

О‘ЗБЕК ТИЛИ ТЕЗАУРУСИДА ОЗИҚ-ОВҚАТ НОМЛАРИНИ СЕМАНТИК ҚИДИРИШ: NLP TEXNOLOGIYALARI VA ALGORITMLAR

Muyassar YO‘LDOSHALIYEVA

o‘qituvchi

Toshkeny shahar Olmazor tumani

15-umumiy o‘rta ta’lim maktabi

Toshkent, O‘zbekiston

Annotatsiya

Maqolada o‘zbek tili tezaurusida oziq-ovqat nomlarini ifodalovchi birliklarni semantik qidirish masalasi tadqiq etilgan. Oziq-ovqat nomlarining leksik-semantik xususiyatlari, ular o‘rtasidagi giperonimik, giponimik, meronimik va sinonimik munosabatlarni elektron tezaurusda aks ettirish imkoniyatlari yoritilgan. Tadqiqotda tabiiy tilni qayta ishlash (Natural Language Processing – NLP) texnologiyalarining tezaurus resurslarini shakllantirish va semantik qidiruv jarayonlarini avtomatlashtirishdagi ahamiyati asoslangan.

Tayanch so‘zlar: NLP, Vector Space Model, normalizatsiya, lemmatizatsiya, tokenizatsiya, TF-IDF(Term Frequency-Inverse Document Frequency), Word Embeddengs (taqsimlangan semantik modellar).

СЕМАНТИЧЕСКИЙ ПОИСК НАИМЕНОВАНИЙ ПРОДУКТОВ ПИТАНИЯ В ТЕЗАУРУСЕ УЗБЕКСКОГО ЯЗЫКА: ТЕХНОЛОГИИ NLP И АЛГОРИТМЫ

Муяссар ЮЛДАШАЛИЕВА

Преподавательница

Общеобразовательная школа №15 Алмазарского района горада Ташкента

Ташкент, Узбекистан

Аннотация

В статье рассматривается проблема семантического поиска единиц, обозначающих наименования продуктов питания и блюд, в тезаурусе узбекского языка. Освещаются лексико-семантические особенности наименований продуктов питания, а также возможности представления гиперонимических, гипонимических, меронимических и синонимических отношений между ними в электронном тезаурусе. Обосновывается значение технологий обработки естественного языка (Natural Language Processing – NLP) для формирования тезаурусных ресурсов и автоматизации процессов семантического поиска.

Ключевые слова: NLP, векторная модель пространства (Vector Space Model), нормализация, лемматизация, токенизация, TF-IDF (Term Frequency–Inverse Document Frequency), Word Embeddings (распределённые семантические модели).

Bugungi kunda raqamli texnologiyalarning jadal rivojlanishi til birliklarini elektron shaklda qayta ishlash, ularning semantik xususiyatlarini modellashtirish hamda foydalanuvchiga zarur axborotni tezkor va aniq taqdim etish imkoniyatlarini kengaytirmoqda. Ayniqsa, tabiiy tilni qayta ishlash (Natural Language Processing – NLP) texnologiyalarining takomillashuvi elektron lugʻatlar, korpuslar, ontologiyalar va tezauruslarni yaratish hamda ulardan samarali foydalanishda muhim ahamiyat kasb etmoqda. Tezauruslar til birliklari oʻrtasidagi semantik munosabatlarni tizimli ravishda aks ettiruvchi muhim lingvistik resurslardan biri boʻlib, ularning elektron shakli zamonaviy axborot-qidiruv tizimlarining lingvistik asosini takomillashtirishga xizmat qiladi.

Oʻzbek tilida oziq-ovqat nomlarini ifodalovchi birliklar leksik-semantik jihatdan rang-barangligi, milliy-madaniy xususiyatlarni aks ettirishi va oʻzaro turli semantik munosabatlarga kirishishi bilan alohida tadqiqot obyekti tashkil etadi. Ushbu birliklarni tezaurusda tizimlashtirishda ularning giperonimik, giponimik, meronimik, sinonimik va boshqa semantik aloqalarini aniqlash muhimdir. Mazkur munosabatlarning elektron tezaurusda toʻgʻri ifodalanishi foydalanuvchiga nafaqat muayyan leksemani topish, balki u bilan semantik jihatdan bogʻliq boshqa birliklarni ham aniqlash imkonini beradi.

