International Journal of Advanced Technology and Natural Sciences ISSN: 2181-144X

DOI: 10.24412/2181-144X-2023-2-23-32

Kenjaev X.B., Toliev X.I.

# NER YONDASHUVI BILAN O'ZBEK TILIDAGI MATNDAN MIQDORLARNI ANIQLASH QOIDALARI

**Kenjayev X.B. –** Mummamad al-Xorazmiy nomidagi Nukus filiali KT kafedrasi assistenti,
**Toliev X.I. –** Muhammad al-Xorazmiy nomidagi TATU TAD kafedrasi doktoranti.

**Annotatsiya.** Ushbu maqolada qoidaga asoslangan nomlangan ob'ektni tanib olish (Named Entity Recognation-NER) asoslari va mavjud usullari qiyosiy tahlil qilingan va qoidaga asoslangan NER ning avzalliklari keltirib o'tilgan. Xususan, NER yordamida matndan miqdor ko'rsatkichlarni ajratib olish masalasi muhokama qilinadi. Qoidalarga asoslangan NER o'lchovlar, foizlar, pul birliklari kabi miqdorlarni aniqlash va chiqarish uchun ko'p qirrali va moslashtirilgan yondashuvni taklif etadi. Lingvistik qoidalar va kalit so'zlarni ishlab chiqish orqali soha mutaxassislari tizimni sohaning o'ziga xos xususiyatlariga moslashtirishi mumkin, bu esa miqdorni aniqlashda aniqlik va moslikni ta'minlaydi. Maqola natijasida o'zbek tilidagi matndan miqdorlarni qoidaga asoslangan NER orqali ajratib olish uchun bir nechta qoidalar taklif etilgan.
**Kalit so'zlar:** NER, qoida, kalit so'z, birlik, miqdor.

# ПРАВИЛА ОПРЕДЕЛЕНИЯ ВЕЛИЧИН ИЗ ТЕКСТА НА УЗБЕКСКОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ ПОДХОДА NER

**Кенжаев Х.Б. –** ассистент кафедры КС Нукуссиский филиал ТУИТ имени Мухаммада аль-Хорезми,
**Толиев Х.И.** – докторант кафедры СПП ТАТУ имени Мухаммада аль-Хорезми.

**Аннотация.** В этой статье рассматриваются основы основанных на принципах NER и основанных на них методов, а также упоминаются преимущества NER на основе правил. В частности, обсуждается вопрос извлечения размера текста с помощью NER. NER на основе правил предлагает универсальный и настраиваемый подход для определения и расчета таких величин, как измерение, проценты и валюта. Разрабатывая лингвистические правила и ключевые слова, специалисты отрасли могут адаптировать систему к отраслевой специфике, обеспечить точность и последовательность количественных оценок. В результате в статье предложено несколько правил, из которых следует извлечь величину из текста на узбекском языке с использованием NER на основе правил.
**Ключевые слова:** НЭР, правило, ключевое слово, единица измерения, количество.

# RULES FOR DETECTING QUANTITIES FROM TEXT IN UZBEK USING THE NER APPROACH

**Kenjaev Kh.B. –** assistant of the department of CS Nukus branch of TUIT named after Muhammad al-Khwarizmi,

**Toliev Kh.I. –** Doctoral student of the Department of SPP TUIT named after Muhammad al-Khwarizmi.

**Abstract.** This article compares the fundamentals of Named Entity Recognition-NER and existing methods and mentions the advantages of rule-based NER. In particular, the issue of extracting quantities from text using NER is discussed. Rule-based NER offers a versatile and customizable approach for defining and calculating quantities such as dimensions, percentages, and currencies. By developing linguistic rules and keywords, industry professionals can tailor the system to industry specifics, ensuring accuracy and consistency in quantitative assessments. As a result of the article, several rules for extracting quantities from text in the Uzbek language using rule-based NER are proposed.
**Key words:** NER, rule, keyword, unit, quantity.

Hozirgi kunda katta hajmdagi raqamli ma'lumotlar (masalan, elektron pochta, ijtimoiy ilovalar, gazetalar va Instagram) turli tillarda mavjud. Bu katta hajmdagi ma'lumotlardan foydali axborotlarni chiqarib olishda ma'lumotlar tuzilma shakllari muhim axamiyat kasb etadi. Odatda katta ma'lumotlar (big data) strukturalangan va strukturalanmagan tuzilmagan shakllarda to'planadi. NLP ning asosiy maqsadi tabiiy tillarda berilagan axborotlar to'plamidan foydali ma'lumotlarni chiqarib olish. Buning natijasida mashina inson tabiiy tillarida ifodalangan ma'lumotlarini tushunishga asos yaratiladi [1]. Hozirda NLP foydalanib savollarga javob beradigan va matnni avtomatik qayta ishlaydigan ko'plab ma'lumot olish tizimlari ishlab chiqilgan [2,3].

Sonlar faktik va to'g'ri ma'lumotlarni yetkazishning asosiy vositasidir. Odatda hujjatlar mazmunini ifodalashda sonli qiymatlardan keng foydalanadi. Sonlarni matnlardan ajratib olish bo'yicha samarali ilmiy-amaliy natijalarga erishilganiga qaramasdan hozirgacha matndan miqdoriy ko'rsatkichlarni ajratib olishning mukammal yechimlari mavjud emas. Tahlil qilingan manbalarning aksariyatida matndan sonlarni to'g'ridan-to'g'ri chiqarib olish usullari taklif etilgan bo'lib, ularning diqqat markazida faqat ilmiy sohalar yotadi [4]. Miqdorni ajratib olish ko'pincha katta tizimning bir qismi. Bu yerda qidirish yoki matnni kiritish jarayonida raqamli axborot birliklarini tizimli aniqlashda undan miqdorlarni chiqarib olish talab qilinadi [5,6,7,8,9]. Matnlarni qayta ishlashda miqdorni aniqlashga kamdan-kam e'tibot qaratilib, bu asosiy maqsadning quyi masalasi sifatida qaralgan. Bunda matndagi miqdor ko'rsatkichlari oddiy raqamlar bilan ifodalangan shakli kabo hususiy hollarigina olingan. Aksariyat tizimlar miqdorlarni o'lchanadigan va metrik birlikka ega bo'lgan raqam sifatida qaragan [4]. Ko'pincha amaliyotda qiymatlarni ifodalashda "ot" iborasi potentsial birlik hisoblanadi (ms: "5 banan"). Bundan tashqari, miqdorlarni yanada mazmunli ifodalashda ularning dinamikasi, o'zgarishi, holat va tegishlilik tushunchalarni qamrab olishi kerak. Misol uchun, "maxsulot ishlab chiqarish xarajat 12% ga pasaydi va sof foyda 20% dan oshdi" gapdagi qiymat/birlik juftligi (12, foiz) turli xil tushunchalar, xarajat bilan foyda bog'liq va bular ma'no jihatdan qarama-qarshi o'zgarishni bermoqda (pasaydi va ortib oshdi).

