OMONIMIYA VA LINGVISTIK TIZIMLARDA OMONIMLARNI ANIQLASH USULLARI

Abjalova Manzura Abdurashetovna
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti
dotsent v.b., filologiya fanlari bo‘yicha falsafa doktori (PhD).
abjalova.manzura@gmail.com

ANNOTATSIYA

Omonimlikni aniqlash tabiiy tilni qayta ishlash (NLP, Natural Language Processing)da dolzarb masalalardan biri hisoblanadi. Mazkur masalada o‘zbek tilidagi matnlarda uchraydigan omonimshakllarni aniqlash, tahlil qilishning bir necha usullari xususida so‘z yuritildi.

Kalit so‘zlar: *omonim shakl, usul, yashirin Markov modeli, N-gramma, so‘z birikmalari modeli.*

ABSTRACT

Determining homonymy is one of the most pressing issues in Natural Language Processing (NLP). There are several ways to identify and analyze homonymous forms in Uzbek texts.

Keywords: *homonym, method, hidden Markov model, N-gram, phrase model.*

АННОТАЦИЯ

Определение омонимии - одна из самых актуальных проблем в обработке естественного языка (NLP). Существует несколько способов выявления и анализа омонимических форм в узбекских текстах.

Ключевые слова: *омоним, метод, скрытая марковская модель, N-грамма, фразовая модель.*

KIRISH

Matnlarni avtomatik qayta ishlash bir necha bosqichga bo‘linadi va ulardan biri morfologik tahlil bosqichi hisoblanadi. Mazkur bosqichda har bir so‘zga morfologik tavsif beriladi: lemma [1]si (asosi), kelishigi, soni, darajasi, nisbati, shaxsi va hk. Morfoanalizning so‘zlarni morfologik teglash vazifasi omonimshakllar bilan murakkablashadi.

Ma’lumki, omonim so‘zlar shakli bir xil, ammo semantikasi turfa xil bo‘lgan leksik birliklar hisoblanadi. Avtomatik qayta ishlashda omonimlik hodisasi quyidagi birliklarda mavjud:

1. so‘z omonimligi – shakldoshlik so‘z asosida bo‘ladi, ya’ni muayyan so‘z bir so‘z turkumi yoki bir necha turkumga mansub ma’no beradi. Masalan:

ot Avval birliklarni, keyin o‘nliklarni qo‘shamiz. ot Kuch – birlikda. ot Og‘irlik birliklari. Til birliklari	modal Suv bor joyda hayot bor. Ruchkang bormi? hisob so‘z Bir necha bor taklif yubordim. *ot* *Bor – kimyoviy element.* fe’l Ishga bormoq. Maktabga bormoq.

Birinchi ustunda “birlik” so‘zi bir so‘z turkumi doirasida omonim hisoblanadi: 1) o‘ngacha bo‘lgan butun son; grammatik ko‘plik aksi (ot); 2) birlashish, hamjihatlik (ot); 3) bir turdagi miqdorlarni o‘zaro baholash uchun qabul qilingan o‘lchov; til qurilishiga xos termin (ot).

Ikkinchi ustunda “bor” so‘zi ikki xil turkumga mansub shakldoshlikni yuzaga keltirgan: 1) mavjud (modal); 2) marta, daf’a, bora (hisob so‘z)

2. Qo‘shimcha omonimligi – muayyan qo‘shimcha vazifasiga ko‘ra qo‘shimchalarning turli guruhiga mansub bo‘ladi. Masalan:

-ki	*ko‘ch**ki**, tep**ki**, turt**ki*** *ust**ki**, ich**ki**, kech**ki***	(ot yasaydi) (sifat yasaydi)
-(i)ng	*uy**ing**, kitob**ing**, ish**ing***	(shakl yasaydi: sintaktik mun.shakli – egalik qo‘shimchasi)
	*ko‘r**ing**, bor**ing**, tayyyorla**ng***	(shakl yasaydi: sintaktik mun.shakli – shaxs-son qo‘shimchasi)

3. Ibora omonimligi – shakli, ya’ni tuzilishi bir xil, ammo semantikasi turlicha bo‘lgan frazeologik birliklar.

qattiq shovqin soldi uyni *boshiga ko‘tarmoq*	yuksak darajada izzat-hurmat qildi onasini *boshiga ko‘tarmoq*

4. Gap omonimligi – muayyan gap ifoda maqsadi yoki mazmuniga ko‘ra farqlanadi.

darak gap Ishni bajarmadim.	so‘roq gap Ishni bajarmadim?

