O‘ZBEK TILI TEZAURUSINI YARATISHDA STANDART ME’YORLAR

Abduraxmonova Nilufar
O‘zbekiston milliy universiteti
Kompyuter lingvistikasi va amaliy tilshunoslik
kafedrasi mudiri, filologiya fanlari doktori, professor.

Xoliyorova Gulruhsor
O‘zbekiston milliy universiteti
Kompyuter lingvistikasi va amaliy tilshunoslik
kafedrasi o‘qituvchisi.

Annotatsiya: Mazkur maqolada sun’iy intellektning so‘z ma’nolarini aniqlash va ma’nolarni bog‘liq shaklarda tushuntirishdagi muammolar tahlilga tortilgan. Maqola tadqiqotchilar tomonidan olib borilgan WordNet va strukturasi, ulardagi mavzuiy bog‘lanishlar, ma’lumotlarni izlash tizimlarida tezaurusning ahamiyati kabi muhim masalalarni ilmiy jihatdan taqqoslashga asoslangan. Tezaurus yaratish uchun lug‘atning mikro va makro sxemasini loyihalashtirishda ingliz tili va boshqa tillar uchun yaratilgan lingvistik me’yorlar o‘zbek tili tezaurusi uchun namuna bo‘lib xizmat qilishi mumkin. Olimlar tomonidan tadqiq etilgan tezaurus turlari va ular uchun tan olingan lingvistik me’yorlar tavsifi bayon etilgan.

Tayanch so‘zlar: tezaurus, WordNet, konseptual masofa, ideografik lug‘at, monolingval tezaurus.

Abstract: This article analyzes the problems of artificial intelligence in determining the meanings of words and explaining the meanings in related forms. The article is based on a scientific comparison of important issues such as WordNet and its structure, thematic links in them, the importance of the thesaurus in information retrieval systems, conducted by researchers. Linguistic standards created for English and other languages can serve as a model for the thesaurus of the Uzbek language when designing the micro and macro scheme of the dictionary for creating a thesaurus. The types of thesauruses researched by scientists and the linguistic norms recognized for them are described.

Keywords: thesaurus, WordNet, conceptual distance, ideographic dictionary, monolingual thesaurus.

Аннотация: В данной статье анализируются проблемы искусственного интеллекта при определении значений слов и объяснении значений в родственных формах. Статья основана на научном сопоставлении таких важных вопросов, как WordNet и его структура, тематические связи в них, значение тезауруса в информационно-поисковых системах, проведенном исследователями. Лингвистические стандарты, созданные для английского и других языков, могут служить моделью тезауруса узбекского языка при проектировании микро- и макросхемы словаря для создания тезауруса. Описаны виды тезаурусов, исследованные учеными, и признанные для них лингвистические нормы.

Ключевые слова: тезаурус, WordNet, концептуальная дистанция, идеографический словарь, одноязычный тезаурус.

Sun’iy intellekt uchun bilimlarni aniqlash, qolaversa matnning og‘zaki va yozma shakldagi mazmunini tushunishda wordnet- tezaurus juda muhim ahamiyat kasb etadi. WordNET tezaurusning bir turi hisoblanadi. E.Agirre WordNet tipidagi lug‘atlardagi har bir so‘z mavzuiy belgilariga ko‘ra ma’no jihatdan sohaviy vektorlarga bog‘lanishini ta’kidlaydi. Uningcha, har bir lug‘atdagi so‘zlar mavzuiy vektor hisoblanib, ularning og‘irligi ikkinchi so‘z ma’nosiga bog‘liqdir [Eneko Agirre1, Enrique Alfonseca2, and Oier Lopez de Lacalle Approximating Hierarchy-Based Similarity for WordNet Nominal Synsets using Topic Signatures -P. 15 / GWC 2004 Second International WordNet Conference, GWC 2004 Brno, Czech Republic, January 20–23, 2004 Proceedings (CD-ROM version)]. Tadqiqot mualliflari ikki metod orqali so‘z ma’nosiga bog‘langanligini ko‘rsatib o‘tishadi:

1. Asosiy so‘z ma’nosining umumiy munosabatlarda foydalanish (sinonim, giponim);
2. Asosiy so‘zning faqay monosemantik munosabatlarda foydalanish.