Nomlangan ob'ektni tanish - NER (Named Entity Recognition) matndan semantik ma'lumot, so'z munosabatlari va ob'ektlarni olishda muhim rol o'ynaydi. Oldingi NER tadqiqot modellari turli shakllarda va o'ziga xos ob'ektlardan ma'lumotlarni oladi [10].

**NER tavsifi**

NER - bu ma'lum ma'lumotlar to'plami yoki korpusdan shaxs ismlari, tashkilotlar, vaqt, joylashuv kabi nomlangan ob'ektlarni aniqlash masalasi bilan shug'ullanadi. Nomlangan ob'ektlar, predmet soha ob'ektlari (tibbiy, oziq-ovqat), korpusda belgilangan nomli ob'ektlar kabilarni o'z ichiga oladi. Masalan:

**Matn:**
Hurshid 2023-yil Oxfordda o'qish uchun 25000$ grant yutib oldi.

**Chiqish:**
Hurshid[Odam] 2023-yil[vaqt] Oxfordda [Tashkilot] o'qish uchun 25000$ [Miqdor son] grant yutib oldi.

1990-yillarda, xususan, Message Understabding Conference (MUC) da NER vazifasi birinchi marta tadqiqotchilar tomonidan kiritilgan va e'tiborga olingan. Ushbu konferensiyada uchta asosiy NER kichik vazifalari aniqlandi: ENAMEX (ya'ni, shaxs, joylashuv va tashkilot), TIMEX (ya'ni, vaqtinchalik ifodalar) va NUMEX (ya'ni, raqamli ifodalar).

NER dan turli maqsadlarda foydlanish mumkin. Shu sababli ham NLP ilovalaridagi NERning roli bir ilovada boshqasidan farq qiladi. NER foydalanadigan NLP ilovalariga quyidagilarni misol qilib keltirish mumkin:

- ✓ Axborot olish (IR-Information Retrieval). IR - kirish so'rovi bo'yicha hujjatlar bazasidan tegishli hujjatlarni aniqlash va olish vazifasi (Benajiba, Diab va Rosso, 2009a). IR NERdan foydalanishning ikkita usuli mavjud: 1) so'rov ichidagi Nomlangan ob'ekt(NE-named Entity) ni tanib olish, 2) hujjatlarni tasniflash uchun hujjatlar ichidagi NE ni tanib olish.
- ✓ Mashina tarjimasi (MT). MT – Bir tabiiy tildagi matnni boshqa tabiiy tilga tarjima qilish vazifasi. NE tarjima komponentining sifati umumiy MT tizimining ish faoliyatini yaxshilaydigan ajralmas qismi hisoblanadi (Babych va Hartli, 2003). O'zbek tilidagi matnni boshqa tillarga tarjima qilishda shaxs nomlari ham tarjima qilinib ketishi mumkin. Masalan, "qahramon" so'zi ingliz tiliga "hero" deb tarjima qilinadi, "Qahramon" so'zi shaxs nomi sifatida kelganda tarjimasiz qolishi kerak.
- ✓ Savol-javob (QA-Question Answering). QA ilovasi IR bilan chambarchas bog'liq, ammo yanada murakkabroq hisoblanadi. QA tizimi savollarni kirish sifatida qabul qilib, buning evaziga qisqa va aniq javoblar beradi. QA tizimlarida NER dan tegishli hujjatlarni aniqlash va keyin to'g'ri javoblarni olish uchun savollar ichidagi NE ni tanib olishda foydalaniladi. NE larni to'g'ri tasniflash ushbu so'rovga javob beradigan hujjatlarn guruhini to'g'ri olishga yordam beradi.
- ✓ Chatbotlar. OpenAI-ning generativ AI, ChatGPT, Google Bard va boshqa chatbotlari foydalanuvchi matnlarida qayd etilgan tegishli ob'yektlarni aniqlash uchun NER modellaridan foydalanadi. Bu ularga foydalanuvchi savolining kontekstini tushunishga yordam beradi va ularning javoblarini yaxshilaydi.
- ✓ Mijozlarni qo'llab-quvvatlash. NER tizimlari mijozlarning fikr-mulohazalarini va shikoyatlarini mahsulot nomi bo'yicha tartibga solishi hamda muayyan mahsulotlar yoki filiallar joylashuvi bo'yicha keng tarqalgan shikoyatlarni aniqlashi mumkin. Bu mijozlarni qo'llab-quvvatlashda kiruvchi so'rovlarga tayyorgarlik ko'rishga, tezroq javob berishga va mijozlarni tegishli qo'llab-quvvatlash stollariga va tez-tez so'raladigan savollar bo'limlariga yo'naltiruvchi avtomatlashtirilgan tizimlarni yaratishga yordam beradi.
- ✓ Moliya. NER rentabellik va kredit riskini tahlil qilish tezligi va aniqligini oshirib, xususiy bozorlar, kreditlar va daromadlar hisobotlaridan raqamlarni tanib oladi. NER shuningdek, ijtimoiy tarmoqlarda va boshqa onlayn postlarda tilga olingan nomlar va kompaniyalarni ajratib ola oladi, bu esa moliyaviy institutlarga aksiyalar narxiga ta'sir ko'rsatishi mumkin bo'lgan o'zgarishlarni kuzatishda yordam beradi.
- ✓ Sog'liqni saqlash. NER vositalari laboratoriya hisobotlari va bemorlarning elektron sog'lig'i qaydlaridan muhim ma'lumotlarni ajratib olishi mumkin, bu esa tibbiyot xodimlariga ish yukini kamaytirish, ma'lumotlarni tezroq va aniqroq tahlil qilish hamda tibbiy yordamni yaxshilashga yordam beradi.
- ✓ Ta'lim. NER talabalar, tadqiqotchilar va professorlarga katta hajmdagi maqolalar va arxiv materiallarini tezda umumlashtirish, shuningdek, tegishli mavzular bo'yicha materiallani topish imkonini beradi.
- ✓ Yangiliklar provayderlari. Yangiliklar provayderlari ko'rib chiqilishi kerak bo'lgan ko'plab maqolalar va ijtimoiy media postlarini tahlil qilish va kontentni muhim

ma'lumotlar va tendentsiyalarga ajratish uchun NERdan foydalanadilar. Bu ularga yangiliklar va dolzarb voqealarni +tezda tushunish va hisobot berishga yordam beradi.

