O‘ZBEK TILI SO‘ZLARINI QOIDAGA ASOSLANGAN USULDA LEMMALASHNING ALGORITM VA MODELINI ISHLAB CHIQISH

Sharipov Maqsud Siddiqovich
Abu Rayhon Beruniy nomidagi Urganch Davlat Universiteti Texnika fanlari nomzodi, dotsent
E-mail:maqsbek72@gmail.com
Sobirov Og‘abek Ollayor o‘g‘li
Abu Rayhon Beruniy nomidagi Urganch Davlat Universiteti Tayanch doktorant
E-mail:sobirovogabek0409@gmail.com

Annotatsiya. Tabiiy tilni qayta ishlash (ing. Natural Language Processing - NLP) so‘nggi yillarda sezilarli yutuqlarga erishmoqda, inson tilini yaxshiroq tushunish va qayta ishlash uchun turli xil algoritm va modellar ishlab chiqilmoqda. Tabiiy tilni qayta ishlashda dastlabki ishlov berishning asosiy vazifalaridan biri lemmalash bo‘lib, so‘zlarni asosiy yoki lug‘at shakliga qisqartirishni o‘z ichiga oladi. Ushbu maqolada biz o‘zbek tilidagi so‘zlarni qoidaga asoslangan yondashuv yordamida lemmalash uchun maxsus moslashtirilgan algoritm va modelni ishlab chiqishni ko‘rib chiqamiz. Lemmalash, ayniqsa, o‘zbek tili kabi boy morfologik tuzilishga ega bo‘lgan tillar uchun muhim bo‘lib, bu jarayonda so‘zlar turli fleksiya va o‘zgarishlarga uchraydi.

Kalit so‘zlar: Tabiiy tilni qayta ishlash, so‘z, lemma, lemmalash, model, algoritm, o‘zbek tili.

Аннотация. Обработка естественного языка (анг. Natural Language Processing - NLP) значительно продвинулась в последние годы, и были разработаны различные алгоритмы и модели для лучшего понимания и обработки человеческого языка. Одной из основных задач предварительной обработки в обработка естественного языка является лемматизация, которая включает в себя сокращение слов до их базовой или словарной формы. В этой статье мы углубимся в разработку алгоритм и модель, специально предназначенных для лемматизации слов в узбекском языке с использованием подхода, основанного на правилах. Лемматизация особенно важна для языков с богатой морфологической структурой, таких как узбекский, где слова могут подвергаться различным флексиям и модификациям.

Ключевые слова: Обработка естественного языка, слово, лемма, леммизация, модель, алгоритм, узбекский язык.

Annotation. Natural Language Processing (NLP) has seen significant advancements in recent years, with various algorithm and models being developed to better understand and process human language. One fundamental preprocessing

task in NLP is lemmatization, which involves reducing words to their base or dictionary form. In this article, we delve into the development of a model and algorithm specifically tailored for lemmatizing words in the Uzbek language using a rule-based approach. Lemmatization is particularly important for languages with rich morphological structures like Uzbek, where words can undergo various inflections and modifications.

Key words: Natural language processing, word, lemma, lemmaization, model, algorithm, Uzbek language.

Kirish.

Hozirgi zamonaviy axborot texnologiyalari davrida tabiiy tillarni qayta ishlash (Natural Language Processing – NLP) sohasidagi ilmiy tadqiqotlar ahamiyati ortib bormoqda. Ayniqsa, so‘z shakllarini lemmalash (ya’ni so‘zning asl ko‘rinishini aniqlash) matnni tahlil qilish, qidiruv tizimlarini takomillashtirish, mashinaviy tarjima va boshqa NLP vazifalari uchun asosiy bosqich hisoblanadi. O‘zbek tili – bu agglyutinativ tillar guruhiga mansub bo‘lib, so‘z yasalishining murakkab qoidalari va ko‘plab grammatik shakllarga ega bo‘lganligi sababli, uning so‘zlarini lemmalash muammosi alohida e’tiborni talab etadi[1], [2].

