

ISSN 2181-922X

LANGUAGE & CULTURE

UZBEKISTON O'ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER
LINGVISTIKASI

2024 Vol. 3 (6)

www.compling.tsuull.uz

UZBEKISTAN

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 3 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti

Bosh muharrir:

Botir Elov

Bosh muharrir o'rinbosari:

Shahlo Hamroyeva

Mas'ul kotib:

Oqila Abdullayeva

Tahrir kengashi

Shuhrat Sirojiddinov (O'zbekiston), Eshref Adali (Turkiya), [Viktor Zaxarov] (Rossiya), Vladimir Benko (Slovakiya), Ayrat Gatiatullin (Tataristan), Rinat Gilmullin (Tataristan), Murat O'rxun (Turkiya), Suyun Karimov (O'zbekiston), Abduvali Qarshiyev (O'zbekiston), Muxammadjon Musayev (O'zbekiston), Kamoliddin Shukurov (O'zbekiston), O'tkir Hamdamov (O'zbekiston), Tal'at Zuparov (O'zbekiston), Bahodir Mo'minov (O'zbekiston), Faxriddin Nurullayev (O'zbekiston), Zulxumor Xolmanova (O'zbekiston), Muqaddas Abdurahmonova (O'zbekiston), Elova Dilrabo (O'zbekiston), Ruhillo Alayev (O'zbekiston), Rasuljon Atamuratov (O'zbekiston), Malika Abdullayeva (O'zbekiston), Mannon Ochilov (O'zbekiston), Xolisa Axmedova (O'zbekiston), Zilola Xusainova (O'zbekiston), Uldona Abdurahmonova (O'zbekiston).

Jurnal haqida ma'lumot

“O'zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro'yxatidagi “O'zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o'qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o'zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to'rt marta chop etiladi.

O'zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e'lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O'zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti. O'zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko'chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz

Website: kompling.tsuull.uz

Alisher Navo'i Tashkent State University of the Uzbek Language and Literature

Chief editor: Botir Elov
Deputy editor-in-chief: Shahlo Hamroyeva
Responsible secretary: Oqila Abdullayeva

Editorial board

Shukhrat Sirojiddinov (Uzbekiston), Eshref Adali (Turkiye), [Viktor Zakharov] (Russia), Vladimir Benko (Slovakia), Ayrat Gatiatullin (Tataristan), Rinat Gilmullin (Tataristan), Murat Orhun (Turkey), Suyun Karimov (Uzbekistan), Abduvali Karshiyev (Uzbekistan), Mukhammadjon Musayev (Uzbekistan), Kamoliddin Shukurov (Uzbekistan), O'tkir Hamdamov (Uzbekistan), Tal'at Zuparov (Uzbekistan), Bahadir Mo'minov (Uzbekistan), Fakhridin Nurullayev (Uzbekistan), Zulkhumor Kholmanova (Uzbekistan), Muqaddas Abdurakhmonova (Uzbekistan), Elova Dilrabo (Uzbekistan), Ruhillo Alayev (Uzbekistan), Rasuljon Atamuratov (Uzbekistan), Malika Abdullayeva (Uzbekistan), Mannon Ochilov (Uzbekistan), Kholisa Akhmedova (Uzbekistan), Zilola Khusainova (Uzbekistan), Uldona Abdurakhmonova (Uzbekistan).

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz

Website: kompling.tsuull.uz

MUNDARIJA

Talat Zuparov

Word2vec metodi orqali matnlarni raqamlashtirish va mashinali o'qitish usullari orqali qayta ishlash.....6

Umidjon Yodgorov

O'zbek tili frazemalarining morfologik shakli va variantlari tadqiqiga an'anaviy va korpus tahlili yondashuvi.....29

Go'zal Erkinjonova

Fe'l so'z turkumiga oid birliklarning leksik-grammatik xususiyati.....49

Botir Elov, Maftuna Baratova

Pos (part of speech) teglash usullari.....62

Zebo Qodirova

Tibbiy atamalarni tur va sinflarga ajratishda ontologik tamoyillar.....78

Gulira'no Nuriddinova

Tabiiy tilni qayta ishlashda eganing modellari.....90

Botir Elov, Maftuna Ahmedova

N-gramlar asosida imloni tuzatish tizimini ishlab chiqish.....101

Mavluda Urazaliyeva

Audiomatnlarni korpusga kiritish muammolari tahlili.....115

CONTENT

Talat Zuparov

Digitization of texts using the word2vec method and processing through machine learning techniques.....27

Umidjon Yodgorov

The traditional and corpus-based approach to the study of morphological forms and variations of uzbek phraseology.....46

Go'zal Erkinjonova

Lexical and grammatical characteristics of verbal word class units.....60

Botir Elov, Maftuna Baratova

Pos (part of speech) tagging methods.....76

Zebo Kodirova

Ontological principles in the division of medical terms into types and classes.....88

Gulira'no Nuriddinova

Models of subject in natural language processing.....100

Botir Elov, Maftuna Ahmedova

Development of a spell correction system based on n-grams.....113

Mavluda Urazaliyeva

Analysis of problems in incorporating audio texts into a corpus.....124

POS (PART OF SPEECH) TEGLASH USULLARI

Botir Elov¹

Maftuna Baratova²

Annotatsiya. Hozirgi zamon lingvistikasi va kompyuter texnologiyalarining rivojlanishi matnni qayta ishlashning avtomatlashtirilgan tizimlarini ishlab chiqishni talab qilmoqda. POS (Part of Speech) teglash matnni tahlil qilish va grammatik jihatdan tasniflash uchun muhim ahamiyatga ega. POS teglash so'zlarning morfologik va sintaktik toifalarini aniqlab, matnni grammatik jihatdan tahlil qilishga yordam beradi.

