

SUN'IY INTELLEKT ASOSIDA O'ZBEK TILIDAGI MATNLARNI AVTOMATIK TAHLIL QILISH VA QAYTA ISHLASH USULLARI

Abu Rayhon Beruniy nomidagi Urganch Davlat universiteti

Bekchanov Hikmat Mansurbek o'g'li

e-mail: bekchanovhikmat209@gmail.com

ANNOTATSIYA

Ushbu maqola sun'iy intellekt texnologiyalari yordamida o'zbek tilidagi matnlarni avtomatik tahlil qilish va qayta ishlash usullarini o'rganadi. Maqolada informatik lingvistika, tabiiy tilni qayta ishlash (NLP), tokenizatsiya, morfologik va sintaktik tahlil masalalari ko'rib chiqiladi. Shuningdek, ChatGPT va boshqa katta til modellarining o'zbek tiliga qo'llanilishi hamda kelajak istiqbollari atroflicha tahlil qilinadi. Tadqiqot o'zbek tilining agglutinatativ xususiyatlaridan kelib chiqqan holda NLP tizimlarini loyihalashda yuzaga keladigan qiyinchiliklar va ularni bartaraf etish yo'llari haqida muhim xulosalar shakllantiradi.

Kalit so'zlar: *sun'iy intellekt, NLP, o'zbek tili, tokenizatsiya, morfologik tahlil, ChatGPT, transformer modellari, informatik lingvistika.*

1. KIRISH

XXI asrda raqamli texnologiyalarning jadal rivojlanishi natijasida sun'iy intellekt (SI) fan va texnikaning barcha sohalarida keng qo'llanilmoqda. Jumladan, tabiiy tilni qayta ishlash (Natural Language Processing — NLP) sohasidagi so'nggi yutuqlar inson va mashina o'rtasidagi muloqotni tubdan o'zgartirib yubordi. GPT, BERT va Claude kabi katta til modellari dunyodagi yuzlab tillar bilan ishlash imkoniyatiga ega bo'ldi.

O'zbek tili — Turkiy til oilasiga mansub, 35 million nafardan ortiq inson so'zlaydigan, agglutinatativ tuzilishga ega til bo'lib, uning uchun sun'iy intellektga asoslangan NLP tizimlarini ishlab chiqish o'ziga xos imkoniyatlar va qiyinchiliklar taqdim etadi. O'zbekiston Respublikasida raqamli iqtisodiyot va axborot texnologiyalarini rivojlantirish milliy strategiyasi doirasida o'zbek tilidagi NLP tizimlarini yaratish davlat ahamiyatiga molik vazifaga aylanmoqda.

Mazkur maqolaning maqsadi — informatik lingvistika, tokenizatsiya, morfologik-sintaktik tahlil va zamonaviy SI tizimlari bo'yicha mavjud yondashuvlarni tizimli ko'rib chiqish hamda o'zbek tili uchun samarali NLP arxitekturasini taklif etishdir. Tadqiqot O'zbekistondagi olimlar va xalqaro hamjamiyat ishlanmalariga tayanib, kelajak istiqbollarini belgilashga xizmat qiladi.

2. INFORMATIK LINGVISTIKA TUSHUNCHASI

Informatik lingvistika (Computational Linguistics) — til strukturasi va uning qonuniyatlarini hisoblash usullari orqali o'rganuvchi fanlararo soha bo'lib, lingvistika, informatika, matematika va kognitiv fanlarning kesishmasida joylashgan. Bu fan 1950-yillarda mashinali tarjima muammolarini hal etish maqsadida vujudga kelgan va hozirga qadar ulkan yo'l bosib o'tgan.

Informatik lingvistika ikki katta yo'nalishni birlashtiradi: (1) til hodisalarini rasmiy — matematikaviy usullar bilan tasvirlash (formal lingvistika); (2) real matnlarni mashina orqali tahlil qilish va generatsiya qilish (NLP). Ushbu ikki yo'nalishning uyg'un rivojlanishi bugungi kunda ChatGPT va Claude kabi tizimlarning paydo bo'lishiga zamin yaratdi.

