TURKIY TILLAR DOIRASIDA NOMLANGAN BIRLIKLARNI ANIQLASHGA OID TADQIQOTLAR TAHLILI

Nazokat O‘rinboyeva
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti magistranti
nnazokat71@gmail.com

ANNOTATSIYA

Ushbu maqolada turkiy tillar, xususan turk, qozoq, qirg'iz va ozarbayjon, tatar, sorani kurd va tojik, uyg’ur tillari uchun nomlangan birliklarni aniqlash (NER – *ingliz tilidan “Named Entity Recognition”*) bo‘yicha mavjud tadqiqotlarni ko'rib chiqish, olib borilayotgan izlanishlar, yaratilgan yangicha modellar, matnlarni teglash jarayonlari, NER tizimining ishlashini baholash uchun foydalaniladigan ma’lumotlar to‘plami va korpusi xususida so‘z boradi.

Kalit so‘zlar: NLP, NE, teglash, indekslash, BiLSTM,F1 ball,API, LSTM, CRF.

ABSTRACT

The article discusses the existing review of Named Entity Recognition (NER) for the Turkic languages, Turkish, Kazakh, Kyrgyz and Azerbaijani, Tatar, Sorani Kurdish and Tajik, Uyghur, on-going research, new models, text tagging processes, NER energy generation processing set and corpus.

Keywords: NLP, NE, tagging, indexing, BiLSTM, F1 score, API, LSTM, CRF.

KIRISH

Nomlangan birliklarni aniqlash (NER) tabiiy tilni qayta ishlashning muhim vazifalaridan biri bo'lib, matnlardagi nomli obyektlarni (shaxslar, tashkilotlar, joylar, sanalar va boshqalar) aniqlash va ularni oldindan belgilangan toifalarga ajratishga qaratilgan. NER turli NLP ilovalarida, jumladan, avtomatik matnni tushunish, mashina tarjimasi, savollarga javob berish va bilimlar bazasini yaratishda keng qo'llaniladi.

Turkiy tillar deganda, Sibirdan Bolqon yarim oroligacha bir chiziq bo‘ylab cho‘zilgan ulkan geografik hududda tarqalgan o‘zbek, uyg‘ur, qozoq, qirg‘iz, qoraqalpoq, saxa (yoqut), tuva, xakas, oltoy, karagas, shor, turkman, ozarbayjon, turk, gagauz, tatar, boshqird, chuvash, qo‘miq, no‘g‘ay, qorachoybolqor, tofalar, chuvash kabi 25 dan ortiq til tushuniladi.

Turkiy tillar, jumladan o'zbek, turk, qozoq, qirg'iz, ozarbayjon, tatar,sorani kurd va tojik, uyg'ur tillari, aglyutinativ tuzilishga ega bo'lib, bu ularning morfologiyasining murakkabligini anglatadi. So'z yasovchi va so'z o'zgartiruvchi qo'shimchalarning ko'pligi so'zlarning turli shakllarining ko'payishiga olib keladi, bu esa NER vazifasini murakkablashtiradi. Shu sababli, turkiy tillar uchun NER tadqiqotlari alohida ahamiyatga ega.

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Turk tilidagi NER tadqiqotlari bir qator ishlarda ko'rib chiqilgan. Turk tilidagi NER tadqiqotlari xususida Dilek Küçük, Adnan Yazic o‘z izlanishlarida [3] “Language-independent IE system” (Cucerzan and Yarowsky, 1999), “Statistical name tagger for Turkish” (Tür et al, 2003), “Person name tagger for financial news texts” (Bayraktar and Taşkaya-Temizel, 2008), “Person mention extractor and a string matching based coreference resolver” (Küçük and Yazıcı, 2008) ilmiy tadqiqot ishlarini misol tariqasida keltirib ushbu sohada tadqiqotlar soni cheklanganligini ta’kidlaydi.

Turk tilida nomlangan birliklarni ajratib olish vazifalari uchun asosiy leksik resurslardan biri Nurettin Demir va Emine Yılmaz tomonidan tuzilgan Türk Adları Sözlüğü [3] (Turk ismlari lug'ati) hisoblanadi. Taxminan 8300 ta ismni o'z ichiga olgan ushbu lug'at turk tilidagi matnlarda shaxs nomlarini aniqlash va tasniflash uchun ishlatiladi. Nomlangan birliklarni, masalan, shaxs nomlarini ajratib olish uchun namunalarga asoslangan usullar, shuningdek, morfologik tahlil qo'llanilishi mumkin.