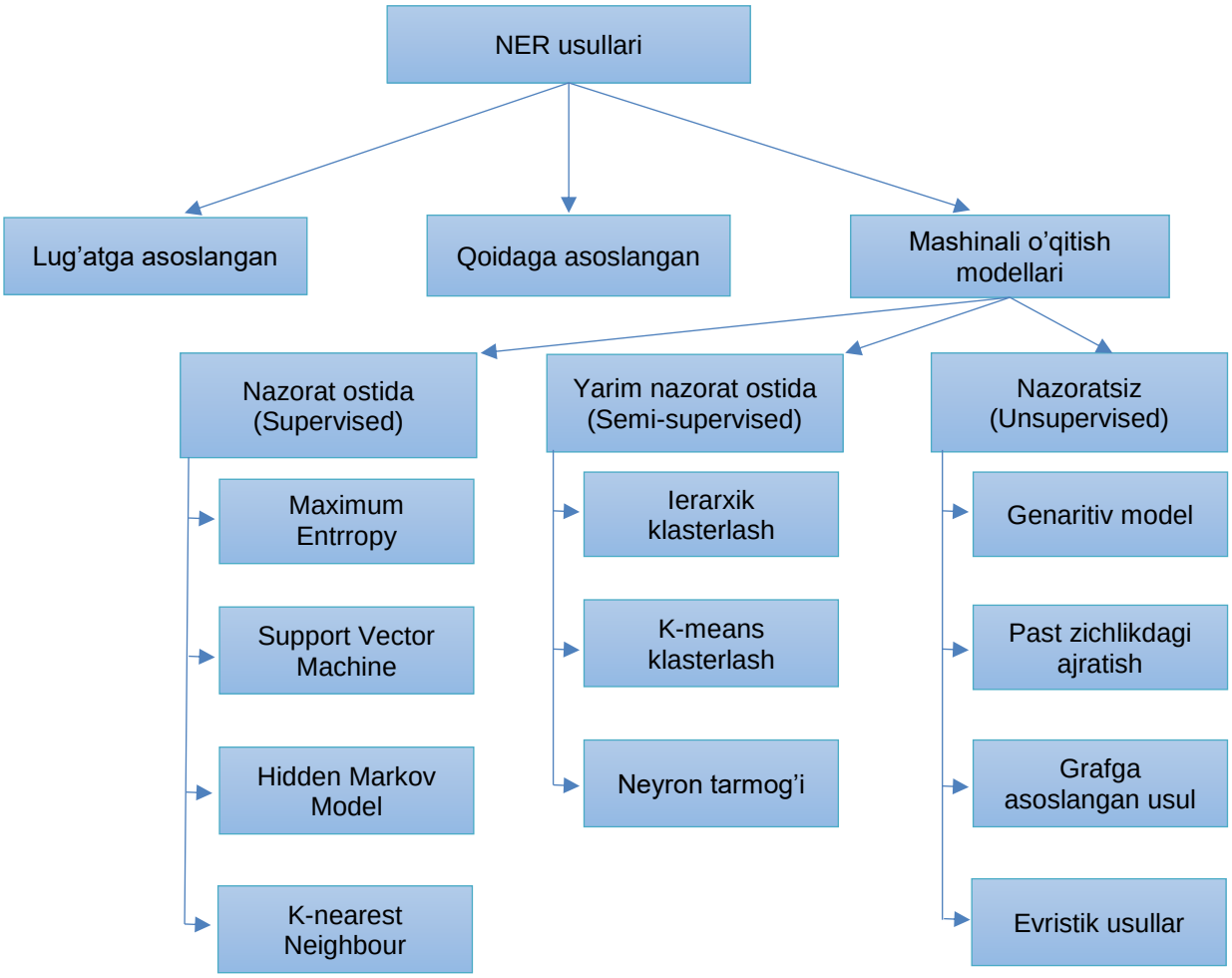
✓ Tavsiya tizimlari. Ko'pgina kompaniyalar o'zlarining tavsiya tizimlari dolzarbligini yaxshilash uchun NER dan foydalanadilar. Misol uchun, Netflix kabi kompaniyalar shaxsiylashtirilgan tavsiyalar berish uchun foydalanuvchilarning qidiruvlarini va ko'rishlar tarixini tahlil qilishda NER-dan foydalanadilar.

✓ Qidiruv tizimlari. NER qidiruv tizimlari uchun juda muhim, Internetda va qidiruvlarda tilga olingan mavzularni aniqlash va toifalarga ajratish. Bu qidiruv tizimlariga mavzularning foydalanuvchi qidiruviga mosligini tushunishga yordam beradi va foydalanuvchilarga aniq natijalarni beradi.

✓ Hissiyot tahlili. NER hissiyotlarni tahlil qilishning asosiy komponentidir. U mahsulot nomlari, brendlar va mijozlar sharhlarida, ijtimoiy media xabarlarida va boshqa strukturalanmagan matnlarda qayd etilgan boshqa ma'lumotlarni tanib oladi. Keyin his-tuyg'ularni tahlil qilish vositasi muallifning kayfiyatini aniqlash uchun ma'lumotlarni tahlil qiladi. NER, shuningdek, so'rov javoblari va shikoyatlarida xodimlarning kayfiyatini tahlil qilish uchun ham ishlatiladi.

**NER usuli**

NERning asosiy uchta usuli mavjud: lug'atga asoslangan yondashuv o'qitishga asoslangan yondashuv (learning based) va qoidalarga asoslangan yondashuv (rule-based).