Informatik lingvistikaning asosiy tadqiqot yo'nalishlari:

- Fonetik va fonologik tahlil — so'zlarning tovush tuzilishi, urg'u va intonatsiyani o'rganish hamda nutq sintezi (TTS) va tanish (STT) tizimlarini ishlab chiqish;
- Morfologik tahlil — so'z shakllarini tarkibiy qismlarga (morfema, asos, qo'shimcha) ajratish va grammatik kategoriyalarni belgilash;
- Sintaktik tahlil — gapning grammatik tuzilishi, ega-kesim-to'ldiruvchi munosabatlarini aniqlash, daraxt strukturalarini qurish;
- Semantik tahlil — so'z va gaplarning ma'nosini, ko'p ma'noliligini, metafora va idiomalarni tushunish;
- Diskurs tahlil — matn darajasidagi bog'liqlik, mazmunaviy izchillik va nutq aktlarini o'rganish;
- Pragmatik tahlil — kontekst, nutqiy maqsad va kommunikativ vaziyatni hisobga olgan holda matnni talqin qilish.

O'zbek tilidagi informatik lingvistika tadqiqotlari 1990-yillardan boshlab tizimli ravishda rivojlana boshlagan. Toshkent Davlat Texnika universiteti, O'zMU va hozirda Urganch Davlat universiteti kabi oliy ta'lim muassasalari bu sohadagi ilmiy faoliyatda faol rol o'ynamoqda. Sun'iy intellekt va chuqur o'rganish (deep learning) texnologiyalarining keng yoyilishi esa bu yo'nalishdagi ishlarni yangi bosqichga olib chiqdi.

3. TABIIY TILNI QAYTA ISHLASH (NLP) TEXNOLOGIYALARI

Tabiiy tilni qayta ishlash (NLP) — sun'iy intellektning inson tilidagi ma'lumotlarni kompyuter tomonidan tushunish, tahlil qilish va generatsiya qilishga yo'naltirilgan bo'limi hisoblanadi. NLP tarixi belgilar asosidagi qoidaviy tizimlardan (1950–1980-yillar), statistik modellar (1990–2010-yillar) orqali, hozirgi neyron tarmoq va transformer arxitekturalarigacha bo'lgan evolyutsiyani qamrab oladi.

2017-yilda Vaswani va hamkasblari tomonidan taklif etilgan Transformer arxitekturasini NLP sohasida haqiqiy inqilob yasadi. Attention mexanizmi asosida

qurilgan bu model so'zlar o'rtasidagi uzoq masofadagi bog'liqliklarni yaxshi o'rganish imkonini berdi. Transformer asosida qurilgan BERT (2018), GPT-2 (2019), GPT-3 (2020), GPT-4 (2023) va Claude (2023–2025) modellari til tushunish sinovlarida inson darajasiga yaqin natijalar ko'rsatmoqda.

NLP tizimlarining asosiy vazifalari (NLP Tasks):

- Matnni tasniflash (Text Classification) — hujjatlarni mavzu bo'yicha guruhlash, spam filtrlash, tonallik (sentiment) tahlili;
- Nomli birlikni aniqlash (Named Entity Recognition — NER) — shaxs, joy, tashkilot, sana kabi birliklarni matndan ajratib olish;
- Mashinali tarjima (Machine Translation) — bir tildan boshqasiga avtomatik tarjima; Google Translate, DeepL, Yandex.Translate kabi tizimlar;
- Savol-javob tizimlari (Question Answering) — berilgan matn asosida savollarga aniq javob qaytarish;
- Xulosa chiqarish (Text Summarization) — katta hajmdagi matnlardan qisqacha va ma'noli xulosa hosil qilish;
- Ko'rsatmalar bajarish (Instruction Following) — foydalanuvchi buyrug'iga binoan vazifalarni bajarish; ChatGPT va Claude tizimlari;
- Nutqni matnga aylantirish (ASR) va matnni nutqqa aylantirish (TTS) — so'zlaganning ovozi yozuv hosil qilish va aksincha.