Tür va boshq. (2003) tomonidan taklif etilgan dastlabki qoidalarga asoslangan tizimdan farqli o‘laroq, hozirgi ko‘rib chiqilayotgan qoidalarga asoslangan tizim va statistik tizim nafaqat shaxs, joylashuv va tashkilot nomlarini, balki raqamli va vaqtinchalik iboralarni ham hisobga oladi. Ikkala tizim ham nomlangan birliklarni ajratib olish uchun turli yondashuvlardan foydalanadi. Statistik tizim 492 821 so‘zdan iborat yangiliklar to‘plamiga o‘rgatilgan bo‘lib, u 37 277 ta nomlangan birlikni (NE) o‘z ichiga oladi. Ushbu tizim taxminan 28 000 so‘zdan (2197 NE) iborat boshqa yangiliklar to‘plamida sinovdan o‘tkazildi va F-o‘lchovda 91,56% eng yaxshi ko‘rsatkichga erishdi. Statistik tizim shartli tasodifiy maydonlar (CRF), yashirin Markov modellari (HMM) yoki chuqur o'rganish usullari kabi mashinaviy o'rganish algoritmlaridan foydalangan holda tilning statistik qonuniyatlarini (masalan, so'z chastotalari va birgalikda uchrashish chastotalari) hisobga oladi. Qoidalarga asoslangan tizim esa 20 131 so'zdan (1591 NE) iborat ma'lumotlar to‘plamida sinovdan o‘tkazildi va F-o‘lchovi 78,7% ga erishdi. Ushbu tizim leksik, grammatik qoidalar kabi aniq belgilangan lingvistik qoidalarga asoslanadi. Olingan natijalarga ko'ra, statistik tizim

qoidalarga asoslangan tizimga nisbatan sezilarli darajada yuqori ko'rsatkichlarga erishgan. Bu statistik tizimning katta hajmdagi o'qitish ma'lumotlaridan foydalanishi va qoidalarga asoslangan tizimga nisbatan ko'proq hisoblash resurslarini talab qiladigan murakkabroq usullarni qo'llashi bilan bog'liq.

Turk tilida nomlangan birliklarni aniqlash (NER) modellari oddiy morfologik analizatorlardan murakkab chuqur o'rganish usullarigacha rivojlandi.

Dastlabki ishlar (morfologik tahlil bilan) Oflazer (1994) tomonidan taklif etilgan morfologik tahlil NER samaradorligini oshirishda muhim rol o'ynadi (Tür va boshq., 2003). Tür va boshq. (2003) modeli 92,73% F1 ballga erishib, keyingi tadqiqotlar uchun asos bo'ldi.

Morfologik tahlilsiz modellar, xususan (Küçük va Yazıcı, 2009; 2012) morfologiyani hisobga olmasdan turli sohalarda sinovdan o'tkazildi va NERning bu holda ham qo'llanilishi mumkinligini ko'rsatdi.

CRF va morfologiya usuli NERda muvaffaqiyatli qo'llanilgandan so'ng, Şeker va Eryigit (2012) morfologik analizator va gazeta ma'lumotlaridan foydalanib 91,94% F1 ballga erishdi. Bu statistik va lingvistik usullarning kombinatsiyasining afzalliklarini ko'rsatadi. Ushbu ish keyinchalik ITU Turk NLP Web Service (Eryigit, 2014) ga integratsiya qilindi.

Chuqur o'rganish (DL)ga asoslangan Turk NERida birinchi yondashuv Kuru, Can va Yuret (2016) tomonidan taklif etildi. Ular belgilar darajasidagi tasvirlardan foydalanib, 91,30% F1 ballga erishdi.

Zamonaviy tadqiqotlardan biri Akkaya va Can (2021) shovqinli ma'lumotlar bilan ishlash qiyinligini ko'rsatib, 67,39% F1 ballga erishdi. So'zlarni ifodalash uchun so'z vektorlari va belgilar darajasidagi xususiyatlar (BiLSTM/CNN yordamida) birlashtiriladi. Gungor va boshq. (2019) "morfologik joylashtirishlar"ni qo'shib, 92,93% F1 ball bilan eng yaxshi natijaga erishdi va boshqa morfologik boy tillar bilan solishtirishni taqdim etdi.

