

e-ISSN 3060-5059

DOI: 10.67227



IJTIMOYIY-GUMANITAR SOHADA ILMIY-INNOVATSION TADQIQOTLAR

ILMIY METODIK JURNAL

Scientific and Innovative Research in
Social and Humanitarian Studies

Vol. 3 • Issue 9

2026



Founded in 2024



publishscience.uz



publish_science_journal@mail.ru



PEER REVIEWED • OPEN ACCESS

KOMPYUTER LINGVISTIKASI VA SAVOL-JAVOB TIZIMLARINING RIVOJLANISH TARIXI: JAHON TAJRIBASI VA O'ZBEK TILI UCHUN XULOSALAR

Po'latova Gulxayyo G'ofurjon qizi

Namangan davlat pedagogika instituti, o'qituvchi (PhD).

e-mail: gpolatova2409@gmail.com ORCID: <https://orcid.org/0009-0008-0158-5916>

Annotatsiya.

Ushbu maqolada kompyuter lingvistikasining shakllanish jarayoni doirasida dastlab mashina tarjimasi va korpus lingvistikasi haqida qisqacha tarixiy ma'lumot beriladi. So'ngra savol-javob tizimlarining rivojlanish bosqichlari ko'rib chiqiladi. Bunda qat'iy qoidalarga asoslangan dastlabki dasturlardan boshlab, neyron tarmoqlarga asoslangan "o'quvchi" modellar va Retrieval-Augmented Generation (RAG) deb nomlanuvchi zamonaviy arxitekturagacha bo'lgan taraqqiyot yo'li tahlil qilinadi. Jahon tajribasi bilan bir qatorda turkiy tillar guruhi hamda boshqa agglyutinatativ tuzilishga ega tillar misolida morfologik jihatdan murakkab tillarga xos yechimlarning qanday sinovdan o'tkazilgani ko'rib chiqiladi. Shuningdek, o'zbek tilida amalga oshirilgan kompyuter lingvistikasi sohasidagi ishlar tizimlashtirilib, ularning savol-javob tizimlarini yaratish uchun qay darajada asos bo'la olishi baholanadi.

Kalit so'zlar: kompyuter lingvistikasi, mashina tarjimasi, korpus lingvistikasi, savol-javob tizimlari, RAG, agglyutinatativ tillar, til modellari, o'zbek tili, tokenizatsiya.

ИСТОРИЯ РАЗВИТИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ И СИСТЕМ ВОПРОС-ОТВЕТ: МИРОВОЙ ОПЫТ И ВЫВОДЫ ДЛЯ УЗБЕКСКОГО ЯЗЫКА

Пулатова Гулхайё Гофуржон кизи

Наманганский государственный педагогический институт, преподаватель (PhD).

Аннотация.

В статье в контексте становления компьютерной лингвистики первоначально представлен краткий исторический обзор развития машинного перевода и корпусной лингвистики. Затем рассматриваются этапы развития вопросно-ответных систем. Анализируется путь от ранних программ, основанных на строгих правилах, через нейросетевые модели «чтения» к современной архитектуре Retrieval-Augmented Generation (RAG). Наряду с мировым опытом на примере тюркских языков и других языков агглютинативного типа рассматриваются подходы, применяемые для решения задач, характерных для морфологически сложных языков. Кроме того, систематизируются исследования в области компьютерной лингвистики узбекского языка и оценивается их потенциал в качестве основы для разработки вопросно-ответных систем.

Ключевые слова: компьютерная лингвистика, машинный перевод, корпусная лингвистика, вопросно-ответные системы, RAG, агглютинативные языки, языковые модели, узбекский язык, токенизация.

THE HISTORY OF THE DEVELOPMENT OF COMPUTATIONAL LINGUISTICS AND QUESTION-ANSWERING SYSTEMS: WORLD EXPERIENCE AND CONCLUSIONS FOR THE UZBEK LANGUAGE

Po'latova Gulxayyo G'ofurjon qizi

Namangan State Pedagogical Institute, Lecturer (PhD).

Abstract.