MANBALAR TAHLILI

Rus tilida omonimlik turlarida so‘zlarning turkumligi bo‘yicha omonimlik, morfologik omonimlik va leksik omonimlik farqlanadi [2]. E’tiborli jihati shundaki so‘zlarni morfologik va leksik omonimligi bo‘yicha guruhlanishi tabiiy tilni qayta ishlashda muhim ahamiyat kasb etadi.

Morfologik omonimlikda bir turkumga mansub bo‘lgan shakldosh so‘zlar lemma (asosi)si turlicha, ammo muayyan shakllaridagina omonimlikni yuzaga keltiruvchi so‘zshakllar e’tiborga olinadi. Masalan:

lemmasi *ter*	lemmasi *terim*
ter*im* – mening terim (-im egalik qo‘shimchasi: I shaxs, birlik)	***terim*** – hosil

Leksik omonimlikda bir lemmaga mansub so‘z turli ma’nolarni beradi:

ot **bog‘** – to‘dalab bog‘langan holat	Bog‘lamoq, bog‘lam. B i r b o g ‘ piyoz
ot **bog‘** – o‘simlik va daraxtlar ko‘p ekilgan joy	bog‘-rog‘, bog‘-bo‘ston; Uzumini ye, b o g ‘ i n i surishtirma.

Omonimshakllarning morfologik va leksik guruhlanishi matnlarni qayta ishlovchi dasturiy ta’minot va tizimlarda lemmatizatsiya va stemming jarayonlari uchun muhim sanaladi [3]. Har ikki texnologiya so‘z yoki so‘zshaklning asosini topishga yo‘naltirilgan bo‘lib,

Ta’kidlash o‘rinliki, omonimlik hodisasi avtomatik qayta ishlash jarayonida eng dolzarb masala hisoblanadi. Shu bois NLPda omonimshakllarni aniqlash va ularni tahlil qilish maxsus o‘rganiladi, hatto bir necha usullar ham ishlab chiqilgan.

METODLAR

Omonimlikni aniqlash metodlarining barchasi ikki guruhga bo‘linadi:

1. Qoidalarga asoslangan usullar. O‘z navbatida, ular quyidagilarga bo‘linadi:

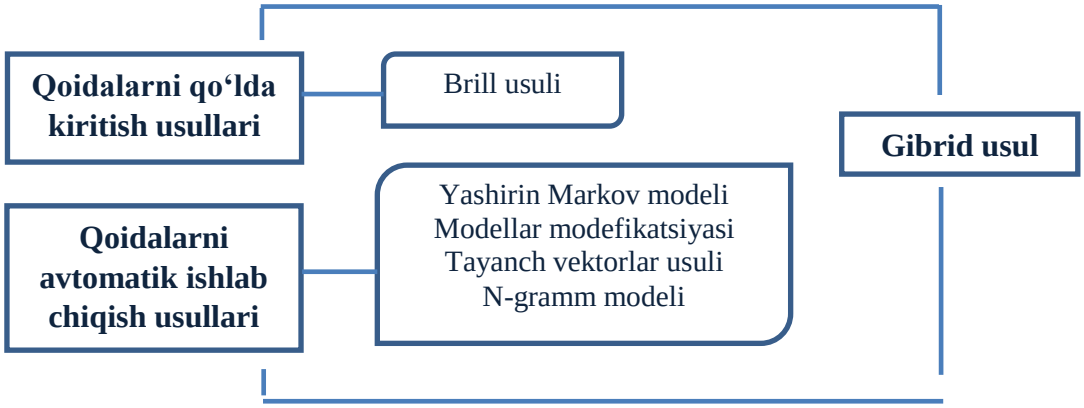
a) Qoidalarni qo‘lda kiritish usullari.

b) Qoidalarni avtomatik ishlab chiqarish usullari.