Kalit so'zlar: *Teg, POS teglash, Part of Speech, NLP, Natural Language Processing, SpaCy, TreeTagger.*

Kirish

Jahon tilshunosligida POS teglashni o'rganishga qaratilgan ko'plab tadqiqotlar olib borilgan. Bunda til morfologiyasi, sintaksisi va tabiiy tilni qayta ishlash (NLP) bo'yicha yetakchi olimlarning ishlari muhim ahamiyatga ega bo'lib, tadqiqotlarni ikki qismga ajratish mumkin:

Nazariy ishlar:

- **Tilshunoslik nazariyasi.** So'z turkumining tuzilishi va uning boshqa so'z turkumlari bilan bog'liqligi tilshunoslik nazariyalarida ko'rib chiqilgan. Masalan, Chomskiy tomonidan ishlab chiqilgan generativ grammatika va Jakendofning kognitiv grammatika nazariyalari gap qismlarining sintaktik va semantik tuzilishini o'rganishda asosiy rol o'ynaydi [Chomsky, 1957].

- **Universal grammatika.** Ba'zi olimlar barcha tillarda o'xshash grammatik qoidalar mavjudligi fikrini ilgari surishgan. Bu nazariyaga asoslanib, so'z turkumlari har bir tabiiy til uchun umumiy

¹Elov Botir Boltayevich – texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti.

E-pochta: elov@navoiy-uni.uz

ORCID: 0000-0001-5032-6648

²Baratova Maftuna Baxtiyor qizi – Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti Kompyuter lingvistikasi fakulteti 1-kurs magistranti.

E-pochta: baratovamaftuna09@gmail.com

andozaga ega bo'lishi mumkin, degan fikr mavjud.

- **Markov zanjirlari va statistik usullar.** Tilni avtomatik tahlil qilishda Markov zanjirlari va statistik usullar so'z turkumlarini aniqlash va tasniflashda ishlatilgan. Bu usul odatda aniq natijalarga erishishga imkon beradi [Gagniuc, 2017].

- **Turkiy tillarning morfologik va sintaktik xususiyatlari.** Turkiy tillar agglutinativ tillar hisoblanadi, ya'ni ularning so'z tuzilishida qo'shimchalar muhim rol o'ynaydi. Ba'zi so'zshakllar turkumi mos o'zak turkumidan farq qilishi mumkin. Masalan, ot so'z turkumiga mansub **"o'yin"** o'zagidan fe'l so'z turkumiga mansub **"o'yna" (o'yin+a)** leksemasi hosil bo'ladi.

- **POS teglashning umumiy standartlari.** So'nggi yillarda tabiiy tillar uchun umumiy POS teglash standartlarini ishlab chiqishga qaratilgan tadqiqotlar olib borilgan. Bu yo'nalish, xususan, Birlashgan Millatlar Tashkilotining Universal Dependencies [De Marneffe va boshqalar, 2021] loyihasi doirasida amalga oshirilgan.

Amaliy ishlar:

- **Mashinali o'qitish texnologiyalari.** POS teglash jarayonida mashinali o'qitish (Machine Learning, ML) algoritmlaridan keng foydalaniladi. Jumladan, MaxEnt (Maksimal Entropiya) [Ratnaparkhi, 1996] va HMM (Yashirin Markov modeli [Ekbal va boshqalar, 2007]) algoritmlari eng ko'p qo'llanilgan usullardan hisoblanadi.

- **Til korpuslari.** Til korpuslariga asoslangan POS teglash ishlari amaliy jihatdan katta ahamiyatga ega. Korpusga asoslangan tadqiqotlarda til birliklarining POS teglarini avtomatik aniqlash uchun keng foydalaniladi. Brown, Penn Treebank kabi katta hajmdagi POS teglangan til korpuslari turli NLP vazifalarini hal qilishga xizmat qiladi. Hozirda o'zbek, turk, qozoq va boshqa turkiy tillarda ko'plab annotatsiyalangan korpuslar yaratilgan bo'lib, bu korpuslar POS teglash uchun asosiy resurs hisoblanadi. Masalan, Turk tilining TNC (Turkish National Corpus) yoki o'zbek tilidagi ToshDO'TAUda ishlab chiqilgan o'zbek tili ta'limiy korpusi bu borada amaliy yordam beradi.

- **Neyron tarmoqlar va hozirgi algoritmlar.** Bugungi kunda POS teglashning zamonaviy tadqiqotlari konvolyutsion va rekurrent neyron tarmoqlar (CNN va RNN) hamda Transformer modellaridan foydalangan holda olib borilmoqda. Bu modellar odatda til birliklarini aniqlashda yuqori aniqlikka ega va ko'plab tillar uchun qo'llaniladi [Ahmed, Singh, Bhattacharyya, 2023]. Ushbu usullar va yondashuvlar hozirda ko'plab NLP tizimlarini ishlab chiqishda qo'llanilib, matnni avtomatik tahlil qilishda yuqori darajadagi samaradorlikka erishgan [Chiche, Yitagesu, 2022].

