LINGVISTIKA
LINGUISTICS

Tabiiy tilni qayta ishlash (nlp)da so‘z turkumlarini teglash masalasi

Manzura Abjalova[1]
Dilrabo Elova[2]

Abstrakt

Kompyuter lingvistikasida tabiiy tilni qayta ishlash ancha murakkab jarayon bo‘lib, unda ijtimoiy tabiatga ega tilning barcha hodisalari, umumiy va xususiy jihatlari, istisnoli holatlari, fonetik, morfologik, leksik, grammatik, semantik va hatto orfoepik xususiyatlarini e’tiborga olish zarur hisoblanadi. Sun’iy intellekt tizimini yaratishga bel bog‘langan ayni damda matn birliklarini raqamli texnologiyalar orqali qayta ishlanishiga erishish muhim natija hisoblanadi. Bunday jarayonda so‘zlarning turkumini aniqlash zarur. Jahon kompyuter lingvistikasidan ushbu lingvo-texnik tahlilning so‘zlar turkumini aniqlash – PoS-tagger, ya’ni so‘z turkumlarini teglash, shuningdek, uning matnlarni avtomatik qayta ishlash jarayoni bosqichi ekanligi ma’lum. Til korpuslarini yaratish uchun boshlangan dastlabki teglash harakatlari bugungi kunga kelib, matn bilan bog‘liq ko‘plab dolzarb masalalar yechimini bermoqda. NLP, ya’ni tabiiy tilni qayta ishlash jarayonida ham so‘z turkumlarini teglash birlamchi vazifa hisoblanib, buning natijasida omonimlikni, ko‘p ma’noli so‘zlar semantikasini aniqlash kabi turli lingvistik noaniqlik matn tarkibida tahlil qilinishiga erishiladi.

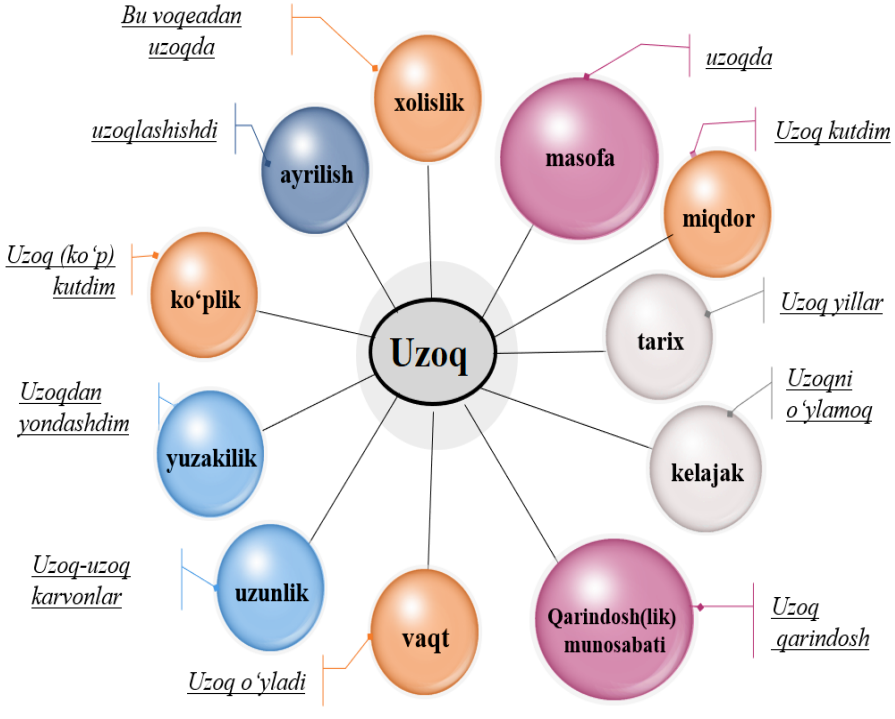
Mazkur maqolada so‘z turkumlarini teglash zarurati, matnlarni kompyuter tahlili jarayonidagi ahamiyati, teglash usullari haqida so‘z yuritiladi.

Kalit so‘zlar: *tabiiy tilni qayta ishlash, teglash, teg, so‘zlar turkumi, formal til, pragmatik xususiyat, korpus, polisemiya, omonimiya.*

[1]*Abjalova Manzura Abdurashetovna* – filologiya fanlari bo‘yicha falsafa doktori (PhD), Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: abjalovamanzura@navoiy-uni.uz
ORCID ID: 0000-0002-1927-2669
[2]*Elova Dilrabo Qudratillayevna* – o‘qituvchi, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: dilrabo@navoiy-uni.uz
ORCID ID: 0000-0002-1927-2968
Iqtibos uchun: Abjalova, M.A., Elova, D.Q. 2021. “Tabiiy tilni qayta ishlash (nlp)-da so‘z turkumlarini teglash masalasi”. *Ozbekiston: til va madaniyat* 1: 6–20.

Kirish

Inson borki, muloqotga oshiqadi, o‘z fikrlari va tuyg‘ularini izhor etishni xohlaydi. Axborot asrida bo‘lajak haqiqatni virtual tasavvur qiling: yoningizdagi robot itga “Men seni yaxshi ko‘raman” (mehr, samimiy munosabat izhor etilyapti) va “Men seni yaxshi ko‘ryapman” (ko‘zning yaxshi ko‘rayotganligini anglatadi) jumlalarini aytdingiz. Har ikki jumlani eshitib it dumini qimirlatadi. Lekin u bu jumlalarni qanday ma’noda qabul qilgan bo‘ladi? Robot it “Men seni yaxshi ko‘raman, Qoravoy” deganingizni eshitganida, u “yaxshi ko‘rmoq” ibora ekanligini anglashi mumkin. Ko‘zlaringizdan, yuzingizda namoyon bo‘lib turgan mehringiz orqali bu ifodaning tuyg‘u ekanligini va unga qandaydir munosabat bildirishini tushunadi. Lekin unga qarata aynan ko‘rish a’zoyingiz (ko‘z)ning shoh pardasi yanada ravshan ko‘ra boshlaganini bildirib aytilgan “Men seni yaxshi ko‘ryapman” (buyam sevinch va hayajon bilan) jumladagi birikmani to‘g‘ri ma’noda tushunmasligi mumkin. Mana shu misol robotni, ya’ni mashinani biz biladigan tilda muloqot qilishni o‘rgatish kerakligini ko‘rsatadi. Yuqoridagi misoldan ayon bo‘lganidek, muayyan so‘zning turli xil kontekstda ishlatilishidagi farqni tushunish juda muhim. Tilning pragmatik xususiyatiga oid ushbu fikrga “uzoq” so‘zi orqali yana bir misol keltiramiz (1-chizma):



1-chizma.