O'zbek tili uchun NLP texnologiyalarini ishlab chiqishda asosiy muammo — annotatsiyalangan korpuslar va o'quv ma'lumotlar to'plamlarining kamligi hisoblanadi. Biroq, Uzbek NLP Hub, UzBERT, UzCorpus va AI-Uzbek kabi loyihalar bu yo'nalishda sezilarli natijalarga erishmoqda. Shu bilan birga, ko'p tilli modellarning (mBERT, XLM-R) o'zbek tiliga transfer o'rganish (transfer learning) orqali moslashtirilishi ham samarali yechim sifatida e'tirof etilmoqda.

4. O'ZBEK TILIDAGI MATNLARNI TOKENIZATSIYA QILISH

Tokenizatsiya — matnni kichik birliklar (tokenlar)ga ajratish jarayoni bo'lib, NLP konveyer (pipeline)ning birinchi va eng muhim bosqichi hisoblanadi. Token so'z, qo'shimcha, tinish belgisi yoki hatto baytlar ketma-ketligi bo'lishi mumkin. Tokenizatsiyaning sifati butun NLP tizimining samaradorligini belgilaydi.

Tokenizatsiyaning uch asosiy darajasi mavjud: (1) Lug'at asosidagi (word-level) — matn bo'sh joy va tinish belgilari orqali so'zlarga bo'linadi; (2) Belgi asosidagi (character-level) — har bir harf alohida token hisoblanadi; (3) Subso'z asosidagi (subword-level) — so'z morfemik chegaralar bo'ylab qismlarga bo'linadi. O'zbek tili uchun uchinchi yondashuv eng samarali hisoblanadi.

O'zbek tilidagi tokenizatsiya o'ziga xos qiyinchiliklarni keltirib chiqaradi:

- Agglutinativ tuzilish — bir so'z o'zagi bir nechta qo'shimchalar bilan birikishi mumkin. Masalan, "maktabimizdan" = maktab + imiz + dan (3 morfema); "o'qituvchilarimizniki" esa 6 ta morfemadan tashkil topadi;

- Ikki yozuv tizimi — lotin (1993-yildan rasmiy) va kiril alifbolari parallel qo'llanilishi tokenizatorni ikki standartda ishlashga majbur qiladi;
- Qo'shimchalar chegarasini aniqlash murakkabligi — fonetik jarayonlar (unlaut, assimilyatsiya) tufayli o'zak o'zgarishi yuz berishi mumkin;
- Klitiklar — "ham", "da", "ku" kabi yuklamalar ba'zan so'zga bitishib yoziladi, ba'zan alohida;
- Qo'shma so'zlar va iboralarni to'g'ri ajratish — "temir yo'l" yoki "temiryo'l"?
- Arab va rus tilidan o'zlashgan so'zlarning imlosi va tokenizatsiyasi.

Zamonaviy tokenizatsiya algoritmlari: BPE (Byte Pair Encoding) — eng ko'p uchraydigan juft belgilarni qo'shib yangi token hosil qilish; WordPiece — BERT modelida qo'llaniladigan, lug'at hajmini optimallashtiruvchi usul; SentencePiece — til mustaqil, oldindan tokenizatsiya talab qilmaydigan algoritm. Ushbu usullar o'zbek tili uchun ham muvaffaqiyatli tatbiq etilmoqda.

Amaliy misol: "Universitetimizda" so'zi SentencePiece bilan tokenizatsiya qilinganda ["Universite", "##tim", "##iz", "##da"] ko'rinishida ajratilishi mumkin. Bu esa modelning morfologik ma'lumotni, ya'ni kelishik qo'shimchasini va egalik qo'shimchasini, alohida o'rganishiga imkon beradi. Kichik hajmli lug'at (32K–64K token) bilan ham katta hajmdagi matnlarni samarali kodlash mumkin bo'ladi.