Yuqoridagi dastlabki metod orqali polisemantik semalar bo‘yicha asosi so‘z ma’nosiga u darajada aniq bo‘lmagan namunalarni to‘plasa, ikkinchi metoddda yanada to‘liqroq namunalarga erishish uchun bu muammolarni chetlab o‘tadi. Ushbu tadqiqotda dastlab so‘zlar lemmatizatsiya qilinadi, so‘ng kontekstga kalit so‘z sifatida ishlatilgan har bir so‘z chastotasiga ko‘ra vektor hosil qilinadi (bag of words).

Biroq shuni ham aytish o‘rinliki, kontekstda uchragan har bir so‘zning chastotasi bo‘yicha xulosa chiqarish to‘liq xulosani bermaydi. Chunki tabiiy til turli nutq pozitsiyasida turli leksik va grammatik qo‘shimchalarni olishi natijasida kutilgan natijaga olib kelmasligi mumkin.

E.Agirre tezaurus tarkibidagi so‘zlarning o‘xshashlik me’yorlarini aniqlash uchun WordNetda mavjud bo‘lgan asosiy standartni iyerarxik bog‘lanish uchun asos qilib oladi [Eneko Agirre1 , Enrique Alfonseca2 , and Oier Lopez de Lacalle Approximating Hierarchy-Based Similarity for WordNet Nominal Synsets using Topic Signatures -P. 15 / GWC 2004 Second International WordNet Conference, GWC 2004 Brno, Czech Republic, January 20–23, 2004 Proceedings (CD-ROM version)]. - Synsetning informatsion kontentiga asoslangan Resnik masofasi metrikasi;

Konseptual masofa deb nomlanuvchi (Conceptual Distance) WordNet iyerxiyasidagi ikki sinset o‘rtasidagi giporonimik bog‘lanishlarning minimal miqdordagi inversiyasi. Yuqorida nomi zikr qilingan olimlarning asosiy yondashuvi shundan iboratki, o‘xshashlik mezoni va uch bosh mezon o‘rtasidagi korrelyasiya sifat o‘lchovi sifatida hizmat qiladi. Kerrelyatsiya quyidagi bosqichlardan iborat bo‘ladi: birinchi bosqichda har bir ot so‘z turkumi uchun simmetrik o‘xshashlik matritsasi mavuiy belgilariga ko‘ra hisoblanadi. Ikki matritsa o‘rtasidagi korrelyatsiya vektorlarga transforma qilinib hisoblanadi. Agar o‘xshashlik juda yaqin bo‘lsa 1, teskarisi bo‘lsa, 0 raqam qo‘yiladi. Shu usul orqali so‘zlarning ma’no jihatdan yaqinligi aniqlanadi[Abdurakhmonova, N.2020,2021,2022].

Tezaurus borasida dunyo tillari bo‘yicha juda ko‘plab ishlar amalga oshirilgan. Har qaysi tilda yaratilgan WordNetni olmaylik ingliz tilida yaratilgan WordNet asosida amalga oshirilgan. Tezaurusni modellashtirishda konseptlarning o‘rni juda muhim sanaladi. Prinston WordNeti turli tillar uchun namuna vazifasini o‘taydi. Ayniqsa, ko‘p tilli semantik resurs sifatida quyidagi jihatlari bilan ahamiyatlidir [Russian WordNet From UML-notation to Internet/Intranet Database Implementation Valentina Balkova 2, Andrey Sukhonogov 1, and Sergey Yablonsky – P.31.]:

- Informatsion aniqlovchi instrumentariy;
- Mualliflarni aniqlovchi instrumentariy;

- Tilni o‘rganishga mo‘ljallangan instrumentariy;
- Tarjima instrumentariysi;
- Qisqartirish;
- Semantik web.