This article provides a brief historical overview of machine translation and corpus linguistics within the development of computational linguistics. It then examines the major stages in the evolution of question-answering systems. The progression from early rule-based programs through neural network-based reading models to the modern Retrieval-Augmented Generation (RAG) architecture is analyzed. Alongside international experience, the article examines approaches developed for morphologically complex languages, with particular reference to Turkic languages and other languages with agglutinative structures. In addition, research in Uzbek computational linguistics is systematized and its potential as a foundation for developing question-answering systems is assessed.

Keywords: computational linguistics, machine translation, corpus linguistics, question-answering systems, RAG, agglutinative languages, language models, Uzbek language, tokenization.

Kirish. Kompyuter lingvistikasi tabiiy tilni avtomatik tahlil qilish va yaratish bilan shug'ullanuvchi fan hisoblanadi. U lingvistika, informatika hamda sun'iy intellekt tutashgan nuqtada shakllangan. Sohaning tarixiga razm solinsa, bir nechta yo'nalish bir vaqtning o'zida, biroq dastlab bir-biridan deyarli mustaqil tarzda rivojlanganini ko'rishimiz mumkin. Bu matnni bir tildan boshqasiga avtomatik o'girish bilan shug'ullanuvchi mashina tarjimasi; katta hajmdagi matn to'plamlarini yig'ish, tartiblash va statistik jihatdan tahlil qilish bilan band bo'lgan korpus lingvistikasi hamda foydalanuvchi bergan savolga hujjatlar yoki bilim bazasi negizida javob qidiruvchi savol-javob tizimlari kabilarni o'z ichiga oladi. Dastlab bu yo'nalishlar bir-biriga bog'liq bo'lmagan holda shakllangan. Keyinchalik ular bir-biri bilan chambarchas aloqaga kirishdi.

O'zbek tili nuqtayi nazaridan mazkur tarixiy yo'lni chuqur o'rganish alohida ahamiyatga ega. Ma'lumki, o'zbek tili agglutinativ tillar qatoriga kiradi. Yangi ma'no yoki grammatik shakl ildizga bir nechta qo'shimchaning ketma-ket birikishi orqali hosil bo'ladi. Natijada bitta leksema o'nlab, hatto yuzlab turli sirt shaklda namoyon bo'lishi mumkin. Bu xususiyat qidiruv mexanizmlariga, tokenizatsiya jarayoniga hamda javoblarni avtomatik baholash uslubiyotiga bevosita ta'sir ko'rsatadi [9]. Xuddi shu sababdan ingliz tili singari flektiv bo'lmagan tillar uchun ishlab chiqilgan yechimlarni o'zbek tiliga hech qanday moslashtirishsiz, o'zgarishsiz tatbiq etish kutilgan natijani bermasligi mumkin.

Maqola bir-biri bilan uzviy bog'liq kompyuter lingvistikasining boshqa ikkita muhim tarmog'i, QA tizimlari misolida jahon amaliyotidagi asosiy bosqichlarni izchil tartibda yoritish, turkiy hamda boshqa agglutinativ tillar tajribasidan kelib chiqib, o'zbek tili uchun amaliy xulosalar shakllantirish, o'zbek tilida mavjud lingvistik resurslar va til modellarini tizimlashtirib, savol-javob tizimini yaratishda aynan qaysi bo'g'inda bo'shliq borligini aniq ko'rsatishga qaratilgan.

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Tadqiqot tanlanma adabiyotlar sharhi shaklida olib borildi. Manbalarni tanlashda sohada burilish nuqtasi bo'lgan, ko'p marta iqtibos keltirilgan asosiy ishlar va turkiy hamda agglutinativ tillarga bevosita taalluqli resurs va tajriba ishlari mezonlaridan foydalanildi. Adabiyotlar sharhi uslubining o'ziga xos afzalligi shundaki, u turli davr va turli yo'nalishga oid manbalarni bir izchil hikoya tarzida, bosqichma-bosqich taraqqiyot sifatida taqdim etish imkonini beradi. Bu esa o'quvchiga sohaning umumiy rivojlanish mantig'ini tezroq va yaxlit tasavvur qilish imkoniyatini yaratadi.