Cetindag'ning maqolasida [4] turk sud-huquqiy matnlari uchun NER teglari quyidagicha:

- **PER (Shaxs):** Odam ismlari, lavozimlar.
- **COU (Sud/Davlat):** Sud nomlaridagi ma'muriy hududlar (sud ierarxiyasi uchun muhim). COU ham "Sud", ham "Davlat" ma'nosida ishlatiladi.
- **ORG (Tashkilot):** Tashkilotlar (to'liq nomlar yoki qisqartmalar).
- **DAT (Sana):** Turli formatdagi sanalar (REF va OFF ob'ektlaridagi sanalar bundan mustasno).
- **LEG (Qonunchilik):** "ayn Kanunun..." ("o'sha qonunning") kabi iboralardagi "ayn" ("o'sha").

- **REF (Ma'lumotnoma):** Hukm chiqarilgan sanalarni o'z ichiga olgan ma'lumotnomalar.
- **OFF (Rasmiy Gazeta):** Nashr etilgan sanalarni o'z ichiga olgan rasmiy gazetalar.

Yana bir tadqiqot ishida turkcha bank hujjatlaridagi nomlovchi birliklarni aniqlashda [6] semantik guruhlar va ularning tasnifi aks etgan jadval:

Kategoriya	Simbol	Ta’rif	Misol
E-mail	EMA	E-mail xabarlari	yonetim@banka.com.tr
Vaqt davomi	ZAM	Vaqt oralig‘i yoki davomi	Yeti kunichida, 1 yillik muddat
Tashkilot	ORG	Tashkilot nomlari	Turkiya Respublikasi Markaziy banki
Sana	TAR		11/05/2020, 13 yanvar
Pul birliklari	PAR	Pul qiymatlari	2 TL, 85 Avro
Maqola raqam(bob)lari	MAD	Huquqiy hijjatlarning boblari	28-modda
Qisqartma	KIS	Nomlar yoki frazalarning qisqaqcha shakli	IBAN, DAB
Qonun nomlari	KAN	Qonun nomlari	Elektron pul institutlari to‘g‘risidagi qonunning 5-moddasi
Ma’lumotnoma	ILG		131161/ Investitsiya raqami A
Farmon	YON	Federal va davlat idoralari tomonidan yaratilgan qoidalar va qoidalar	Banklarning kredit operatsiyalari to‘g‘risidagi nizom

Turk tilida nomlangan birliklarni aniqlash (NER) vazifasi uchun Twitter ma'lumotlariga asoslangan korpus [7] ham yaratilgan. Ushbu ma'lumotlar to'plami, asosan, qidiruv so'rovlariga xos bo'lgan qisqa matn ketma-ketliklarini qayta ishlash uchun mo'ljallangan bo'lib, 97 428 ta qidiruv so'rovini va 294 100 ta so'zni o'z ichiga oladi. Taxminan 100 mingta ma'lumot birligidan iborat ushbu korpus tozalash va belgilash jarayonlari orqali shakllantirilgan bo'lib, turk tilidagi NER tadqiqotlari uchun muhim manba hisoblanadi.

Qozoq tili murakkab morfologik tuzilishga ega bo'lgan agglyutinativ tildir. Bunda har bir so'z o'zagi yuzlab, hatto minglab yangi so'zlarni hosil qilishi mumkin. Bu esa nomlangan

birliklarni avtomatik aniqlash (NER) jarayonida ma'lumot yetishmasligi muammosini keltirib chiqaradi. Qozoq tili uchun Tolegen va boshqalar (2016) [9] tomonidan "Qozoq NER korpusi" (KNC) yaratilib, unga izoh berilgan va CRF (Shartli Tasodifiy Maydonlar) usuli yordamida morfologik (6 ta xususiyat) hamda so'z turkumi (4 ta xususiyat) kabi omillarning ta'siri chuqur o'rganilgan. Olingan natijalar shuni ko'rsatdiki, modelning ishlashini morfologik xususiyatlardan samarali foydalanish orqali sezilarli darajada yaxshilaydi. Natijada, CRFga asoslangan NER tizimi Qozog'iston test ma'lumotlarida 89,81% F1 ballga erishgan. Keyinchalik o'tkazilgan tadqiqotda CRFga asoslangan ushbu NER tizimi [aniq chuqur o'rganish modeli nomi ko'rsatiladi, masalan: BiLSTM-CRF, BERT] chuqur o'rganish modeliga qarshi asosiy me'yoriy ko'rsatkich sifatida qo'llanilgan. Ushbu keyingi tadqiqotda 25 xil birlik sinfi uchun jami 112 702 ta jumla va 136 333 ta izohni o'z ichiga olgan ma'lumotlar to'plamidan foydalanilgan.

