

Компьютер лингвистикада tibbiyot hujjatlari va ularning lingvistik xususiyatlarining ahamiyati

Sharipova Ra'no Akmalovna
sharipovarano2004@gmail.com
Magistratura talabasi,
Buxoro davlat universiteti

Annotatsiya Zamonaviy sog'liqni saqlash tizimlarida tibbiy hujjatlar klinik qaror qabul qilish, davolash jarayonini rejalashtirish va boshqarish, shuningdek ilmiy-tadqiqot faoliyatini tashkil etishda muhim axborot manbasi hisoblanadi. Raqamli tibbiyot va elektron sog'liqni saqlash tizimlarining jadal rivojlanishi natijasida tibbiy hujjatlar hajmi, shakli va mazmuniy murakkabligi sezilarli darajada oshib bormoqda. Biroq ushbu hujjatlarning lingvistik xususiyatlari, jumladan ularning ko'pincha strukturasi bo'lishi, terminologik jihatdan yuqori zichlikka ega ekanligi, qisqartmalar va professional birliklarning keng qo'llanilishi hamda kontekstga kuchli bog'liqligi ularni avtomatik qayta ishlash jarayonida jiddiy muammolarni yuzaga keltiradi. Bu holat, ayniqsa resurslari cheklangan tillar, xususan o'zbek tili uchun dolzarbdir. Mazkur maqolada o'zbek tilidagi tibbiy hujjatlarning lingvistik xususiyatlari kompyuter lingvistikasi nuqtayi nazaridan tizimli tahlil qilinadi. Tadqiqot natijalari tibbiy matnlarning asosiy muammolarini aniqlash, ularni avtomatik qayta ishlash imkoniyatlarini baholash va o'zbek tilidagi tibbiy hujjatlar uchun maxsus soha-yo'naltirilgan til modellarini ishlab chiqish zarurligini ilmiy jihatdan asoslashga qaratilgan. Ushbu yondashuvlar kelgusida milliy sog'liqni saqlash axborot tizimlarida samarali, xavfsiz va ishonchli tibbiy NLP yechimlarini yaratishga xizmat qiladi va amaliy tatbiq etilishiga zamin yaratadi.

Kalit so'zlar Tibbiy hujjatlar, lingvistik xususiyatlar, o'zbek tili, kompyuter lingvistikasi, tibbiy NLP

Значение медицинских документов и их лингвистических особенностей в компьютерной лингвистике

Шарипова Раъно Акмаловна
sharipovarano2004@gmail.com
Магистрант,
Бухарский государственный университет

Аннотация В современных системах здравоохранения медицинская документация является важным источником информации для принятия клинических решений, планирования и управления процессом лечения, а также организации научно-исследовательской деятельности. В результате стремительного развития цифровой медицины и электронных систем здравоохранения объем, форма и содержательная сложность медицинских документов существенно возрастают. Однако лингвистические особенности таких документов, в частности их часто неструктурированный характер, высокая терминологическая насыщенность, широкое использование сокращений и профессиональных единиц, а также сильная контекстуальная зависимость, создают серьезные трудности в процессе их автоматической обработки. Данная проблема особенно актуальна для языков с ограниченными ресурсами, в том числе узбекского языка. В настоящей статье лингвистические особенности

медицинских документов на узбекском языке системно анализируются с позиций компьютерной лингвистики. Результаты исследования направлены на выявление основных проблем медицинских текстов, оценку возможностей их автоматической обработки и научное обоснование необходимости разработки специализированных, ориентированных на предметную область языковых моделей для медицинской документации на узбекском языке. Предлагаемые подходы создают основу для разработки эффективных, безопасных и надежных решений медицинского NLP в национальных информационных системах здравоохранения и их практического внедрения.