Mashinani tilga o‘qitish uchun nima qilmoq kerak?

Ma’lumki, deyarli barcha mamlakat xalq ta’limi maskanlarida boshlang‘ich sinflaridanoq o‘quvchi gapni o‘qiydi va undagi ot, sifat, son, fe’l, ravish, olmosh so‘z turkumlarini aniqlaydi. Mana shu so‘z turkumlari **tegi** hisoblanadi. So‘z turkumlarini teglash (ingliz tilida bu *part-of-speech tagging* (*POS tagging* yoki *PoS tagging* yoxud POST), rus tilida *частеречная разметка* deyiladi) matnni avtomatik qayta ishlash bosqichi bo‘lib, uning vazifasi matnda qo‘llangan so‘z (shakl)larning turkumi va grammatik xususiyatlarini aniqlash hisoblanadi. Shu vazifasi bilan POS-tagging matnni avtomatik tahlil qilishning dastlabki bosqichlaridan biri sanaladi.

Korpus lingvistikasida so‘z turkumlarini teglash, grammatik kategoriyalarni teglash va so‘zlarni toifalashda noaniqliklarni bartaraf etish uchun so‘zni faqat uning lug‘atdagi shakliga asoslanib emas, balki matn (jumla)dagi ifodasi bo‘yicha uning turkumlik tegi va jumla (xatboshi, ibora)da boshqa so‘zlar bilan birikish imkoniyatini hisobga olish muhim sanaladi. Gap bo‘laklari teglarini identifikatsiyalash bir muncha qiyin jarayon. Sababi o‘zbek tilidagi jamiki so‘zlarni universal holda 12 turkum doirasida teglash imkoniyati yo‘q. So‘z uning jumla tarkibida reallashish holati va N-gramm [Abjalova 2020, 73-77] so‘zlarning semantik valentligiga binoan polifunksional bo‘lishi mumkin. Masalan: “Shifoxonaga bemorni keltirishdi” va “Shifoxonaga bemor odamni keltirishdi” jumlalardan birinchisida bemor so‘zi turkumlik belgisi (kim? so‘rog‘iga javob berayotgan tushum kelishigidagi so‘z)ga ko‘ra ot turkumi, 2-jumlada esa (qanday? so‘rog‘iga javob beryapti) sifat turkumi vazifasidagi so‘z hisoblanadi. O‘zbek tili izohli lug‘atida mavjud 11 000 o‘zlashma so‘zlardan 66 ta xuddi shunday polifunksional so‘zlar aniqlandi [O‘zbek tilining izohli lug‘ati 2006].

Bir so‘zning kontekstda ifodalanishiga ko‘ra turli turkumga xoslanishi so‘z turkumlari (ST) teglari uchun umumiy parametrni belgilash imkonini bermaydi. Bu holat korpus uchun ST teglarini qo‘lda bajarish imkonsizligini ko‘rsatadi. Shuningdek, yangi kontekstlarning yuzaga kelishi va tilda neologizmlarning paydo bo‘lishi teglashtirish jarayonining davomiyligini ko‘rsatadi. Shu bois til korpusida ST teglashtirishda mashinali kodlashtirishga tayaniladi. Rus tilidagi matnlarni qayta ishlashda so‘z turkumlarini teglashtirish uchun *Yandex* qidiruv tizimida *Mystem* morfologik analizatori, *Tree Tagger’*da rus tili *utuliti* [Abjalova 2020,157], Python dasturlash tilida yaratilgan NLTK dasturiy kutubxona mavjud.

So‘z turkumlarini teglash tarixi

PoS-tagging bo‘yicha tadqiqotlar korpus tilshunosligi bilan chambarchas bog‘liq. Matnni kompyuterda tahlil qilish uchun ingliz tilining birinchi yirik korpusi 1960-yillarning o‘rtalarida Genri Kucher va V. Nelson Frensis tomonidan Braun Universitetida ishlab chiqilgan. Shu bois mazkur til korpusi “Braun korpusi” nomiga ega bo‘lgan.

Brown korpusida so‘z turkumlarini teglash uchun ko‘p yillar davomida inson tomonidan so‘zlar va ularning turkumlari ro‘yxat qilingan. Qoidalarga asoslangan dastlabki harakatlar Green va Rubin dasturi yordamida qo‘lda tuzilgan yirik ro‘yxat asosida amalga oshirildi. Bunda, asosan, lingvistik birliklarning ketma-ket doimiy joylashuvi e’tiborga olingan. Masalan, artikldan keyin ot turkumi (noun) keladi, fe’l (verb) turkumi emas. Bunday qat’iy qoidalarga asoslangan dastur 70% to‘g‘ri ishladi. Uning natijalari bir necha bor qo‘lda tekshirildi va tuzatildi, keyinchalik foydalanuvchilar tomonidan ham tuzatishlarni yuborish imkoniyati yaratildi, natijada 70-yillarning oxiriga kelib teglashtirish (yoxud markirovka) deyarli mukammal bo‘ldi.

