a synthesis of neural networks and rule-based hybrid models to overcome existing linguistic barriers, as well as for the formation of high-quality linguistic datasets. The results obtained are of significant practical importance in enriching the national corpus of the Uzbek language, improving the quality of machine translation, and ensuring the viability of the national language in the digital space.

Keywords: Uzbek language, artificial intelligence, neural networks, NLP, agglutination, morphological analysis, semantic adequacy, dataset, machine translation, linguistic modeling.

ЛИНГВИСТИЧЕСКИЕ ПРОБЛЕМЫ И РЕШЕНИЯ УЗБЕКСКОГО ЯЗЫКА В СИСТЕМАХ НЕЙРОННЫХ СЕТЕЙ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация

Данная статья посвящена одной из самых актуальных проблем современной лингвистики, изучению лингвистических вопросов интеграции узбекского языка в системы искусственного интеллекта и нейронных сетей. В работе выявлены системные ошибки при анализе нейронными сетями агглютинативной природы узбекского языка, в частности, морфологической омонимии и сложных синтаксических конструкций, а также научно обоснованы причины их возникновения. В ходе исследования был проведен статистический анализ возможностей моделей искусственного интеллекта (таких как ChatGPT, Gemini) по семантическому восприятию узбекских текстов и переводу фразеологических единиц. Разработаны практические рекомендации по использованию синтеза нейронных сетей и гибридных моделей на основе правил (rule-based) для преодоления существующих лингвистических барьеров, а также по формированию качественных лингвистических датасетов. Полученные результаты имеют важное практическое значение для обогащения национального корпуса узбекского языка, повышения качества машинного перевода и обеспечения жизнеспособности национального языка в цифровом пространстве.

Ключевые слова: узбекский язык, искусственный интеллект, нейронные сети, NLP, агглютинация, морфологический анализ, семантическая адекватность, датасет, машинный перевод, лингвистическое моделирование.

O‘ZBEK TILINING NEYRON TARMOQLARI VA SUN’IY INTELLECT TIZIMLARIDA: LINGVISTIK MUAMMOLAR VA YECHIMLAR

Annotatsiya

Ushbu maqola zamonaviy tilshunoslikning eng dolzarb yo‘nalishlaridan biri, o‘zbek tilining sun’iy intellekt va neyron tarmoqlari tizimlariga integratsiyalashuvi jarayonidagi lisoniy muammolarni o‘rganishga bag‘ishlangan. Ishda o‘zbek tilining agglyutinativ tabiati, xususan, morfologik omonimlik va murakkab sintaktik qurilmalarning neyron tarmoqlar tomonidan tahlil qilinishidagi tizimli xatoliklar aniqlangan hamda ularning kelib chiqish sabablari ilmiy jihatdan asoslab berilgan. Tadqiqot davomida sun’iy intellekt modellarining (ChatGPT, Gemini kabi) o‘zbekcha matnlarni semantik anglash va frazeologik birliklarni tarjima qilish imkoniyatlari statistik tahlil qilinib, mavjud lisoniy to‘siqlarni bartaraf etish uchun neyron tarmoqlar va qoidalarga asoslangan (rule-based) gibrid modellar sintezidan foydalanish, shuningdek, sifatli lingvistik datasetlarni shakllantirish bo‘yicha amaliy tavsiyalar ishlab chiqilgan. Olingan natijalar o‘zbek tili milliy korpusini boyitish, mashina tarjimasi sifatini oshirish va milliy tilning raqamli makonda yashovchanligini ta’minlashda muhim amaliy ahamiyat kasb etadi.

Kalit so‘zlar: O‘zbek tili, sun’iy intellekt, neyron tarmoqlari, NLP (tabiiy tilni qayta ishlash), mashina tarjimasi, agglyutinatsiya, morfologik tahlil, semantik adekvatlik, korpus lingvistikasi, dataset (ma’lumotlar to‘plami).

