SAVOL–JAVOB TIZIMINING LINGVISTIK VA TEXNOLOGIK MODELI

Umarova Nozimaxon Qobil qizi
Kompyuter lingvistikasi mutaxassisligi
1-bosqich magistranti
nozimaumarova10302000@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur tezisda o‘zbek tilida huquqiy savol-javob tizimini yaratish uchun lingvistik va texnologik model taklif etiladi. Huquqiy matnlar rasmiylik, normativlik, terminologik aniqlik, murakkab sintaktik tuzilma va mantiqiy izchillik kabi xususiyatlari bilan oddiy matnlardan farq qiladi. Shu sababli bunday matnlar bilan ishlovchi savol-javob tizimi faqat kalit so‘zga asoslangan qidiruv bilan cheklanmasligi, balki, savolni lingvistik tahlil qilish, huquqiy terminlarni aniqlash, matn segmentlarini ajratish, semantik moslikni baholash va javobni manbaga tayangan holda shakllantirish bosqichlarini o‘z ichiga olishi zarur. Tadqiqotda lingvistik tahlil, BM25 asosidagi qidiruv, transformer modellarga asoslangan semantik moslashtirish va qoidalarga asoslangan yondashuv komponentlarini birlashtiruvchi model taklif etiladi.

Kalit so‘zlar: *savol-javob tizimi, huquqiy NLP, lingvistik model, texnologik model, BM25, semantik moslik, o‘zbek tili.*

Abstract. This paper proposes a linguistic and technological model for developing a legal question-answering system in the Uzbek language. Legal texts differ from general texts due to their formality, normativity, terminological precision, complex syntactic structure, and logical consistency. Therefore, a question-answering system operating on such texts should not rely solely on keyword-based retrieval. Instead, it must incorporate multiple stages, including linguistic analysis of the query, identification of legal terms, segmentation of text, evaluation of semantic similarity, and generation of answers grounded in legal sources. The study presents a hybrid model that integrates linguistic processing, BM25-based retrieval, transformer-based semantic matching, and rule-based

verification components. This approach aims to improve both the accuracy and reliability of legal question-answering systems in the Uzbek language.

Keywords: *question-answering system, legal NLP, linguistic model, technological model, BM25, semantic similarity, Uzbek language.*

Kirish. So‘nggi yillarda tabiiy tilni qayta ishlash sohasida savol-javob tizimlari muhim amaliy yo‘nalishlardan biriga aylandi. Bunday tizimlar foydalanuvchining tabiiy tilda bergan savolini qabul qilib, katta hajmdagi matnlar ichidan tegishli javobni topish va uni tushunarli shaklda taqdim etishga xizmat qiladi. Ayniqsa, huquqiy sohada bunday tizimlarga ehtiyoj yuqori, chunki qonun hujjatlari, kodekslar va normativ matnlar hajman katta, mazmunan murakkab hamda maxsus terminologiyaga boy bo‘ladi. LegalAI tadqiqotlarida huquqiy ma’lumotlarning asosiy qismi matn shaklida bo‘lishi, shu sababli huquqiy vazifalarning ko‘pi NLP texnologiyalariga tayanishi qayd etiladi [13:218-223]. Huquqiy savol-javob tizimlarining asosiy vazifasi foydalanuvchi savolini tushunish, tegishli huquqiy manbani aniqlash va unga asoslangan javobni shakllantirishdan iborat. Bunday tizimlar oddiy qidiruv tizimlaridan farqli ravishda faqat hujjatlarni topish bilan cheklanib qolmay, foydalanuvchining savol niyatini aniqlash, savolga mos huquqiy segmentni tanlash va javobni normativ manba bilan bog‘lashni talab qiladi. Huquqiy savol-javob bo‘yicha tadqiqotlarda ham ushbu yo‘nalish huquqiy axborotni avtomatik izlash, tahlil qilish va foydalanuvchiga mos javob berishning muhim shakli sifatida baholanadi [10:7-11].