Ushbu korpusdan son-sanoqsiz so‘zlar chastotasi va ularning turkumlarini o‘rganish uchun foydalanilgan va boshqa ko‘plab tillarda shunga o‘xshash teglarni shakllantirish hamda rivojlantirishga sababchi bo‘lgan. Yaxshi teglangan ta’limiy korpuslar til modelini sinovdan o‘tkazish va takomillashtirish uchun qimmatli manbadir. Matnlar korpusi tilshunosga grammatik qoidalarni ishlab chiqishda e’tibordan chetda qolgan lingvistik va nutqiy vaziyatlarni ko‘rsatib (yoxud eslatib) turadigan tabiiy til manbasidir.

Bir muncha vaqt so‘z turkumlarini teglash tabiiy tilni qayta ishlash (NLP)ning ajralmas qismi deb hisoblangan, sababi ayrim holatlar mavjudki, so‘zning turkumini uning (ayniqsa, polifunksional va ko‘p ma’noli so‘zlarning) semantikasini, hattoki kontekst pragmatikasini tahlil qilmasdan aniqlab bo‘lmaydi. Kompyuter dasturlarida esa ularning mukammalligini ta’minlovchi semantik va pragmatik tahlil bosqichlarini yaratish ancha murakkab jarayon hisoblanadi.

So‘z turkumlarini teglash zarurati

ST teglashtirish usullarini ko‘rib chiqishdan oldin, avvalo, STni teglash nima uchun kerak va ulardan qayerda foydalanilishiga to‘xtalamiz (2-chizma).

2-chizma.

Eng muhimi, ST teglari tabiiy tilni qayta ishlash (Natural Language Processing / NLP) uchun eng birlamchi zaruriy lingvistik element hisoblanadi, shu bois STni teglash NLPda turli xil muammolarni soddalashtirish uchun dastlabki shart sifatida amalga oshiriladi.

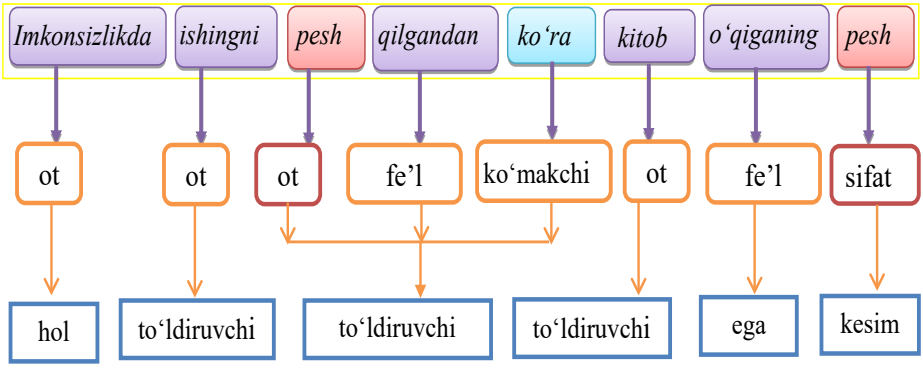
Tilning fonetik, leksik, morfologik, sintaktik va semantik sathlarini o‘zida qamragan universal lingvistik axborot tizimi (korpus)da aynan morfotahlilni amalga oshirish uchun ham so‘z turkumlarini teglash zarur hisoblanadi. Shunda istalgan so‘z (shakl)-ga o‘quvchi uchun to‘g‘ri tahlil natijasi til korpusida taqdim etish imkoniyati bo‘ladi. Til korpusi mavjud bo‘lmagan holda ko‘pgina tillarda so‘zlarning turlanish va tuslanish jarayonlarida lingvistik ta’minotga kiritilgan morfologik lug‘atlarga murojaat qilinadi [Abjalova 2020, 38-39; 155].

O‘zbek tilida 50 000 dan ortiq leksemalar mavjud bo‘lib, korpus va lingvistik kompyuter dasturlari bazasida har bir leksemaning turkumi aniqlab berilishi muhim masala hisoblanadi. Shunday so‘zlar mavjudki, turkumlik belgisini o‘zida namoyon qilmaydi yoxud gap tarkibida kontekstual ma’nosi o‘quvchini chalg‘itadi. Masalan, “... test sinovlaridan o‘tkazish yuzasidan shaxsan javobgarligi belgilab qo‘yilsin” (https://lex.uz/acts/-4276890), “Shaxsan o‘zim keldim”, “Shaxsan bajardim”, “Bular hammasi lotincha yoki lotinchaga yaqin so‘zlar. Men, shaxsan, shunday deb bilaman” (A.Qahhor. Adabiyot muallimi) jumlalaridagi *shaxsan* yasama so‘zining turkumini aniqlash mushkul. Ba’zi o‘rinlarda o‘zlik olmoshi o‘rnida qo‘llanilayotgan leksema (olmosh), ba’zi hollarda ravish leksema tarzida namoyon etadi. Bu holatda so‘zning turkumini aniqlashda turkumlarning kategorial xususiyatlariga murojaat etiladi. Ular 4 ta: semantik, sintaktik, morfologik va so‘z yasalishi xususiyatlaridir.

Ma’lumki, o‘zbek tilida 12 so‘z turkumi (mustaqil so‘z turkumlari: ot, fe’l, sifat, ravish, son, olmosh; yordamchi so‘z turkumlari: bog‘lovchi, ko‘makchi, yuklama; alohida olingan so‘zlar turkumi:

modal, taqlid, undov)ga so‘z yasovchi qo‘shimchalarning qo‘shilishi natijasida 4 so‘z turkumi: ot, fe’l, sifat, ravish yasaladi. Aniqlangan yasovchi qo‘shimchalar (337 ta: ot yasovchi qo‘shimcha 114 ta, fe’l yasovchi 58 ta, sifat yasovchi 117 ta, ravish yasovchi qo‘shimcha 48 ta) [Abjalova 2020,122-123] sirasida **-an** ravish yasovchi affiks hisoblanadi. Ushbu parametrdan kelib chiqib xulosalash mumkinki, ot turkumiga mansub “shaxs” so‘ziga **-an** yasovchi qo‘shimchasi birikishi natijasida yasama ravish hosil qilingan: shaxs (Ot) ~ {-an} => shaxsan.