Kompyuter yordamida tarjima qilish g'oyasi ilk bor 1949-yilda amerikalik olim W. Weaver tomonidan tayyorlangan memorandumda ilmiy jihatdan asoslab berildi. U Ikkinchi jahon urushi davrida shifrlarni yechish jarayonida qo'llanilgan kriptografik usullarni tabiiy tilni tarjima qilishga tatbiq etish mumkinligini taxmin qildi [1; 15–23-b.]. Aynan shu g'oya keyingi o'n yilliklar davomida mashina tarjimasi sohasining rivojlanish yo'nalishini belgilab berdi. Dastlab qoidaviy, so'ngra statistik va hozirgi kunda esa neyron tarmoqqa asoslangan yondashuvlar aynan shu poydevor ustida qad ko'tardi. Katta hajmdagi, tizimli tarzda tuzilgan matn to'plamlarini yaratish g'oyasi 1960-yillarda paydo bo'ldi. Braun universitetidagi amerikalik olimlar G. Kuchera va W. Frensis taxminan bir million so'zdan iborat, turli janrlarni muvozanatli tarzda qamrab olgan birinchi yirik elektron korpusni, ya'ni mashhur Braun korpusini yaratib, uni statistik metodlar yordamida tahlil qilishdi [2; 1-b.]. Bu ish o'zidan keyingi barcha korpus lingvistikasi va statistik NLP tadqiqotlari uchun uslubiy namuna bo'lib xizmat qildi. Yuqorida sanab o'tilgan ikki yo'nalish kompyuter lingvistikasining boshqa nihoyatda muhim tarmoqlaridandir.

Ilk savol-javob dasturlaridan biri hisoblangan Baseball 1961-yilda yaratilgan bo'lib, u beysbol musobaqalari haqidagi ma'lumotlar bazasiga tabiiy tilda kiritilgan savollarga javob bera olgan [3; 219–224-b.]. Bunday tizimlar juda tor predmet sohasida samarali ishlagan va qo'lda tuzilgan qat'iy qoidalarga to'liq tayangan, shu sababli ularni boshqa mavzu yoki sohaga ko'chirish deyarli imkonsiz edi.

Qoidaviy dasturlar bilan neyron davr o'rtasida muhim oraliq bosqich mavjud. 1999-yildan boshlab AQSh Milliy Standartlar va Texnologiyalar Instituti (NIST) tashkil etgan TREC konferensiyasining savol-javob yo'nalishi savol-javob tizimlarini yagona metodika asosida, katta hajmdagi yangilik matnlari to'plamida solishtirish imkonini berdi. Bu jarayon sohada birinchi marta ommaviy, takrorlanuvchi baholash madaniyatini shakllantirdi [11]. Shuningdek, xitoy-amerikalik tadqiqotchi D. Chen va hammualliflari 2017-yilda taqdim etgan DrQA tizimida ochiq domenli savol-javob jarayoni ikki alohida bosqichga ajratildi. Birinchi bosqichda bigram xeshlash va TF-IDF statistikasi asoslangan qidiruvchi modul Wikipedia hajmidan mos maqolalarni tanlaydi. Ikkinchi bosqichda esa ko'p qatlamli takrorlanuvchi neyron tarmoq tanlangan matndan aniq javob bo'lagini ajratib oladi [4; 70–79-b.]. Bunday "o'quvchi" modellarning o'qitilishi ko'p jihatdan bir yil oldin taqdim etilgan SQuAD kabi yirik miqyosli matnni o'qib tushunish benchmarklariga bog'liq bo'lgan. Bu to'plam Wikipedia maqolalari asosida 100 mingdan ortiq savol-javob juftligini o'z ichiga olib, javobni aynan manba matndan ajratib olish vazifasini standartlashtirdi va shu orqali keyingi barcha neyron o'quvchi modellar uchun umumiy o'lchov mezoniga aylandi [14; 83–92-b.].

Britaniyalik tadqiqotchi P. Lewis va hammualliflari tayyor seq2seq generator modelni Wikipedia matnlaridan tuzilgan zich indeksdan olingan parchalar bilan birlashtirib, model parametrlarida saqlanadigan bilimga tashqi, dinamik ravishda yangilanadigan bilim manbaini qo'shdilar [5; 59–74-b.].