Ushbu tadqiqotda [3] zamonaviy turk tilining birinchi to'liq fe'llar leksikoni taqdim etilgan bo‘lib, TDK lug‘atining chuqur tahlili natijasida turk tilida 4700 ta oddiy fe’l mavjudligi aniqlangan va fe’llar turli morfologik xususiyatlarga ko‘ra qo‘shimcha ilovalarda keltirilgan. Ushbu tadqiqot [4] Somali tili uchun matnlarni lemmatizatsiya qilish bo‘yicha birinchi bo‘lib, yozma Somali matnlari uchun qoidaga asoslangan lemmatizator va leksikondan foydalanishni taklif etadi. Ushbu tadqiqot [5] o‘zbek tili uchun lemmatizatsiya algoritmini yaratishga qaratilgan bo‘lib, cheklangan holatlar mashinasi yordamida so‘z affikslarini olib tashlash va so‘zning lug‘atda mavjud lemmalarini aniqlash jarayonini o‘z ichiga oladi.

Maqolada o‘zbek tili so‘zlarini qoidaga asoslangan usulda lemmalashning modeli va algoritmi ishlab chiqish masalasi tadqiq etiladi. Muammoning dolzarbligi shundaki, mavjud lemmalash yechimlari ko‘pincha umumiy tillar (masalan, ingliz yoki rus tillari) uchun ishlab chiqilgan bo‘lib, O‘zbek tilining o‘ziga xos lingvistik xususiyatlarini to‘liq hisobga olmaydi. Bundan tashqari, O‘zbek tilida so‘z yasalishi va so‘z shakllarini tahlil qilishda morfologik qoidalarni aniq modellashtirish algoritmik jihatdan qiyin masalalardan biridir.

Tadqiqotning asosiy maqsadi – o‘zbek tilining morfologik qoidalarini hisobga olgan holda, so‘z shakllarini lug‘aviy asosga (lemmaga) keltirish uchun izchil model va algoritmni taklif etishdir. Bu jarayonda so‘z yasalishi qonuniyatlari, suffikslar, prefikslar va so‘z ildizlarini tahlil qilish usullari ko‘rib chiqiladi.

ISSN 3030-3907

Metodologiya. O‘zbek tilidagi so‘zlarni morfologik tahlil qilish orqali ularning lemmasini (asl shaklini) aniqlash uchun qoidaga asoslangan usul va algoritmni ishlab chiqish uchun quyidagi metodik yondashuvlar qo‘llanilgan.

- **Yuklama qo‘shimchasini aniqlash va olib tashlash** – Agar so‘z tarkibida yuklama *(-mi, -chi* kabi) qo‘shimchasi bo‘lsa, dastlab ular ajratib olinadi.
- **So‘z turkumini aniqlash** – So‘z yordamchi yoki oraliq so‘z bo‘lsa, lemma sifatida qaytariladi.
- **Shaxs-son va zamon qo‘shimchalarini tahlil qilish** – Fe’llardagi shaxs-son *(-man, -san*) va zamon *(-di, -gan*) qo‘shimchalari qirqib tashlanadi.
- **Mayl va ravishdosh qo‘shimchalarini olib tashlash** – Fe’llarning mayl *(-ay, -ey*) va ravishdosh (*-a, -ib*) shakllarini tahlil qilish orqali asosiy fe’l shakliga qisqartiriladi.
- **Otning sintaktik shakllarini qirqish** – Kelishik (*-ni, -ga*) va egalik (*-im, -ing*) qo‘shimchalari ajratilib, otning asosiy shakli olinadi.
- **Ko‘plik shakllarini tahlil qilish** – (*-lar*) qo‘shimchasining semantik ahamiyatiga qarab, u lemma tarkibida saqlanishi yoki olib tashlanishi belgilanadi.