Ключевые слова *Медицинская документация, лингвистические особенности, узбекский язык, компьютерная лингвистика, медицинский NLP*

The importance of medical documents and their linguistic features in computational linguistics

Sharipova Rano Akmalovna
sharipovarano2004@gmail.com
Master's Student,
Bukhara State University

Annotation *In modern healthcare systems, medical documentation serves as a crucial source of information for clinical decision-making, planning and management of treatment processes, as well as for the organization of research activities. As a result of the rapid development of digital medicine and electronic healthcare systems, the volume, format, and semantic complexity of medical documents have increased significantly. However, the linguistic characteristics of such documents, including their frequent lack of structure, high terminological density, extensive use of abbreviations and professional expressions, and strong context dependency, pose serious challenges for automatic processing. This issue is particularly relevant for low-resource languages, including Uzbek. This article provides a systematic analysis of the linguistic features of Uzbek-language medical documents from the perspective of computational linguistics. The findings aim to identify the key challenges inherent in medical texts, assess the feasibility of their automatic processing, and provide a scientific justification for the development of specialized, domain-oriented language models for Uzbek medical documentation. These approaches contribute to the creation of effective, secure, and reliable medical NLP solutions for national healthcare information systems and support their practical implementation.*

Keywords *Medical documents, linguistic characteristics, Uzbek language, computational linguistics, medical NLP*

Kirish

O'zbekistonda raqamli texnologiyalar, sun'iy intellekt va til resurslarini rivojlantirish

masalalariga davlat miqyosida alohida e'tibor qaratilmoqda. Garchi kompyuter lingvistikasi sohasini bevosita tartibga soluvchi maxsus

qonunlar hali to'liq shakllanmagan bo'lsa-da, sun'iy intellekt va raqamli til resurslari bilan bog'liq umumiy huquqiy mexanizmlar bosqichma-bosqich joriy etilmoqda. Mazkur jarayon kompyuter lingvistikasi texnologiyalarini tizimli, mas'uliyatli va milliy manfaatlariga mos tarzda rivojlantirish uchun muhim normativ asos yaratmoqda.

Bu hujjatlar nafaqat umumiy AI siyosatini belgilaydi, balki til resurslari va etik me'yorlarni shakllantirishda ham yo'naltiruvchi rol o'ynaydi.

1. "Sun'iy intellektni qo'llash orqali yuzaga keladigan munosabatlarni tartibga soluvchi qonun" – O'zbekiston Respublikasi Oliy Majlisi Qonunchilik palatasi tomonidan 15-aprel 2025-yilda birinchi o'qishda qabul qilingan va keyinchalik uchinchi o'qishda ma'qullangan qonun loyihasi sun'iy intellektni huquqiy tartibga solish uchun asosiy hujjat sifatida shakllanmoqda. U qonunchilik darajasida AI tushunchasini belgilaydi, davlat siyosatining asosiy yo'nalishlarini va maxsus vakolatli organning vazifalarini ko'zda tutadi. Shuningdek, sun'iy intellektdan foydalangan holda yaratilgan axborot resurslarini belgilash (markirovka) va shaxsiy ma'lumotlarga noqonuniy ishlov berishga qarshi javobgarlikni belgilashni maqsad qiladi.
2. "Sun'iy intellektni qo'llash orqali yuzaga keladigan munosabatlarni tartibga soluvchi qonun"ning uchinchi o'qishda ma'qullanishi – 12-avgust 2025-yilda O'zbekiston Respublikasi Oliy Majlisi tomonidan ushbu qonun loyihasi uchinchi o'qishda ham ma'qullangan bo'lib, unda AI tizimlaridan foydalanishda inson huquqlari, shaxsiy ma'lumotlarni himoya qilish va axborot tizimlarida AI qo'llash uchun umumiy huquqiy asoslar belgilandi. Bu hujjat AI texnologiyalarini boshqarish bo'yicha normativ bazaning shakllanishida katta qadam hisoblanadi.
3. "AI ishlab chiqish va qo'llash etikasining tamoyillari" (Rules and principles of ethics

in the development and use of AI) – 3-dekabr 2025-yilda Prezident qarori bilan tasdiqlanishi rejalashtirilgan hujjat sun'iy intellektni ishlab chiqish jarayonida axloqiy tamoyillar – mas'uliyat, shaffoflik, adolat, diskriminatsiyaga qarshi me'yorlar, ma'lumotlarni himoya qilish va inson huquqlarini hurmat qilishni belgilaydi. Bu hujjat AI tizimlarining insonxohlik tamoyillariga mos ravishda ishlab chiqilishi uchun muhim etik mexanizm hisoblanadi.