1-jadval. Savol-javob tizimlari taraqqiyotining to'rtta asosiy bosqichi

Bosqich	Davr	Vakil ish	Asosiy cheklov
Qoidaviy	1960-y.	Baseball	Predmet sohasi tor, boshqa mavzuga ko'chirish qiyin
Gibrid - statistik	1999–2011	TREC QA, IBM Watson	Ko'plab qo'lda yozilgan modulga tayanadi, chuqur o'qitish yo'q
Neyron	2016–	DrQA, SQuAD	Javob faqat matndan ajratiladi,

o'quvchi	2019		komponentlar alohida o'qitiladi
RAG	2020– hozir	DPR, RAG	Natija retriever va generator sifatiga qattiq bog'liq

Amerikalik tadqiqotchi Jonathan H. Clark va hammualliflari tomonidan yaratilgan TyDi QA to'plami 11 ta tipologik jihatdan bir-biridan farq qiluvchi tilni, jumladan, fin va koreys kabi agglyutinatativ tillarni ham o'z ichiga oladi [6; 54–70-b.]. Bu to'plamning muhim afzalligi shundaki, savollar boshqa tildan tarjima qilinmagan, balki har bir til uchun to'g'ridan-to'g'ri o'sha tilda tuzilgan. Bu esa tarjima qilingan savollarga xos bo'lgan tilning o'ziga xos xususiyatlarini yo'qotib qo'yish muammosining oldini oladi.

Turkiy tillar oilasi ichida o'zbek tili uchun eng yaqin va o'rganishga arzigulik namuna KazQAD to'plamidir. Qozog'istonlik olim R. Yeshpanov va hamkasblari ushbu to'plamni deyarli 6 ming noyob savol, 12 mingga yaqin o'rin darajasidagi relevantlik bahosi hamda qozoq tilidagi Wikipedia'dan olingan 800 mingdan ortiq o'rinlik korpus asosida shakllantirishdi [7; 145–156-b.].

Turkiy tillar oilasidan tashqarida turgan, ammo tipologik jihatdan yaqin bo'lgan fin-ugor tillari oilasiga mansub venger tili ham yuqori darajada agglyutinatativ tuzilishga ega. Budapeshtlik tadqiqotchi A. Novak va hammualliflari venger tili uchun MILQA nomli savol-javob benchmark bazasini yaratishdi. Ular ingliz tilidagi SQuAD 2.0 tamoyillariga asoslangan bo'lsa-da, ha-yo'q savollari, bir nechta matn bo'lagidan iborat javoblar hamda hisob-kitob talab qiluvchi savollar kabi qo'shimcha savol turlarini kiritish orqali uni kengaytirishdi [16; 188–198-b.].

Turkiyalik tadqiqotchi C. Toraman va hammualliflari turk tili uchun beshta xil tokenizatsiya strategiyasini bir-biri bilan qiyoslab chiqishdi. Tadqiqot natijalariga ko'ra, morfologik segmentatsiyaga asoslangan tokenizator BPE va WordPiece kabi keng tarqalgan usullar bilan bemaolol raqobatlasha oldi. Ayniqsa, lug'at hajmi kattalashtirilgan sharoitda morfologik va so'z darajasidagi usullar nisbatan ko'proq foyda keltirdi [8; 1–21-b.].

Xitoyda tadqiqot olib borayotgan tadqiqotchilar Alim Murat va Samat Ali morfologik jihatdan boy uchta turkiy til, ya'ni uyg'ur, o'zbek va qirg'iz tillari uchun yagona teglash tizimi asosida MultiPOS_ukg nomli maxsus ma'lumot to'plamini yaratgan. Bunda o'zakning prefiks va suffiksilarini chuqur o'rganuvchi hamda ko'p boshli diqqat mexanizmiga asoslangan POS teglash usulini taklif qilishdi [15; 66495–66504-b.].

O'zbekistonlik U. Salayev qo'shimchalarning to'liq ro'yxatiga tayangan UzMorphAnalyser vositasini taklif qildi. Ushbu tizim so'zning negizini aniqlaydi, uning lemmasini topadi, to'liq morfologik ma'lumotni ajratib beradi hamda morfo-fonetik istisno holatlarni ham hisobga oladi [9].