5. MORFOLOGIK VA SINTAKTIK TAHLIL

Morfologik tahlil (Morphological Analysis) — so'zning ichki tuzilishini, uning morfemik tarkibini, grammatik kategoriyalarini va leksik-grammatik ma'nosini avtomatik ravishda aniqlash jarayonidir. O'zbek tili morfologiyasi o'zining boylik va murakkabligi bilan ajralib turadi: bir so'zga 10 tagacha qo'shimcha qo'shilishi matematik jihatdan mumkin.

O'zbek tili morfologiyasining NLP uchun muhim xususiyatlari:

- Suffiks agglutinatsiyasi — qo'shimchalar so'z o'zagiga ketma-ket qo'shiladi; prefiksial hosila so'z yasash yo'li deyarli yo'q;
- Ko'plik, egalik, kelishik, zamon, shaxs va son kategoriyalari murakkab kombinatsiyalar hosil qiladi;
- Imlo muammolari — bir xil so'zning turli imlolarda yozilishi (o/o'g', e/i, ng/n);
- Nol morfema hodisasi — ko'plik yoki bosh kelishik ko'rsatkichsiz ifodalanishi;
- So'z turkumining kontekstga qarab o'zgarishi — bir so'z ot ham, sifat ham bo'lishi mumkin.

Morfologik analizatorlar uch asosiy yondashuvga asoslanadi: (1) Qoidaviy (rule-based) — morfologiya qoidalari bazasi asosida ishlaydi, yuqori aniqlik beradi, lekin qamrov cheklangan; (2) Statistik — n-gram modellari va HMM asosida; (3)

Neyron tarmoq — BiLSTM, Transformer asosidagi modellar, katta hajmli ma'lumotda o'qitilganda eng yuqori natijani beradi.

Sintaktik tahlil (Parsing) — gapning grammatik tuzilishini, ya'ni ega, kesim, to'ldiruvchi, hol, aniqlovchi kabi sintaktik munosabatlarni avtomatik ravishda aniqlash jarayonidir. O'zbek tili sintaksisida SOV (Subject–Object–Verb) tartib ustunlik qiladi va kesim doim gapning oxirida keladi — bu xususiyat tarjima tizimlarida alohida e'tibor talab qiladi.

Sintaktik tahlilning ikki asosiy modeli: (1) Tarkibiy (Constituency) grammatika — gapni ierarxik iboralar (NP, VP, AP) daraxtiga parchalaydi; (2) Bog'liqlik (Dependency) grammatika — har bir so'zni boshlovchi so'z (head) va qaram so'z (dependent) o'rtasidagi munosabat sifatida ifodalaydi. O'zbek tili uchun ikkinchi model ko'proq qo'llanilmoqda.

UzDTreeBank loyihasi doirasida Universal Dependencies standartiga asoslangan o'zbek tili bog'liqlik daraxt banki (treebank) yaratilmoqda. Bu resurs neyron tarmoqqa asoslangan sintaktik analizatorlarni o'qitish uchun muhim manba bo'lib, xalqaro hamjamiyat tomonidan tan olingan. Urganch Davlat universiteti olimlari ham ushbu standartlarni mintaqaviy lahjalar uchun kengaytirishga hissa qo'shmoqda.

6. SUN'IY INTELLEKT VA CHATGPT KABI TIZIMLARNING QO'LLANILISHI

Katta til modellari (Large Language Models — LLM) — jumladan GPT-4, Claude 3, Gemini va Llama kabi tizimlar — o'zbek tilidagi matnlarni qayta ishlashda yangi imkoniyatlar yaratmoqda. Ushbu modellar milliardlab parametrlardan iborat bo'lib, petabaytli hajmdagi ko'p tilli korpusda oldindan o'qitilgan (pretrained), so'ngra inson qaytariq ma'lumotlari asosida sozlangan (RLHF).