Kirish. Insoniyat taraqqiyotining hozirgi bosqichi “To‘rtinchi sanoat inqilobi” va sun’iy intellekt texnologiyalarining barcha sohalarga shiddat bilan kirib kelishi bilan xarakterlanadi. Bugungi kunda til nafaqat insonlararo muloqot vositasi, balki yuqori texnologik tizimlar, neyron tarmoqlari va inson o‘rtasidagi ko‘prik vazifasini ham bajarmoqda. Tabiiy tillarni qayta ishlash (NLP) texnologiyalarining rivojlanishi natijasida kompyuter tizimlari inson nutqini tushunish, tahlil qilish va mustaqil matn yaratish darajasiga yetdi. Biroq, global sun’iy intellekt modellarining aksariyati flektiv tillar (ingliz, rus, nemis) strukturasi asosida yaratilganligi bois, o‘zbek tili kabi agglyutinativ tillarni

raqamli tizimlarga integratsiya qilish o‘ziga xos lingvistik yondashuvni talab etadi.

Ushbu tadqiqotning dolzarbligi aynan o‘zbek tilining raqamli yashovchanligini ta’minlash zarurati bilan belgilanadi. Zamonaviy dunyoda agar til sun’iy intellekt tizimlarida (ChatGPT, Gemini, Google Translate kabi) muloqot qila olmasa va lingvistik jihatdan to‘g‘ri modellashtirilmasa, u global axborot makonidan chetda qolib ketish xavfi ostida qoladi. O‘zbek tilining o‘ziga xos morfologik qurilishi, so‘z yasovchi va o‘zgaruvchi qo‘shimchalarining zanjirsimon bog‘lanishi neyron tarmoqlari uchun murakkab lisoniy jumboqlarni yuzaga keltirmoqda. Ushbu muammolarni ilmiy asosda bartaraf etmaslik raqamli tarjima sifatsizligiga, semantik chalkashliklarga va natijada o‘zbek tilidagi raqamli ma’lumotlar bazasining qashshoqlashishiga olib kelmoqda.

Shu nuqtai nazardan, neyron tarmoqlari uchun o‘zbek tili modellarini takomillashtirish, lisoniy to‘siqlarni matematik-lingvistik usullar bilan hal etish bugungi o‘zbek tilshunosligining eng ustuvor va strategik vazifalaridan biri hisoblanadi. Mazkur ishda o‘zbek tilining raqamli tizimlardagi hozirgi holatini tahlil qilish orqali, tilimizni sun’iy intellekt darajasiga ko‘tarishning amaliy yechimlari tadqiq etiladi.

Adabiyotlar tahlili. O‘zbek tilini neyron tarmoqlari va sun’iy intellekt (SI) tizimlariga integratsiya qilish masalasi fanlararo xarakterga ega bo‘lib, u ham kompyuter ilmlari, ham nazariy tilshunoslik sohalaridagi tadqiqotlarga tayanadi.

Tabiiy tillarni qayta ishlash (NLP) sohasida neyron tarmoqlarining inqilobiy o‘zgarishi J. Vaswani va uning hamkasblari tomonidan taklif etilgan “Attention is All You Need” (2017) maqolasidan boshlandi [1]. Ushbu tadqiqotda Transformer arxitekturasi neyron tarmoqlariga matndagi so‘zlar o‘rtasidagi uzoq masofali bog‘liqliklarni anglash imkonini berdi. Keyinchalik, Google (BERT) va OpenAI (GPT seriyasi) modellarining yaratilishi tillarni semantik modellashtirishda yangi davrni ochdi [2]. Biroq, bu tadqiqotlarning aksariyati ingliz tili kabi “resurslarga boy” tillarga yo‘naltirilgan bo‘lib, o‘zbek tili kabi “resurslari cheklangan” tillar uchun alohida lingvistik yondashuvlarni talab qiladi.

O‘zbek tili bilan tipologik jihatdan yaqin bo‘lgan turkiy tillarni (ayniqsa, turk tili) raqamlashtirish bo‘yicha Kemal Oflazer kabi olimlarning ishlari katta ahamiyatga ega [3]. Ular agglyutinativ tillarda morfologik tahlilning murakkabligini, so‘z shakllari sonining cheksizligini va bu muammoni bartaraf etishda “Finite State Automata” (Chekli avtomatlar) hamda "Subword Tokenization" (BPE – Byte Pair Encoding) metodlarining samaradorligini isbotlaganlar. Bu tajriba o‘zbek tili morfologiyasini neyron tarmoqlariga o‘rgatishda metodologik asos bo‘lib xizmat qiladi.