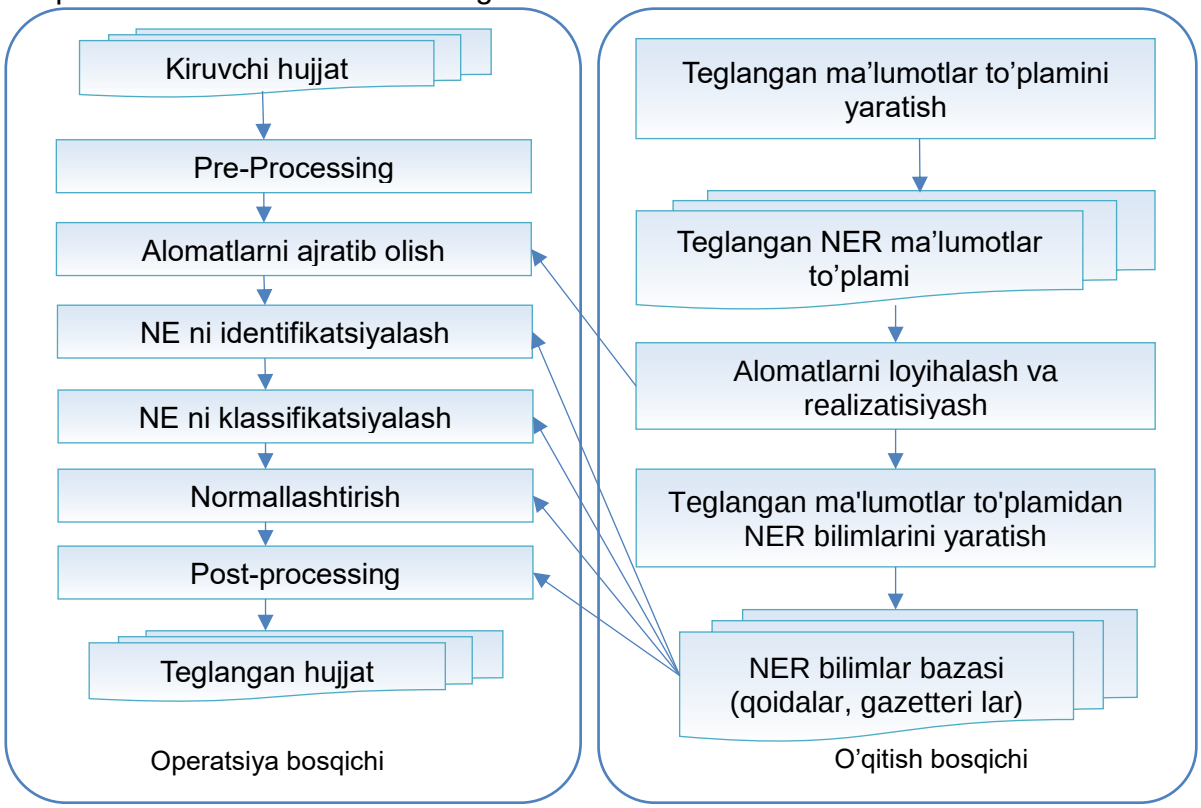

**1-rasm. Nomlangan ob'ektni tanib olish(NER) usullari.**

**Lug'atga asoslangan NER** hujjatlardan ma'lumot olish uchun ishlatiladi, chunki u tanib olingan termlar bo'yicha ID ma'lumotlarini taqdim etishi mumkin. Ushbu usul termlarni moslashtirish orqali Nomlangan ob'ektlarni aniqlaydi. Lug'atga asoslangan yondashuvlar noto'g'ri ijobiy tanib olish va yangi nashr etilgan nomlarni qamrab oladigan yagona

resursning yo'qligi kabi cheklovlarga ega. Bu usul yuqori aniqlik darajasiga ega bo'lsada, lekin u faqat lug'atga kiritilgan Nomlangan ob'ektlar(Named Entities) nigina taniy oladi.

**O'qitishga asoslangan NER yondashuvlari.** Mashinali o'qitish (ML) bu tizim qarorlarini yanada samaraliroq qiladigan murakkab qoliblar (patterns) va algoritmlar asosida avtomatik o'rganish bilan bog'liq. O'qitishga asoslangan yondashuvlar 3 turga bo'linadi : nazorat ostida, yarim nazorat ostida, nazoratsiz. O'rganish algoritmini tanlash NER tizimlari uchun ham muhimdir.

**Nazorat ostida o'qitish** usullari belgilangan o'quv ma'lumotlar to'plami (training data set) yoki korpusidan foydalangan holda mashinani o'qitishga asoslangan hamda ko'rinmas ma'lumotlar to'plami yoki korpus uchun natijalarni bashorat qiladi. Bunda NER tasniflash muammosi sifatida qaralib, unda belgilangan o'quv ma'lumotlar to'plami kirish sifatida beriladi. Ushbu hujjatlarda nomlari ko'rsatilgan ob'ektlarning namunalari inson ekspertlari tomonidan aniqlanadi. Tasniflash algoritmi 1) namunalarni umumlashtirishda, 2) nomli ob'ektlarning nusxalarini aniqlashda, 3) yangi hujjatga qo'llash mumkin bo'lgan qoidalar to'plamini aniqlashda. nazorat ostida o'qitish yondashuviga asoslangan NER tizimining kontseptual arxitekturasini 3-rasmdagi kabi ko'rsatish mumkin.




**2-rasm. Nazorat ostidagi o'qitishga asoslangan NER tizimining kontseptual arxitekturasi**

Tegishli alomatlarni tanlash yoki belgilarni to'plash nazorat ostidagi NER tizimlarida muhim vazifadir. Belgialar (Labels) o'qitish modellarini yaratishda muhim rol o'ynaydi. Ushbu modellar timsollarni (pattern) taniy oladi va ma'lumotlar to'plamini to'g'ri tasniflaydi. Turli tadqiqotchilar bir nechta nazoratsiz o'rganish usullaridan foydalananlar. Juimladan, Yashirin Markov Modeli (HMM-Hidden Markow Model) NER asosidagi tizimlar, Yordam Vektor mashinasi (SVM-Support vector machine) NER asosidagi tizimlar, Maksimal Entropiya Markov Modeli NER asosidagi tizimlar va boshqalar [ 6].