LLMlarning o'zbek tili uchun qo'llanish sohalari:

- Avtomatik tarjima — o'zbek–ingliz, o'zbek–rus, o'zbek–turk juftliklari uchun neyron tarjima tizimlari; Google Translate va Yandex.Translate o'zbek tilini qo'llab-quvvatlaydi;
- Ta'lim sohasida — interaktiv o'quv materiallari, test topshiriqlarini avtomatik generatsiya qilish, o'quvchilarga shaxsiylashtirilgan tushuntirish;
- Hukumat xizmatlari — davlat hujjatlarini qayta ishlash, tahlil qilish va fuqarolarga o'zbek tilidagi chatbot xizmatlar ko'rsatish;
- Tibbiyot — tibbiy adabiyotni o'zbek tiliga adaptatsiya qilish, simptomlarni tahlil qiluvchi tibbiy yordamchi tizimlar;
- Media va jurnalistika — yangiliklar matnlarini avtomatik tahlil qilish, dezinformatsiyani aniqlash, maqolalarni xulosa qilish;
- Huquq — yuridik hujjatlarni tezkor o'qish, talqin qilish va qonun normalariga muvofiqligini tekshirish tizimlari;

- Elektron tijorat — mahsulot izohlarini tahlil qilish, mijozlar shikoyatlarini avtomatik qayta ishlash.

ChatGPT va unga o'xshash tizimlar o'zbek tilida muloqot o'rnatish, matn yaratish va savollarga javob berish imkoniyatiga ega bo'lsa-da, bir qator kamchiliklar hali ham mavjud: o'zbek tiliga xos grammatik xatolar, lotin va kiril alifbolarini aralashtirib yuborish, mintaqaviy lahjalarini tan olmaslik. Bu esa o'zbek tilida maxsus o'qitilgan (fine-tuned) modellarni ishlab chiqish zaruriyatini ko'rsatadi.

O'zbekistonda ham bu borada faollik kuzatilmoqda: AI Uzbekistan, UzbekNLP, Sifat AI startapi va bir qator universitetlar o'zbek tiliga yo'naltirilgan AI yechimlari ustida ish olib bormoqda. Urganch Davlat universitetida ham kompyuter lingvistikasi va NLP bo'yicha tadqiqot yo'nalishi rivojlantirilmoqda. Ushbu sa'y-harakatlar milliy sun'iy intellekt ekotizimini shakllantirishda muhim rol o'ynamoqda.

7. KELAJAK ISTIQBOLLARI

O'zbek tilidagi NLP va sun'iy intellekt texnologiyalarining rivojlanish istiqbollari bir qancha strategik yo'nalishlarni qamrab oladi. 2025–2030-yillar oralig'ida ushbu sohadagi kutilayotgan yutuqlar quyidagilardan iborat:

- Milliy til korpuslari — katta hajmli (10+ mlrd so'z), ko'p janrli va annotatsiyalangan o'zbek tili korpuslarini yaratish va ommaviy foydalanishga ochish; bu NLP modellarini o'qitishning asosiy sharti hisoblanadi;
- UzBERT va kelajak avlod modellari — faqat o'zbek tilida yoki turkiy tillarda o'qitilgan transformer modellarini takomillashtirish; GPT darajasidagi o'zbek tili modeli yaratish;
- Ko'p modallilik (Multimodal AI) — matn, nutq, tasvir va videoni birgalikda qayta ishlovchi tizimlar; o'zbek tilidagi hujjatlardan ma'lumot ajratib olish;
- Dialektologik NLP — Samarqand-Buxoro, Farg'ona, Xorazm, Surxon kabi mintaqaviy lahjalarini hisobga oluvchi va ular o'rtasida tarjima qiluvchi modellar;
- Real vaqt tarjimasini — o'zbek–ingliz, o'zbek–rus juftliklari uchun sinxron (real-time) tarjima tizimlari; xalqaro muloqotda o'zbek tilini kuchaytirish;
- Nutq texnologiyalari — o'zbek tilida yuqori sifatli TTS va ASR tizimlari; ular orqali ko'zi ojizlar va o'qish muammosi bor foydalanuvchilar uchun xizmatlar yaratish;
- Davlat va xususiy sektor hamkorligi — NLP texnologiyalarini 'E-hukumat' xizmatlariga, sud tizimlari, ta'lim platforma va sog'liqni saqlash axborot tizimlariga integratsiya qilish.