- **Avtomatlashtirilgan POS teggerlar.** Tabiiy tillar uchun avtomatik POS teggerlar yaratish jarayonlari rivojlanmoqda. Misol uchun, Stanford NLP, spaCy va boshqa mashhur NLP paketlarida tabiiy tillar uchun maxsus modellar va teggerlar mavjud bo'lib, ular asosida turli tillardagi matnlarda POS teglash jarayoni amalga oshiriladi.

Asosiy qism

Annotatsiya - POS teglashning mohiyati va vazifasi

Texnologik taraqqiyot bilan birgalikda tilni avtomatik tahlil qilish texnologiyalari rivojlanib bormoqda. Bunda tabiiy tilni qayta ishlash (Natural Language Processing, NLP) tizimlarining muhim qismi hisoblangan POS (Part of Speech) teglash tamoyillari dolzarb hisoblanadi.

Teg – Kompyuter yordamida matn tahlilini amalga oshirish jarayonini tezlashtirish va osonlashtirishga xizmat qiluvchi shartli belgi yoki maxsus kod. Teglar bir necha turlarga bo'linadi: semantik teg, sintaktik teg va grammatik teg. Grammatik teg, shuningdek, PoS (Part of speech) tegging nomi bilan ham mashhur [Elov, Xudoyberganov, 2024].

POS teglash – so'zlarni matn ichida ularning morfologik va sintaktik xususiyatlariga qarab tasniflash jarayonidir. Ushbu jarayon matnni avtomatik tahlil qilishda va tilni qayta ishlashda asosiy bosqichlardan biri hisoblanadi [Jurafskiy, Martin, 2020].

Masalan: *Xayriyatki[[MD]] qurilish[[VB]] bo'yicha[[II]] muovinin[[N]] senga[[P]] o'xshagan[[VB]] ammamning buzog'i[[IB]] emas[[VB]] teglariga ega bo'ladi.*

POS teglash tabiiy tilni qayta ishlashning asosiy qismi bo'lib, mashinalarni tarjima qilish, ovozli buyruqlarni tanish va qidiruv tizimlarini takomillashtirish uchun zarurdir.

POS teglash tizimlarining turlari

1. Qo'lda teglash: Lingvistlar tomonidan amalga oshiriladigan an'anaviy usul.

2. Avtomatik teglash: Kompyuter dasturlari yordamida amalga oshiriladi. Avtomatik teglashda statistik usullar va mashinali o'rganish modellaridan foydalaniladi.

3. Gibrid usullar: Qoidaviy va statistik usullarni birlashtiradi [Bird va boshqalar, 2009].

Teglashning qo'llanilish sohasi

• **Tilni avtomatik tarjima qilish:** Google Translate kabi vositalar.

- **Chatbotlar:** Mijozlarga xizmat ko'rsatishda.
- **Til o'rganish dasturlari:** Grammatikani o'rganishda yordam.

- **Qidiruv tizimlari:** Teglash yordamida qidiruvni aniqlashtirish [Manning, Schutze, 1999].

- **POS Teglash algoritmlari va texnologiyalari**

- **Qoidaviy (Rule-Based) yondashuvlar.**

Qoidaviy yondashuvlar tilda oldindan tuzilgan grammatik qoidalarga asoslanadi. Masalan, o'zbek tilida asosan gaplar ega (ot[[N]]) bilan boshlanib, kesim (fe'l[[VB]]) bilan tugallanadi. Ba'zi inversiyaga uchragan gaplar bundan mustasno.

- **Statistik yondashuvlar**

Statistik metodlar (masalan, HMM – Hidden Markov Model) katta hajmdagi matn korpuslaridan o'rganilgan ehtimollik ma'lumotlariga asoslanadi. Ushbu yondashuv 90% dan yuqori aniqlikka erishishga yordam beradi.

- **Mashinali o'rganish asosidagi algoritmlar**

POS teglashda mashinali o'rganish modellaridan foydalanish zamonaviy texnologiya hisoblanadi. Jumladan, support vector machines (SVM) yoki neyron tarmoqlar asosidagi modellar qo'llaniladi [Mikolov, 2013].

- **Dasturiy vositalar**

- **NLTK:** Ingliz tili uchun keng qo'llaniladi.

NLTK (Natural Language Toolkit) NLTK Python asosidagi ochiq manba kutubxona bo'lib, tabiiy tilni qayta ishlashda keng qo'llaniladi. U matnni tahlil qilish, tokenlash, POS (Part of Speech) teglarni aniqlash, sintaktik tahlil va semantik tahlil qilish kabi ko'plab funksiyalarni amalga oshiradi. Bu vosita ilmiy izlanishlar va o'quv loyihalar uchun juda mos keladi. NLTK yangi foydalanuvchilar uchun foydali bo'lgan ko'plab o'rnatilgan ma'lumotlar to'plamlarini o'z ichiga oladi, masalan, Brown Corpus va WordNet kabi.

- **SpaCy:** Ko'p tilli POS teglash uchun samarali vosita.

SpaCy ko'p tilli tabiiy tilni qayta ishlash uchun zamonaviy va samarali vositadir. U tezkor ishlash uchun optimallashtirilgan va korporativ loyihalar uchun mos. SpaCy yordamida tokenlash, POS teglarni aniqlash, obyektlarni tanish (NER), matn tahlili va hatto mashina o'rganish modellarini yaratish mumkin. U o'zining boy integratsiyasi va ishlash tezligi bilan taniqli.