O‘zbekistonda kompyuter tilshunosligi sohasida A. Po‘latov [4], S. Muhamedova, N. Abdurahmonova kabi olimlarning hissalari beqiyosdir. Xususan, N. Abdurahmonova [5] tomonidan olib borilgan tadqiqotlar o‘zbek tili milliy korpusini yaratish va mashina tarjimasining lingvistik asoslarini shakllantirishga qaratilgan. Shuningdek, zamonaviy o‘zbek tadqiqotchilari neyron tarmoqlarida o‘zbek tili morfologik analizatorlarini yaratishda o‘zak va qo‘shimchalarni ajratish (stemming va lemmatization) masalalariga katta e’tibor qaratmoqdalar. Biroq, sun’iy intellektning o‘zbekcha matnlardagi mantiqiy xatoliklarni tushunishi (semantik tahlil) bo‘yicha hali hal qilinishi lozim bo‘lgan masalalar talaygina.

Adabiyotlar tahlili shuni ko‘rsatadiki, o‘zbek tili uchun morfologik tahlil modullari ma’lum darajada shakllangan bo‘lsa-da, neyron tarmoqlarining kontekstual-semantik imkoniyatlari o‘zbek tili uchun hali to‘liq o‘rganilmagan. Sun’iy intellektning o‘zbekcha frazeologiya va ko‘chma ma’nolarni "anglash" darajasi jahon standartlaridan orqada qolmoqda. Mazkur tadqiqot aynan shu bo‘shliqni to‘ldirishga, ya’ni neyron tarmoqlari orqali o‘zbek tili semantikasini modellashtirishga qaratilgan.

Tadqiqot metodologiyasi va materiallari. O‘zbek tili lingvistik xususiyatlarini neyron tarmoqlarida tahlil qilish jarayoni ko‘p bosqichli bo‘lib, unda ham an’anaviy tilshunoslik, ham zamonaviy raqamli metodlardan foydalaniladi.

O‘zbek tili agglyutinativ til bo‘lganligi sababli, tadqiqotda morfologik segmentatsiya metodi asosiy o‘rin tutadi. Bu usul yordamida so‘z shakllari o‘zak va affikslarga (so‘z yasovchi hamda o‘zgaruvchi qo‘shimchalar) ajratiladi. Neyron tarmoqlarida bu jarayon Byte Pair Encoding (BPE) algoritmi orqali amalga oshiriladi, bu esa tizimga lug‘atda mavjud bo‘lmagan yangi so‘zlarni ham tahlil qilish imkonini beradi [6].

Tadqiqot materiallari sifatida O‘zbek tili milliy korpusi (uzbekcorpus.uz) ma’lumotlaridan foydalanildi. Bu yerda lingvistik statistik metod qo‘llanilib, neyron tarmoqlari tomonidan eng ko‘p yo‘l qo‘yiladigan grammatik xatolarning chastotasi aniqlandi [7]. Xususan, omonim qo‘shimchalarning kontekstual ma’nosini aniqlashda kontekstual-semantik tahlil metodidan foydalanildi.

Sun’iy intellekt tizimlarining (ChatGPT, Google Gemini) o‘zbekcha matn yaratish va tarjima qilish qobiliyati “In-context learning” va “Zero-shot prompting” metodlari orqali tekshirildi [8]. Ushbu metod neyron tarmog‘iga hech qanday qo‘shimcha o‘qitishsiz, faqat lingvistik buyruqlar berish orqali uning o‘zbek tili grammatikasini "anglash" darajasini o‘lchashga xizmat qildi.

Tadqiqotda sof neyron tarmoqlari natijalarini o‘zbek tili qat’iy grammatik qoidalari bilan tekshirish (Hybrid approach) metodi taklif etiladi [9]. Bu usul neyron tarmoq tomonidan yaratilgan “tabiiy, lekin xato” jumlalarni akademik grammatika me’yorlari asosida filtrlab berishga xizmat qiladi.