O'zbekistonlik N. Bobojonova va hamkasblari ikkita monolingual o'zbek tilidagi BERT modelini so'z turkumlarini avtomatik belgilash vazifasida sinab ko'rishdi hamda shu jarayonda o'zbek tili uchun birinchi ochiq, UPOS standartida teglangan benchmark to'plamini yaratishdi [10; 287–293-b.].

O'zbek tili uchun BERT arxitekturasiga asoslangan dastlabki monolingual modellardan biri bo'lmish UzBERT modeli 2021-yilda taqdim etilgan. Keyinroq o'zbekistonlik tadqiqotchi E. Quriyozov va hammualliflari BERTbek nomli, turli hajm va sifatdagi matn manbalari asosida o'qitilgan bir necha BERT modelini taklif qilishdi hamda ularni sentiment tahlili, ko'p yorliqli mavzu klassifikatsiyasi va nomlangan birlik (NER) vazifalarida ko'p tilli mBERT bilan solishtirib, sinovdan o'tkazishdi. Natijalar BERTbek modellarining barcha uchta vazifada ham mBERT'dan ustun ekanligini ko'rsatdi [17; 33–44-b.].

Shuningdek, matn generatsiya qiluvchi yirik til modellari sohasida o'zbek tili hamon kam resursli til sifatida qaraladi. X. Yusupxo'jayev yaqinda 0,6–2 milliard parametrlilik yettita ochiq manbali modelni o'zbekcha tarjima, matnni tushunish va matn generatsiya qilish vazifalarida sinab ko'rdi. Natijalarga ko'ra, sinalgan modellarning aksariyati izchil bo'lmagan yoki tillarni aralashtirib yuboradigan natija berdi. Faqat bitta Kimi-K2-Instruct modeli amaliy jihatdan qoniqarli natijaga erishdi [18; 67–73-b.].

Shunga yaqin mavzuda M. Najmiddinov zamonaviy yirik til modellarning o'zbekcha matn generatsiyasidagi tipik xatolarini tizimli tahlil qildi va qo'shimchalar tartibi, unlilar uyg'unligi, kelishik-shaxs mosligi hamda "Sen" va "Siz" farqlanishi kabi grammatik sohalaridagi xatolarning umumiy xato hajmining uchdan bir qismidan ortig'ini tashkil etishini aniqladi. Muallifning xulosasiga ko'ra, subword darajasidagi tokenizatsiya va inglizcha me'yorlarga asoslangan arxitekturalar o'zbek tilining morfologik boyligini to'liq aks ettira olmaydi [19].

2-jadval. Savol-javob tizimlari uchun ahamiyatli o'zbekcha resurslar

Resurs	Nimani beradi	Nimani bermaydi
UzMorphAnalyser	Stemming, lemmatizatsiya, morfologik belgilar (91%+ aniqlik)	Semantik qidiruv va javob shakllantirish
BBPOS / UPOS to'plami	Kontekstli POS teglash (lotin va kirill), ochiq benchmark	QA yoki retrieval baholash imkoniyati
BERTbek va boshqa generativ LLM'lar	Matnni tushunish va generatsiya qilish uchun asosiy model salohiyati	QA-retrieval vazifasiga moslashtirish va shu yo'nalishda baholash
Ochiq domenli QA to'plami	Ko'rib chiqilgan manbalarda uchramadi	Mavjud bo'shliq sifatida qayd etiladi

Yuqorida sanab o'tilgan to'rt bosqich bir-birini butunlay inkor etmasdan, aksincha, oldingi bosqichning yutuqlari

ustiga qurilib borganini alohida ta'kidlash lozim. Masalan, TREC va Watson davrida shakllangan ommaviy baholash madaniyati keyinchalik SQuAD kabi benchmarklarda, so'ngra zamonaviy RAG tizimlarida ham davom ettirildi. Qidiruv mexanizmi esa TF-IDF kabi statistik usullardan DPR singari zich vektorli qidiruvga almashtirilgan. Shuningdek, javob shakllantirish bosqichi matndan ajratib olishdan erkin generatsiyaga o'tgan. Bu kuzatuvga ko'ra, kelajakda o'zbek tili uchun QA tizimi qurilganda ham bosqichma-bosqich yondashuv oqilona strategiya bo'lib xizmat qilishi mumkinligini ko'rishimiz mumkin.