- **Treetagger:** Ko'p tilli matn tahlili uchun

TreeTagger ko'p tilli matnlarni qayta ishlash uchun mo'ljallangan kuchli vosita bo'lib, uni asosan POS teglarni aniqlash

va ma'lumotlarni tahlil qilishda foydalaniladi. U nafaqat ingliz tilida, balki ko'plab boshqa tillarda ham ishlaydi. TreeTagger tilshunoslik tadqiqotlari va ko'p tilli tahlillar uchun foydali hisoblanadi. Bu dastur ochiq manba sifatida taqdim etiladi va ko'plab tillar uchun model kutubxonalarini o'z ichiga oladi.

Lingvistik teglash tavsiflovchi yoki analitik belgilarni til ma'lumotlari bilan bog'lashni o'z ichiga oladi. Strukturlanmagan (qayta ishlanmagan) ma'lumotlar matnli yoki har qanday manba yoki janrdan olingan yoki vaqt funksiyalari (audio, video va/yoki fiziologik yozuvlar) shaklida bo'lishi mumkin. Teglar barcha turdagi transkripsiyalarni (fonetik xususiyatdan nutqiy qurilmagacha), POS teglash, ma'no belgilari, sintaktik tahlil, NER obyektlar, semantik rol belgilari, vaqt va hodisalarni aniqlash, so'zlarning sintaktik zanjirlarini, nutq darajasini o'z ichiga olishi mumkin [Elov, Xudoyberganov, 2024].

Nutq qismini belgilash orqali kayfiyat yoki his-tuyg'ularni aniqlash.

Nutq qismini belgilashning yana bir imkoniyati — matndagi kayfiyat yoki his-tuyg'ularni aniqlashdir. Kayfiyat tahlili — bu muallifning muayyan mavzu yoki shaxsga nisbatan umumiy munosabati yoki fikrini (ijobiy, salbiy yoki neytral) aniqlash jarayoni. His-tuyg'u tahlili esa, muallifning quvonch, qayg'u, g'azab yoki qo'rquv kabi hissiy holatlarini aniqlash jarayonidir. Nutq qismini belgilash orqali matndagi kayfiyat yoki his-tuyg'ularni aniqlashning bir necha usuli mavjud. Bir usul — ma'lum nutq qismlarini kayfiyat yoki his-tuyg'u ko'rsatkichlari sifatida ishlatishdir. Masalan, sifat yoki ravish sifatida belgilangan so'zlar ko'pincha hissiy mazmunga ega bo'ladi va matndagi kayfiyat yoki his-tuyg'ularni aniqlash uchun ishlatilishi mumkin. Misol uchun, "baxtli", "xafa", "jahldor" yoki "qo'rqinchli" kabi so'zlar ko'pincha his-tuyg'ularni ifodalaydi va ularning mavjudligi matn muallifining kayfiyatini yoki his-tuyg'ularini anglatishi mumkin. Yana bir usul — matndagi kayfiyat yoki his-tuyg'ularni aniqlash uchun kontekst va sintaktik tuzilishdan foydalanishdir. Masalan, "emas" yoki "hech qachon" kabi inkor so'zlar bilan birga ishlatilgan so'zlar matnning kayfiyatini o'zgartirishi mumkin. Masalan, "Men baxtli emasan" iborasi salbiy kayfiyatni bildiradi, "Men baxtliman" iborasi esa ijobiy kayfiyatni bildiradi. Shuningdek, "juda" yoki "nihoyatda" kabi kuchaytirgichlar kayfiyat yoki his-tuyg'u intensivligini oshirishi, "biroz" kabi kuchsizlantiruvchilar esa kamaytirishi mumkin. Aniq nutq qismlaridan va sintaktik tuzilishdan foydalanish bilan bir qatorda, matnning umumiy kayfiyat yoki his-

tuyg'usini ham aniqlash mumkin. Masalan, asosan ijobiy kayfiyatda yozilgan matn ijobiy his-tuyg'uni ifodalashi mumkin, asosan salbiy kayfiyatda yozilgan matn esa salbiy his-tuyg'uni ifodalashi mumkin. Bunga matndagi ijobiy va salbiy so'zlarning umumiy taqsimotini tahlil qilish orqali erishish mumkin. Nutq qismini belgilash orqali matndagi kayfiyat yoki his-tuyg'ularni aniqlash imkoniyatlariga qaramay, bu usulning cheklovlari mavjudligini ta'kidlash lozim. Bir cheklov shundaki, kayfiyat va his-tuyg'ular ko'pincha subyektiv bo'lib, odamdan odamga farq qilishi mumkin. Bu esa, xuddi shu matnni turli odamlar turlicha talqin qilishi mumkinligini anglatadi, shuning uchun matndagi kayfiyat yoki his-tuyg'ularni nutq qismini belgilash orqali aniq aniqlash qiyin bo'lishi mumkin. Yana bir cheklov shundaki, kayfiyat va his-tuyg'ular faqat so'zlar orqali emas, balki boshqa belgilar orqali ham ifodalanadi. Ovozning ohangi, yuz ifodalari va tana harakatlari ham kayfiyat va his-tuyg'ularni ifodalashi mumkin, ammo bular nutq qismini belgilash orqali aniqlanmaydi. Bu esa nutq qismini belgilash usuli, ayniqsa, gapirilgan matnlarda matndagi kayfiyat yoki his-tuyg'ularni aniq aniqlay olmasligini anglatadi. Xulosa qilib aytganda, nutq qismini belgilash matnda ifodalangan kayfiyat yoki his-tuyg'ularni aniqlashda qo'llanilishi mumkin bo'lsa-da, bu usul mukammal emas va uning cheklovlari mavjud. Nutq qismini belgilash orqali kayfiyat yoki his-tuyg'ularni tahlil qilishda bu cheklovlarni hisobga olish, shuningdek, nutq qismini belgilash natijalarini to'ldirish uchun boshqa usullar, masalan, kontekstual tahlil yoki og'zaki bo'lmagan belgilarni qo'llash zarur. Turli tillar va yozish uslublariga qarab nutq qismini belgilash tizimlarining samaradorligi qanday o'zgaradi?