Qozoq tilida nomlangan birliklarni aniqlash (NER) uchun zamonaviy mashina ta'limi modellari ishlab chiqilib, eng yaxshi model test ma'lumotlarida 97,22% F1 ballga erishdi. NER — yarim tuzilgan yoki tuzilmagan matnlardagi nomlangan birliklarni aniqlash va ularni toifalarga ajratish vazifasi bo'lib (Nadeau & Sekine, 2007), NERlar shaxslar, tashkilotlar, joylar (Sang & Meulder, 2003), miqdorlar, pul birliklari, sanalar kabi ifodalarni o'z ichiga oladi (Chinchor, 1998). NER tabiiy tillarni qayta ishlashda, jumladan, matnni tushunish, tarjima, savollarga javob berish va bilimlar bazasini yaratishda qo'llaniladi (Cheng & Erk, 2020; Babych & Hartley, 2003; Aliod va boshq., 2006; Etzioni va boshq., 2005). 1995-yildan beri NERga ilmiy, biznes (Schon va boshq., 2019) va mudofaa (Han va boshq., 2020) sohalarida qiziqish ortib bormoqda. Grishman va Sundheim (1996) televideniye yangiliklaridan 112 702 ta jumlalik (86 246 ta noyob) ma'lumotlar to'plamini yaratdi. Bu ma'lumotlar to'plami qozoq tilida so'zlashuvchi ikki tilshunos tomonidan izohlanib, eng katta qozoq NER izohli korpusini tashkil etdi. Izohlash ko'rsatmalari 25 ta NER turiga izoh berish qoidalarini va misollarini o'z ichiga oladi. Qozog'iston NERni avtomatlashtirish uchun yaratilgan to'rtta modeldan eng yuqori aniqlikdagi F1 balli 97,22% ga yetdi.

"KyrgyzNER" loyihasining asosiy maqsadi qirg‘iz matnlaridagi nomlangan birliklarni aniq aniqlash va tasniflashga qodir NER ma’lumotlar to‘plamini yaratishdan iborat. Qirg‘iz tili uchun qo‘lda izohlangan NER ma’lumotlar to‘plamini yaratish orqali turli qirg‘iz NLP (Tabiiy Tilni Qayta Ishlash) ilovalarining samaradorligini oshirish va qirg‘iz tili texnologiyalarining rivojlanishiga hissa qo‘shish ko‘zda tutilgan. Ma’lumotlar to‘plami 24.kg saytidan olingan 1500 ta yangilik maqolasini o‘z ichiga oladi. Ushbu ma’lumotlar to‘plami NER modelini o‘qitish va baholash uchun

asos bo‘lib xizmat qiladi. Izohlangan matnlar fayllarda CoNLL2003 formatida saqlangan. Qirg‘iz tilidagi matnlar ko‘pincha o‘ziga xos qiyinchiliklarga, jumladan, murakkab morfologiya va turli grammatik tuzilmalarga ega. Ushbu loyiha qirg‘iz tili uchun eng mos modelni aniqlash maqsadida turli zamonaviy NER arxitekturalari va usullarini o‘rganadi va sinovdan o‘tkazadi. Modelning samaradorligi standart ko‘rsatkichlar, o‘zaro tekshirish usullari va mavjud NER tizimlari bilan solishtirish orqali sinchkovlik bilan baholanadi hamda uning samaradorligi va asosiy modellarga nisbatan rivojlanishi ko‘rsatiladi.