MUHOKAMA

Yuqorida keltirilgan xalqaro tajribadan o'zbek tili uchun bir qator xulosalarni ayta olamiz. Bunda milliy, ya'ni o'zbek tilida to'plangan QA benchmarkini yaratish eng samarali dastlabki qadam bo'lishi mumkin. Qozoq tilidagi KazQAD hamda venger tilidagi MILQA tajribasi shuni ko'rsatadiki, o'qitish uchun zarur bo'lgan ma'lumotlarni mashina tarjimasida yordamida olish mumkin bo'lsa-da, ishonchli baholash tizimini qurish uchun asl, mahalliy sharoitda to'g'ridan-to'g'ri yig'ilgan savollar talab etiladi. Hamda tokenizatsiya usulini tanlash faqat eksperimental tekshiruv yordamida aniqlanishi lozim. Turk tili misoli morfologik segmentatsiyaning muayyan foyda berishi mumkinligini ko'rsatadi. Biroq bunday natija lug'at hajmi va aniq bajariladigan vazifaga bog'liq bo'lib qoladi. Bu taxmin o'zbekcha generativ LLM'lar ustida o'tkazilgan so'nggi tadqiqot bilan ham tasdiqlanadi. Shuningdek, mavjud modellarning aksariyati subword darajasidagi tokenizatsiya tufayli o'zbek tilining qo'shimchalar tizimini to'g'ri aks ettira olmaydi [19]. Shu nuqtayi nazardan, UzMorphAnalyser asosida qurilgan segmentatsiya usulini oddiyroq, an'anaviy tokenizatsiya usullari bilan solishtirib ko'rish o'zbek tili uchun ilmiy jihatdan o'rinni va foydali bo'lardi. Kelajak tadqiqotlar uchun o'zbek tiliga maxsus moslashtirilgan morfologik tokenizator yaratish alohida vazifa sifatida qo'yilishi maqsadga muvofiq; bu maqolaning ilmiy xulosasini yanada mustahkamlaydi. Bundan tashqari, baholash metodikasini o'zbek tili morfologiyasiga moslashtirish zarurati tug'ilishi mumkin. Bir xil mazmundagi javob turlicha grammatik shaklda ifodalanishi mumkinligi sababli, an'anaviy aniq moslik (Exact Match — EM) ko'rsatkichi sun'iy ravishda past chiqishi ehtimoldan xoli emas. Shu bois semantik o'xshashlik metrikalari (ROUGE, BLEU yoki BERTScore)ni qo'llash tavsiya etiladi.

Amaliy nuqtayi nazardan, yuqorida keltirilgan xulosalarni hayotga tatbiq etish uchun bir necha tomonlama harakat talab etiladi. Bunda akademik muassasalar va davlat tashkilotlari o'rtasida hamkorlik o'rnatilib, o'zbek tilida savollar va javoblarni yig'ish bo'yicha maxsus loyiha ishga tushirilishi maqsadga muvofiq bo'ladi. Bunday loyihada KazQAD va MILQA tajribasidagi kabi savollar tarjima qilinmasdan, aynan o'zbek tilida so'zlashuvchi ishtirokchilardan to'g'ridan-to'g'ri yig'ilishi lozim. Bundan tashqari, mavjud UzMorphAnalyser va BBPOS kabi vositalarni, shuningdek, BERTbek va boshqa mavjud o'zbekcha modellarni retriever yoki generator sifatida QA vazifasida sinovdan o'tkazish, ularning real ta'sirini o'lchash zarur. Shuningdek, baholash metodikasini ishlab chiqishda faqat aniq moslik (EM) ko'rsatkichiga tayanmasdan, semantik o'xshashlikni hisobga oluvchi qo'shimcha mezonlarni — masalan, ROUGE, BLEU yoki vektor embedding asosidagi BERTScore kabi metrikalarni — joriy etish zarur. Bu o'zbek tilining boy morfologiyasi sharoitida javoblarni yanada adolatli va ishonchli baholash imkonini beradi.

