

WORDNET ASOSIDA SO‘Z MA’NOSINI ANIQLASH

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO‘TAU

Muxammadyakubova Mo‘mina Esojon qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
saidakadirova002@gmail.com
ToshDO‘TAU

Annotatsiya. Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasida so‘zning aniq ma’nosini aniqlash – ya’ni, Word Sense Disambiguation (WSD) – dolzarb va murakkab vazifalardan biridir. Matn kontekstida bitta so‘z turli ma’nolarga ega bo‘lishi mumkinligi sababli, qaysi ma’no aynan shu kontekstga mos kelishini aniqlash talab etiladi. *WSD* nafaqat tilni avtomatik tushunishda, balki tarjima, axborot qidiruvi va boshqa ko‘plab amaliy sohalarida ham muhim rol o‘ynaydi. Hozirgi kunda usullar asosan statistik yondashuvlar va mashinani o‘rganish algoritmlariga tayansa-da, ular katta hajmdagi belgilanganda ma’lumotlarga ehtiyoj sezishi va tilning murakkab semantik munosabatlarini to‘liq aks ettira olmasligi sababli cheklovlarga ega. Shu sababli, so‘nggi yillarda WordNet kabi boy leksik resurslardan, xususan uning bilimga asoslangan imkoniyatlaridan samarali foydalanish g‘oyasi ilgari surildi. Ushbu maqola WordNetning leksik semantik tuzilishi, uning asosiy tarkibiy qismlari, bilimga asoslangan yondashuvi va bu yondashuv asosida so‘z ma’nosini aniqlash uchun ishlab chiqilgan yangi algoritmik metodlarni tahlil qiladi.

Abstract. In the field of Natural Language Processing (NLP), determining the exact meaning of a word - that is, Word Sense Disambiguation (WSD) - is one of the current and complex tasks. Since a single word can have different meanings in the context of a text, it is necessary to determine which meaning corresponds to this context. WSD plays an important role not only in automatic language understanding, but also in translation, information retrieval, and many other practical areas. Although currently the methods are mainly based on statistical approaches and machine learning algorithms, they have limitations due to the need for large amounts of data and the inability to fully reflect the complex semantic relationships of the language. Therefore, in recent years, the idea of effectively using rich lexical resources such as WordNet, in particular its knowledge-based capabilities, has been put forward. This article analyzes the lexical semantic structure of WordNet, its main components, knowledge-based approach, and new algorithmic methods developed based on this approach for determining the meaning of words.

Абстрактный. В области обработки естественного языка (NLP) определение точного значения слова, то есть устранение неоднозначности слов (WSD), является одной из актуальных и сложных задач. Поскольку одно слово может иметь разные значения в контексте текста, необходимо определить, какое значение соответствует этому контексту. WSD играет важную роль не только в автоматическом понимании языка, но и в переводе, поиске информации и многих других практических областях. Хотя в настоящее время методы в основном основаны на статистических подходах и алгоритмах машинного обучения, они имеют ограничения из-за необходимости больших объемов данных и невозможности в полной мере отразить сложные семантические связи языка. Поэтому в последние годы выдвигается идея эффективного использования богатых лексических ресурсов, таких как WordNet, в частности его возможностей, основанных на знаниях. В данной статье анализируется лексико-семантическая структура WordNet, ее основные компоненты, подход, основанный на знаниях, и новые алгоритмические методы, разработанные на основе этого подхода для определения значения слов.

Kalit so‘zlar: *WordNet, Semantik tahlil, So‘z ma’nosi aniqlash, Leksik-semantik tarmoq, Ontologiya, Synset, NLP, Ma’no munosabatlari.*

Kirish. Tabiiy tilga ishlov berish sohasida so‘z ma’nosini aniqlash masalasi sohaning dolzarb masalalaridan biri. So‘z ma’nosini avtomatik aniqlashda turli yondashuvlardan foydalanish mumkin.