Nutq qismini belgilash tizimlari samaradorligiga ta'sir qiluvchi omillar

1. Tilning grammatika va sintaksis murakkabligi: Maqsad tilining grammatikasi va sintaksisining murakkabligi POS belgilash tizimlari samaradorligiga ta'sir qiladi. Masalan, nemis yoki rus tillari kabi murakkab infleksiya tizimiga ega tillarda, so'zlarning ko'plab infleksion shakllari mavjudligi sababli, to'g'ri belgilash qiyin bo'lishi mumkin. Aksincha, ingliz yoki ispan tillari kabi soddaroq infleksiya tizimiga ega tillarni belgilash nisbatan osonroqdir.

2. Tilning standartlashganligi va izchillik darajasi: Ingliz tili kabi barqaror va yaxshi aniqlangan grammatik qoidalarga ega tillarni belgilash nisbatan osonroqdir. Boshqa tomondan, xitoy tili kabi ko'proq kontekstga bog'liq grammatikaga ega tillarda so'zlarni to'g'ri belgilash qiyinroq bo'lishi mumkin.

3. Morfologik murakkablik darajasi: Ba'zi tillar, masalan, arab yoki ibroniy tillari, keng ko'lamli tuslanish shakllariga ega bo'lib, bu tillarda fe'l va otlarni to'g'ri belgilash qiyin bo'lishi mumkin. Aksincha, fransuz yoki italyan kabi soddaroq morfologiyaga ega tillar nisbatan osonroq belgilanishi mumkin.

4. Matnning yozish uslubi: Rasmiy yoki ilmiy uslubda yozilgan matnlarni belgilash, odatda, ko'proq standart va bashoratli til ishlatilishi sababli osonroq. Aksincha, og'zaki yoki norasmiy uslubdagi matnlar sleng, jargon va standart bo'lmagan so'zlardan foydalaniganligi uchun qiyinroq belgilanishi mumkin.

5. Standart bo'lmagan yoki noaniq so'zlar mavjudligi: Ko'p ma'noli yoki bir nechta nutq qismiga ega so'zlarni to'g'ri belgilash qiyinroq bo'lishi mumkin. Bu, ayniqsa, ingliz tili kabi katta lug'at boyligiga ega va so'z yasash imkoniyati yuqori bo'lgan tillarda qiyinchilik tug'diradi.

6. Tayyorlangan ma'lumotlar sifati va hajmi: Katta va xilma-xil belgilangan ma'lumotlar to'plamlarida o'rgatilgan tizimlar yangi matnlar va tillarga yaxshi umumlashishi mumkin. Ammo noto'g'ri belgilangan yoki maqsad til yoki uslubni yetarli darajada aks ettirmaydigan ma'lumotlardan o'rgatilgan tizimlar samarasiz ishlashi mumkin. Xulosa qilib aytganda, nutq qismini belgilash tizimlarining samaradorligi turli tillar va yozish uslublariga qarab sezilarli darajada o'zgaradi. Grammatika va sintaksisning murakkabligi, standartlashganlik darajasi, morfologik murakkablik, yozish uslubi, shuningdek, standart bo'lmagan yoki noaniq so'zlarning mavjudligi bu tizimlarning aniqligiga ta'sir ko'rsatishi mumkin. Bundan tashqari, belgilangan ma'lumotlar sifat va miqdori tizimlarning umumlashish qobiliyatiga ta'sir qilishi mumkin. Nutq qismini belgilash yordamida qidiruv tizimini optimallashtirish (SEO)ni qanday yaxshilash mumkin?

Nutq qismini belgilash (POS tagging) — bu matndagi so'zlarga tegishli nutq qismini aniqlash va belgilash jarayoni. Bu, odatda, matnning sintaksisi va grammatikasini tahlil qiluvchi tabiiy tilni qayta ishlash algoritmlari yordamida amalga oshiriladi va har bir so'zning gapdagi rolini aniqlaydi. Nutq qismini belgilashdan qidiruv tizimini optimallashtirish (SEO) maqsadlarida foydalanishning bir usuli matndagi eng muhim kalit so'zlarni aniqlashdir. SEOda kalit so'zlar — bu odamlar onlayn ma'lumot qidirishda ishlatadigan aniq termin va iboralar. Ushbu kalit so'zlarni mazmuningizda belgilab, ularni nishonga olish orqali veb-saytingizning qidiruv tizimlari natijalarida ko'rinishini va reytingini oshirishingiz mumkin. SEO maqsadlarida