4. "AI sohasida davlat strategiyasi" (Strategy for the Development of Artificial Intelligence Technologies until 2030) – Prezident tomonidan tasdiqlangan strategik hujjat bo'lib, 2024-yilda sun'iy intellekt texnologiyalarini rivojlantirish bo'yicha uzoq muddatli yo'l xaritasi sifatida qabul qilingan. Ushbu strategiya AI ekotizimini shakllantirish, milliy til resurslarini yaratish va raqamli kompetensiyalarni oshirishga yo'naltirilgan.
5. Milliy til korpusi va raqamli til resurslarini shakllantirish tashabbuslari O'zbekistonning til resurslarini raqamlashtirishga qaratilgan tashabbuslar bo'yicha milliy til korpusi va boshqa resurslar yaratilmoqda. Bu jarayon o'zbek tilida ilmiy tadqiqotlar, NLP modellar va kompyuter lingvistikasi uchun zarur infratuzilmaviy asos bo'lib xizmat qiladi (masalan, Milliy til korpusi haqida loyiha va platformalar). Bu hujjatlar bevosita qonun bo'lmasa ham, tillarni raqamlashtirish va resurslar yaratish bo'yicha davlat darajasidagi qarorlar sifatida ilmiy maqolalarda qayd etilishi mumkin.

Bugungi kunda sog'liqni saqlash tizimlarining raqamlashtirilishi natijasida katta hajmdagi tibbiy matnlar shakllanmoqda. Klinik yozuvlar, tashxis xulosalari, epikrizlar, laboratoriya natijalari hamda dori vositalariga oid hujjatlar shifokorlar va tibbiyot mutaxassislari faoliyatining ajralmas qismiga

aylangan. Ushbu hujjatlar nafaqat amaliy tibbiyotda, balki tibbiy axborot tizimlari, sun'iy intellekt va qaror qabul qilishni qo'llab-quvvatlovchi tizimlarda ham muhim ahamiyat kasb etadi.

Tabiiy tilni qayta ishlash (Natural Language Processing – NLP) texnologiyalarining rivojlanishi tibbiy hujjatlarni avtomatik tahlil qilish imkoniyatlarini kengaytirmoqda. Biroq tibbiy matnlar umumiy til matnlariga nisbatan murakkab bo'lib, ular yuqori darajadagi terminologik zichlik, qisqartmalar, kontekstga bog'liq ma'nolar va notekis sintaktik tuzilmalarga ega. Bu holat ayniqsa o'zbek tili kabi resurslari cheklangan tillar uchun dolzarb muammo hisoblanadi.

Mazkur maqolaning maqsadi – o'zbek tilidagi tibbiy hujjatlarning lingvistik xususiyatlarini aniqlash va ularning til modellari yordamida avtomatik qayta ishlashga ta'sirini ilmiy jihatdan tahlil qilishdan iborat.

Adabiyotlar sharhi

Tibbiy hujjatlar tilini qayta ishlash masalasi xalqaro miqyosda kompyuter lingvistikasi va sun'iy intellektning muhim yo'nalishlaridan biri sifatida shakllangan. Tadqiqotlar shuni ko'rsatadiki, tibbiy matnlar umumiy til korpuslariga nisbatan sezilarli darajada murakkab bo'lib, ularning lingvistik tabiati maxsus yondashuvlarni talab etadi.

Tibbiy matnlarni qayta ishlash bo'yicha asosiy konsept va muammolarni tizimli ko'rsatgan keng qamrovli sharhlar mavjud. NLM (National Library of Medicine (AQSh Milliy tibbiyot kutubxonasi)) tomonidan tashkil etilgan va klinik NLP rivojiga bag'ishlangan ishlar tibbiy matnlarning strukturasizligi, kontekstga bog'liqligi va terminologik murakkabligini dastlabki va asosiy muammolar sifatida belgilaydi; shuningdek, klinik NLP tadqiqotlari uchun annotatsiyalangan korpuslar va soha resurslarning zarurligi qayd etilgan (NLM).