Tezaurus ideografik lug‘at sifatida tilning barcha jihatlarini inobatga oladi. Masalan, rus tili korpusi uchun quyidagi lug‘atlar kiritilgan:

1) rus tili gGrammatik lug‘ati;
2) kompyuter lug‘ati;
3) geografik nomlar lug‘ati;
4) kishi nomlari, ptaronim va familiyalar lug‘ati;
5) biznes lug‘ati;
6) huqushunoslik lug‘ati;
7) jargon so‘zlar lug‘ati
8) izohli lug‘at.
9) Rus tili Russicon tezaurusi
10) Rus tili Orfografik lug‘ati

Lingvistik ma’lumotlar bazasi sifatida barcha lug‘atlar *text* fayl sifatida kiritilgan bo‘lib, har bir leksik birlik so‘z t*urkumi, kelishik shakli, jinsi, son, zamon, shaxs-son, daraja, nisbat, mayl, aspect, shakl, tur, o‘timli-o‘timsizligi, jonli yoki jonsiz* ekanligiga ko‘ra teglangan. Mualliflar rus tili tezaurusini yaratishda EuroWordNet tizimidagi namunadan faydalangan:

- Birlashgan yondashuv (merge approach): monolingval leksik resursdan taksonomiyani qurish, so‘ngra bilingval lug‘atdan foydalanib, moslashtirish jarayonini amalga oshirish (mapping);
- Kengaytirilgan yondashuv: bilingval lug‘atlardan foydalanib, ingliz tili sinsetiga so‘zlarni bevosita moslashtirish.

2007-yildagi holatiga rus tili tezaurusidagi (WordNet.ru) [Russian WordNet From UML-notation to Internet/Intranet Database Implementation Valentina Balkova 2, Andrey Sukhonogov 1, and Sergey Yablonsky – P.31.] so‘zlar statistikasi:

Russian WordNet Word Report					
Total	Noun	Verb	Adj	Adv	Other
111749	44751	27997	20736	4997	13268

Olimlar tezaurusning konseptual modelini ishlab chiqishda bir nechta yondashuvlar bo‘yicha ishlarni ilmiy-amaliy jihatdan amalga oshrishga harakat qilganlar.

Axborot qidiruv tizimi uchun tezaurusning ahamiyati katta. Foydalanuvchi tomonidan yuboriladigan so‘rov orqali olinadigan ma’lumotlar semantik to‘rga, ya’ni tezaurusga bog‘langan bo‘ladi. So‘zlar va konseptlar o‘rtasidagi leksik va semantik munosabatlarni formal vositalar orqali ifodalash mumkin [Edward A.F. , Nutter J.T. etc. Building a large thesaurus for information retrieval // - P. 101]. Lingvistik lug‘atlarda so‘zlarning leksik sematik jihatdan munosabatlari sinonim, antonim, omonim kabi munosabatlari bilan cheklanib qoladi, biroq axborot qidiruvi uchun so‘zlarning taksonomik va iyerarhik munosabatlarining ham berilishi katta ahamiyatga ega. J. Edward va boshqalar tomonidan olib borilgan tadqiqotlarda axborot qidiruv tizimi uchun katta hajmdagi tezaurus yaratishda mashina o‘qiy oladigan lug‘atlar uchun matnlarni qayta ishlash metodidan foydalanib, maqolalarga berilgan izohlar analiz qilingan. Semantik tarmoqdagi

munosabatlar bilan olingan ma’lumotlar birlashtirilgan. So‘ng Smart va CODER eksperimental qayta ishlash tizimidan foydalanib, tezaurusdagi birliklar tekshirilgan.

J.Aitchison, A.Gilchrist va D.Bawden kabi olimlar tomonidan tezaurus modellari bo‘yicha izlanishlar olib borilgan. Ularning “Thesaurus construction and use: a practical manual” [J.Aitchison, A.Gilchrist va D.Bawden Thesaurus construction and use: a practical manual (4-edtition), Taylor & Francis e-Library, 2005. -202 p.] nomli qo‘llanmasida tezaurusga doir ilmiy tadqiqotlar tahlilga tortilgan. Unga ko‘ra axborot qidiruv tizimida tezaurus 4 xil imkoniyatga ega bo‘lishi mumkin:

- tezaurus qidiruv va indeksatsiya uchun foydalaniladi;
- tezaurus indeksatsiyada foydalaniladi, qidiruvda foydalanilmaydi;
-tezaurus qidiruvda foydalaniladi, indeksatsiyada foydalaniladi;
-tezaurus har ikkisida ham foydalanilmaydi.