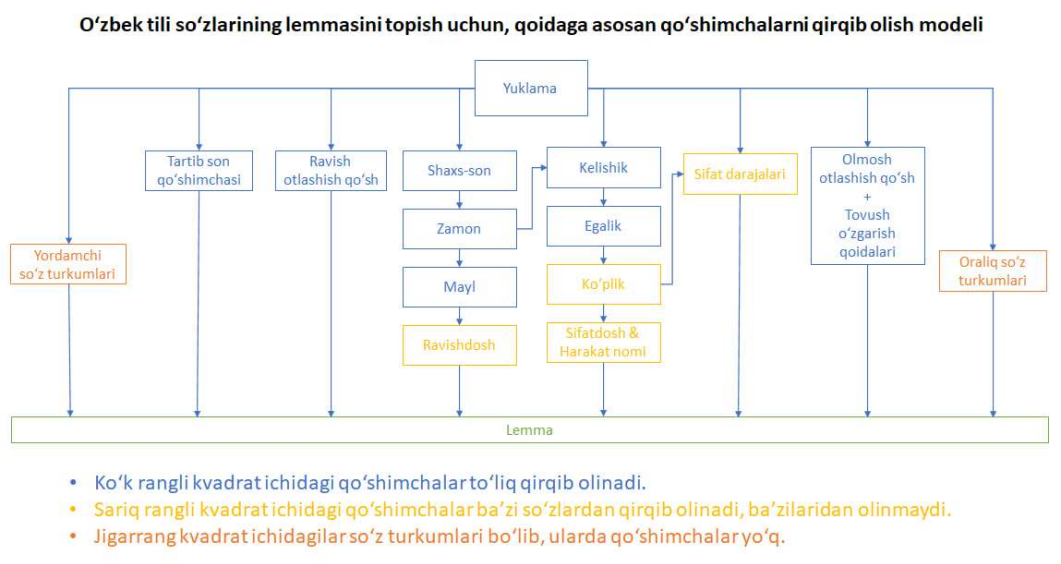
Ushbu metodlar asosida algoritm ishlab chiqilgan, ushbu algoritm [1-rasm] so‘zni bosqichma-bosqich tahlil qiladi. Har bir bosqichda so‘z tarkibidagi morfemalar (asos va qo‘shimchalar) ajratiladi, so‘ngra lemma aniqlanadi. Agar so‘z yordamchi yoki oraliq so‘z turkumlariga tegish bo‘lsa, uning o‘zi lemma sifatida chiqariladi. Algoritm quyidagi ketma-ketlikda ishlaydi:

1. So‘z tarkibida yuklama qo‘shimchasi bo‘lsa qirqib olinadi, aks holda keyingi qadamga o‘tiladi
2. Agar so‘z yordamchi yoki oraliq so‘z turkumlariga tegishli bo‘lsa 7-qadamga o‘tiladi, aks holda keyingi qadamga o‘tiladi
3. Agar so‘z tarkibida shaxs-son yoki zamon qo‘shimchalari bo‘lsa qirqib olinadi, aks holda keying qadamga o‘tiladi
4. Agar mayl qo‘shimchalari bo‘lsa mayl qo‘shimchalari qirqib olinadi keyin ravishdosh qo‘shimchalardan ba’zilari qoida asosida qirqib olinadi, aks holda keying qadamga o‘tiladi.
5. Agar otning sintaktik shakl yasovchi qo‘shimchalari (kelishik, egalik) bo‘lsa qirqib olinadi keyin boshqa so‘z turkumlari (sifat, olmosh, ravish, son) otlashgan holatda qoldirib 7-qadamga o‘tiladi, aks holda keying qadamga o‘tiladi
6. Otning ko‘plik shakli –lar qo‘shimchasi bo‘lsa qoidaga ko‘ra ba’zi holatlarda olib tashlanadi ba’zi holatlarda so‘z tarkibida qoldiriladi keyin harakat nomi va sifatdosh qo‘shimchalari ham shunday qilinadi va 7-qadamga o‘tiladi
7. So‘z lemma sifatida chiqariladi.

ISSN 3030-3907

Taklif etilayotgan metodologiya o‘zbek tilining morfologik murakkabligini hisobga olgan holda, so‘zlarni lemmalashning izchil va qoidaga asoslangan algoritm va modelini taqdim etadi. Algoritmning asosiy ustunligi – u faqatgina statistik usullarga emas, balki tilning chuqur grammatik tuzilishiga asoslanganligidadir.