- Bilimga asoslangan yondashuvlar
- Mashinali o‘qitishga asoslangan yondashuvlar
- Gibrid yondashuvlar

Bilimga asoslangan yondashuv – oldindan belgilangan lingvistik qoidalar yordamida so‘zning turli kontekstlardagi ma’nolari aniqlanadi. Ushbu yondashuv ham o‘z navbatida qoidaga asoslangan, *WordNet* yoki *Tezaurusga* asoslangan algoritmlarga bo‘linadi.

Tabiiy tilning qat’iy qoidalari asosida so‘z ma’nosini aniqlash mumkin, ammo bu usuldan foydalanish barcha noaniq so‘zlarni aniqlash imkonini bermaydi. Tilning *WordNet* tizimi yordamida ham so‘z ma’nosini aniqlash mumkin.

WordNet – tabiiy tildagi so‘zlar va ularning o‘zaro semantik munosabatlarini tizimli shaklda aks ettiruvchi yirik leksik ma’lumotlar bazasidir. Ushbu resurs so‘zlarning ma’nolari, sinonimlik, antonimlik va iyerarxik bog‘lanishlarini aniqlik bilan ifodalaydi. Tabiiy tilning *WordNet* tizimi tilning semantik strukturasi

chuqurroq o'rganish, kontekstual tahlilni amalga oshirish va morfologik xususiyatlarni (qo'shimchalar, so'z yasash tizimi va boshqalar) qamrab oladi.

Tilning semantik strukturasi so'zlar o'rtasidagi murakkab bog'lanishlarni aniqlaydi va ularni iyerarxik tarzda tasniflaydi. *Kontekstual tahlil* esa so'zning turli kontekstlarda qanday ma'nolarda qo'llanilishini o'rganish. *Morfologik moslashuv* o'zbek tilining boy qo'shimcha tizimi va so'z yasash xususiyatlarini hisobga olgan holda leksik ma'lumotlarni shakllantirish demakdir.

WordNetning tizimi bir qancha qismlarga bo'linadi. Quyida uning asosiy tarkibiy qismlari keltirilgan:

a) Sinonim to'plamlari bu *Synsets* deb ham nomlanadi. Har bir *Synset*, ya'ni sinonimlar to'plami – bitta aniq semantik tushunchani ifodalaydi. Masalan, o'zbek tilida “go'zal”, “chiroyli”, “jozibali”, “maftunkor” kabi so'zlar bitta guruh sifatida ko'rib chiqilishi mumkin. *Synset* to'plamlari yordamida semantic aniqlik hamda kontekstual moslashuvchanlik masalalarini yechish mumkin.

b) Umumiy va xos atamalar (Giperonim va Giponim). Bu munosabatlar so'zlar o'rtasidagi iyerarxik bog'lanishni ifodalaydi:

Giperonim: Kengroq tushunchani ifodalaydi, masalan, “transport vositasi”. Giponim: Ushbu umumiy tushunchaga kiritilgan aniqroq tushunchalarni bildiradi. Masalan, “avtomobil”, “velosiped”, “avtobus”. Demak Giponim so'zlar giperonimlarni tarkibiy qismi bo'ladi. Giponim va giperonimlar so'zlarning semantik munosabatlarini yanada aniqroq aks ettirishga yordam beradi.

c) Qismlar va butunlik munosabatlari (Meronim va Holonim). Bu bog'lanishlar obyektning qismlari bilan uning butunligi o'rtasidagi aloqani ifodalaydi:

Meronimlar obyektning alohida qismlarini aks ettiradi. Masalan, “uy” so'zida “eshik”, “deraza”, “tom”. Holonimlar esa qismlardan butunlikni qayta tiklaydi, ya'ni, qismlar asosida to'liq obyekt tuzilishini aniqlashga imkon yaratadi.

d) Qarama-qarshi ma'nolar. Antonim– so'zlarning qarama-qarshi yoki zidd ma'nolarini ifodalaydi. Masalan, “yuqori” va “pastki”, “katta” va “kichik” kabi tushunchalar. Antonim so'zlar yordamida *semantik chegaralarni belgilash, kontekstual tahlil* qilish kabi masalalarni yechishda yordam berishi mumkin.