nutq qismini belgilashdan foydalanish uchun, avval veb-saytingiz yoki blog post matnini tahlil qilib, eng muhim kalit soʻzlarni aniqlashingiz kerak. Buni qoʻlda yoki mazmuningiz mavzusiga asoslangan tegishli kalit soʻzlarni topishda yordam beruvchi vosita orqali amalga oshirish mumkin. Maqsad kalit soʻzlaringizni aniqlagandan soʻng, nutq qismini belgilash yordamida bu soʻzlarning qanday nutq qismi ekanligini aniqlashingiz va ular matnda qanday ishlatilganini koʻrishingiz mumkin. Masalan, agar kalit soʻz ot boʻlsa, u SEO maqsadlari uchun feʼl yoki sifatga qaraganda koʻproq ahamiyatga ega boʻlishi mumkin. Buning sababi, odatda, otlar koʻproq maʼno anglatadi va foydalanuvchilar tomonidan qidiruv soʻzi sifatida ishlatilishi ehtimoli yuqori. Bundan tashqari, nutq qismini belgilash orqali har bir kalit soʻzning matndagi grammatik rolini aniqlashingiz mumkin. Masalan, agar kalit soʻz gapning egasi sifatida ishlatilgan boʻlsa, u toʻldiruvchi yoki aniqlovchi sifatida ishlatilgandan koʻra koʻproq ahamiyatga ega boʻlishi mumkin. Bu gapning egasi odatda gapning asosiy diqqat markazi boʻlib, qidiruv soʻzi sifatida ishlatilishi ehtimoli yuqori. SEO maqsadlarida muhim kalit soʻzlarni aniqlash uchun nutq qismini belgilashdan foydalanishning yana bir usuli kalit soʻzning matnda uchrash chastotasi va kontekstini tahlil qilishdir. Matnda koʻproq yoki muhim joylarda uchraydigan kalit soʻzlar SEO uchun kamroq yoki kamroq muhim joylarda ishlatiladiganlarga qaraganda koʻproq ahamiyatli boʻlishi mumkin. Shuningdek, nutq qismini belgilash yordamida kalit soʻzlaringiz ishlatiladigan sintaktik soʻzlarni aniqlash mumkin. Masalan, agar kalit soʻz maxsus ibora yoki gap tarkibida ishlatilgan boʻlsa, u boshqa kontekstda ishlatilgandan koʻra SEO uchun koʻproq ahamiyatli boʻlishi mumkin. Umuman olganda, nutq qismini belgilash SEO maqsadlarida matndagi eng muhim kalit soʻzlarni aniqlash uchun foydali vosita boʻlishi mumkin. Kalit soʻzlarning nutq qismlari, grammatik rollari, chastotasi va kontekstini tahlil qilish orqali, ularning mazmuningizda qanday ishlatilganligini yaxshiroq tushunishingiz va ularni qidiruv tizimlari uchun optimallashtirishingiz mumkin. Bu SEOning faqat bir jihati boʻlsa-da, nutq qismini belgilash qidiruv tizimlarida yuqori oʻrin egallash ehtimoli yuqori boʻlgan yanada samarali va nishonga olingan kontent yaratishga yordam beradi. Kalit soʻzning nutq qismi SEO uchun uning samaradorligiga qanday taʼsir qiladi? Kalit soʻzlar qidiruv tizimini optimallashtirish (SEO)ning muhim tarkibiy qismidir, chunki ular qidiruv tizimlariga veb-sahifa mazmunini tushunishga va uni mos ravishda reytinglashga yordam beradi. Kalit soʻzning nutq qismi SEO uchun uning samaradorligiga sezilarli taʼsir koʻrsatadi,

chunki bu uning kontentda qanday ishlatilishini va qidiruv tizimlari tomonidan qanday talqin qilinishini aniqlaydi. Avvalo, ingliz tilidagi turli nutq qismlarini tushunish muhim. Bularga otlar, fe'llar, sifatlar, ravishlar, olmoshlar va predloglar kiradi. Ushbu nutq qismlarining har biri gapda o'ziga xos vazifani bajaradi va kontentning ma'nosi va tuzilishiga ta'sir ko'rsatadi. Masalan, otlar odamlar, joylar, narsalar yoki tushunchalarni nomlash uchun ishlatiladi, fe'llar esa harakatlarni, holatlarni yoki hodisalarni ifodalaydi.