Mazkur tadqiqot ishi [10] kam resursli tillardan biri bo‘lgan **ozarbayjon tili** uchun NER tizimini ishlab chiqishga bag‘ishlangan. Tadqiqot doirasida qoidalarga asoslangan va mashinaviy o‘qitishga asoslangan ikki xil model ishlab chiqilib, ularning ishlash ko‘rsatkichlari tanish va notanish ma’lumotlar to‘plamlarida solishtiriladi hamda eng maqbul yondashuv aniqlandi. Qoidalarga asoslangan yondashuvda asosiy usul sifatida statistik ma’lumotlardan foydalaniladi va u muayyan natijalarni ta’minlaydi — har ikki ma’lumotlar to‘plami uchun taxminan 70% F1 ko‘rsatkichi qayd etilgan. Mashinaviy o‘qitishga asoslangan yondashuv quyidagi uchta modelni qamrab oladi:

1. Spacy kutubxonasidan foydalangan holda noldan o‘qitilgan CNN modeli eng yuqori natijalarni har bir test ma’lumotlar to‘plami uchun 90% dan ortiq F1 ko‘rsatkichi namoyish etdi.

2. Natijalarni qiyoslash maqsadida oldindan o‘qitilgan ko‘p tilli Spacy modeli ham qo‘llanildi. Ushbu modelning test ma’lumotlarida 50% dan kam ko‘rsatkichga erishgani, NER modelini o‘qitishda domenning ahamiyatini tasdiqlaydi.

3. Qayta o‘qitilgan ko‘p tilli model mavjud ko‘p tilli modelni yangi o‘quv ma’lumotlar to‘plami asosida qayta o‘qitish orqali hosil qilingan.

Nomlangan birliklarni aniqlash (NER) texnologiyasi tatar tilidagi matnlarda nomlangan birliklarni avtomatik aniqlash va klassifikatsiya qilish orqali turli sohalarda muhim rol o'ynaydi. Nomlangan birliklarni ajratib olish algoritmi n-grammlarni solishtirish g'oyasiga asoslanadi. Taqqoslash “Tugan Tel” tatar milliy korpusi hajmi doirasida amalga oshiriladi va shu bilan natijalarning aniqligini oshiradi.[14]

NER qidiruv tizimlarining aniqligi va samaradorligini oshirib, foydalanuvchilarga kerakli ma'lumotlarni tezroq topish imkonini beradi. Korxonalarga ijtimoiy tarmoqlardagi tatar tilidagi kontentni tahlil qilish, brendlar, mahsulotlar va shaxslar haqidagi fikr-mulohazalarni aniqlash orqali jamoatchilik fikrini o'rganishga yordam beradi. Masalan, yangi mahsulot muhokamasini tahlil qilishda NER mahsulot va raqobatchilar nomlari, shuningdek, mijozlar fikrlarini aniqlaydi.

NER tarjima jarayonida nomlangan birliklarni to'g'ri aniqlash orqali madaniy jihatdan to'g'ri va aniq tarjimani ta'minlaydi. NER tizimlarini yaratishda qoidalarga asoslangan, statistik (CRF) va chuqur o'rganish (RNN, Transformer) usullari qo'llaniladi. Tatar tilining aglyutinativ tuzilishi (so'zlarga ko'plab qo'shimchalar qo'shilishi) NER uchun qiyinchiliklar tug'diradi. Shu sababli, tatar tili uchun NER tizimlarini yaratishda tilning morfologik xususiyatlarini hisobga oluvchi maxsus usullar, masalan, morfologik tahlil va lemmalashtirishdan foydalanish muhimdir. Samarali modellar yaratish uchun yetarli miqdordagi belgilangan ma'lumotlar (matn korpuslari)ning mavjudligi ham muhim ahamiyatga ega.

Lingvanex nafaqat mashina tarjimasi xizmati, balki u turli tillar, jumladan tatar tili uchun ham nomlangan birlikni aniqlash (NER) texnologiyasini taklif etadi. Lingvanex'ning tatar NER yechimining afzalligi zamonaviy neyron tarmoqlar texnologiyalaridan foydalangan holda tatar tilidagi matnlarda nomlangan birliklarni yuqori aniqlik bilan aniqlaydi. NER texnologiyasi turli sohalarda, jumladan, ma'lumot qidiruvi, ijtimoiy media tahlili, kontentni mahalliylashtirish, chat-botlar yaratish va boshqa sohalarda qo'llaniladi. Lingvanex NER yechimi API orqali boshqa ilova va tizimlarga oson integratsiya qilinadi hamda foydalanuvchi ma'lumotlarining xavfsizligini ta'minlash uchun qat'iy standartlarga amal qiladi.