**Yarim nazorat ostida o'qitish** usullari kam hajmdagi belgilangan namuna ma'lumotlaridan foydalanadi. Keyin namunalarni katta hajmdagi teglashtirilmagan ma'lumotlar yoki korpus bilan birlashtiradi. Misol uchun turli hayvonlarning ba'zi fotosuratli ma'lumotlar to'plami (data set) mavjud bo'lsin. Bunda ayrim hayvon rasmlari mushuk, it, sigir

kabi nomlangan, lekin ko‘pchiligida nomlari (ma’lumotlar) belilmagan. Bu yerda belgilanmagan ma’lumotlarni eng yaxshi bashorat qilish uchun nazorat ostida va nazoratsiz usullar qo‘llanilishi mumkin. NER uchun yarim nazorat ostida o'rganish "bootstrapping" usuli ishlatiladi.

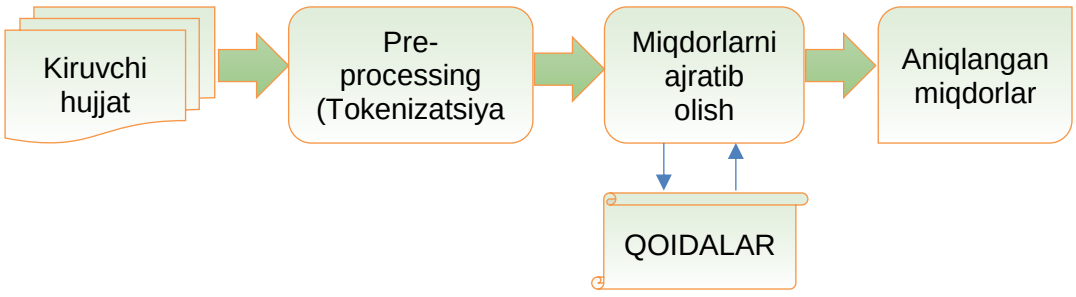
**Nazoratsiz o'qitish** usullari tasniflanmagan yoki belgilanmagan ma'lumotlardan foydalaniladi. Nazoratsiz o'qitish qaror qabul qilish uchun belgilanmagan ma'lumotlarga qo'llaniladi. Nazoratsiz o'qitishning ikkita asosiy yondashuvi: klasterlash va assotsiatsiya. Klasterga asoslangan yondashuvi kontekst o'xshashligi asosida nomlangan ob'ektlarini aniqlash uchun taqsimot statistikasidan foydalaniladi. Assotsiatsiya qoidalari yondashuvi katta ma'lumotlar to'plami yoki korpus ichida nomlangan ob'ektlari yoki qoidalarni topishda qo'llaniladi. Nazoratsiz klasterlash turli tillarda NER uchun qo'llaniladi [3].

Tasniflashning nazorat ostidagi o’qitish yondashuvlari bilan bog'liq asosiy muammo bu ularning katta, reprezentativ va yuqori sifatli belgilangan o'quv ma'lumotlar to'plamining mavjudligiga bog'liqligidir. Odatda, bunday belgilangan ma'lumotlar to'plamlari mutaxassislar tomonidan yaratiladi. Bu belgilangan ma'lumotlar to'plamini yaratish ancha qimmat, vaqt talab qiluvchi va ba’zan xatoliklarga yo'l qo'yilishi mumkin. Nazoratsiz o’qitish yondashuvlari NE (Named entity) hodisalarini avtomatik ravishda aniqlash uchun belgilanmagan hujjatlar to'plami bilan ishlaydi. Ulardan biri odatda ma'lum bir NE turidagi kichik berilgan ro'yxatidan boshlanadi (masalan, ma'lum shaxs nomlarining kichik ro'yxati) va ushbu ro'yxatga qo'shimchalarni topishga harakat qiladi. Bu berilgan misollarni qo'llashda umumiy timsollarni aniqlash va ushbu timsollarga mos keladigan har qanday so'z nomzod NE hodisasi ekanligini taxmin qilish orqali amalga oshiriladi.

Nazoratsiz yondashuvlar odatda tasniflash vazifalarida qo'llaniladigan yarim nazorat ostida yondashuvlardan biroz farq qiladi. Bu yerda, dastlabki ro'yxatga qo'shimcha ravishda, belgilanmagan misollarning katta to'plami ham bo’ladi; masalan, NERda tegishli otlarning kelishi belgisiz misollar sifatida ko'rib chiqilishi mumkin va har bir NE turi sinfga mos keladi. Klassifikator dizayni ushbu belgilanmagan misollarni hisobga oladi.

**Qoidalarga asoslangan yondashuvlar** mutaxassislar tomonidan ishlab chiqilgan qoidalar to'plamidan iborat. Qoidalar sintaktik-leksik qoidalar qoliplariga, lingvistik va predmet-sub’ekt bilimlarga asoslanadi [11]. Yetarli aniqlik va yuqori samaradorlikka erishish uchun qoidaga asoslangan NER va tasniflash tizimlari predmet sohaning o'ziga xos xususiyatlarini hisobga olgan holda ishlab chiqiladi. Ushbu tizimlar ba'zi cheklovlarga ega bo'lsada, ular qimmat portativ emas va predmet sohaga bog'liq. Bunda predmet soha bilimlari va dasturlash qobiliyatlari uchun inson tajribasi talab qilinadi. Qoidalarga asoslangan NER tizimlari faqat bitta predmet soha uchun mo'ljallangan va boshqa predmet sohalarga ko'chirilmaydi.

Odatda, o’zbek tilida so’z qo’shimchlar orqali ko’plab yangi so’zlar yaratiladi. Shu va o’qitiladigan korpus kamligi sababli o’zbek tilidagi sonlarni ajratib olishda qoidalarga asoslangan NERdan foydalanish samarali hisoblanadi. Masalan, o’zbek tilidagi “shahardagilarning” so’zi ingliz tilidagi sinonimi quyidagicha: “*people who live in the city*”. Qoidalarga asoslangan NERning afzallik jihati shundan iboratki, ma’lum soha uchun tabiiy tildagi har bir o’zgachaliklarga ekspertlar qoidalar yaratishi mumkin.