Sifatlar otlar yoki olmoshlarni tavsiflaydi, ravishlar esa fe'llarni, sifatlarini yoki boshqa ravishlarni aniqlaydi. Olmoshlar ot yoki boshqa olmoshlar o'rniga ishlatiladi, predloglar esa ot va olmoshlarni gapning boshqa so'zlariga bog'laydi. SEO nuqtayi nazaridan, kalit so'zning nutq qismi uning samaradorligiga sezilarli ta'sir ko'rsatishi mumkin. Masalan, agar kalit so'z ot bo'lsa, u kontentning asosiy e'tibori sifatida ishlatilganda samaraliroq bo'lishi mumkin. Buning sababi shundaki, qidiruv tizimlari aniq mavzu yoki predmeti atrofida to'plangan kontentga ko'proq e'tibor beradi va kalit so'z sifatida ot ishlatilishi kontent mavzusini aniqlashga yordam beradi. Boshqa tomondan, agar kalit so'z fe'l yoki sifat bo'lsa, u kontentning asosiy e'tiborini tavsiflash yoki aniqlash uchun ishlatilganda samaraliroq bo'lishi mumkin. Bu nutq qismlari kontentga qo'shimcha kontekst va tafsilotlarni qo'shishga yordam beradi, bu esa qidiruv tizimlariga kontentni aniqroq tushunishga yordam beradi. Bundan tashqari, kalit so'zlarni turli shakllarda (masalan, birlik yoki ko'plik, hozirgi yoki o'tgan zamon) ishlatish ham SEO uchun ularning samaradorligiga ta'sir qilishi mumkin. Masalan, agar kontent kalit so'zning bir nechta hodisalari haqida bo'lsa, uni ko'plik shaklida ishlatish samaraliroq bo'lishi mumkin; agar u hozirgi hodisalar yoki tendensiyalar haqida bo'lsa, hozirgi zamonda ishlatilishi samaraliroq bo'lishi mumkin. Shuni ham ta'kidlash kerakki, kalit so'zning nutq qismi SEO uchun uning samaradorligiga kontentning tuzilishi va tashkiloti bilan bog'liq holda ta'sir ko'rsatishi mumkin. Masalan, kalit so'zning gapning egasi sifatida ishlatilishi unga ko'proq urg'u berishi va qidiruv tizimlariga ko'proq sezilarli bo'lishiga yordam beradi. Xuddi shunday, kalit so'z gapning kesimi sifatida ishlatilganda unga ko'proq kontekst qo'shilishi va kontent uchun muhimroq bo'lishi mumkin. Umuman olganda, kalit so'zning nutq qismi uning SEO uchun samaradorligida muhim rol o'ynaydi. Nutq qismlarini va ularning kontentning ma'nosi va tuzilishiga qanday ta'sir ko'rsatishini tushunish orqali kalit so'zlardan qanday foydalanishni optimallashtirishingiz va kontentingizning qidiruv

tizimi natijalarida yuqori o'rin egallash ehtimolini oshirishingiz mumkin. Kalit so'zni tanlashda va kontentda qanday ishlatishda uning nutq qismi va samaradorligini hisobga olish SEO maqsadlarida eng yaxshi natijalarga erishish uchun muhim.

POS teglashda so'zlarni teglash uchun standartlar va belgilash tizimlari lingvistika va tabiiy tilni qayta ishlash (NLP) doirasida maxsus ishlab chiqilgan bo'lib, ushbu standartlar turli til va ilovalar uchun moslashtiriladi. POS teglashda turli xil standartlar bor. Shuningdek teglash jarayonida taggestlar sonlar, harflar va sonlar va harflar bilan belgilanishi mumkin.

Tatar tilidagi nutq qismi teglari Apertiumning morfologik teggeri bilan izohlangan tatar aralash korpusida qo'llaniladi. **Tatar Mixed corpus** tegsetlar tizimida 29 ta teg mavjud.

1-chizma

ag	Part of speech	Example
abbr	Abbreviation	Аббревиатура
adj	Adjective	Прилагательное
adv	Adverb	Наречие
cm	comma	,
cnjadv	Adverbial conjunction	Наречие-союз
cnjcoo	Coordinating conjunction	Сочинительный союз
cnjsub	Subordinating conjunction	Подчинительный союз
cop	Copula	Копула
det	Determiner	Детерминатив
ideo	Ideophone	Звукоподражательное слово
ij	Interjection	Междометие
n	Noun	Существительное
np	Proper noun	Имя собственное
num	Numeral	Числительное
post	Postposition	Послелог

postadv	Postadverb	Посленаречие
prn	Pronoun	Местоимение
sent	sentence marker	. ? !
v	Verb	Глагол
vaux	Auxiliary verb	Вспомогательный глагол
apos	apostrophe	'
guio	hyphen	-
lpar	left Parenthetical marker	(
lquot	left Quote marker	“, «
mod_ass	Assertive modal particle	бит
mod_ind	Indefinite modal particle (expresses doubt)	дыр
qst	Modal question particle	микән
rpar	right Parenthetical marker)
rquot	right Quote marker	”, »

Turkish part-of-speech tagset

2-chizma

ag	Part of speech	example
Noun	noun	zaman
Punc	punctuation	, .
Adj	adjective	olan, büyük
Verb	verb	ise, vardır
Conj	conjunction	ve, da
Adv	adverb	olarak, en
Num	number	1

Postp	postposition	için, ile
Det	determiner	bu, her
Pron	pronoun	kendi, onun
Ques	question	mi, mu
Interj	interjection	ey, be
Dup	dup	pırlı, tefek

Tajik part-of-speech tagset tizimi 16 toifaga ko'ra 01 dan 16 gacha raqamlarni belgilagan.

3-chizma

ag	Description	Example (lemma+tag)
01	nouns	сол:01
02	adjectives	нав:02
03	numerals	як:03
04	pronouns	он:04
05	verbs	аст:05
06	infinitives	кардан:06
07	adjectival participles	карда:07
08	adverbial participles	ситез:08
09	adverbs	пас:09
10	prepositions	ба:10
11	postpositions	қатй:11
12	conjunctions	аммо:12
13	particles	низ:13
14	interjections	а:14
15	onomatopoeia	шарр:15
16	numeratives	нафар:01, нафар:16

Xulosa

Dunyo tilshunosligida til korpusi matnlari ustida o'tkazilgan turli tajribalar ko'p va ular shuni ko'rsatadiki, o'zak ma'lumotlarini sintaktik vazifa bilan birlashtirish morfologik jihatdan boy til uchun POS teglash natijasini yaxshilaydi, bu esa NLP vazifasining hal qilish samaradorligini oshirishga xizmat qiladi. Turkiy tillar korpuslaridagi ba'zi teglar tizimi jadvali keltirildi. Bunda tillarda tegsetlar nomlanishi va sonlari farqli bo'lib, ularning tashkil qilinishi tillarning tuzilishiga bog'liq. Tillarda yaratilayotgan korpuslar ko'p mehnat va vaqt talab qiladi va aynan mana shu korpuslarning ahamiyati va qiymatini tilda ishlab chiqilgan tegsetlar oshiradi.