Kam resursli tillar uchun tabiiy tilni qayta ishlash (NLP) vositalarini, xususan, NER tizimlarini yaratishda bir qator qiyinchiliklar mavjud. Ushbu muammolarni hal qilish uchun vositalar, ma'lumotlar va ekspertlar (tilshunoslar va maslahatchilar)dan innovatsion tarzda foydalanish zarur. Sorani kurd va tojik tillari uchun ishlab chiqilgan NER tizimlari bunga misol bo'la oladi. Ushbu tizimlar minimal qo'lda annotatsiya talab qilgan holda, lingvistik bilimlar va vazifa xususiyatlarini muvaffaqiyatli integratsiya qiladi. Asosiy maqsad – tillar o'rtasidagi yashirin qonuniyatlarni aniqlash va ularni boshqa tillarga ko'chirishdan iborat. Bu maqsadda quyidagi metodlar [15]qo'llaniladi:

- Shovqinli joylashuv ro'yxatlariga asoslangan nazoratsiz sinf induksiyasi oldindan belgilangan sinflarsiz, ma'lumotlardagi o'xshashliklarga ko'ra sinflarni avtomatik ravishda aniqlaydi. "Shovqinli joylashuv ro'yxatlari" ma'lumotlardagi xatoliklar yoki noaniqliklarni anglatadi, tizim esa bunday "shovqin"ga chidamli bo'lishi kerak.
- Ekspert baholarini hisobga oluvchi morfologik induksiyasi so'zlarning morfologik tuzilishini (o'zaklar, qo'shimchalar) avtomatik ravishda tahlil qiladi. Ekspertlarning morfologiya bo'yicha bilimlari tizimga kiritilishi mumkin, bu esa aniqlikni oshirishga yordam beradi.

Ushbu yondashuv kam resursli tillar uchun NER tizimlarini yaratishda muhim ahamiyat kasb etadi. U mavjud inson

resurslaridan samarali foydalanish va ma'lumotlardagi yashirin qonuniyatlarni avtomatik ravishda aniqlash orqali bunday tillar uchun sifatli NLP vositalarini yaratishga imkon beradi.

Uyg'ur NER tadqiqotlari ko'pincha qo'lda yaratilgan xususiyatlarga bog'liq bo'lgan CRF va gibrid usullardan foydalanadi [16, 17, 18]. Bunday xususiyatlarni to'plash qiyinligi sababli, ushbu ishda uyg'ur NER uchun ikki tomonlama LSTM va CRF qatlamiga asoslangan, pastki so'z va belgi ko'rinishidagi neyron tarmoq arxitekturasi taklif etiladi.

Mavjud tadqiqotlar asosan alohida nomlangan birlik turlariga qaratilgan (masalan, shaxslar [17], joylar [18], tashkilotlar [19], vaqt/son/miqdor [20]). Taklif etilayotgan yondashuv esa ularni bir vaqtda aniqlaydi va qo'lda ishlangan xususiyatlar o'rniga so'z va belgi ko'rinishlaridan foydalanadi.

Uyg'ur NER [21] uchun o'quv korpusining yo'qligi sababli, umumiy mashina tarjimasi va birlik chegarasi belgilashga asoslangan teglashni boshqacha usulini taklif etiladi. Bunda xitoycha-uyg'urcha parallel korpusdan foydalanib, xitoycha jumlalardagi nomlangan birlik teglari olinadi, birlashtiriladi va neyron mashina tarjimasi orqali uyg'ur tiliga o'tkaziladi. Keyin uyg'ur o'zak va affikslarini birlashtirish usuli orqali yuqori sifatli korpus yaratiladi. Ushbu korpusda o'qitilgan uyg'ur NER tizimi asosiy tizimdan 3,79% ga yuqori natijaga erishdi.

XULOSA

Xulosa qilib aytganda, turkiy tillarning aglyutinativ tuzilishi NER vazifasini jiddiy qiyinchiliklarga duchor qiladi. Ushbu muammoni hal qilish uchun morfologik tahlilga asoslangan maxsus usullar, maxsus lingvistik resurslar va zamonaviy mashinaviy o'qitish usullarini moslashtirish zarur. Kelgusidagi tadqiqotlar ushbu yo'nalishlarga qaratilishi va turkiy tillar uchun yuqori sifatli NER tizimlarini yaratishga xizmat qilishi lozim.