**2-rasm. Qoidalarga asoslangan NER yordamida miqdorlarni aniqlash jarayoni**

**Miqdorni ifodalash**

Umumiy holda sanaladigan yoki o'lchanadigan har qanday son miqdor hisoblanadi. Bu miqdorni quyidagi komponentlar $(v; u; ch; cn)$ bilan ifodalash mumkin [5]:

1. *Qiymat* $(v)$: Haqiqiy son, qiymatlar diapazoni, kattalik, ko'plik yoki davomiylikni tavsiflaydi. Masalan, "*avtomobil 0 dan 72 km/soatgacha tezlashdi*", bunda *v*=(0;72) diapazon; *"avtomobil 72 km/soat tezlikka tezlashdi",* bu bitta qiymatga ega *v*=72. Qiymatlar turli kattalikda bo'lib ba'zan kasr sonlarni ham o'z ichiga oladi. Masalan, *" daromadining 1/5 qismi",* bunda *v*=0,2.

2. *Birlik* $(u)$: muayyan kattalikni miqdoran baholash uchun asos. O'zbek tilida birliklar odatda sondan keyin kelib, u sonning o'lchov birligini ko'rsatadi. O'lchov birliklar nomi og'irlik (kg), uzunlik (m), suyuqliq (l) kabi juda ko'p. Masalan, "U bozordan 2 kg olma oldi." gapida o'lcham birilik kilogramm (*u=kg*). Bunga qo'shimcha o'zbek tilidagi *ta, nafar, dona* kabi miqodor so'zlar ham birliklarga kiradi.

3. *O'zgarish (ch)*: Miqdor qiymatining o'zgartiruvchisi, bu qiymat qanday o'zgarib borayotganini tavsiflaydi, masalan, *"taxminan 35$"* taxminiylikni tavsiflaydi. O'zgartirish uchun asosan 4 toifa bor: *= (teng), ∼ (taxminan), > (ko'proq), < (kamroq)*. Bu toifalar miqdor chegaralarini tavsiflaydi. Mavjud o'zgarish turlariga yana ikkita toifani kiritish mumkin: o'sish va pasayish; yuqoriga va pastga. Masalan, "Alochi o'quvchilar soni o'tgan yilgidan 8% ga oshdi", bunda *ch=oshdi*.

4. *Tushuncha(konsepsiya) (cn)*: Tushunchalar o'lchanadigan xususiyatlar yoki qiymat nazarda tutilgan yoki ta'sir qiladigan ob'ektlardir. Masalan, "Jami mahalla aholisining 180 nafarini qariyalar tashkil etadi." gapda 180 miqdor *cn=qariyalar* ning sonini ifodalaydi; "BMW Group 200 million dollar sarmoya kiritmoqda" gapda investitsiya *cn=BMW Group* tomonidan amalga oshirilgan. Ayrim hollarda kontseptsiya gapning turli qismlarida keladi, masalan, "iPhone 11 64 GB xotiraga ega." gapda *cn=iPhone 11, xotira*.

**Miqdorni ajratib olish**

Yuqoridagi 4 ta komponenta orqali matnlardan miqdorlar chiqarib olinadi. Bunda kiruvchi ma'lumot matndagi gaplar bo'lsa, chiquvchi ma'lumot aniqlangan miqdorlardir. Misol uchun *"Nur mahallasida jami 1238 nafar aholi istiqomat qiladi. Aholi soni o'tgan yilgidan 10% ga oshdi"*. Ushbu misolda miqdorlar quyidagicha bo'ladi:

1-gap: ⟨v=1238, u=nafar, cn=jami aholi⟩

2-gap: ⟨v=10, u=foiz, ch=oshdi, cn=aholi soni⟩

**Tokenizatsiya**

Matn tahlilida qaralayotgan masalaning o'ziga xis xususiyatidan kelib chiqib, ko'pchilik hollarda maxsus so'z tokenizatsiyasi amalga oshiriladi. Bu o'zgacha tokenizator qiymatlar va birliklardagi ajratuvchi belgilarni farqlaydi va so'zlar orasidagi bo'linishlanni foydalanmaydi. Masalan, *"Spark 8 soniyada 0 dan 80 km/soat tezlikka chiqadi."* gapda oddiy tokenizator *km/s → (km, /, h)* shaklida tokenlaydi. Yana bir misol, tinish belgilarini o'z ichiga olgan raqamli token (masalan, *2:33E-3*)

**Qiymat, birlik va o'zgarishlarni aniqlash**

Tokenlashtirilgan matn bog'liqlikni tahlil qilish daraxti va POS teglariga asoslangan qoidalar to'plamiga mos keladi. Qoidalar *qiymat, birlik va o'zgarish* bilan bog'liq tokenlarni topishga mo'ljallangan. *Qiymat/birlik* juftliklari ko'pincha turli xil gaplarda raqam/ot, raqam/belgi yoki son/sifatlar to'plamidir. Diapazonlarda qoidalar yanada murakkablashadi, chunki pastki va yuqori chegaralar "dan...gacha" yoki "oraliq" kabi kalit so'zlar bilan aniqlanadi. O'zgarishlar ko'pincha raqamga bevosita aloqador bo'lib, uning qiymatini sifat yoki fe'llar o'zgartiradi.

**Kontseptsiyani aniqlash qoidalari**

Kontseptsiya(tushuncha)lar quyidagi qoidalar asosida aniqlanadi:

1. Agar *nafar, ta, dona* kabi hisob so'zlar sonlardan keyin va hech qanday qo'shimcha qo'shilmasdan kelsa va son oldidan *jami, umumiy* kabi jamlovchi so'zlar kelsa,

aniqlangan son oldidagi jamlovchi so'z va hisob so'zdan keyin kelgan so'z va gapdagi fe'l tushuncha sifatida olinadi. Masalan, "Bizning mahallada jami 1280 nafar aholi yashaydi", bunda *cn=(jami aholi yashaydi)*.

2. Agar *nafar, ta, dona* kabi hisob so'zlar sonlardan keyin va hech qanday qo'shimcha qo'shilmasdan kelsa va gapda kesim bo'lmasa, son oldidagi so'z va so'zlar tushuncha sifatida olinadi. Masalan, "Bizning mahalladagi jami aholi soni 1280 nafar.", bunda *cn=( jami aholi soni)*.