Tezaurus yaratishning bir nechta usullari mavjud. Yuqorida tadqiqotchilar tomonidan keltirilgan tezaurus turlariga qarab ularning qurilish arxitekturasi va kontenti ham talablari ko‘p tilli yoki bir tilli tezaurus uchun standartlar yaratiladi. Dastlab 1974-yilda Xalqaro standartlashtirish tashkiloti tomonidan bir tilli tezaurus uchun yaratilgan. Keyingilari 1985-yillarda nashr qilingan. 1967-yilda nashr qilingan Injineriya va ilmiy terminlar Amerika tezaurusi alifbo tartibida tuzilgan. Monolingval tezaurus uchun milliy standartlar Britaniya standartini o‘z ichiga olgan. 1959-yildan 1993-yilgacha bo‘lgan evolyutsion yo‘riqnomasi va tezaurus tuzish tamoyillarining shakllanish bosqichlari Krooks va Lankaster tomonidan tadqiq etilgan [J.Aitchison, A.Gilchrist va D.Bawden Thesaurus construction and use: a practical manual (4-edtition), Taylor & Francis e-Library, 2005. – P. 25.].

J.Aitchison fikricha tezaurus yaratishning dastlabki bosqichlaridan biri bu uning tarkibiga kiritiladigan birliklarning grammatik shakli, talaffuzi, ko‘plik shakli, qisqartma yoki qo‘shma shaklda ekanligi tashkil etadi. Ikkinchi bosqichda u yoki bu konseptga tegishli konseptni ifodolovchi ikki yoki undan ortiq sinonimlarning mosini tanlash masalalari o‘rganiladi. Uchinchi bosqichda muayyan turga kiruvchi terminlarni, jumladan o‘zlashma so‘zlar, slenglar, tijorat nomlar va atoqli otlarni aniqlash kabilar bajariladi[Abdurakhmonova, N.2020,2021,2022]. To‘rtinchi, tezaurusning xususiy maqsadidan kelib chiqqan holda terminlarning turli kontekstda u yoki bu ma’noda qo‘llanishini cheklash uchun ayrim qoidalar yaratiladi. Ushbu cheklov omonimlarni aniqlash imkonini beradi. Ingliz tili terminologiyasida controlled language nazorat qilingan tilde indekslangan termin tanlangan (preffered yoki non preferred term) termin sifatida qo‘llanadi. Tanlangan termin berilgan konseptni ifodalashda ishlatiladi. U deksriptor yoki kalit so‘z deb ham nomlanadi. Tanlanmagan termin sinonim yoki kvazi sinonim sifatida ham qo‘llanadi. Ushbu birliklar indekslash uchun emas, balki tanlangan terminga mos birlikni aniqlashga yordam beradi.

Mualliflarning qayd etishicha, indekslangan terminlar orqali ifodalangan konseptlar umumiy kategoriyalarge tegishli bo‘ladi. Ushbu kategoriyalar fasatlar analizida fundamental kategoriyalardan iborat bo‘ladi. Ular konkret birliklar, mavhum konseptlar va atoqli otlardan iborat bo‘ladi.

Konkret birliklar:

Narsalar va fizik qismlar:

BINOLAR

QAVATLAR

OROLLAR
Materiallar
SEMENT
YOG‘OCH
ALYUMIN
Abstrakt konseptlar
Faoliyat va voqealik
BOSHQARUV
URUSHLAR
MAROSIMLAR
Abstrakt birliklar, narsalar, materiallar, faoliyatlar
QONUN
NAZARIYA
KUCHLANISH
SAMARADORLIK
Sohalar va fanlar
FIZIKA
METEROLOGIYA
PSIXOLOGIYA
O‘lchov birliklari
KILOGRAM
METR
Atoqli otlarga tegishli Individual birliklar
NIGERIA
INSON HUQUQLARI KOMISSIYASI

Shu tarzda standarrt birliklarning kategoriyalari keltiriladi. Shuningdek, indekslangan terminlarga ot va otli frazalar keltiriladi. Otli frazaga kiruvchi umumiy shakllar sifat frazadan iborat bo‘ladi.