Anʼanaviy kalit soʻz asosidagi qidiruv tizimlarida foydalanuvchi soʻrovi va qidiruv natijasi oʻrtasidagi semantik bogʻliqlik har doim ham yetarli darajada taʼminlanmaydi. Shu sababli tezaurusda semantik qidiruv algoritmlaridan foydalanish muhim ahamiyatga ega. Bunda NLP texnologiyalari asosida lemmatizatsiya, morfologik tahlil, semantik munosabatlarni aniqlash va oʻzaro bogʻlangan birliklarni qidirish jarayonlarini avtomatlashtirish mumkin [1].

Matnda oʻzbek tili tezaurusida oziq-ovqat nomlarini ifodalovchi birliklarni semantik qidirishning lingvistik va algoritmik asoslari, NLP texnologiyalaridan foydalanish imkoniyatlari hamda semantik qidiruvni tashkil etish tamoyillari tahlil qilinadi. Tadqiqotning asosiy maqsadi oziq-ovqat nomlari bilan bogʻliq leksik birliklarni semantik munosabatlar asosida izlash va ularni elektron tezaurus tizimida samarali aks ettirish modelini taklif etishdan iborat.

Oziq-ovqat nomlarini tezaurus shaklida tizimlashtirish jarayoni nafaqat lugʻaviy birliklarni jamlash, balki ular oʻrtasidagi semantik munosabatlarni aniqlash va modellashtirishni talab etadi. Shu sababli, zamonaviy axborot tizimlarida semantik qidiruv algoritmlari va tabiiy tilni qayta ishlash (NLP) texnologiyalaridan foydalanish muhim ahamiyat kasb etadi. Semantik qidiruv algoritmlari va tabiiy tilni qayta ishlash (NLP) texnologiyalari inson tilini chuqurroq tushunishga asoslangan qidiruv tizimlarini yaratadi. Bu tizimlar kalit soʻzlarga asoslangan oddiy qidiruvdan farqli oʻlaroq, foydalanuvchining niyatini va kontekstini aniqlashga qaratilgan boʻlib, NLP yordamida soʻzlar va iboralar orasidagi munosabatlarni tushunadi. Semantik qidiruvda asosiy texnologiyalar sifatida NLP, bilim grafiklari, sunʼiy intellekt (AI) va mashinani oʻrganish (ML) qoʻllaniladi [10]. Ular maʼlumotlarni yanada aniq va kontekstga mos natijalar bilan taʼminlaydi. Masalan, “osh” soʻzi qidirilganda tizim nafaqat aynan shu birlikni, balki uning turlarini, sinonimlari va bogʻliq tushunchalarini ham chiqaradi. “Shoʻrva” qidirilganda unga yaqin boʻlgan “suyuq taomlar” kategoriyasini ham chiqarib beradi. Agar foydalanuvchi “goʻshtli taomlar” deb yozsa, tizim semantik jihatdan barcha taomlarni ajratib beradi. Bu esa tezaurusni oddiy lugʻatdan farqli ravishda bilimlar tizimi sifatida ishlashini taʼminlaydi [15].

NLP texnikalari tokenizatsiya, soʻrovga berilgan soʻzni tanib olish, munosabatlarni ajratish va chuqur oʻrganishga asoslangan matn ifodalash usullarini oʻz ichiga oladi. Akademik jurnallar uchun yaratilgan semantik qidiruv tizimlari soʻz birikmalarining vektorli taqdimoti va ontologiyalar yordamida soʻrovlarni aniqroq moslashtirish imkonini beradi, bu esa ilmiy tadqiqotlarni samarali topishda yordam beradi. Shuningdek, huquqiy maʼlumotlar bazalarida semantik qayta ishlash matnni tahlil qilishda metadata ajratishdan koʻra kengroq imkoniyatlar yaratadi, ammo polisemik soʻzlar va noaniqliklar kabi muammolar mavjud. Semantik qidiruv algoritmlari soʻzlarning sinonimlari, kontekstdagi munosabatlari va boshqa lingvistik xususiyatlarini hisobga oladi, shuning uchun ular murakkab va noaniq soʻrovlarni ham yaxshiroq tushunadi [4]. Zamonaviy yondashuvlarda Word2Vec kabi vektorli taqdimotlar va chuqur oʻrganishga

asoslangan modellar (masalan, entence-BERT) qo'llanilib, katta hajmdagi ma'lumotlardan tez va aniq natijalar olish imkonini beradi. Bu texnologiyalar internet qidiruvi, mahsulot qidiruvi, ilmiy maqolalar izlash kabi ko'plab sohalarda qo'llanilmoqda va foydalanuvchilarga yanada samarali axborot topishda yordam bermoqda.