So‘zlarning lemmasini aniqlashda qo‘shimchalar qirqib olinishi kerak, biroq shunday holatlar borki, ba’zi qo‘shimchalar olib tashlanadi, ba’zilari esa so‘z tarkibida qoldiriladi. So‘z o‘sha qo‘shimcha bilan birgalikda lemma sifatida qaytariladi. O‘zbek tilini morfologik tahlil qilish orqali aniqlangan qoidalar va yuqorida keltirilgan algoritm asosida o‘zbek tili so‘zlarining lemmasini topuvchi model [1-rasm] ishlab chiqildi.



1-rasm. Lemmalash modeli.

Muhokama. Ishlab chiqilgan algoritm va model O‘zbek tilining morfologik jihatdan murakkab tuzilishini hisobga olgan holda, so‘zlarni lemmalash masalasiga tizimli yechim taklif etadi. Algoritm bosqichma-bosqich tahlil usulini qo‘llab, so‘z tarkibidagi har bir qo‘shimchani turi va funksiyasiga ko‘ra ajratadi. Masalan, yuklama qo‘shimchalari (-mi, -chi) birinchi bosqichda, shaxs-son va zamon qo‘shimchalari (-man, -di) keyingi bosqichlarda tahlil qilinadi. Bu jarayon O‘zbek tilidagi so‘z yasalishining izchilligini aks ettirib, lemmalashning aniqligini ta’minlaydi. Model esa morfologik tahlilni vizual va rangli kodlash orqali aniqroq tushunish imkonini beradi. Ko‘k rang bilan belgilangan qo‘shimchalar (masalan, zamon shakllari) to‘liq qirqib tashlanadi, sariq rangdagilar (ko‘plik -lar) esa kontekstga qarab ishlov beriladi.

Modelning lingvistik to‘plamlar va fonetik qoidalar bilan mustahkamlanishi uning amaliy qo‘llanilishini kengaytiradi. Kelajakda ushbu yondashuvni korpuslar asosida kengaytirib, O‘zbek tilini qayta ishlash sohasidagi boshqa vazifalar (masalan, morfologik belgilash, sintaktik tahlil) uchun ham qo‘llash mumkin.

ISSN 3030-3907

Xulosa. Tadqiqot natijalari O‘zbek tilini qayta ishlash sohasidagi ilmiy ishlar uchun asos bo‘lib xizmat qiladi, shuningdek, milliylashtirilgan dasturiy mahsulotlar yaratishga imkon yaratadi.

FOYDALANILGAN ADABIYOTLAR.

[1] M. Sharipov, D. Samandarova, and O. Sobirov, “O‘zbek tili matnlarini morfologik tahlil qilish uchun lingvistik manba yaratish va lemmalash,” in *“KOMPYUTER LINGVISTIKASI: MUAMMOLAR, YECHIM, ISTIQBOLLAR” Xalqaro ilmiy-amaliy konferensiya*, Toshkent: http://compling.navoiy-uni.uz, 2024, pp. 55–59.

[2] B. Elov, S. Khamroeva, and Z. Khusainova, “O‘ZBEK TILIDAGI TURLI TUZILISHLI SO‘ZLARNI LEMMALASH USULLARI,” pp. 13–19, Jan. 2023.

[3] M. Nakipoglu and A. Untak, “A complete verb lexicon of Turkish based on morphemic analysis,” pp. 221–280, Jan. 2008.

[4] S. A. Mohamed and M. A. Mohamed, “Lexicon and Rule-based Word Lemmatization Approach for the Somali Language,” 2023.

[5] M. Sharipov and O. Sobirov, “Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language,” in *CEUR Workshop Proceedings*, V. J. and K. B., Eds., CEUR-WS, 2022, pp. 154 – 159. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146112590&partnerID=40&md5=e1080c39d101c0e351cfed1a8228d391

ISSN 3030-3907