Masalan,
AYOL ISHCHILAR
BOSMA NASHRLAR
SHAHAR JAMOAT JOYLARI

Standartlarga ko‘ra, garchi ayrim tezauruslarda vaqt, holat, o‘lcham, hajm, vaziyat va xususiyat kabil sifatlar kiritilgan bo‘lsa-da, indekslanuvchi terminlar sirasiga sifatlar qabul qilinmagan. ANSI standartlarida sifatlarni alohida holatda ko‘rsatish tavsiya qilinmaydi.

Olimlar ushbu fikrga namuna sifatida quyidagilarni keltirishadi:
<PORTABLE=> portable typewriter
USE TYPWRITER and AUXILIARY>
RECTANGULAR=> Rectangular windows
USE WINDOWS AND RECTANGULAR>

Shuningdek, ayrim standartlarda ularni otli birikmalarning tarkibi sifatida keltirish o‘rinli deb qaralgan:
MINIATURE SIZE
PORTABE DEVICES kabi.

Ravish so‘z turkumiga tegishli birliklar, ayniqsa very, highly kabilar tezaurusga kiritilmaydi, ular faqat muayyan birikmaning tarkibida keltiriladi: very large scale integration

Tezaurusda fe’l so‘z turkumi ham kiritilmaydi, uning o‘rniga u hosil qilgan ot, fe’l ishtirokidagi otli birikma, gerundiy kabilar kiritiladi: COMMUNICATION (COMMUNICATE emas), ADMINISTRATION (ADMINISTER emas), WALKING (WALK emas).

Ingliz tilida yaratilgan tezauruslarda artikllar ham kiritilmaydi.

Xulosa o‘rnida shuni e’tirof o‘rinliki, o‘zbek tili tezaurusi uchun yaratiladigan ma’lumotlar bazasi tabiiy tilni qayta ishlash jarayoni uchun, qolaversa, axborot qidiruv tizimi uchun moslanganligi kelgusida sun’iy intellekt texnologiyalarida tadqiqotlar uchun muhim manba bo‘lib xizmatt qiladi.

Foydalanilgan adabiyotlar:

1. Eneko Agirre1, Enrique Alfonseca2, and Oier Lopez de Lacalle Approximating Hierarchy-Based Similarity for WordNet Nominal Synsets using Topic Signatures -P. 15 / GWC 2004 Second International WordNet Conference, GWC 2004 Brno, Czech Republic, January 20–23, 2004 Proceedings (CD-ROM version)
2. Edward A.F., Nutter J.T. etc. Building a large thesaurus for information retrieval // - P. 101
3. J.Aitchison, A.Gilchrist va D.Bawden Thesaurus construction and use: a practical manual (4-edtition), Taylor & Francis e-Library, 2005. -202 p.
4. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021, January). Uzwordnet: A lexical-semantic database for the uzbek language. In *Proceedings of the 11th Global Wordnet conference* (pp. 8-19).
5. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 73-75). IEEE.
6. Abdurakhmonova, N., & Ismailov, A. S. (2022). APPLYING WEB CRAWLER TECHNOLOGIES FOR COMPILING PARALLEL CORPORA AS ONE STAGE OF NATURAL LANGUAGE PROCESSING. In *СОВРЕМЕННАЯ ФИЛОЛОГИЯ. СОЦИАЛЬНАЯ И НАЦИОНАЛЬНАЯ ВАРИАТИВНОСТЬ ЯЗЫКА И ЛИТЕРАТУРЫ* (pp. 22-27).
7. Abduraxmonova, N., & Abduvaxobov, G. I. (2021). O ‘quv lug ‘atini tuzishning nazariy metodologik asoslari. *СЎЗ САНЪАТИ ХАЛҚАРО ЖУРНАЛИ*, *103*.
8. Sulevmanov, D., Gatiatullin, A., Prokopyev, N., & Abdurakhmonova, N. (2020, November). Turkic morpheme web portal as a platform for turkology research. In *2020 International Conference on Information Science and Communications Technologies (ICISCT)* (pp. 1-5). IEEE.
9. Russian WordNet from UML-notation to Internet/Intranet Database Implementation Valentina Balkova2, Andrey Sukhonogov1, and Sergey Yablonsky – P.31.
10. Russian WordNet from UML-notation to Internet/Intranet Database Implementation Valentina Balkova2, Andrey Sukhonogov1, and Sergey Yablonsky – P.34.