Tabiiy tilni qayta ishlash (NLP) – bu sun'iy intellekt sohasining bir bo'lagi bo'lib, kompyuter va inson o'rtasidagi vositachilik vazifasini bajarishni zimmasiga olgan texnologiyalar to'plamidir. NLP yordamida kompyuterlar matn yoki nutqni tushinish, tahlil qilish, his-tuyg'ularni aniqlash, so'zlarni tokenlarga ajratish, grammatik belgilash (part-of-speech tagging), nomlangan obyektlarni aniqlash (named entity recognition) kabi vazifalarni bajaradi. So'nggi yillarda chuqur o'rganish (deep learning) va transformer modellarning rivojlanishi NLP tizimlarining samaradorligini sezilarli darajada oshirdi; masalan, BERT va boshqa pre-trained modellar tilni yanada chuqurroq anglash imkonini beradi. NLP texnologiyalari ko'plab sohalarda qo'llaniladi: mashina tarjimasi, nutqni tanib olish, chat-botlar, matn tasnifi va tahlili kabi. Zamonaviy NLP tizimlari klassik leksikon asosidagi usullar bilan bir qatorda neyron tarmoqlar va e'tibor mexanizmlarini ham o'z ichiga oladi, bu esa ko'p tilli va kontekstga mos tushishni yaxshilaydi. Shuningdek, NLP ilm-fan va sanoatda katta hajmdagi matn ma'lumotlarini avtomatik qayta ishlashda muhim rol o'ynaydi va kelajakda yanada rivojlanishiga yo'l ochdi. Tabiiy tilni qayta ishlash texnologiyalari oziq-ovqat nomlarini qayta ishlashda bir nechta bosqichlarni amalga oshiradi. Dastlab matnlar ichidan biz oziq-ovqat nomlariga qidiruv berganimizda uni alohida birliklarga ajratib beradi. Bu jarayon tokenizatsiya bo'lib, u haqida I bobda batafsil aytganmiz. Masalan, “qovurilgan kartoshka” so'zini “qovurilgan”+” kartoshka” kabi birliklarga ajratib beradi. Yoki “pishirilgan” so'zidan “pishirmoq” so'zini ajratib beradi. Ikkalasini ham asosi bir xil, lekin ular ifodalagan ma'no ikki xil. Keyingi jarayon lemmatizatsiya bo'lib, “nonni”, “nondan”, “nonimizga”, “noningdan” kabi so'zlarni boshlang'ich asos shakli “non”ni topib beradi [8]. Normalizatsiya – matndagi ortiqcha so'zlarni olib tashlaydi, harflarni katta-

kichikni bir xil shaklda ifodalab berishda ishlatiladi [8]. *Bugun tandirda non va somsa yopdik*. Gapda non va somsa soʻzlarini bogʻlab kelgan va soʻzi hech qanday maʼno ifodalamaydigan ikkinchi darajali boʻlak boʻlgani uchun non va somsa soʻziga qidiruv berilganda bu soʻz ishtirok etmaydi. Bu birliklar stop soʻzlar deyiladi. Gastronomik birliklarni ifodalashda morfologik tahlil ham ahamiyatga ega. Ammo barcha oziq-ovqatlar ot soʻz turkumiga kirgani tufayli tahlil jarayonida qaysi kelishik shaklida ekanligi va qoʻshimchalar tasnifi masalalariga ahamiyat beriladi. Semantik tahlil jarayonida oziq-ovqat nomlarini maʼno munosabatini aniqlashim uchun kerak. Bundan tashqari hududiy taomlarni yaxlit jihatdan bogʻlash uchun NER(Named Entity Recognition) ham muhim. U oziq-ovqatlarni qidiruvga berganda maxsus nomlar bilan atalgan atoqli oziq-ovqatlarni ajratib beradi. Masalan, “Samarqand oshi”, “Fargʻonacha osh”, “Shodibek palov” kabilardir [11].