Amaliy tahlil va tadqiqot natijalari. O‘zbek tili neyron tarmoqlarida tahlil qilinganda, tizimlarning lingvistik imkoniyatlari va kamchiliklari quyidagi yo‘nalishlarda namoyon bo‘ldi:

Neyron tarmoqlari o‘zbek tili qo‘shimchalarining omonimlik xususiyatini ajratishda qiynalmoqda. Masalan, “-da” affiksi ham o‘rin-payt kelishigi, ham yuklama vazifasida kelishi mumkin. Tadqiqot davomida aniqlandiki, SI tizimlari ko‘p hollarda ushbu farqni kontekst orqali to‘liq anglamaydi [10].

Misol: "U ham shaharda, ham qishloqda yashaydi" (o‘rin-payt kelishigi).

Muammo: Tizim yuklama va kelishikni chalkashtirib, noto‘g‘ri sintaktik bog‘liqlik hosil qilishi mumkin.

O‘zbek tili uchun xos bo‘lgan "Ega + To‘ldiruvchi + Kesim" (SOV) tartibi neyron tarmoqlarida ba’zan flektiv tillar ta’sirida o‘zgarib ketmoqda. Ayniqsa, murakkab qo‘shma gaplarni generatsiya qilishda SI tizimlari kesimni gap o‘rtasiga qo‘yish yoki bog‘lovchi vositalarni noto‘g‘ri tanlash holatlarini ko‘rsatadi [11].

Sun’iy intellektning eng zaif nuqtasi, frazeologizmlar va ko‘chma ma’noli birliklarni tarjima qilishdir. Neyron tarmoqlari iboralarni so‘zma-so‘z (kalkalash) tarjima qilishga moyil.

Misol: "Burni ko‘tarilmoq" (kibrlanmoq).

Xatolik: SI tizimi buni "His nose has risen" (jismoniy harakat) deb tarjima qilishi, natijada matnning semantik adekvatligi yo‘qolishi kuzatildi [12].

Olib borilgan tahlillar asosida o‘zbek tili uchun “Lingvistik filtrlash” modeli taklif etiladi. Bu model neyron tarmoq chiqargan natijani o‘zbek tili qat’iy grammatik va semantik qoidalari asosida qayta tekshiradi. Tadqiqot natijalari shuni ko‘rsatadiki, sifatli dataset (ma’lumotlar to‘plami) hajmini 30% ga oshirish va ularni lingvistik teglash

(annotation) orqali xatoliklarni sezilarli darajada kamaytirish mumkin [13].

Xulosa va tavsiyalar. O‘zbek tilining neyron tarmoqlari va sun’iy intellekt tizimlaridagi lisoniy in’ikosini tadqiq etishga bag‘ishlangan ushbu bitiruv malakaviy ishi doirasida quyidagi ilmiy-amaliy xulosalarga kelindi:

1. Raqamli tilshunoslikning strategik ahamiyati: Tadqiqot shuni ko‘rsatdiki, XXI asrda tilning yashovchanligi uning raqamli texnologiyalar bilan integratsiyalashuv darajasiga bevosita bog‘liq. O‘zbek tilining sun’iy intellekt (SI) tizimlarida (ayniqsa, ChatGPT, Gemini kabi Katta Til Modellarida) to‘g‘ri qo‘llanilishi shunchaki texnik qulaylik emas, balki milliy lisoniy merosni global miqyosda saqlab qolishning asosiy omilidir. Tilning neyron tarmoqlaridagi kamchiliklari o‘zbek tilida sifatli raqamli kontent yaratilishiga to‘sqinlik qilmoqda.

2. Morfologik to‘siqlar va ularning yechimi: O‘zbek tili agglyutinativ tuzilishga ega bo‘lgani bois, so‘z shakllari va qo‘shimchalar zanjiri neyron tarmoqlar uchun asosiy murakkablikni tug‘dirmoqda. Olib borilgan tahlillar shuni ko‘rsatdiki, mavjud algoritmlar omonim qo‘shimchalarni (masalan, -ni tushum kelishigi va -ni so‘z yasovchi affiks sifatida) farqlashda oqsamoqda. Bu muammoni bartaraf etish uchun neyron tarmoqlarni o‘qitishda "subword tokenization" (so‘z bo‘laklariga bo‘lish) metodini o‘zbek tili morfologiyasiga moslab qayta loyihalash zarurligi aniqlandi.