Friedman va hamkorlari klinik matnlarning asosiy xususiyatlari sifatida strukturasiz yozuvlar, qisqartmalar, terminologik zichlik va kontekstga kuchli

bog'liqlikni ko'rsatadi. Ularning tadqiqotlarida klinik hujjatlar ko'pincha to'liq gaplardan iborat bo'lmasligi, balki fragmentar iboralar shaklida ifodalanishi ta'kidlanadi. Bu holat avtomatik sintaktik tahlil va semantik interpretatsiyani murakkablashtiradi (Friedman, 1995).

So'nggi yillarda transformer arxitekturasiga asoslangan til modellari tibbiy NLP sohasida sezilarli natijalarga erishdi. Lee va boshqalar tomonidan taklif etilgan BioBERT modeli tibbiy terminologiya bilan boyitilgan korpuslarda oldindan o'qitish orqali umumiy BERT modeliga nisbatan yuqori samaradorlikni namoyon etdi. (Lee J et al. 2020) Shuningdek, ClinicalBERT modeli klinik yozuvlar bilan ishlashda kontekstual tushunishni yaxshilashga qaratilgan (Alsentzer et al., 2019).

Domain-adaptatsiya (umumiy maqsad uchun o'qitilgan til modelini muayyan soha matnlariga moslashtirish jarayoni) va oldindan o'qitilgan modellarning samarasi. Soha uchun mo'ljallangan oldindan o'qitilgan transformer modellar (domain-adapted PLM) – BioBERT, ClinicalBERT kabi yondashuvlar tibbiy vazifalarda an'anaviy BERT modellaridan ancha yuqori natija ko'rsatgan. BioBERT asosan bio-matnlarga (ilmiy maqolalar, PubMed) pre-training orqali soha tilini yaxshiroq egallagani aniqlangan; ClinicalBERT esa klinik yozuvlar ustida fine-tuning va pretraining orqali kasalxona yozuvlariga moslashtirilgan embeddinglar taqdim etadi. Ushbu ishlar domain-adaptatsiyaning muhimligini va tibbiy matnlar uchun maxsus pretraining zarurligini ko'rsatadi.

Past-resursli tillar va ularning xususiyatlari. So'nggi yillarda past-resursli (low-resource) tillar bo'yicha olib borilgan tadqiqotlar – ham metodologik, ham amaliy jihatdan – til resurslari, annotatsiya standartlari va terminologik normallashtirishning yetishmasligi modellarning samaradorligini pasaytirishini ko'rsatadi. (Adelani et al., 2021) O'zbek tiliga xos vaziyatda morfologik va sintaktik xususiyatlari, turli dialekt va yozuv variantlari hamda terminlarning ruscha va

lotincha shakllari boshqacha qiyinchiliklar tug'diradi; shu sababli umumiy inglizcha modellar bevosita o'zbek tibbiy matnlarida yaxshi ishlamaydi (Tsarfaty et al., 2020).

O'zbek tilidagi korpus va resurslar holati. O'zbekiston va turkiy olimlar tomonidan so'nggi yillarda o'tkazilgan korpus yaratuvchi loyihalar (milliy va ta'limiy korpuslar, morfologik/sintaktik teglar) o'zbek NLP fondini boyitmoqda, lekin tibbiy soha uchun maxsus annotatsiyalangan korpuslar va ontologiyalar yetarli emas. CEUR va mahalliy ishlarda morfologik va sintaktik teglar yaratish bo'yicha tajriba va tavsiyalar mavjud; shuningdek milliy korpus yaratish bo'yicha hisobot va metodologik maqolalar chiqmoqda – bular kelajakda tibbiy korpuslar yaratishga asos bo'la oladi.