Huquqiy matnlar oddiy matnlardan farqli ravishda murakkab sintaktik tuzilma, maxsus terminologiya, normativ mazmun va qat’iy mantiqiy izchillikka ega. Shu sababli huquqiy savol-javob tizimlarida lingvistik model va texnologik model o‘zaro bog‘liq holda ishlashi zarur. Lingvistik model savol va huquqiy matn tarkibidagi til birliklarini tahlil qiladi, texnologik model esa qidiruv, semantik moslashtirish va javobni shakllantirish jarayonlarini amalga oshiradi. Huquqiy NLP

bo‘yicha so‘nggi tadqiqotlarda ham huquqiy matnlarni avtomatik qayta ishlashda matn tasnifi, axborot ajratish, huquqiy qidiruv, savol-javob tizimlari va izohlanuvchanlik kabi vazifalar alohida ilmiy yo‘nalish sifatida ko‘rsatiladi [1:3-5]. O‘zbek tilida huquqiy savol-javob tizimini yaratish masalasi yanada murakkab hisoblanadi. Chunki o‘zbek tili agglyutinativ til bo‘lib, grammatik ma’nolar asosan qo‘shimchalar orqali ifodalanadi. Natijada bir huquqiy termin turli grammatik shakllarda kelishi mumkin. Bunday holat savol va huquqiy matn o‘rtasidagi yuzaki leksik moslikni kamaytiradi. Kam resursli va morfologik jihatdan murakkab tillar uchun huquqiy QA tizimlarini yaratishda korpus, savol-javob juftliklari va semantik moslikni aniqlash alohida ahamiyat kasb etadi [11:77-79].

Mazkur tezisning maqsadi o‘zbek tilida huquqiy savol-javob tizimini yaratish uchun lingvistik va texnologik komponentlarni birlashtirgan modelni taklif etishdan iborat. Ushbu model savolni dastlabki qayta ishlash, lingvistik tahlil, huquqiy korpusdan qidiruv, semantik moslashtirish, qoida-asoslangan tekshiruv va javobni shakllantirish bosqichlarini o‘z ichiga oladi.

Asosiy qism. Huquqiy savol-javob tizimining lingvistik va texnologik modeli bir necha o‘zaro bog‘langan bosqichlardan tashkil topadi. Bunday tizimda lingvistik model savol va huquqiy matnni tushunish uchun asos yaratadi, texnologik model esa ushbu lingvistik natijalarni qidiruv, moslashtirish va javob shakllantirish jarayonida qo‘llaydi. Shu sababli huquqiy QA tizimining samaradorligi faqat texnik algoritmning kuchiga emas, balki til birliklari qanchalik to‘g‘ri aniqlanishi va huquqiy matnning mazmuniy tuzilmasi qanchalik puxta modellashtirilishiga ham bog‘liq.

Texnologik modelni loyihalashda NLP pipeline yondashuvi asosiy metodologik asos sifatida olinadi. Pipeline jarayonida NLP masalasi bir necha kichik bosqichlarga ajratiladi: ma’lumot yig‘ish, matnni tozalash, dastlabki qayta ishlash, xususiyatlarni shakllantirish, modellashtirish, baholash, amaliy joriy etish hamda

modelni monitoring qilish [6:1-2]. Huquqiy savol-javob tizimi uchun bu ketma-ketlik mehnat qonunchiligi matnlarini korpusga jamlash, ularni modda va bandlar bo‘yicha segmentlash, savollarni normalizatsiya qilish, qidiruv va semantik moslashtirish modullarini birlashtirish hamda javobning manbaga muvofiqligini tekshirish imkonini beradi.

Avvalo, tizim foydalanuvchi savolini dastlabki qayta ishlashdan boshlanadi. Bu bosqichda savoldagi ortiqcha belgilar olib tashlanadi, imlo va grafik variantlar normallashtiriladi, so‘zlarning asosiy shakllari aniqlanadi. O‘zbek tilida bu bosqich juda muhim, chunki bitta leksik birlik turli qo‘shimchalar orqali ko‘plab shakllarda namoyon bo‘lishi mumkin. Masalan, “xodim”, “xodimning”, “xodimga”, “xodimlar”, “xodimlarning” shakllari alohida ko‘rinishda bo‘lsa-da, ular bir asosiy tushuncha bilan bog‘liq. Agar tizim bu shakllarni yagona leksik birlik sifatida taniy olmasa, savol bilan huquqiy matn orasidagi moslikni aniqlash qiyinlashadi. Shu sababli lemmatizatsiya, morfologik normalizatsiya va terminlarni standart shaklga keltirish lingvistik modelning asosiy vazifalaridan biri hisoblanadi. O‘zbek tilida NLP tizimlarini rivojlantirishda tilga mos transformer modellar muhim ahamiyatga ega. UzBERT modeli o‘zbek tilidagi matnlarni oldindan o‘qitilgan BERT arxitekturasi asosida qayta ishlash imkonini beruvchi dastlabki yondashuvlardan biri sifatida taqdim etilgan [9:1-5]. Keyingi tadqiqotlarda BERTbek modeli ham o‘zbek tili uchun maxsus tayyorlangan pretrained language model sifatida ishlab chiqilgan bo‘lib, u o‘zbekcha matnlarni semantik tahlil qilish, tasniflash va boshqa NLP vazifalarida qo‘llash uchun muhim asos yaratdi [7:56-59]. Bu modellar huquqiy savol-javob tizimlarida savol va normativ matn o‘rtasidagi semantik yaqinlikni aniqlashda foydalanilishi mumkin.