XULOSA

Kompyuter lingvistikasi mashina tarjimasida, korpus lingvistikasi va savol-javob tizimlari kabi bir necha yo'nalishda parallel tarzda rivojlanib keldi. Savol-javob tizimlari esa o'z navbatida qat'iy qoidalarga asoslangan dastlabki dasturlardan neyron tarmoqli o'quvchi modellarga, so'ngra RAG arxitekturasiga qadar izchil taraqqiy etdi. Qozoq, venger va turk tillari misolida ko'rib chiqilgan tajriba mahalliy resurs yaratishning hamda tilga xos modellashtirishning naqadar muhim ahamiyat kasb etishini yaqqol namoyish etadi.

O'zbek tilida bugungi kunga kelib morfologik tahlil, so'z turkumini belgilash hamda umumiy maqsadli til modellari bo'yicha ochiq va sifatli vositalar mavjud. Biroq ochiq domenli savol-javob to'plami ko'rib chiqilgan manbalar orasida uchramadi. Bundan tashqari, mavjud til modellari esa hali maxsus QA vazifasida sinovdan o'tkazilmagan. Shu sababdan, eng amaliy va zarur keyingi qadam sifatida qozoq tilidagi KazQAD va venger tilidagi MILQA tajribasiga tayanib, bevosita o'zbek tilida yig'ilgan, ishonchli QA benchmarkini yaratish hamda mavjud o'zbekcha til modellarini shu benchmarkda sinovdan o'tkazish taklif etiladi. Bunday to'plam kelajakda o'zbek tili uchun RAG asosidagi zamonaviy savol-javob tizimlarini qurishning ishonchli asosiga aylanishi mumkin.

FOYDALANILGAN ADABIYOTLAR

1. Weaver W. Translation // *Machine Translation of Languages: Fourteen Essays* / ed. by W. N. Locke, A. D. Booth. – Cambridge, MA: The Technology Press of MIT; New York: John Wiley & Sons, 1955. – P. 15–23. – URL: <https://aclanthology.org/www.mt-archive.info/srch/genmisc-50.htm> ACL Anthology
2. Kučera H., Francis W. N. *Computational Analysis of Present-Day American English*. – Providence, RI: Brown University Press, 1967. – 424 p. – URL: <https://books.google.com/books?id=Gb55AAAAIAAJ> Google Books
3. Green B. F., Wolf A. K., Chomsky C., Laughery K. Baseball: An Automatic Question-Answerer // *Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference*. – 1961. – P. 219–224. – DOI: 10.1145/1460690.1460714. – URL: <https://doi.org/10.1145/1460690.1460714> DOI
4. Chen D., Fisch A., Weston J., Bordes A. Reading Wikipedia to Answer Open-Domain Questions // *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vol. 1: Long Papers*. – Vancouver, Canada, 2017. – P. 1870–1879. – DOI: 10.18653/v1/P17-1171. – URL: <https://aclanthology.org/P17-1171/> ACL Anthology
5. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // *Advances in Neural Information Processing Systems*. – 2020. – Vol. 33. – P. 9459–9474. – URL: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>