Bu sohaning mavzuyimizga bogʻliq jihatlari koʻp. Kelajakda tezaurusning koʻp tilli shaklda rivojlantirish kun tartibimizga qoʻyilgan. Katta hajmdagi matnlarni tarjima qilish jarayonida oziq-ovqat nomlarini tanishi, uni tarjima qilishda ekvivalenti yoki transkripsiyasini qoʻyishda NLPning ahamiyati katta. Tarjima jarayonida semantik algoritmlar yordamida tarjima qilishda matn tarkibida kelgan birikma holatdagi taomlarni oʻz maʼnosida kelib chiqishi ham ahamiyatlidir. Taomlar tarkibidagi *tovuq shoʻrva, ilik shoʻrva, qatlama somsa, qatiqli moshxoʻrda* kabi birikma holatidagi taomlarni tarjima qilishda ularni semantik jihatdan tanib olish muhimdir [15].

Biz gastronomik birliklar tezaurusini yaratishda semantik qidiruv algoritmlarining modellaridan foydalanamiz. *Vector Space Model*- bunda har bir ovqat nomi vector koʻrinishiga oʻtkazib olinadi [2]. Oʻxshashliklarni kosinus burchagi orqali aniqlab olamiz. Masalan, osh va palov soʻzlari oʻxshash, ularning vektorlari yaqin, ammo shoʻrvaning vektori uzoq.

TF-IDF(Term Frequency-Inverse Document Frequency)-modeli boʻlib, matndagi muhim soʻzlarni ajratib beradi. Bu modelni qoʻllashdan oldin tezaurusga kiritilayotgan baza tarkibidan asosiy soʻzlarni ajratib yoki belgilab qoʻyishimiz

kerak, shundagina bu model o'z ta'siriga ega bo'ladi. Masalan, "osh" so'zi yuqori chastotali, lekin u umumiy ma'noda, "Farg'onacha osh" esa aniqlovchi birlik sifatida namoyon bo'ladi.

Word Embeddengs (taqsimlangan semantik modellar)- ushbu model eng muhim zamonaviy yondashuv shakli bo'lib, u:

Word2Vec

FastText

GloVe

kabi turlaridan iborat [5]. Bu modellar asosan, so'zlarning yaqinligini o'rganib bizga kontekstdagi semantik yaqin ma'nodagi so'zlar qatorini aniqlashda yordam beradi. Masalan, *somsa, manti, chuchvara* bir biriga yaqin so'zlar, ya'ni barchasi-ichiga qiyma solib xamirdan tayyorlanadigan taom. Yoki *moshkichiri, moshxo'rda, moshli palov* so'zlari bo'lib, bu qatorlaning tarkibi tayyorlanishi turlicha, lekin har birining takibida mosh mahsuloti borligini ifodalaydi. Bu model hech qachon bizga sho'rva va lag'monni birga chiqarib bermaydi, sababi ikkisinin bir-biriga yaqinlik jihati umuman yo'q.

O'zbek tili agglyutinativ tillar guruhiga kirgani uchun so'z asosi qo'shimcha qo'shilishi natijasida o'zgarish hodisasi ko'p uchraydi. Mana shu jarayonda FastText biz uchun juda muhim. Chunki u bizga taomlarga qo'shilib kelgan qo'shimchali shakllarni hisobga olib ishlaydi.

Zamonaviy NLP tizimlarida Transformer asosidagi modellar ham mavjud.

BERT

mBERT

GPT arxitekturasini

Bular matnni chuqur tahlil qilib, matn tarkibidagi ko'p ma'noli so'zlarni ajratib beradi. Masalan, *osh* va *osh qilmoq* so'zlari o'rtasidagi ma'noga aniqlik kiritib beradi. Bittasi ot, ikkinchisi fe'lni ifodalaganini shu model asosida bilinadi.

Maqsadga erishish uchun semantik qidiruv tizimi arxitekturasida ham muhim ahamiyatga ega.