Soha-adaptatsiyalangan modellar (BioBERT/ClinicalBERT va ularning ochiq implementatsiyalari) COVID-19 va boshqa klinik misollarda NER, QA va klinik ma'lumotni ajratishda sezilarli yutuq ko'rsatdi; shu tajribalar O'zbek tilida ham pretraining + fine-tuning yondashuvini qo'llash samarali bo'lishini taxmin qilishga asos beradi. Bundan tashqari, past-resursli sharoit uchun transfer learning, multilingual training va subword tokenizatsiyaga asoslangan strategiyalar tavsiya etiladi

Adabiyotlar tahlili orqali quyidagilar aniqlandi:

- o'zbek tibbiy matnlarining aniq lingvistik razmerini o'lchash uchun annotatsiyalangan, turli janrlarni qamragan korpus mavjud emas;
- statistik va neyron modellarni qo'llash uchun terminologik lug'at va qisqartmalar jamlanmasining yetishmasligi;
- vaqt, voqea va sabab-oqibat munosabatlarini aniqlovchi modellar uchun pragmatik va diskurs resurslarining zaif.

Ushbu bo'shliqlarni to'ldirish maqsadida maqolamizda tavsiya qilingan metodologik qadamlar (korpus yaratish, annotatsiya

sxemalari, terminologik xaritalash) amaliy va ilmiy jihatdan dolzarb.

Biroq ushbu modellar asosan ingliz tilidagi yirik annotatsiyalangan korpuslarga tayangan holda ishlab chiqilgan. Adelan va hamkorlari tomonidan olib borilgan tadqiqotlarda past resursli tillarda NLP vazifalarining asosiy muammolari sifatida ma'lumotlar tanqisligi, terminologik nomuvofiqlik va lingvistik normalarning yetarli darajada shakllanmaganligi ko'rsatiladi. Bu holat o'zbek tili uchun ham dolzarb hisoblanadi.

Turkiy tillar kontekstida olib borilgan ayrim tadqiqotlar tibbiy terminologiyaning ko'p variantlilik va o'zlashma birliklarning ustunligini qayd etadi. O'zbek tilidagi tibbiy matnlarda lotincha va ruscha terminlarning parallel qo'llanilishi, shuningdek, qisqartmalarning standartlashmaganligi kuzatiladi. Bu esa til modellari uchun leksik va semantik noaniqliklarni yuzaga keltiradi.

O'zbek tili bo'yicha mavjud ilmiy ishlarda asosan umumiy NLP vazifalari – morfologik tahlil, mashinali tarjima va matn tasnifi masalalari yoritilgan. Tibbiy hujjatlarning lingvistik xususiyatlariga bag'ishlangan maxsus tadqiqotlar esa juda cheklangan. Shu sababli, o'zbek tilidagi tibbiy hujjatlarni lingvistik jihatdan tizimli tahlil qilish va ularni til modellari nuqtayi nazaridan baholash muhim ilmiy bo'shliqni to'ldirishga xizmat qiladi.

Mazkur maqola aynan shu bo'shliqni qisman to'ldirishni maqsad qilib, o'zbek tilidagi tibbiy hujjatlarning lingvistik xususiyatlarini xalqaro tibbiy NLP tadqiqotlari kontekstida tahlil qiladi.

Tadqiqot metodologiyasi

Mazkur tadqiqotda sifat (qualitative) tahlil metodologiyasi qo'llanildi. Tahlil uchun ochiq manbalardan olingan o'zbek tilidagi tibbiy hujjatlar tanlab olindi. Jumladan:

- dori vositalariga oid yo'riqnomalar;
- tibbiy tavsiyalar va klinik ko'rsatmalar;
- shifokorlarning bemor haqidagi kundalik yozuvlari va kasallik tarixlari;

- tibbiy tekshiruvlar yakunida beriladigan xulosalar;
- tibbiyot sohasidagi ilmiy maqolalar;
- klinik tadqiqot hujjatlari va statistik hisobotlar;
- anonimlashtirilgan klinik yozuv namunalari.