WordNetning eng katta afzalliklaridan biri – uning semantik munosabatlarni grafik shaklida aks ettirishi hisoblanadi. Har bir so'z yoki sinonim to'plami o'zaro bog'lanib, ular orasidagi aloqalar orqali ifodalanadi.

WordNet yordamida so'z ma'nosini aniqlash. So'z ma'nosini aniqlash tabiiy tilni tushunish va ishlov berishda muhim bosqichlardan biridir. Bitta so'zning turli kontekstlarda turlicha ma'nolarga ega bo'ladi.

1-jadval: Omonimlar va ularning ma'nolari

Omonim so'zlar	So'z turkumlari	So'z ma'nolari
olchin	ot	O'zbek xalqining etnik sos-taviga kirgan urug'lardan birining nomi.2 Shu uruqqa mansub kishi.
		Uzunlik o'lchov birligi; ar-shin (q.) va gaz (q.) qiymatlariga teng.
Ombir	ot	Dastasi uzun katta yog'och to'qmoq
		ombur.
		Devor tagidan suv o'tishi uchun qoldirilgan tuynuk; quvur.

WordNet yordamida so'z ma'nosini aniqlashda bir nechta algoritmlardan foydalanish mumkin:

Leacock-Chodorow o'lchovi: WordNetdagi so'zlar orasidagi masofani va iyerarxik bog'lanishlarni hisobga olib, ikki so'z orasidagi masofani aniqlaydi. Ushbu o'lchov yordamida ularning semantik yaqinligi baholanadi. Ikkita so'z orasidagi semantik o'xshashlik quyidagi formula yordamida aniqlanadi.

$$sim(c_1, c_2) = -\log \frac{lenght(c_1, c_2)}{2 \times maxDepth}$$

Bu yerda, $sim(c_1, c_2)$ – c_1 va c_2 so'zlarning o'xshashlik koeffitsiyenti, $lenght(c_1, c_2)$ – c_1 va c_2 so'zlar orasidagi masofa, $maxDepth$ – butun ontologik daraxtdagi eng katta chuqurlik, ya'ni ildiz (root) tugunidan eng chuqur joylashgan tugungacha bo'lgan masofa yoki balandlik. Bu parametr normalizatsiya uchun kerak. WordNet'da **tushunchalar**, synsetlar shaklida tashkil qilingan, ularning har birining giperonimi mavjud. **maxDepth**ni aniqlash uchun har bir synset uchun ildizgacha bo'lgan yo'l uzunligi hisoblanadi. Semantik o'xshashlikni aniqlashning yana bir usuli Wu-Palmer o'lchovidir.

Wu-Palmer o'lchovi: So'zlar ierarxiyasidagi eng past umumiy domenanta (common ancestor) ga nisbatan har bir so'zning chuqurligi hisobga olinadi. Natijada, ikki so'z orasidagi semantik o'xshashlik aniqlanib, ularning qanchalik yaqin yoki uzoq ekanligi o'lchanadi.

Formula:

$$sim(c_1, c_2) = \frac{2 \times Depth(LCS)}{Depth(c_1) + Depth(c_2)}$$

c_1, c_2 -taqqoslanayotgan ikkita tushuncha

$Depth(c_1), Depth(c_2) - c_2$ va c_1 tushunchalarning WordNet iyerarxiyasidagi chuqurligi, ya'ni ildizdan boshlab nechinchi qatlamda joylashgani

$Depth(LCS)$ – “Lowest Common Subsumer” — c_2 va c_1 tushunchalar uchun umumiy bo'lgan eng chuqur (yaqin) umumlashtiruvchi tushunchaning chuqurligi.