REFERENCES

1.Ö.Bayraktar,T.Taşkaya. Person name extraction from Turkish Financial news text using local grammar based approach. *In Proceedings of the International Symposium on Computer and Information Sciences*, 2008.

2.S.Cucerzan, D.Yarowsky. Language independent named entity recognition combining morphological and contextual evidence. *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

3. Demir, N., & Yılmaz, E. Türk Adları Sözlüğü. *Türk Dil Kurumu Yayınları*. 2000.

4.Dilek K., Adnan Y. Named Entity Recognition Experiments on Turkish Texts. *Flexible Query Answering Systems, 8th International Conference*, 2009.
5.C.Çetindag, B.Yazıcıoglu, A.Koç. Named entity recognition in Turkish legal texts. Cambridge university press. *Natural Language Engineering 1–28*, 2022.
6. O.Kabasakal , A.Mutlu. Named Entity Recognition in Turkish Bank Documents. *Kocaeli Journal of Science and Engineering*. 4(2): (2021) 86-92
7.B.C¸arık, R.Yeniterzi. A Twitter Corpus for Named Entity Recognition in Turkish. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.Marseille, 2022. 4546–4551
8. Seker G., Eryigit G. Initial explorations on using crfs for turkish named entityrecognition. *In: COLING 2012, 24th International Conference on ComputationalLinguistics, Proceedings of the Conference: Technical Papers, 8-15* December 2012,Mumbai, India. pp. 2459–2474
9.Tolegen, G., Toleu, A., Zheng, X.: Named entity recognition for kazakh using conditional random fields. *In: Proceedings of the 4-th International Conference onComputer Processing of Turkic Languages TurkLang 2016*. pp. 118–127.
10. Khundova N. Named Entity Recognition for the Azerbaijani Language. *IEEE 15th International Conference on Application of Information and Communication Technologies (AICT)*, 2021.
11.Ibiyev K., Attila N. Using Zero-Shot Transfer to Initialize azWikiNER, a Gold Standard Named Entity Corpus for the Azerbaijani Language. *Lecture Notes in Computer Science*, 2021.
12.D.Küçük, A.Yazic. Named Entity Recognition Experiments on Turkish Texts. *Flexible Query Answering Systems, 8th International Conference*, 2009.
13.O.Bayraktar, T.Temizel. Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach. *Conference: Computer and Information Sciences*, 2008.
14.Nevzorova O., Mukhamedshin D.,Galieva A. Named Entity Recognition in Tatar : Corpus-Based Algorithm. 2018
15. Patrick Littell, Kartik Goyal, David Mortensen, Alexa Little, Chris Dyer, Lori Levin.Named Entity Recognition for Linguistic Rapid Response in Low-Resource Languages: Sorani Kurdish and Tajik.
16.Rozi, A.; Zong, C.; Mamateli, G.; Mahmut, R.; Hamdulla, A. Approach to recognizing Uyhgur names basedon conditional random fields. *J. Tsinghua Univ.* 2013, 53, 873–877.6.
17. Tashpolat, N.; Wang, K.; Askar, H.; Palidan, T. Combination of statistical and rule-based approaches forUyghur person-name recognition. *Acta Autom. Sin*. 2017, 43, 653–664.7.

18.Maimaiti, M.; Abiderexiti, K.; Wumaier, A.; Yibulayin, T.; Wang, L. Uyghur location names recognition basedon conditional random fields and rules. *J. Chin. Inf. Process.* 2017, 31, 110–118

19.Maihefureti Rouzi, M.; Aili, M.; Yibulayin, T. Uyghur organization name recognition based on syntactic andsemantic knowledge. *Comput. Eng.* Des. 2014, 35, 2944–2948.20.

20.Halike, A.; Wumaier, H.; Yibulayin, T.; Abiderexiti, K.; Maimaiti, M. Research on recognition and translationof Chinese-Uyghur time and numeral and quantifier. *J. Chin. Inf. Process.* 2016, 30, 190–200.

21.Saimaiti, A., Van, L. va Yibulayin, T. Learning Subword Embedding to Improve Uyghur Named-Entity Recognition // *Information*. – 2019. – J. 10, № 4. – B. 139.