Matnlar quyidagi lingvistik mezonlar asosida tahlil qilindi:

- leksik xususiyatlar;
- terminologik birliklar;
- sintaktik tuzilmalar;
- kontekst va diskurs darajasi.

Tahlil natijalari tibbiy til modellarini ishlab chiqishda hisobga olinishi zarur bo'lgan lingvistik omillarni aniqlashga qaratildi.

Tibbiyot hujjatlarining lingvistik tahlili

O'zbek tilidagi tibbiy hujjatlar mazmuni, funksiyasi va yozilish shakliga ko'ra bir necha asosiy guruhlariga ajratilishi mumkin. Ushbu tasnif tibbiy matnlarni avtomatik qayta ishlash jarayonida muhim ahamiyat kasb etadi, chunki har bir hujjat turi o'ziga xos lingvistik xususiyatlarga ega.

Birinchi guruhga klinik hujjatlar kiradi. Bular anamnez, epikriz, tashxis xulosalari, shifokor kundalik yozuvlari kabi matnlardir. Ushbu hujjatlar ko'pincha strukturasi, fragmentar va telegrafik uslubda yoziladi. Gaplar to'liq bo'lmasligi, fe'l shakllarining tushib qolishi va nom birikmalarining ustunligi bu turdagi matnlar uchun xosdir.

Ikkinchi guruhni laboratoriya va instrumental tekshiruv natijalari tashkil etadi. Bu hujjatlar raqamlar, qisqartmalar va maxsus belgilarga boy bo'lib, tabiiy til elementlari bilan rasmiy belgilar qorishmasidan iborat bo'ladi. Lingvistik jihatdan bunday matnlar gibrid xarakterga ega bo'lib, ularni qayta ishlash maxsus yondashuvlarni talab etadi.

Uchinchi guruhga dori vositalari yo'riqnomalari va tibbiy tavsiyalar kiradi. Ushbu matnlar nisbatan tartibli, me'yoriy sintaksisga ega bo'lsa-da, yuqori terminologik zichlik va murakkab sintaktik konstruktsiyalar bilan ajralib turadi.

Tibbiy hujjatlarining muhim xususiyatlaridan biri ularning qat'iy grammatik va stilistik me'yorlarga bo'ysunmasligidir. Klinik yozuvlar, anamnez va epikrizlarda ko'pincha to'liq gaplar emas, balki qisqa iboralar, punktlar yoki telegrafik yozuvlar uchraydi. Masalan, tashxis bo'limida fe'l ishtirokisiz nom birikmalari ustunlik qiladi. Bunday fragmentarlik avtomatik gap aniqlash, sintaktik tahlil va bog'lanishlarni aniqlash jarayonlarini murakkablashtiradi. Umumiy til modellari odatda to'liq gaplarga moslashtirilgan bo'lgani sababli, tibbiy hujjatlar bilan ishlashda ularning samaradorligi pasayadi.

Morfologik va sintaktik xususiyatlar

O'zbek tilidagi tibbiy hujjatlarining morfologik tahlili shuni ko'rsatadiki, ot turkumiga mansub so'zlar ustunlik qiladi. Fe'llar ko'pincha buyruq-istak yoki noaniq shaklda qo'llaniladi, ba'zi hollarda esa umuman tushib qoldiriladi. Bu holat tibbiy matnlarning nominal xarakterga ega ekanligini ko'rsatadi.

Sintaktik jihatdan tibbiy hujjatlarda murakkab gaplar kam uchraydi. Aksincha, qisqa, bo'laklangan, ba'zan grammatik jihatdan tugallanmagan konstruktsiyalar ustunlik qiladi. Bu esa avtomatik sintaktik tahlil modellarining aniqligini pasaytiradi

Terminologik birliklar va ularning shakllanishi

Tibbiy hujjatlarda terminlar yuqori chastotada qo'llaniladi. Ushbu terminlar ko'pincha lotincha, ruscha yoki xalqaro shaklda bo'lib, ularning o'zbek tilidagi muqobillari har doim ham yagona standartga ega emas. Natijada bir tushuncha turli variantlarda ifodalanishi mumkin. Bu holat til modellarida leksik tarqoqlikni kuchaytiradi va terminlarni normalizatsiya qilish zaruratini keltirib chiqaradi. Ayniqsa kasallik nomlari, dori vositalari va anatomik atamalar bu jihatdan muammoli hisoblanadi.