Agar tabiiy tilning WordNet tizimi mavjud bo'lsa yoki tildagi so'zlarning sinset to'plamlari tuzilgan bo'lsa, uni quyidagicha aniqlash mumkin. O'zbek tilidagi so'zlarning sinset to'plamlaridan foydalanib aniqlash mumkin. Yuqorida ko'rsatilgan algoritmlar WordNet bilimiga asoslangan yondashuvi yordamida turli so'zlarning ma'nolarini kontekstga mos ravishda farqlash uchun ishlatiladi. Formulalar biror bir dasturlash tilida kodlash orqali hisoblanadi. Nlp masalalarini hal qilish uchun Python eng qulay til hisoblanadi.

```
from nltk.corpus import synset as wn

def get_wupalmer_similarity (w1, w2, pos='n'):

    sin1 = wn.Synsets(w1, pos=pos)

    sin2 = wn.Synsets(w2, pos=pos)

    if not sin1 or not sin2:

        return None

    max_sim = 0

    for s1 in sin1:

        for s2 in sin2:

            sim = s1.wup_similarity(s2)

            if sim and sim > max_sim:

                max_sim = sim

    return max_sim
```

Izlanishlar davomida o'zbek tilidagi leksik omonimlarning semantik to'plamlari tuzilib json formatga joylandi. Hosil qilingan dataset ustida algoritmlardan foydanildi va 1-jadvaldagi kabi natijalar olindi.

2-jadval: Algoritmning aniqlik darajasi

Algoritm	Aniqlik (%)	F1-O'lchov
Lesk	72.5	68.3
Wu-Palmer	78.9	74.1
Leacock-Chodorow	81.2	79.5

Xulosa. Ushbu maqolada WordNetning leksik ma'lumotlar bazasi, uning asosiy tarkibiy qismlari, bilim grafigi yondashuvi va so'z ma'nosini ajratib ko'rsatish (WSD) jarayoni batafsil yoritildi. WordNet bilim grafigi yordamida ishlab chiqilgan yangi algoritmik yondashuv, Leacock-Chodorow, Wu-Palmer va Lesk kabi metodlar asosida so'zlar orasidagi semantik o'xshashlikni aniqlash imkonini beradi. Eksperimental natijalar ushbu yondashuvning yuqori aniqligi va samaradorligini tasdiqladi. Jahon miqyosidagi izlanishlar va taqqoslashlar asosida, WordNet bilimga asoslangan yondashuvi istiqbolli yo'nalish hisoblanadi, chunki u tarjima, axborot qidiruvi va sun'iy intellekt tizimlarida keng qo'llanilishi mumkin.

Kelajakdagi rivojlanish yo'nalishlari, jumladan, transformer modellar bilan integratsiya, ko'p tilli qo'llab-quvvatlash, katta hajmdagi ma'lumotlar ustidagi eksperimentlar va gibrid tizimlar yaratish, ushbu metodologiyani yanada mustahkamlash va kengaytirishga qaratilgan. Natijada, tilni qayta ishlash tizimlarining samaradorligi oshib, foydalanuvchilarga aniq, ishonchli va mantiqiy xizmatlar taqdim eta oladi. Ushbu maqola WordNet bilimga asoslangan yondashuvi yordamida so'z ma'nosini ajratib ko'rsatishning yangi yondashuvini ko'rsatib, tilshunoslik va amaliy dasturlarda yanada ilg'or natijalarga erishish uchun mustahkam poydevor yaratadi.

Foydalanilgan adabiyotlar:

1. Arindam Chatterjee, Salil Rajeev Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya, “IndoWordNet va Word Sense Disambiguation uchun Asboblarni Tanishtiruvchi”, http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/18_Tools%20for%20IndoWordNet%20&%20WSD.pdf
2. C. Fellbaum, “WordNet: An Electronic Lexical Database”, MIT Press, 1998. Lokesh Nandanwar, Kalyani Mamulkar, “Supervised, Semi-Supervised and Unsupervised WSD Approaches: An Overview”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064
3. Kuhoo Gupta, Manish Shrivastava, Smriti Singh va Pushpak Bhattacharyya, “Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi”, COLING/ACL-2006, Sidney, Avstraliya, iyul, 2006.



<https://pdfs.semanticscholar.org/1daa/370563deb1e9f5772d6ada1edfb4ec4cd7b4.pdf>

4. Banerjee va Pederson, “An adapted Lesk algorithm for Word Sense Disambiguity using Word Net”,
<http://www.cs.cmu.edu/~banerjee/Publications/cicling2002.pdf>

5. Sinha M. va boshqalar, “Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity”,
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.3892&rep=rep1&type=pdf>