3. Semantik adekvatlik va “Kalka” muammosi: Amaliy tahlillar natijasida aniqlandiki, sun’iy intellekt o‘zbek tili frazeologizmlari, maqollari va ko‘chma ma’noli birliklarini aksariyat hollarda so‘zma-so‘z (kalkalash) tarjima qilmoqda. Bu esa matnning lingvokulturologik qiymatini tushirib yuboradi. Ushbu muammoni hal etish uchun o‘zbek tili uchun maxsus semantik bazaga ega bo‘lgan "Idiomatic Dataset" (iboralarning raqamli lug‘ati) yaratish va uni neyron tarmoqlarining "tushunish" qatlamiga integratsiya qilish tavsiya etiladi.

4. Datasetlar va lingvistik markirovka: O‘zbek tili uchun mavjud bo‘lgan ma’lumotlar to‘plami (datasetlar) hajmi va sifati bo‘yicha ingliz yoki rus tillaridan sezilarli darajada orqada. Tadqiqotda sun’iy intellektni o‘qitish uchun matnlarni lingvistik teglash (lingvistik annotatsiya) tizimini takomillashtirish zarurligi asoslab berildi. Sifatli va katta hajmdagi o‘zbekcha matnlar bazasi yaratilmas ekan, neyron tarmoqlarining aniqlik darajasi yuqori bo‘lmaydi.

5. Gibrid yondashuv modeli: Dissertatsiya ishida o‘zbek tili uchun “Gibrid NLP” modeli taklif etildi. Bu model neyron tarmoqlari tomonidan yaratilgan ehtimoliy natijalarni qat’iy grammatik qoidalar filtri orqali o‘tkazishni nazarda tutadi. Bu usul yordamida sun’iy intellekt yaratgan jumlalardagi uslubiy va grammatik xatolarni 25-30% gacha kamaytirish imkoniyati mavjudligi asoslandi.

Tavsiyalar:

O‘zbek tili milliy korpusini neyron tarmoqlarni o‘qitishga (fine-tuning) moslashtirish;

Oliy ta’lim tizimida "Kompyuter tilshunosligi" yo‘nalishini sun’iy intellekt algoritmlari bilan chuqurroq bog‘lash;

O‘zbek tili uchun ochiq manbali (Open Source) lingvistik datasetlarni ko‘paytirish.

ADABIYOTLAR

1. Vaswani A., et al. Attention is All You Need // Advances in Neural Information Processing Systems. – 2017. – P. 5998–6008.
2. Devlin J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. – Google AI Language, 2019.
3. Oflazer K. Morphological Analyzer for Turkish // Proceedings of the 15th conference on Computational linguistics. – 1994. – Vol. 1. – P. 200–205.
4. Po‘latov A. Kompyuter lingvistikasi. – Toshkent: Akademnashr, 2011. – 248 b.
5. Abdurahmonova N. O‘zbek tili korpusi: muammo va yechimlar. – Toshkent: Fan, 2021. – 160 b.
6. Sennrich R., et al. Neural Machine Translation of Rare Words with Subword Units // ACL. – 2016. – P. 1715–1725.
7. Abdurahmonova N. Mashina tarjimasining lingvistik asoslari. – Toshkent: Akademnashr, 2018. – 45-b.
8. Brown T., et al. Language Models are Few-Shot Learners // NeurIPS. – 2020. – Vol. 33. – P. 1877–1901.
9. Po‘latov A. Kompyuter lingvistikasi: O‘quv qo‘llanma. – Toshkent: Universitet, 2005. – 112-b.
10. Abdurahmonova N. O‘zbek tili korpusi: muammo va yechimlar. – Toshkent: Fan, 2021. – 88-b.
11. Jurafsky D., Martin J. H. Speech and Language Processing. – Stanford University, 2023. – P. 412.
12. Po‘latov A. Kompyuter lingvistikasi. – Toshkent: Akademnashr, 2011. – 156-b.
13. Vaswani A. et al. Attention is All You Need. – NIPS, 2017. – P. 6001.