3. Agar yuqopridagi *nafar, ta, dona* kabi hisob so'zlar sondan keyin kelib *-i, -ini, -sini* qo'shimchalari qo'shilgan bo'lsa u holda tushuncha sifatida sondan keyin kelgan so'z yoki so'zlar olinadi. Agar gapda fe'l bo'lsa, fe'l ham tushuncha sifatida olinadi. Masalan, "Bizning mahallada yashovchi fuqarolarning 23 *nafar**ini*** ka'm ta'minlanganlar tashkil qiladi", bunday holatda *cn=(ka'm ta'minlanganlar(tashkil qiladi))*

4. Agar yuqopridagi *nafar, ta, dona* kabi hisob so'zlar sondan keyin kelib *-ni* qo'shimchalari qo'shilgan bo'lsa u holda tushuncha sifatida sondan oldin kelgan so'z yoki so'zlar olinadi. Agar gapda fe'l bo'lsa, fe'l ham tushuncha sifatida olinadi. Masalan, "Bizning mahallada ka'm ta'minlanganlar 23 nafar**ni** tashkil qiladi", bunday holatda *cn=(ka'm ta'minlanganlar(tashkil qiladi))*

5. Ba'zan gapda sondan keyin kelgan hisob so'zlarga *–dan* qo'shimchasi qo'shilib, undan keyin *ko'proq, kamroq* kabi o'zgarish so'zlar keladi. Bunday holatda o'zgarish so'zlardan keyin kelgan so'z yoki so'zlar tushuncha sifatida olinadi. Masalan, "Mahallamizdagi fuqarolarning 780 nafardan ko'prog'i oliy ma'lumotli", bu holatda *cn=(oliy ma'lumotli).*

6. Agar gapda sonlar uyushiq bo'laklar tarkibida kelsa, unda gapdagi kesim har bir son uchun tushuncha sifatida olinadi. Masalan, "Mahallada jami 1336 ta kirish, 251 ta chiqish hujjatlari, 63 ta buyruq va 226 ta farmoyish ro'yxatga olindi.". Bunda tushunchalar *cn=(kirish ro'yxatga olindi), cn=(chiqish hujjatlari ro'yxatga olindi), cn=(63 ta buyruq ro'yxatga olindi), cn=(226 ta farmoyish ro'yxatga olindi).*

7. Agar gapda umumiydan qismni ajratib ko'rsatish ma'nosida sonlar ishlatilsa *-dan* qo'shimchasi qo'shilgan so'zdan keying so'zlar bitta gap sifatida undan oldingi so'zlar alohida gap sifatida qaraladi. Bunda *-dan* qo'shimchasi qo'shilgan so'z odatda ikki tomonda ham tushunch sifatida olinadi. Masalan, "Mahalladagi 700 nafar yoshlardan 581 tasi sport to'garaklariga qatnashadi". Bunda *cn=yoshlar, cn=yoshlar to'graklariga qatnashadi*

Shuni ta'kidlab o'tish zarur tushunchalar sifatida olinayotgan so'z yoki so'zlar kalit so'zlar bazasiga oldindan kiritilgan bo'lishi kerak.

**Baholash:**Baholash ko'rsatkichlari Nomlangan ob'ektni tanib olish (NER) tizimlari, shu jumladan qoidalarga asoslangan NER tizimlarining ishlashini baholash uchun muhim vositadir. Ushbu ko'rsatkichlar tizim matn ma'lumotlaridagi nomli ob'ektlarni qanchalik to'g'ri aniqlayotganini aniqlashga yordam beradi. Quyida ko'p qo'llaniladigan NER baholash ko'rsatkichlari keltirilgan:

**Aniqlik (Precision):** Aniqlik to'g'ri aniqlangan ob'ektlarning NER aniqlagan ob'ektlarning umumiy soniga nisbatini hisoblash orqali tizimining aniqligini o'lchaydi. U "Tizim belgilagan barcha ob'ektlardan qanchasi to'g'ri edi?" degan savolga javob beradi.

$$Precision = \frac{TP}{TP + FP}$$

Rost ijobiy (TP – true positive): to'g'ri aniqlangan ob'ektlar soni.

Yolg'on ijobiy (FP - false positive): tizim tomonidan aniqlangan, lekin haqiqatda mavjud bo'lmagan ob'ektlar soni (noto'g'ri ob'ektlar).

Yuqori aniqlik qiymati tizim noto'g'ri ob'ektlarni belgilamaslikda yaxshi ekanligini ko'rsatadi, lekin u ba'zi ob'ektlarni o'tkazib yuborishi mumkin.

**Sezuvchanlik (Sensitivity):** Sensitivity tizim qobiliyatini to'g'ri aniqlangan ob'ektlarni ularning umumiy soniga nisbatini hisoblash orqali o'lchaydi. U "Barcha haqiqiy ob'ektlardan nechtasini tizim aniqladi?" degan savolga javob beradi.

$$Sensitivity = \frac{TP}{TP + FN}$$

Yolg'on salbiy (FN - False Negatives): rostdan mavjud bo'lgan, lekin tizim tomonidan aniqlanmagan ob'ektlar soni (o'tkazib yuborilgan ob'ektlar).

Yuqori Sensitivity qiymati ob'ektlarning ko'pini tanib olishda tizim yaxshi ekanligini ko'rsatadi, lekin ko'proq noto'g'ri ijobiy natijalar berishi mumkin.

**F1 ball:** F1 ball aniqlik va sezuvchanlik ning garmonik o'rtacha ko'rsatkichidir. Bu aniqlik va sezuvchanlik o'rtasidagi muvozanatni ta'minlaydi. U ayniqsa noto'g'ri ijobiy va noto'g'ri salbiylarni hisobga oladigan bitta ko'rsatkichni hisoblashda foydali hisoblanadi. U ko'pincha NER tizimlarini solishtirish va baholash uchun ishlatiladi.

$$F1 = 2 * \frac{(precision * Sensitivity)}{precision + Sensitivity}$$

F1 balli 0 dan 1 gacha o'zgarib turadi va yuqoriroq qiymat NER tizimining yaxshiroq ishlashini ko'rsatadi.