Umuman olganda, so‘z o‘ta murakkab hodisa sifatida talqin qilinib, ayni vaqtda ham til birligi, ham nutq birligi bo‘lishi ta’kidlanadi. Til birligi bilan nutq birligining teng kelib qolishi holati, asosan, o‘zgarmaydigan turkumlarga kuzatiladi. Bizga ma’lum bo‘lgan polifunksional, ko‘p ma’noli va omonim so‘zlar tabiati esa jiddiy tadqiqot va amaliy kuzatuv jarayonlarini talab etadi.



3-chizma.

“Imkonsizlikda ishingni pesh qilgandan ko‘ra kitob o‘qiganing pesh” gapida (3-chizma) so‘zlarning turkumlarga teglanishi natijasida gap bo‘laklarining aniqlanishi va “pesh” omonim so‘zining ketma-ket kelgan so‘zlar (pesh qilmoq (Ot + F); o‘qiganing pesh (F + Sif) va joylashuv o‘rniga ko‘ra (gap so‘nggida ot-kesim => Sifat) turkumi aniqlanishini ko‘rish mumkin.

So‘z turkumi har qanday so‘z birikmasi (3-chizmada “pesh qilgandan ko‘ra” – vositali to‘ldiruvchi uch turkumdagi so‘zlar birikuvidan iborat) va gap strukturasida turli sintaktik funksiyalarni bajarishi mumkin. Har bir so‘z turkumida asosiy, birlamchi sintaktik funksiya mavjud bo‘ladi. Birlamchi sintaktik funksiya so‘z turkumining leksik ma’nosidan kelib chiqadi va bu ma’noning o‘ziga xos transpozitsiyasi sifatida gavdalanadi [Ganiyeva 2019, 277].

ST teglash uchun lingvistik bazada so‘zlar va ularning turkumlari ko‘rsatilgan ro‘yxatning kiritilishi kifoya emas. Yuqoridagi so‘z turkumini aniqlash holatidagi kabi izchillikning yo‘qolishi yoxud bir shaklga ega polifunksional, omonim yoki ko‘p manoli so‘zlarning gapda ifodalagan turkumini topish hatto mutaxassis tilshunosni ham fikr yuritishga, izlanishga undaydi. Shuningdek, o‘zbek tilidagi ko‘pgina so‘zlar muayyan turkumga mansubligi aniqlanmagan. Har bir tabiiy tilda mavjud bunday muammolar e’tiborga olinib STni teglashda bir necha usullarga tayaniladi.

So‘z turkumlarini teglash usullari

Aksariyat hollarda so‘z turkumlarini teglashda quyidagi usul (metod, algoritm)larga asoslaniladi:

• qoidalarga asoslangan usul.

• stoxastik (yoxud statistik) usul.

Qoidalarga asoslangan PoS teglar.

Eng azaliy teglash usullaridan biri bu qoidalarga asoslangan POS-teglash sanaladi. Bunda, asosan, Brill usuli qo‘l keladi [Bril 1992]. Qoidalarga asoslangan teggerlar har bir so‘zni teglash uchun lug‘at yoxud leksikadan foydalanadilar. Agar so‘zda (polifunksional, omonim, ko‘p ma’noli so‘zlar nazarda tutilmoqda) bir nechta teglar bo‘lsa, unda qoidalarga asoslangan teggerlar gapdagi so‘zning turkumlik tegini to‘g‘ri aniqlash uchun qo‘lda yozilgan qoidalardan foydalanadi. Yanada aniq teglarni berishda so‘zning lingvistik xususiyatlarini undan oldingi va keyingi so‘zlarni tahlil qilish orqali qoidalarga asoslanib belgilash orqali ham amalga oshirilishi mumkin. Masalan, qaratqich kelishigidagi ismga mansub so‘zdan so‘ng kelgan lingvistik birlik egalik qo‘shimchasini olgan ot turkumidagi so‘z hisoblanadi. Masalan, mening kitobim, akamning uyi, Salimaning ko‘ylagi kabi. Demak, bunday holda so‘zning ot turkumligi o‘zining oldida kelayotgan qaratqich kelishigidagi ism orqali belgilanadi. Ingliz tilidan misolga e’tibor qaratamiz: oldingi so‘z artikl bo‘lsa, u holda undan keyin kelayotgan so‘z ot turkumiga oid leksik birlik sanaladi. Masalan, *an egg, a book, the train, the windows* kabi.

PoS teglaridagi bunday holatlar qoidalar shaklida kodlanadi.

Ushbu qoidalar quyidagilarni tashkil etishi mumkin:

1. Lingvistik me’yorlarga asoslangan qoidalar. Tilning orfografik qoidalariga asoslangan yuzlab qoidalar umumiy, xususiy va istisnoli qoidalar bazasi tarzida shakllantiriladi [Abjalova 2020].

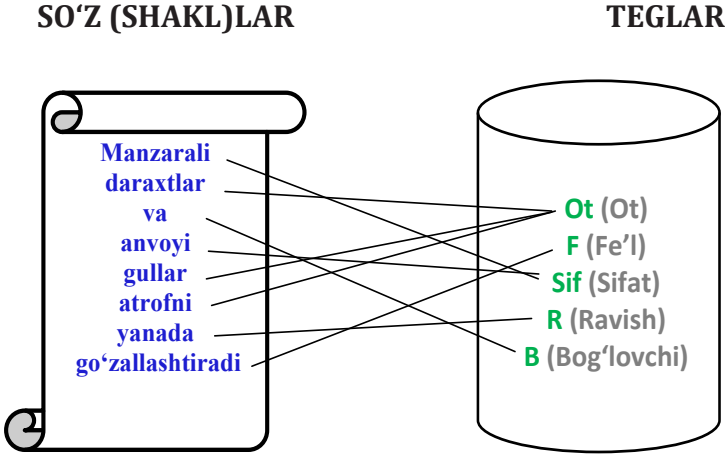
2. Kontekstli shablon qoidalari, ya’ni gap tarkibidagi ko‘chma ma’noga ega so‘zning muntazam ravishda konnotativ ma’noda

qo‘llanishi dastur xotirasida saqlanadi, natijada keyingi jarayonlarda o‘sha konnotativ so‘z bilan bog‘liq noaniqliklar bartaraf etiladi.