Foydalanuvchi so'rovi

NLP orqali tahlil;

semantik interpretatsiya;

tezaurus bilan solishtirish;

natijani reytinglash;

foydalanuvchiga chiqarish.

Tadqiqot natijalari shuni ko'rsatadiki, semantik qidiruv algoritmlari va NLP texnologiyalari oziq-ovqat nomlari tezaurusini oddiy lug'atdan intellektual, semantik bog'langan tizimga aylantirib beradi. Foydalanuvchi so'rovini chuqur tushunadi. Madaniy va semantik bog'liqliklarni ochib beradi. An'anaviy qidiruv tizimlari til birliklarining faqat tashqi shakliga asoslangan holda ishlashi sababli murakkab semantik munosabatlarni to'liq aks ettira olmaydi. Semantik yondashuv esa, aksincha so'zlarning ichki mazmuni, kontekstual bog'liqligi va foydalanuvchi so'rovining haqiqiy maqsadini aniqlash orqali ancha mukammal natijalarni taqdim etadi. Semantik qidiruv algoritmlari va NLP texnologiyalarini oziq-ovqat nomlari tezaurusiga integratsiyasi natijasida tezaurusning ilmiy va amaliy ahamiyati ortadi. Tezaurus nafaqat lingvistik tadqiqotlar uchun balki raqamli lug'atlar, tajima tizimlari va intellektual qidiruv platformalari uchun ham muhim ilmiy-amaliy manba bo'lib xizmat qiladi.

FOYDALANILGAN ADABIYOTLAR:

1. Абдурахмонова Н.З., Эшмаматова М.Х. Компьютер лингвистикаси: назарий ва амалий асослар. – Тошкент: ТДЮУ, 2021. – 180 б.
2. Gulyamova Sh. O‘zbek tilida elektron lug‘at yaratishning leksik-semantik asoslari. – Samarqand: SamDU, 2022. – 140 b.
3. Махмудов К. Ўзбек тансиқ таомлари. — Тошкент: Меҳнат, 1989. – 344 б.
4. Одилова Г. К. Глутоник дискурс асослари. –Тошкент: Mumtoz so‘z, 2020. – 287 б.
5. Пўлатов А.Қ. Компьютер лингвистикаси асослари. – Тошкент: Фан, 2009. – 220 б.
6. Гулова З.А. Концепт «еда» в русском и польском языках: Автореф. дисс. ... канд. филол. наук. – Душанбе, 2013. – 24 с.
7. Ibratova A. Sh. O‘zbek tilida non va non mahsulotlari bilan bog‘liq lug‘aviy birliklar tadqiqi. Filologiya fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi avtoreferati. –Samarqand: TATU Samarqand filiali, 2023. – 51 b.
8. Раҳматуллаев Ш. Тилшунослик терминларининг изоҳли луғати. – Тошкент: Ўзбекистон Миллий энциклопедияси, 2005.
9. Савельева О.Г. Концепт «еда» как фрагмент языковой картины мира: лексико семантический и когнитивно-прагматический аспекты (на материале русского и английского языков): Автореф. дисс. ... канд. филол. наук. – Краснодар, 2006. – 25 с.
10. Xoliyorova G‘.G. O‘zbek tili electron tezaurusini modellashtirishning leksikografik tadqiqi: filol.fan.bo‘yicha falsafa doktori (PhD) dis.avtoreferati. –Toshkent, 2025. –56 b.
- 11.Худаярова М. Ўзбек тилидаги таом номларининг лингвистик таҳлили. Филол.ф.н. дис... автореферати. – Тошкент, 2008. –56 б.
- 12.Ўзбек тилининг изоҳли луғати: 6 жилдлик. –Тошкент: Ғафур Ғулом номидаги нашриёт-матбаа ижодий уйи, 2023.

13. Ўзбек тили изоҳли луғати: 6 жилдли. 3-жилд/таҳрир ҳайъати:
Э.Бегматов ва бошқ. –Тошкент: Ўзбекистон миллий энциклопедияси
Давлат илмий нашриёти, 2022. –240 б.
14. Yo‘ldoshaliyeva M. Oziq-ovqat nomlarini ifodalovchi birliklarning
tezaurusda berilishi. mag.diss., 2026. 90 b.