O‘zbek tilida bunday dastlabki qayta ishlash bosqichida tokenizatsiya va morfologik tahlil bilan bir qatorda stemming ham muhim o‘rin tutadi. Agglyutinativ tillar uchun o‘zak va qo‘shimcha chegarasini aniqlash, fonetik-morfologik

o‘zgarishlarni hisobga olish hamda OOV so‘zlar bilan ishlash masalalari alohida muammo sifatida ko‘rsatilgan [5:6-9]. Shu bois texnologik model savoldagi huquqiy terminlarni faqat sirtqi shaklda emas, balki ularning lemma, grammatik rol va huquqiy tushuncha sifatidagi funksiyasi bo‘yicha ham qayta ishlashi kerak.

Lingvistik tahlil bosqichida savolning turi ham aniqlanishi kerak. Huquqiy savollar mazmuniga ko‘ra faktual, normativ, izohli yoki protsessual xarakterga ega bo‘lishi mumkin. Masalan, “Xodim yillik ta’til olish huquqiga egami?” savoli normativ mazmunga ega bo‘lib, unda foydalanuvchi muayyan huquqiy normani bilishni istaydi. “Mehnat shartnomasi nima?” savoli esa izohli xarakterga ega bo‘lib, tushuncha mazmunini aniqlashga qaratilgan. “Ish beruvchi xodimga necha kunlik ta’til berishi kerak?” kabi savollar esa faktual va normativ belgilarni birlashtiradi. Savol turini aniqlash keyingi qidiruv strategiyasini belgilaydi, chunki har bir savol turi turlicha javob shaklini talab qiladi.

Keyingi bosqichda savol tarkibidagi asosiy huquqiy birliklar ajratiladi. Bunday birliklarga “xodim”, “ish beruvchi”, “mehnat shartnomasi”, “ta’til”, “ish vaqti”, “majburiyat”, “huquq”, “javobgarlik” kabi terminlar kiradi. Ushbu birliklar tizim uchun asosiy semantik signal vazifasini bajaradi. Masalan, foydalanuvchi “ish beruvchi xodimni ogohlantirmasdan ishdan bo‘shata oladimi?” deb so‘rasa, tizim “ish beruvchi”, “xodim”, “ogohlantirish”, “ishdan bo‘shatish” kabi birliklarni ajratishi va ularni mehnat qonunchiligining tegishli normalari bilan bog‘lashi lozim.

Texnologik modelning birinchi muhim bosqichi huquqiy korpusdan qidiruvni amalga oshirishdir. Bu jarayonda BM25 kabi statistik qidiruv algoritmlaridan foydalanish mumkin. BM25 axborot qidiruv tizimlarida keng qo‘llanadigan model bo‘lib, hujjatdagi so‘z chastotasi va hujjatlar to‘plamidagi tarqalish darajasiga asoslangan holda moslikni baholaydi [12:333-389]. Huquqiy QA tizimida BM25 dastlabki nomzod segmentlarni topish uchun qulay, chunki u katta hajmdagi huquqiy korpusdan savolga yaqin bo‘lgan moddalar, bandlar yoki paragraflarni tez ajratib

beradi. Biroq BM25 faqat yuzaki leksik moslikka asoslanadi. Bu esa o‘zbek tilidagi huquqiy matnlar uchun yetarli bo‘lmasligi mumkin. Chunki bir xil huquqiy mazmun turli grammatik va sintaktik shakllarda ifodalanadi. Masalan, foydalanuvchi “xodim ta’tilga chiqishi mumkinmi?” deb so‘rasa, huquqiy matnda bu mazmun “xodimga har yili mehnat ta’tili beriladi” yoki “yillik asosiy ta’til olish huquqi kafolatlanadi” shaklida kelishi mumkin. Bunday holatda faqat kalit so‘zga asoslangan qidiruv to‘liq natija bermasligi ehtimoli bor. Shu sababli BM25 natijalaridan keyin semantik moslashtirish bosqichini qo‘llash zarur.