Qoidalarga asoslangan metodga binoan so‘z turkumlarini teglash ikki bosqichda amalga oshiriladi:

Birinchi bosqichda PoS-tegger (izohli, imlo, morfologik yoxud orfografik) lug‘at(lar)ga tayanadi. U lug‘at yordamida har bir so‘zning turkum(lar)ini aniqlab chiqadi.

Ikkinchi bosqichda polifunksional yoki omonim so‘zlarning turkumlari qo‘lda yoziladi va bunday so‘zlarni jumla tarkibidagi vazifasini aniqlash qoidalarining katta ro‘yxati ishlab chiqiladi.



4-chizma. Gapdagi har bir so‘zga so‘z turkumining belgilanishi.

Qoidalarni avtomatik generatsiya qilish metodining yaqqol namunasi – amerikalik lingvist Erik Brill metodidir [Brill 1995]. Ish uslubi quyidagicha:

1. Boshlash: Har bir so‘z o‘sha so‘zning tez-tez qo‘llaniladigan tegi bilan bog‘lanishi lozim. Noma’lum so‘zlar ot so‘z turkumi sifatida qabul qilinadi. Mazkur bosqichdan nafaqat o‘rganish jarayoni boshlanadi, balki omonimlarni bartaraf etish metodi ham harakatga keladi.

2. Tez-tez uchraydigan xatolik uchun o‘zgarish (qayta ishlash) qoidasini yaratish kerak.

3. Istalgan minimal xatoga erishguncha ikkinchi bosqichni takrorlash lozim.

O‘tkazish (qayta ishlash) qoidalarida “eski teg, yangi teg, shart” to‘plamlari o‘z ifodasini topadi va qoidada belgilangan shart bajarilganda eski teg yangisi bilan almashtiriladi. Ushbu metodning kamchiligi qoidalar sonining ko‘payishi bilan aniqlik daraja-

sining pasayishida kuzatiladi [Brill 1992], bu Paretoning prinsipiga mos keladi: 80% harakat 20% natijani ta'minlaydi. Shu bilan birga, prinsip aksincha jarayonda ham ishlaydi: boshlash bosqichining faqat bitta qadamini bajarish natijasida omonimlikni bartaraf etishda yuqori aniqlikka erishish mumkin. SinTagRus korpusida o'tkazilgan test natijalaridan ma'lum bo'ldiki, ushbu metod har bir so'zning gapdagi o'rnini 97,4%, morfologik xususiyatlarning to'liq majmuini 87,6% aniqlik bilan topish imkonini beradi [Abjalova 2020,73].

Qoidalarga asoslangan ST teglari xususiyatlari

Qoidalarga asoslangan PoS teglari quyidagi xususiyatlarga ega:

– Ushbu tegerlar bilimga asoslanadi.

– Qoidalar qo'lda yaratiladi.

– Axborotlar qoidalar shaklida kodlanadi.

– Qoidalar cheklangan bo'ladi. Raqamli texnologiya uchun cheksizlik mavhumlikni ifodalaydi, "va shu kabilar", "..." (ko'p nuqta), "va hokazo" kabi birikma va belgilar ro'yxatning davomiyligini emas, balki noaniqligini bildiradi. Shu bois kompyuter lingvistikasida bu kabi noaniqliklarga yo'l qo'yilmaydi, balki har bir lingvistik birlikning tabiati, xususiyatlari dasturiy ta'minot bazasida aniq ko'rsatilishi talab etiladi.

–Tilni modellashtirishda teglashtirish qoidasiga asoslaniladi.

Stoxastik teglash usuli.

Mazkur usul chastota yoki ehtimollik (statistika)ka asoslanadi. Shu bois ayrim manbalarda statistik yoxud ehtimolikka asoslangan usul tarzida tushuntiriladi. Ayon bo'lganidek, oddiy stoxastik teglashda ST teglari uchun quyidagi metodlardan foydalaniladi:

1. Chastotali yondashuv.

Ushbu yondashuvda stoxastik tegerlar so'zning matnda ma'lum bir teg bilan uchrashi ehtimoli asosida grammatik noaniqliklarni bartaraf etadi. Shuni ham aytish mumkinki, o'rganilayotgan to'plam (matn qismi)da muayyan so'z bilan tez-tez qo'llaniladigan teg o'sha so'zning noaniqligiga ma'lumot berishga yordamchi tegdir. Masalan, qo'llanilish darajasiga binoan tushum kelishigidagi so'zdan so'ng kelgan lingvistik birlik fe'l turkumiga mansub hisoblanadi: kitobni o'qimoq, ismni yozmoq, mamlakatni aylanib chiqmoq => {Wni ~ V} kabi. Teglash jarayonidagi bunday yondashuvning asosiy muammosi, u tabiiy tilda birikuvchanligi bo'lmagan teglar ketma-ketligini keltirib chiqarishi mumkin. Masalan, "Do'stimga kitobni sovg'a qilish uchun sotib oldim" misolida sovg'a so'zi o'ng tomondan

“qilish uchun” birikmasi bilan birikuvchanlikka ega, ammo chapdagi so‘z bilan (kitobni) na grammatik, na semantik jihatdan birika oladi. Bunday hollarda teglashtirishda birikuvchanlikka ega bo‘lmagan so‘zlar nomutanosibligiga asoslanish natijasida morfoanaliz jarayonida noto‘g‘ri ma’lumot yuzaga keladi.