6. Clark J. H., Choi E., Collins M., Garrette D., Kwiakowski T., Nikolaev V., Palomaki J. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages // *Transactions of the Association for Computational Linguistics*. – 2020. – Vol. 8. – P. 454–470. – DOI: 10.1162/tacl_a_00317. – URL: <https://aclanthology.org/2020.tacl-1.30/> ACL Anthology
7. Yeshpanov R., Efimov P., Boytsov L., Shalkarbayuli A., Braslavski P. KazQAD: Kazakh Open-Domain Question Answering Dataset // *Proceedings of LREC-COLING 2024*. – Torino, Italy, 2024. – P. 9645–9656. – URL: <https://aclanthology.org/2024.lrec-main.843/> ACL Anthology
8. Toraman C., Yilmaz E. H., Şahinuç F., Özcelik O. Impact of Tokenization on Language Models: An Analysis for Turkish // *ACM Transactions on Asian and Low-Resource Language Information Processing*. – 2023. – Vol. 22, No. 4. – Article 116. – P. 1–21. – DOI: 10.1145/3578707. – URL: <https://doi.org/10.1145/3578707> OpenMETU
9. Salaev U. UzMorphAnalyser: A Morphological Analysis Model for the Uzbek Language Using Inflectional Endings // *AIP Conference Proceedings*. – 2024. – Vol. 3244, No. 1. – Article 030058. – DOI: 10.1063/5.0241461. – URL: <https://doi.org/10.1063/5.0241461> PyPI
10. Bobojonova L., Akhundjanova A., Ostheimer P. S., Fellenz S. BBPOS: BERT-Based Part-of-Speech Tagging for Uzbek // *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)*. – Abu Dhabi, UAE, 2025. – P. 287–293. – URL: <https://aclanthology.org/2025.loreslm-1.23/> ACL Anthology
11. Voorhees E. M., Tice D. M. The TREC-8 Question Answering Track // *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*. – Athens, Greece: ELRA, 2000. – URL: <https://aclanthology.org/L00-1018/> ACL Anthology
12. Ferrucci D., Brown E., Chu-Carroll J., Fan J., Gondek D., Kalyanpur A. A., Lally A., Murdock J. W., Nyberg E., Prager J., Schlaefel N., Welty C. Building Watson: An Overview of the DeepQA Project // *AI Magazine*. – 2010. – Vol. 31, No. 3. – P. 59–79. – DOI: 10.1609/aimag.v31i3.2303. – URL: <https://doi.org/10.1609/aimag.v31i3.2303> ResearchGate
13. Karpukhin V., Oğuz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W. Dense Passage Retrieval for Open-Domain Question Answering // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. – 2020. – P. 6769–6781. – DOI: 10.18653/v1/2020.emnlp-main.550. – URL: <https://aclanthology.org/2020.emnlp-main.550/> ACL Anthology
14. Rajpurkar P., Zhang J., Lopyrev K., Liang P. SQuAD: 100,000+ Questions for Machine Comprehension of Text // *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. – Austin, Texas, 2016. – P. 2383–2392. – DOI: 10.18653/v1/D16-1264. – URL: <https://aclanthology.org/D16-1264/> ACL Anthology
15. Murat A., Ali S. Low-Resource POS Tagging With Deep Affix Representation and Multi-Head Attention // *IEEE Access*. – 2024. – Vol. 12. – P. 66495–66504. – DOI: 10.1109/ACCESS.2024.3395454. – URL: <https://doi.org/10.1109/ACCESS.2024.3395454> ResearchGate
16. Novák A., Novák B., Zombori T., Szabó G., Szántó Z., Farkas R. A Question Answering Benchmark Database for Hungarian // *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*. – Toronto, Canada, 2023. – P. 188–198. – DOI: 10.18653/v1/2023.law-1.19. – URL: <https://aclanthology.org/2023.law-1.19/> ACL Anthology
17. Kuriyozov E., Vilares D., Gómez-Rodríguez C. BERTbek: A Pretrained Language Model for Uzbek // *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*. – Torino, Italy, 2024. – P. 33–44. – URL: <https://aclanthology.org/2024.sigul-1.5/> ACL Anthology
18. Yusupkhujayev K. Benchmarking Pre-Trained Open-Source Large Language Models for Uzbek: Evaluating Performance in a Low-Resource Setting Across Translation, Comprehension, and Generation // *Annali d'Italia*. – 2025. – No. 71. – P. 67–73. – DOI: 10.5281/zenodo.17223973. – URL: <https://zenodo.org/records/17223973> zenodo.org
19. Solidjonov D., Najmiddinov M. Investigating Linguistic Errors in Large Language Model Generation of Uzbek Text // *Cogent Arts & Humanities*. – published online 2025; Vol. 13, Issue 1, 2026. – Article 2600519. – DOI: 10.1080/23311983.2025.2600519. – URL: <https://doi.org/10.1080/23311983.2025.2600519> Tandfonline