**Tartiblilik (Accuracy):** Tartiblilik - bu to'g'ri aniqlangan ob'ektlarning ma'lumotlar to'plamidagi ob'ektlarning umumiy soniga nisbatini hisoblaydigan ko'rsatkich. Tartiblilik umumiy ko'rsatkich bo'lsa-da, matnning aksariyat qismida ob'ektlar mavjud bo'lmagan nomutanosib ma'lumotlar to'plami bilan ishlashda u noto'g'ri bo'lishi mumkin.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Eng mos ko'rsatkichni tanlash NER ilovangizning aniq maqsadlari va talablariga bog'liq. Misol uchun, agar noto'g'ri pozitivlarni minimallashtirish kerak bo'lsa, aniqlik (precision) ga e'tibor qaratish muhim, agar iloji boricha ko'proq ob'ektlarni tanib olish kerak bo'lsa, sezuvchanlik (sensitivity) muhimroq hisoblanadi. Muvozanatli ishlash o'lchovi kerak bo'lganda, F1 ball yaxshi tanlovdir.

**Xulosa.**

Strukturalanmagan matnlardan miqdorlarni aniq ajratib olish katta ahamiyatga ega hisoblanadi. Moliyaviy tahlil, ilmiy tadqiqotlar, elektron tijorat, sog'liqni saqlash yoki raqamli ma'lumotlarga ega har qanday soha uchun miqdorlarni ajratib olish muhim vazifadir.

Qoidalarga asoslangan NER matndan miqdorlarni aniqlash va olish uchun mustahkam va moslashuvchan tizimni taklif etadi. Uning afzalliklari moslashuvchanlik va shaffoflik hisoblanadi. Shu bilan birga mashinani o'qitishga asoslangan yondashuvlarga nisbatan kamroq resurs talab qiladi. Muayyan qoidalar va qoliblarni ishlab chiqish orqali soha mutaxassislari tizimni o'z ma'lumotlarining o'ziga xos xususiyatlariga moslashtirishi mumkin, bu esa miqdorni olishda aniqlik va moslikni ta'minlaydi.

Qoidalarga asoslangan NER ning miqdorlarni chiqarishdagi asosiy kuchli tomonlaridan biri uning butun sonlardan o‘nli kasrlar, kasrlar va eksponential belgilargacha bo‘lgan turli sonlarni tanib olish qobiliyati hisoblandi. Bundan tashqari, u tegishli o'lchov birliklarini boshqarishi mumkin, bu miqdorlar ifodalagan kontekstni har tomonlama tushunishga imkon beradi.

Qoidalarga asoslangan miqdorlarni ajratib olish tizimlarini baholash ularning samaradorligini ta'minlash uchun juda muhimdir. Aniqlik, sezuvchanlik va F1 reytingi kabi ko'rsatkichlar aniqlik va to'liqlikni o'lchash uchun qimmatli mezon bo'lib xizmat qiladi. Xatolarni tahlil qilish asosida qoidalarni doimiy ravishda takomillashtirish va moslashtirish vaqt o'tishi bilan tizim ish faoliyatini yaxshilash uchun muhim qadamdir.

Xulosa qilib aytganda, qoidalarga asoslangan NER tabiiy tilni qayta ishlashda muhim vosita hisoblanadi. Uning matnli ma'lumotlardan miqdorlarni olishdagi qobiliyati turli sohalarda ma'lumotlarga asoslangan qarorlar qabul qilish uchun yordam beradi. Matnli ma'lumotlarning hajmi va murakkabligi o'sishda davom etar ekan, qoidalarga asoslangan NER matn ichidagi raqamli ma’lumotlarni tanib olishda muhim yechimlardan biri bo’lib qoladi.

**Foydalanilgan adabiyotlar ro‘yxati:**


**1.** Shah, D. N., and H. Bhadka. 2017. A survey on various approaches used in named entity recognition for Indian languages. International Journal of Computer Application 167 (1):11–18. doi:10.5120/ijca2017913878.

**2.** L.A.Pizzato ,D.Molla , C.Paris, Pseudo relevance feedback using named entities for question answering, in: Proceeding soft he 2006 Australian Language Technology Workshop, ALTW-2006,2006,pp.89–90

**3.** Sazali, S. S., Rahman, N. A., & Bakar, Z. A. (2016). Information extraction: Evaluating named entity recognition from classical Malay documents. 2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP). doi:10.1109/infrkm.2016.7806333

**4.** Luca Foppiano, Laurent Romary, Masashi Ishii, and Mikiko Tanifuji. 2019. Automatic identification and normalisation of physical measurements in scientific literature. In Proceedings of the ACM Symposium on Document Engineering 2019, Berlin, Germany, September 23-26, 2019, pages 24:1–24:4. ACM.

**5.** Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

**6.** Tongliang Li, Lei Fang, Jian-Guang Lou, Zhoujun Li, and Dongmei Zhang. 2021. AnaSearch: Extract, Retrieve and Visualize Structured Results from Unstructured Text for Analytical Queries. In *WSDM’ 21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 906–909. ACM

**7.** Sunita Sarawagi and Soumen Chakrabarti. 2014. Opendomain Quantity Queries on Web Tables: Annotation, Response, and Consensus Models. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’14, New York, NY, USA - August 24 - 27, 2014*, pages 711–720. ACM.

**8.** Somnath Banerjee, Soumen Chakrabarti, and Ganesh Ramakrishnan. 2009. Learning to Rank for Quantity Consensus Queries. In *Proceedings of the 2 nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 243–250. ACM.

**9.** Arun S. Maiya, Dale Visser, and Andrew Wan. 2015. Mining Measured Information from Text. In *Proceedings of the 38th International SIGIR Conference on Research and Development in Information Retrieval*, pages 899–902. ACM.

**10.** Gürkan, A. T., B. Özenç, I. Çam, B. Avar, G. Ercan, and O. T. Yıldız. 2017. A new approachfor named entity recognition. 2nd international conference on computer science and engineering 474–79. doi: 10.1109/UBMK.2017.8093439

**11.** Ben Abacha, A., Zweigenbaum, P.: Medical entity recognition: a comparaison of semantic and statistical methods. In: Proceedings of BioNLP 2011 Workshop, pp. 56–64. Association for Computational Linguistics, Portland, June 2011. http://www.aclweb.org/anthology/W11-0207