2. Teglarning ketma-ketligi ehtimoli yoxud n-gramma usuli.

Stoxastik usulning mazkur yondashuvi tegger berilgan teglar ketma-ketligining qo‘llanilish ehtimolini hisoblaydi. Ketma-ketlik o‘lchovi, ya’ni n (bigram – ikki element ketma-ketligi, trigram – uch ketma-ket teg, 4 gram – to‘rt teg ketma-ketligi) teglarga asoslangani uchun bu yondashuv N-gramma usuli ham deyiladi. N-gramma – matnlarga avtomatik ishlov berishda keng qo‘llaniladigan matematik hisob vositasidir. O‘zbek kompyuter lingvistikasida S.Rizayev harf birikmalarini *bigramm, trigramm* terminlari bilan ifodalagan [Rizayev 2006,18].

Yashirin Markov modeli stoxastik usulda faol qo‘llaniladi. 1960-yillarda L.E. Baum va uning hamkasblari tomonidan ishlab chiqilgan [Baum, Sell 1968, 211-227] mazkur usul statistik jarayonda yuzaga keladigan barcha variantlar ehtimolligini hisobga olishga yordam beradi. Masalan, ma’lum bir matnda ot turkumiga oid so‘zlar bog‘lovchiga nisbatan tez-tez va ko‘p uchrasa unda ayni kontekstda mavjud omonim katta ehtimollik bilan bog‘lovchi emas, ot turkumiga oid so‘z bo‘ladi, keyingi ehtimollikda bog‘lovchi sifatida hisobga olinadi. Kontekstni tavsiflash uchun N-grammadan foydalaniladi. N-gramma – so‘zlar yoki teglar kabi N-identifikator elementlarning ketma-ketligini ifodalaydi.

Yashirin Markov modellari termodinamika, statistik mexanika, fizika, kimyo, iqtisodiyot, moliya, signallarni qayta ishlash, axborot nazariyasi, nutqni qayta ishlash, husnixat, imo-ishoralarni tanib olish, [Starner, Pentland 1995] so‘z turkumlarini teglash va bioinformatikada keng qo‘llaniladigan statistik model hisoblanadi [Li 2003; Ernst 2012].

So‘z turkumlarini stoxastik teglash usuli xususiyatlari

Stoxastik PoS-tegerlar quyidagi xususiyatlarga ega:

– Mazkur teglashtirish teglarning ketma-ket qo‘llanish darajasi bo‘yicha ehtimolligiga asoslanadi.

– Ta’lim korpusi talab qilinadi.

– Korpusda mavjud bo‘lmagan so‘zlar uchun hech qanday ehtimollik bo‘lmaydi.

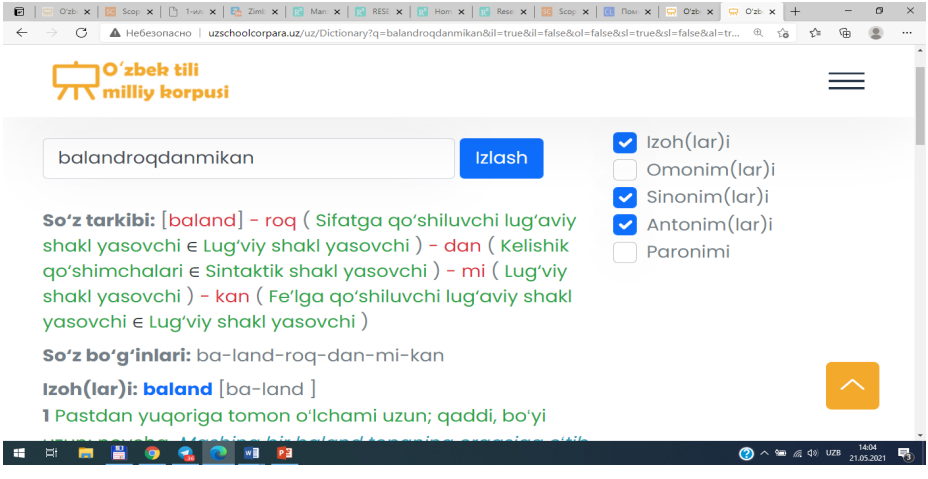
– Ta’lim korpusidan tashqari boshqa tur til korpusidan ham

foydalanish mumkin.

– Eng oddiy ST teglash usuli, chunki bu usulda til korpusidagi faol tarzda ketma-ket qo‘llangan teglarni tanlab oladi.

Amaliy natija

Bir necha yillar davomida olib borilgan tadqiqotlar va 2020-2021-yillardagi amaliy sa’y-harakatlar natijasida Toshkent davlat o‘zbek tili va adabiyoti universiteti Axborot texnologiyalari hamda Amaliy tilshunoslik va lingvodidaktika kafedralari hamkorligida AM-FZ-201908172 raqamli “O‘zbek tili ta’limiy korpusini yaratish” mavzusidagi amaliy loyiha doirasida O‘ZBEK TILI MILLIY va TA’LIMIY KORPUSLARI yaratildi. Bugungi kunda mazkur korpusda mosfoanalizator (avtomatik morfologik tahlil), sinonimizator (qidiruvga yozilgan so‘zga uning ma’nodoshlarini taqdim etish dasturi), shuningdek, so‘zni bo‘g‘inlarga ajratish, izoh (lar)ini taqdim etish va antonimlarini ko‘rsatish imkoniyatlari yaratilgan.