Semantik moslashtirish bosqichida transformer modellaridan foydalaniladi. BERT modeli matnni ikki tomonlama kontekst asosida ifodalashi va so‘zlarning ma’nosini kontekstga bog‘liq holda aniqlashi bilan savol-javob tizimlari uchun muhim texnologik asos hisoblanadi [3:171-176]. Huquqiy QA tizimida transformer modeli savol va huquqiy segmentni vektor ko‘rinishida ifodalab, ular orasidagi semantik yaqinlikni baholashi mumkin. Natijada tizim faqat aynan bir xil so‘zlar uchrashiga emas, balki savol mazmuni bilan normativ matn mazmuni o‘rtasidagi yaqinlikka tayanadi. Huquqiy savol-javob tizimida matn segmentatsiyasi ham muhim o‘rin tutadi. Huquqiy hujjatlar ko‘pincha uzun, murakkab va ko‘p qatlamli bo‘ladi. Shu sababli butun hujjatni emas, balki mazmunan yaxlit bo‘lgan kichik segmentlarni qidiruv va semantik moslashtirish jarayoniga kiritish samaraliroq hisoblanadi. Huquqiy hujjatlarni semantik segmentlarga ajratish bo‘yicha tadqiqotlarda matnni faktlar, argumentlar, norma, masala, qaror va boshqa ritorik rollar asosida tahlil qilish huquqiy axborotni qayta ishlashni yengillashtirishi ko‘rsatilgan [8:153-159]. Bu yondashuv savol-javob tizimi uchun ham foydali, chunki foydalanuvchi savoliga javob ko‘pincha butun hujjatda emas, balki muayyan semantik segmentda joylashgan bo‘ladi.

Semantik moslashtirishda so‘zlarni vektor ko‘rinishida ifodalash usullari ham muhim nazariy asos bo‘la oladi. Word2Vec, GloVe va FastText kabi modellar

so‘zlar o‘rtasidagi semantik yaqinlikni sonli vektorlar orqali ifodalashga xizmat qiladi, biroq Word2Vec morfologik shakllarni alohida birlik sifatida ko‘rishi va OOV so‘zlar bilan ishlashda cheklanishi mumkin [4:225-227]. Shuning uchun huquqiy QA tizimi uchun klassik embedding yondashuvlarini o‘zbek tiliga mos lemmatizatsiya va transformer asosidagi kontekstli ifodalar bilan birlashtirish maqsadga muvofiqdir.

Bundan tashqari, huquqiy matnlarda murakkab gaplar alohida e’tibor talab qiladi. Bitta normativ gap ichida shart, sabab, qarama-qarshilik, tanlov, maqsad yoki istisno kabi bir nechta munosabatlar ifodalanishi mumkin. Masalan, “agar xodim uzrli sabablarga ko‘ra ishga kelmasa, ish beruvchi uni intizomiy javobgarlikka tortishdan oldin holatni aniqlashi lozim” kabi gapda shart, sabab va majburiyat munosabatlari mavjud. LeGen tadqiqotida huquqiy gaplarning murakkab diskurs tuzilmasini generativ modellar orqali ajratish, ya’ni shart, sabab, qarama-qarshilik va tanlov munosabatlarini aniqlash muhim vazifa sifatida ko‘rsatilgan [2:1-17]. Bunday yondashuv huquqiy QA tizimida savolga mos normativ birlikni aniqroq topishga yordam beradi.

Taklif etilayotgan modelda yana bir muhim bosqich qoidaga asoslangan tekshiruvdir. Huquqiy sohada javobning shunchaki mantiqan o‘xshash bo‘lishi yetarli emas, u tegishli qonun hujjati, modda yoki band bilan bog‘langan bo‘lishi kerak. Shu sababli tizim javobni shakllantirishdan oldin topilgan segmentning savolga haqiqatan mosligini, unda tegishli huquqiy norma mavjudligini va javobning manbaga tayanganligini tekshiradi. Bu bosqich LegalAI tizimlari uchun zarur bo‘lgan ishonchlilik va izohlanuvchanlikni ta’minlashga xizmat qiladi.

Umuman olganda, taklif etilayotgan model quyidagi ketma-ketlikda ishlaydi:

Savol → dastlabki qayta ishlash → lingvistik tahlil → huquqiy korpusdan qidiruv → semantik moslashtirish → qoida-asoslangan tekshiruv → javob

Birinchi bosqichda savol normallashtiriladi. Ikkinchi bosqichda savolning turi, terminlari, morfologik shakllari va asosiy huquqiy birliklari aniqlanadi. Uchinchi bosqichda BM25 yordamida dastlabki huquqiy segmentlar tanlanadi. To‘rtinchi bosqichda transformer model yordamida ushbu segmentlar savol bilan semantik jihatdan solishtiriladi. Beshinchi bosqichda javobning huquqiy norma bilan bog‘liqligi tekshiriladi. Yakunda foydalanuvchiga aniq, qisqa va manbaga tayangan javob taqdim etiladi.