Tibbiy terminologiya o'zbek tilida asosan o'zlashma birliklar hisobiga boyigan. Lotin va rus tillaridan kirib kelgan terminlar ko'pincha asl shaklida yoki fonetik moslashtirilgan holda ishlatiladi. Natijada terminlarning bir nechta

variantda qo'llanishi kuzatiladi. Masalan, bir xil kasallik nomi turli hujjatlarda turlicha ifodalanishi mumkin. Bu holat avtomatik termin aniqlash va normallashtirish jarayonlarini murakkablashtiradi. Shu sababli tibbiy til modeli ishlab chiqishda terminologik lug'atlar va moslashtirish mexanizmlaridan foydalanish muhim hisoblanadi.

Tibbiy hujjatlar faqat lingvistik emas, balki pragmatik jihatdan ham o'ziga xosdir. Matn muallifi va o'quvchisi ko'pincha professional mutaxassislar bo'lgani sababli, ko'plab ma'lumotlar oshkor etilmaydi va umumiy bilimga tayangan holda beriladi. Bu holat yashirin ma'no (implicit meaning)ning ustunligiga olib keladi. Til modellari uchun esa bunday yashirin axborotni aniqlash katta muammo tug'diradi. Shu bois tibbiy NLP tizimlari kontekstual bilim bazalari bilan integratsiyalashgan bo'lishi zarur.

Tibbiy hujjatlarda terminlar yuqori chastotada qo'llaniladi. Ushbu terminlar ko'pincha lotincha, ruscha yoki xalqaro shaklda bo'lib, ularning o'zbek tilidagi muqobillari har doim ham yagona standartga ega emas. Natijada bir tushuncha turli variantlarda ifodalanishi mumkin.

Bu holat til modellarida leksik tarqoqlikni kuchaytiradi va terminlarni normalizatsiya qilish zaruratini keltirib chiqaradi. Ayniqsa kasallik nomlari, dori vositalari va anatomik atamalar bu jihatdan muammoli hisoblanadi.

Kontekstga kuchli bog'liqlik va yashirin axborot

Tibbiy hujjatlarda ko'plab ma'lumotlar bevosita ifodalanmaydi, balki professional bilimga tayangan holda tushuniladi. Shifokorlar uchun ravshan bo'lgan ayrim tushunchalar matnda ochiq bayon etilmasligi mumkin.

Natijada, til modellari yashirin axborotni aniqlashda qiynaladi. Bu esa tibbiy hujjatlar uchun kontekstual tushunishga ega maxsus modellarni ishlab chiqishni talab qiladi.

Ushbu tadqiqot davomida aniqlangan lingvistik xususiyatlar o'zbek tilidagi tibbiy hujjatlarni avtomatik qayta ishlash jarayonida yuzaga keladigan asosiy muammolarni ilmiy

jihatdan asoslab beradi. Avvalo, strukturasizlik va sintaktik fragmentarlik tibbiy matnlarning umumiy til modellariga mos kelmasligini ko'rsatadi. Transformer arxitekturasiga asoslangan zamonaviy til modellari kontekstual bog'lanishlarni yaxshi o'rganishga qodir bo'lsa-da, tibbiy hujjatlardagi to'liq bo'lmagan gaplar va telegrafik yozuvlar ularning imkoniyatlarini cheklaydi.

Terminologik zichlik va leksik variativlik masalasi esa til modelini o'qitishda maxsus terminologik resurslardan foydalanish zarurligini ko'rsatadi. O'zbek tilida tibbiy terminlarning yagona standartga ega emasligi model uchun semantik tarqoqlikni keltirib chiqaradi. Xalqaro tajribada bu muammo soha yo'naltirilgan korpuslar va ontologiyalar yordamida hal etilgan bo'lsa, o'zbek tili uchun bunday resurslar endigina shakllanmoqda.