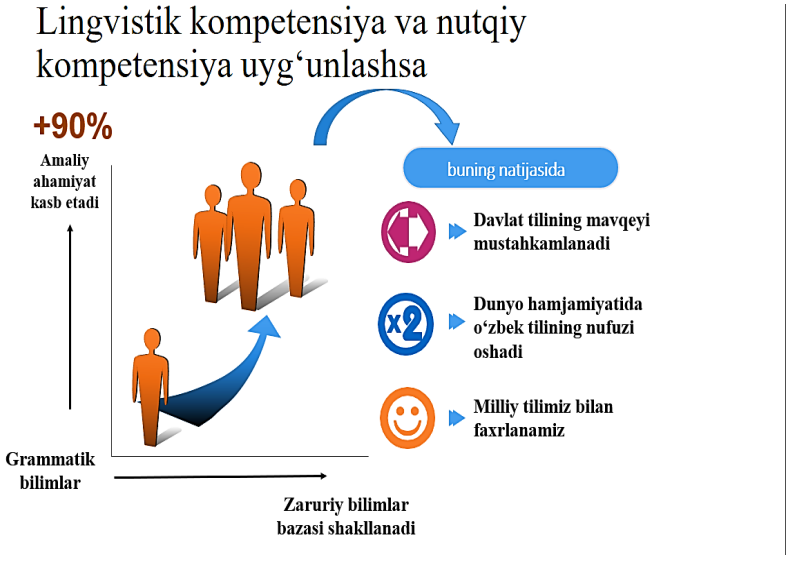


5-rasm. O‘zbek tili Milliy korpusida “balandroqdanmikan” so‘z shaklining morfologik tahlil (“So‘z tarkibi” qismida) natijasi.

Xulosa tarzida aytish joizki, tabiiy tilni qayta ishlash (NLP)-da so‘z turkumlarini teglash algoritmlarni yaratish avtomatik tahlil, morfoanaliz va tarjimon dasturlari uchun birlamchi shart amalga oshirilgan sanaladi, natijada matnlarning grammatik jihatdan sifatli tahlil qilinishiga erishiladi. Bugungi kunda neyrotarmoqqa asoslangan sun’iy intellekt tizimida qoidalarga asoslangan va stoxastik usullardan gibrid tarzda foydalanilmoqda.

Lingvistik bilimlar kompyuter dasturlari va til korpuslari bazasi uchun eng zarur ma’lumot manbai hisoblanadi. Lingvistik

prosessorda lisoniy ma’lumotlar bazasini shakllantirishda, kundalik hayotda til bilimlaridan foydalanish ko‘nikmalariga asoslanilsa, bunday vaziyatlar adabiy til me’yorlariga aylantirilsa, kompyuter dasturlarining tahlil imkoniyati mutaxassis darajasida mukammallikka yetadi. Shu bois lingvistik kompetensiya va nutqiy kompetensiya doimiy ravishda bir-birini taqozo etadi. Aynan so‘z (asosan, polifunksional, ko‘p ma’noli, omonim so‘zlar)ning turkumini belgilashda ham maqola avvalida “uzoq” so‘zi misolidagi kabi so‘zning pragmatik va kontekstual ma’nolari hamda qo‘shimcha semalariga asoslanilsa, raqamli texnologiya dasturlari va tizimlarining amaliy ahamiyati yanada oshadi (6-rasm).



6-rasm.

Umuman olganda, bugungi kunda lingvistik bilimlar tildan amalda foydalanish imkoniyatiga asoslanib mukammallashtirilsa, nazariy manbalarga tayanilib, kompyuter dasturlari va tizimlari uchun formal o‘zbek tili yaratilsa, kelajakda barcha turdagi (o‘zbek tarjimon dasturi, nutqni tanish, o‘zbek tili morfoanalizatori, kompyuter muloqoti kabi) dastur hamda elektron tizimlarining yuzaga kelishiga zamin yaratiladi.

Adabiyotlar

Abjalova, M. 2020. *Tahrir va tahlil dasturlarining lingvistik modullari.* Toshkent: Nodirabegim.

Abjalova, M., Yuldashev, A. 2021. “Methods for Determining Homonyms in Linguistic Systems”. *ACADEMICIA: An International Multidisciplinary*

Research Journal 11 (2): 700-715. DOI: 10.5958/2249-7137.2021.00522.X.

Abjalova, M. 2021. "O'zbek tili Milliy korpusida so'zshakllarni leksikografik baza asosida qidiruv imkoniyatlari". *Kompyuter lingvistikasi: muammo, yechim, istiqbollar*, 12-17. Toshkent: ToshDO'TAU.

Abjalova, M. 2021. "O'zbek tili milliy korpusida sinonimayzer yoxud sinonimizatorni yaratish masalasi". *O'zbek Milliy va ta'limiy korpuslarining yaratishning nazariy hamda amaliy masalalari*, 38-40. Toshkent: ToshDO'TAU.

O'zbek tilining izohli lug'ati. 2006. 80 000 dan ortiq so'z va so'z birikmasi. A. Madvaliyev tahriri ostida. 5 jildli. Toshkent.

Brill, E. 1995. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging". *Computational Linguistics* 21: 543-565. http://acl.ldc.upenn.edu/J/J95/J95-4004.pdf.

Brill, E. 1992. "A simple rule-based part of speech tagger". *Proceedings of ANLC* 154.

Baum, L. E.; Sell, G. R. 1968. "Growth transformations for functions on manifolds". *Pacific Journal of Mathematics* 27 (2): 211-227.

Ganiyeva, Dildora. 2019. "Мазмуний синкретизм ва полифункционаллик". *NamDU ilmiy axborotnomasi* 6: 275-278.

Rizayev, S. 2006. *O'zbek tilshunosligida lingvostatistika asoslari.* Toshkent: Fan.

Starner, Thad, Pentland, Alex. 1995. *Real-Time American Sign Language Visual Recognition From Video Using Hidden Markov Models.* Master's Thesis, MIT, Program in Media Arts.

Li, N; Stephens, M. 2003. "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". *Genetics* 165 (4): 2213-2233. doi:10.1093/genetics/165.4.2213.

Ernst, Jason; Kellis, Manolis. 2012. "ChromHMM: automating chromatin-state discovery and characterization". *Nature Methods* 9 (3): 215–216. doi:10.1038/nmeth.1906. PMC 3577932. PMID 22373907.