Foydalanilgan adabiyotlar

- Bird, S., Klein, E., Loper, E. Natural Language Processing with Python. O'Reilly Media, 2009.
- Chomsky N. Syntactic Structures. Mouton. Aspects of the Theory of Syntax. MIT Press, 1965.
- Elov B., Xudoyberganov N. O'zbek tili korpusi matnlarini pos teglash // Xalqaro ilmiy-amaliy konferensiya: "Kompyuter lingvistikasi: muammolar, yechim, istiqbollar", 2024.
- Gagniuc, Paul A. Markov Chains: From Theory to Implementation and Experimentation. USA, NJ: John Wiley & Sons., 2017.
- <https://cyberleninka.ru>
- <https://github.com/Step2AI/Uzbek-tilidagi-Suniy-Intellekt-manbalari>.
- <https://github.com/Step2AI/Uzbek-tilidagi-Suniy-Intellekt-manbalari>.
- <https://ieeexplore.ieee.org/abstract/document/10286619>
- <https://ieeexplore.ieee.org/abstract/document/10286792>
- https://link.springer.com/chapter/10.1007/978-3-031-27199-1_4
- <https://nlp.stanford.edu/>
- <https://translate.google.com/?hl=ru>
- <https://uznatcorpara.uz/>
- <https://www.tnc.org.tr/>
- <https://www.ubuy.uz/uz/product/ETFHTZRC-natural-language-processing-a-quick-introduction-to-nlp-with-python-and-nltk-step-by-step-tutorial-f>
- <https://www.ubuy.uz/uz/product/ETFHTZRC-natural-language-processing-a-quick-introduction-to-nlp-with-python-and-nltk-step-by-step-tutorial-f>

- Jurafsky, D., Martin, J. H. Speech and Language Processing. Pearson, 2020.
- Manning, C., Schutze, H. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
- Mikolov, T. et al. "Efficient Estimation of Word Representations in Vector Space", 2013.
- Treetagger: A Tool for POS Tagging. Institute for Computational Linguistics, 2001.

POS (PART OF SPEECH) TAGGING METHODS

Botir Elov¹

Maftuna Baratova²

Abstract. The development of modern linguistics and computer technologies requires the development of automated text processing systems. POS (Part of Speech) tagging is important for text analysis and grammatical classification. POS tagging helps to analyze text grammatically by identifying morphological and syntactic categories of words.

Keywords: *Tag, POS tagging, Part of Speech, NLP, Natural Language Processing, SpaCy, TreeTagger.*

References

Elov B., Xudoyberganov N. O'zbek tili korpusi matnlarini pos teglash // Xalqaro ilmiy-amaliy konferensiya: "Kompyuter lingvistikasi: muammolar, yechim, istiqbollar", 2024.

Chomsky N. Syntactic Structures. Mouton. (1965). Aspects of the Theory of Syntax. MIT Press.

Gagniuc, Paul A. (2017). Markov Chains: From Theory to Implementation and Experimentation. USA, NJ: John Wiley & Sons

Jurafsky, D. & Martin, J. H. (2020). Speech and Language Processing. Pearson.

Manning, C. & Schutze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

Treetagger: A Tool for POS Tagging. (2001). Institute for Computational Linguistics.

Mikolov, T. et al. (2013). "Efficient Estimation of Word Representations in Vector Space".

<https://cyberleninka.ru>

<https://www.ubuy.uz/uz/product/ETFHTZRC-natural-language-processing-a-quick-introduction-to-nlp-with-python-and-nltk-step-by-step-tutorial-f>

¹*Elov Botir Boltayevich* – doctor of philosophy in technical sciences, docent. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.

E-mail: elov@navoiy-uni.uz

ORCID: 0000-0001-5032-6648

²*Baratova Maftuna Baxtiyor qizi* - Master of degree. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.

E-mail: baratovamaftuna09@gmail.com

<https://github.com/Step2AI/Uzbek-tilidagi-Suniy-Intellekt-manbalari>.

<https://www.tnc.org.tr/>

<https://uznatcorpara.uz/>

<https://nlp.stanford.edu/>

<https://cyberleninka.ru>

<https://translate.google.com/?hl=ru>

<https://www.ubuy.uz/uz/product/ETFHTZRC-natural-language-processing-a-quick-introduction-to-nlp-with-python-and-nltk-step-by-step-tutorial-f>

<https://github.com/Step2AI/Uzbek-tilidagi-Suniy-Intellekt-manbalari>.

<https://ieeexplore.ieee.org/abstract/document/10286792>

https://link.springer.com/chapter/10.1007/978-3-031-27199-1_4

<https://ieeexplore.ieee.org/abstract/document/10286619>

Jurnal 2017-yil 26-oktyabrda O'zbekiston Respublikasi Matbuot va axborot agentligi tomonidan 0936-raqam bilan ro'yxatdan o'tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc) dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko'chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga **.**.*-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. "Cambria" garniturasida.
Shartli b.t. 7,51. Nashr b.t. 7,62.

"O'zbekiston: til va madaniyat" jurnali tahririyatida tayyorlandi va sahifalandi.
"YASHNOBOD NASHR" bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.