Bunday integratsiyalashgan yondashuv o‘zbek tilidagi huquqiy savol-javob tizimlari uchun muhim afzalliklarga ega. Birinchidan, u o‘zbek tilining agglyutinativ xususiyatini hisobga oladi. Ikkinchidan, huquqiy terminlarning turli grammatik shakllarini yagona semantik birlik sifatida qayta ishlash imkonini beradi. Uchinchidan, qidiruv va semantik moslashtirish bosqichlarini birlashtirish orqali javob aniqligini oshiradi. To‘rtinchidan, qoida-asoslangan tekshiruv javobning huquqiy ishonchliligini kuchaytiradi.

Xulosa. Xulosa qilib aytganda, o‘zbek tilida huquqiy savol-javob tizimini yaratishda lingvistik va texnologik modelni birgalikda ishlab chiqish zarur. Lingvistik model savolning mazmuniy tuzilishini, huquqiy terminlarni, morfologik shakllarni va sintaktik bog‘lanishlarni aniqlashga xizmat qiladi. Texnologik model esa qidiruv, semantik moslashtirish, huquqiy tekshiruv va javobni shakllantirish jarayonlarini amalga oshiradi.

Taklif etilgan model o‘zbek tilining agglyutinativ xususiyati, huquqiy diskursning normativligi va huquqiy matnlarning murakkab sintaktik tuzilishini hisobga oladi. Mazkur yondashuv kelgusida mehnat qonunchiligiga oid huquqiy chatbotlar, elektron maslahat tizimlari va sun’iy intellektga asoslangan huquqiy yordamchi platformalarni yaratish uchun nazariy va amaliy asos bo‘lib xizmat qilishi mumkin.

Foydalanilgan adabiyotlar ro‘yxati

1. Ariai F., Demartini G. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges // ACM Computing Surveys. – 2025.

2. C. R. Chaitra, Kulkarni S., Sagi S.R.A.V., Pandey S., Yalavarthy R., Chakraborty D., Upadhyay P. LeGen: Complex Information Extraction from Legal Sentences using Generative Models // Proceedings of the Natural Legal Language Processing Workshop. – 2024.

3. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019.

4. Elov B. Matnli ma’lumotlarni Word2Vec, GloVe va FastText chuqur o‘rganish usullari vositasida qayta ishlash // Raqamli Transformatsiya va Sun’iy Intellekt. – 2024. – Vol. 2. – Issue 5.

5. Elov B.B., Hamroyeva Sh.M., Abdullayeva O.X., Husainova Z.Y., Xudayberganov N.U. Agglyutinativ tillar uchun POS teglash va stemming masalasi (turk, uyg‘ur, o‘zbek tillari misolida) // O‘zbekiston: til va madaniyat. – 2023. – № 2.

6. Elov B.B., Khamroeva Sh.M., Xusainova Z.Y. The pipeline processing of NLP // E3S Web of Conferences. – 2023. – Vol. 413. – Article 03011.

7. Kuriyozov E., Vilares D., Gómez-Rodríguez C. BERTbek: A Pretrained Language Model for Uzbek // Proceedings of SIGUL / LREC-COLING. – 2024.

8. Malik V., Sanjay R., Guha S.K., Hazarika A., Nigam S., Bhattacharya A., Modi A. Semantic Segmentation of Legal Documents via Rhetorical Roles // Proceedings of the Natural Legal Language Processing Workshop. – 2022.

9. Mansurov B., Mansurov A. UzBERT: Pretraining a BERT Model for Uzbek // arXiv preprint. – 2021. – arXiv:2108.09814.

10. Martinez-Gil J. A survey on legal question-answering systems // Computer Science Review. – 2023. – Vol. 48. – Article 100552.

11. Rakhimova D., Turarbek A., Karyukin V., Sarsenbayeva A., Alieyev R. Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation // Computers. – 2025. – Vol. 14. – No. 9. – Article 354.

12. Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval. – 2009. – Vol. 3. – No. 4.

13. Zhong H., Xiao C., Tu C., Zhang T., Liu Z., Sun M. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – 2020.