Aniqlangan lingvistik xususiyatlar til modelini loyihalashda asosiy omillar sifatida qaralishi lozim. Strukturasiz matnlar modelni oldindan o'qitish bosqichida maxsus korpuslardan foydalanishni talab qiladi.

Terminologik zichlik esa soha yo'naltirilgan maxsus tokenizatsiya strategiyalarini qo'llash zarurligini ko'rsatadi. Qisqartmalar bilan ishlash uchun kontekstga asoslangan aniqlash mexanizmlari joriy etilishi lozim.

O'zbek tibbiy NLP uchun metodologik tavsiyalar

Tadqiqot natijalari asosida o'zbek tilidagi tibbiy hujjatlar bilan ishlash uchun quyidagi metodologik tavsiyalarni berish mumkin:

- tibbiy matnlar uchun alohida korpus yaratish;
- terminologik lug'at va ontologiyalar ishlab chiqish;
- annotatsiya sxemalarini standartlashtirish;
- til modellarini soha bo'yicha moslashtirish (fine-tuning).

Mazkur tadqiqotning ilmiy ahamiyati shundaki, u o'zbek tilidagi tibbiy hujjatlarni lingvistik jihatdan tizimli ravishda yoritadi. Amaliy jihatdan esa mazkur ish natijalari

tibbiy axborot tizimlari, elektron sog'liqni saqlash platformalari va tibbiy sun'iy intellekt tizimlarini ishlab chiqishda qo'llanilishi mumkin.

Xulosa

Mazkur maqolada o'zbek tilidagi tibbiy hujjatlarning lingvistik xususiyatlari kompyuter lingvistikasi nuqtayi nazaridan tizimli ravishda tahlil qilindi. Tadqiqot natijalari tibbiy hujjatlarning strukturasizligi, terminologik zichligi, qisqartmalarning keng qo'llanilishi va kontekstga kuchli bog'liqligi ularni avtomatik qayta ishlashda asosiy muammolar ekanligini ko'rsatdi.

Tadqiqotning ilmiy yangiligi shundaki, unda o'zbek tilidagi tibbiy hujjatlar birinchi

marta til modeli ishlab chiqish nuqtayi nazaridan kompleks lingvistik tahlil qilindi. Olingan natijalar o'zbek tili uchun maxsus tibbiy korpus yaratish, annotatsiya sxemalarini ishlab chiqish va soha yo'naltirilgan til modellarini o'qitish uchun metodologik asos bo'lib xizmat qiladi.

Amaliy jihatdan mazkur ish natijalari tibbiy axborot tizimlari, klinik hujjatlarni avtomatik tahlil qilish, tibbiy qarorlarni qo'llab-quvvatlovchi tizimlar hamda tibbiy chat-botlar yaratishda qo'llanilishi mumkin. Kelgusida tadqiqotni kengaytirish annotatsiyalangan tibbiy korpus yaratish va eksperimental til modellarini ishlab chiqish orqali amalga oshirilishi rejalashtirilmoqda.

Adabiyotlar ro'yxati:

1. Adelani, D. I. et al. (2021). *MasakhaNER: Named Entity Recognition for African Languages*. ACL.
2. Alsentzer, E. et al. (2019). *Publicly Available Clinical BERT Embeddings*. NAACL.
3. Friedman, C, Johnson, S. B, Forman B. (1995). *A general natural-language text processor for clinical radiology*. J Am Med Inform Assoc.
4. Kun.uz. (2025). *O'zbekistonda sun'iy intellektni huquqiy tartibga solish masalalari*.
5. Lee, J. et al. (2020). *BioBERT: a pre-trained biomedical language representation model*. Bioinformatics.
6. National Library of Medicine (NLM). *Clinical Natural Language Processing*. Bethesda, MD.
7. Tsarfaty, R. et al. (2020). *Cross-lingual transfer learning for morphologically rich languages*. TACL.