Qurbonova, M., Abjalova, M. va boshq. 2021. *O'zbek tili o'zlashma so'zlarining urg'uli lug'ati.* O'quv-uslubiy lug'at. Toshkent: Nodirabegim.

https://en.wikipedia.org/wiki/Hidden_Markov_model.

https://www.freecodecamp.org/news/an-introduction-to-part-of-speech-tagging-and-the-hidden-markov-model-953d45338f24.

https://uzjournals.edu.uz/namdu/vol1/iss6/46.

https://coderlessons.com/tutorials/akademicheskii/obrabotka-estestvennogo-iazyka/pometka-chasti-rechi-pos.

https://habr.com/ru/post/125988.

https://ru.wikipedia.org/wiki/Частеречная_разметка.

https://en.wikipedia.org/wiki/Part-of-speech_tagging#:~:text=In%20corpus%20linguistics%2C%20part%2Dof,its%20definition%20and%20its%20context.

https://ru.wikipedia.org/wiki/Часть_речи.

The Task of Tagging Part of Speech in Natural Language Processing

Manzura Abjalova[1]
Dilrabo Elova[2]

Abstract

Natural language processing is a complex process that requires consideration of all phenomena, general and particular aspects, exceptions, phonetic, morphological, lexical, grammatical, semantic and even orthoepic features of a social language. At the same time, since the creation of an artificial intelligence system is an important result, it is important to achieve text processing units using digital technologies. In this process, it is necessary to determine the parts of speech of words. In computational linguistics, it is known that this linguistic and technical analysis is a stage of marking words – PoS-tagger, that is, tagging a part of speech of words, and is also considered the main stage of automatic text processing. Initial attempts to create tags in creating a language corpus solved many of the most pressing text problems today. In natural language processing (NLP), word labeling is also a primary concern, which leads to the analysis of various linguistic ambiguities in the text, such as the definition of homonymy and semantics of ambiguous words.

This article discusses about part-of-speech tagging words, its significance in the process of computer analysis of texts, and the methods of tagging.

Key words: *natural language processing, tagging, tag, part-of-speech, formal language, pragmatic feature, corpus, polysemy, homonymy, PoS-tagging.*

References

Abjalova, M. 2020. *Tahrir va tahlil dasturlarining lingvistik modullari.* Toshkent: Nodirabegim.

Abjalova, M., Yuldashev, A. 2021. “Methods for Determining Homonyms in Linguistic Systems”.*ACADEMICIA: An International Multidisciplinary*

[1]*Manzura A. Abjalova* – Doctor of Philosophy in Philology (PhD), Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: abjalovamanzura@navoiy-uni.uz
ORCID ID: 0000-0002-1927-2669
[2]*Dilrabo Q. Elova* – teacher, Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: dilrabo@navoiy-uni.uz
ORCID ID: 0000-0002-1927-2968
For citation: Abjalova, A.A., Elova, D.Q. 2021. “The Task of Tagging Part of Speech in Natural Language Processing”. *Uzbekistan: Language and Culture* 1: 6–20.

Research Journal 11 (2): 700-715. DOI: 10.5958/2249-7137.2021.00522.X.

Abjalova, M. 2021. "O'zbek tili Milliy korpusida so'zshakllarni leksikografik baza asosida qidiruv imkoniyatlari". *Kompyuter lingvistikasi: muammo, yechim, istiqbollari*, 12-17. Toshkent: ToshDO'TAU.

Abjalova, M. 2021. "O'zbek tili milliy korpusida sinonimayzer yoxud sinonimizatorni yaratish masalasi". *O'zbek Milliy va ta'limiy korpuslarining yaratishning nazariy hamda amaliy masalalari*, 38-40. Toshkent: ToshDO'TAU.

O'zbek tilining izohli lug'ati. 2006. 80 000 dan ortiq so'z va so'z birikmasi. A. Madvaliev tahriri ostida. 5 jildli. Toshkent.

Brill, E. 1995. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging". *Computational Linguistics* 21: 543-565.

Brill, E. 1992. "A simple rule-based part of speech tagger". *Proceedings of ANLC* 154.

Baum, L. E.; Sell, G. R. 1968. "Growth transformations for functions on manifolds". *Pacific Journal of Mathematics* 27 (2) 211-227.

Ganiyeva, Dildora. 2019. "Mazmunij sinkretizm va polifunkcionallik". *NamDU ilmiy axborotnomasi* 6: 275-278.

Rizayev, S. 2006. *O'zbek tilshunosligida lingvostatistika asoslari.* Toshkent: Fan.

Starner, Thad, Pentland, Alex. 1995. *Real-Time American Sign Language Visual Recognition From Video Using Hidden Markov Models.* Master's Thesis, MIT, Program in Media Arts.

Li, N; Stephens, M. 2003. "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". *Genetics* 165 (4): 2213-2233. doi:10.1093/genetics/165.4.2213.

Ernst, Jason; Kellis, Manolis. 2012. "ChromHMM: automating chromatin-state discovery and characterization". *Nature Methods* 9 (3): 215–216. doi:10.1038/nmeth.1906. PMC 3577932. PMID 22373907.

Qurbonova, M., Abjalova, M. va boshq. 2021. *O'zbek tili o'zlashma so'zlarining urg'uli lug'ati.* O'quv-uslubiy lug'at. Toshkent: Nodirabegim.

https://en.wikipedia.org/wiki/Hidden_Markov_model.

https://www.freecodecamp.org/news/an-introduction-to-part-of-speech-tagging-and-the-hidden-markov-model-953d45338f24.

https://uzjournals.edu.uz/namdu/vol1/iss6/46.

https://coderlessons.com/tutorials/akademicheskii/obrabotka-estestvennogo-iazyka/pometka-chasti-rechi-pos.

https://habr.com/ru/post/125988.

https://ru.wikipedia.org/wiki/Chasterechnaja_razmetka.

https://en.wikipedia.org/wiki/Part-of-speech_tagging#:~:text=In%20corpus%20linguistics%2C%20part%2Dof,its%20definition%20and%20its%20context.

https://ru.wikipedia.org/wiki/Chast'_rechi.