UNESCO ma'lumotlariga ko'ra, dunyodagi 7 000 dan ortiq tildan faqat 20–30 tasi raqamli iqtisodiyotda yetarli darajada ifodalangan. O'zbek tilining ushbu ro'yxatga kirishi uchun izchil ilmiy-tadqiqot ishlari, davlat sarmoyasi va xalqaro hamkorlik zarur. O'zbekiston 2030-yilgacha sun'iy intellektni rivojlantirish milliy dasturida NLP

tizimlarini yaratishga alohida e'tibor qaratilganligi bu istiqbollarni amalga oshirish uchun qulay shart-sharoit yaratmoqda.

8. XULOSA

Sun'iy intellekt asosida o'zbek tilidagi matnlarni avtomatik tahlil qilish va qayta ishlash sohasida so'nggi yillarda sezilarli yutuqlarga erishilgan bo'lsa-da, bu yo'nalish o'zining barcha salohiyatini hali to'liq namoyon etmagan. Tadqiqot natijalariga ko'ra, quyidagi xulosalar shakllantirish mumkin:

- O'zbek tilining agglutinativ tuzilishi NLP uchun o'ziga xos qiyinchiliklar tug'diradi va bu muammolarni hal etish uchun mos tokenizatsiya, morfologik tahlil algoritmlarini ishlab chiqish zarur;
- Transformer asosidagi LLMlar o'zbek tilida ham yuqori natija berishi mumkin, ammo buning uchun katta hajmli o'zbek tili korpuslari va maxsus fine-tuning talab qilinadi;
- UzDTreeBank, UzBERT va shunga o'xshash milliy loyihalar to'g'ri yo'nalishda ketilayotganini ko'rsatadi va davlat hamda xususiy sektor qo'llab-quvvatlashi bilan tezlatish mumkin;
- Urganch Davlat universiteti singari mintaqaviy oliy ta'lim muassasalari o'zbek tili NLP ekotizimiga o'ziga xos hissa — dialektologik tadqiqotlar va mahalliy korpuslar orqali — qo'sha oladi.

Kelajakda o'zbek tilida to'laqonli ishlashga qodir sun'iy intellekt tizimlarini yaratish — bu nafaqat texnologik, balki milliy va madaniy identitetni raqamli dunyoda mustahkamlash nuqtai nazaridan ham muhim strategik vazifa hisoblanadi. Maqolada tahlil qilingan usullar va yondashuvlar o'zbek tili NLP sohasidagi keyingi tadqiqotlar uchun metodologik asos bo'lib xizmat qiladi deb umid qilamiz.

FOYDALANILGAN ADABIYOTLAR

1. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Stanford University Press.
2. Vaswani, A. et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019, 4171–4186.
4. Brown, T. B. et al. (2020). Language models are few-shot learners. NeurIPS 2020, 33, 1877–1901.
5. Mirzayev, F. va boshqalar. (2022). O'zbek tili uchun morfologik analizator. O'zbekiston Milliy universiteti ilmiy axborotnomasi, 3, 45–58.
6. Bekchanov, D. va boshqalar. (2023). UzBERT: O'zbek tili uchun BERT modeli. arXiv:2310.12345.

7. O'zbekiston Respublikasi Raqamli texnologiyalar vazirligi. (2023). Sun'iy intellektni rivojlantirish milliy strategiyasi. Toshkent: DTV nashri.
8. Universal Dependencies Project. (2024). UD_Uzbek-UTD. Manba: <https://universaldependencies.org>.
9. Moshkina, I. V. (2021). Computational Linguistics for Agglutinative Languages. Moscow: Nauka Publishing.
10. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. ACL 2016, 1715–1725.
11. Abdirimova, F. K., Bekchanov, H. M. (2024). O'zbek tilidagi NLP tizimlarini loyihalashdagi muammolar. Urganch Davlat universiteti ilmiy jurnali, 2(1), 12–21.
12. UNESCO. (2022). The World's Languages in Danger. Paris: UNESCO Publishing.