2. Statistikaga asoslangan usullar.

Ushbu guruhlarning har birining o‘ziga xos afzalliklari va kamchiliklari mavjud. Bunday vaziyatlarda tez-tez sodir bo‘ladiganidek, ikkala guruhning xususiyatlarini

(va afzalliklarini) bir usulda birlashtirish avval erishilgan natijalarga qaraganda yaxshiroq natijani ko‘rsatishi mumkin. Bunday usul gibrid usuli deb nomlanadi.



1-sxema. Omonimlikni aniqlash usullari.

Mazkur usullarga tayananuvchi tizimlar o‘z navbatida quyidagi guruhni tashkil etadi:

1. Qo‘lda yaratilgan qoidalarga asoslangan tizimlar.

2. Ehtimoliy modellar asosida yaratilgan va tavsiflangan korpuslarga tayanadigan tizimlar.

3. Ehtimollik modellari va qoidalarga asoslangan gibrid tizimlar.

Omonimiyani aniqlash uchun har bir so‘zshaklni “tasniflash” kerak, ya’ni uning lemmasi, so‘z turkumi va bir tegga birlashuvchi morfologik xususiyatlar to‘plami bilan bog‘lab qo‘yiladi.

Yashirin Markov modeli Baum L.E. va uning hamkasblari tomonidan ishlab chiqilgan [6] mazkur model omonimlikni aniqlashning statistik metodi statistik jarayonda yuzaga keladigan barcha variantlar ehtimolligini hisobga olishga yordam beradi. Masalan, ma’lum bir matnda ot turkumiga oid so‘zlar bog‘lovchiga nisbatan tez-tez va ko‘p uchrasa unda ayni kontekstda mavjud omonim katta ehtimollik bilan bog‘lovchi emas, ot turkumiga oid so‘z bo‘ladi, keyingi ehtimollikda bog‘lovchi sifatida hisobga olinadi. Kontekstni tavsiflash uchun N-grammadan foydalaniladi. N-gramma – matnlarga avtomatik ishlov berishda keng qo‘llaniladigan matematik hisob vositasidir. O‘zbek kompyuter lingvistikasida S.Rizayev harf birikmalarini bigramm, trigramm terminlari bilan ifodalagan [5].

N-gramma – so‘zlar yoki teglar kabi N-identifikator elementlarning ketma-ketligini ifodalaydi. Ikki element ketma-ketligi – bigramma, uch element ketma-ketligi esa trigramma, deyiladi. Masalan, *old qo‘shimcha+ot* holati bigrammaga misol bo‘ladi.

Omonimlikni aniqlashning oddiy statistik metodi va boshqa shu kabi metodlarning tavsifini keltirish uchun quyidagi usullar ishlatiladi:

– w_i – jumladagi i-o‘rinda joylashgan so‘z, t_i – ushbu so‘zning identifikatori (tegi).

– $D_{(w)} = \{t_1^w, t_2^w, \ldots, t_k^w\}$ w so‘zining barcha mumkin bo‘lgan belgilar majmui. Ushbu ma’lumotlarni morfologik lug‘at yordamida olish mumkin. Agar so‘z lug‘atda bo‘lmasa uni Brill usulida bajarilganidek, ot so‘z turkumi sifatida hisoblash mumkin, ammo lingvistik ta’minot ishonchli bo‘lishi uchun barcha mumkin bo‘lgan teglarni qo‘yib chiqish kerak.

– C – korpusdagi muayyan holatlar soni (n-gramm). Bunda $C(t)$ – t teglar soni; va $C(t_1,t_2)$ – bigrammalar soni (t_1,t_2).

– $C_t(w,t)$ – t tegli w so‘zlar soni.

– $F(w,t)$ – w so‘zida t tegi mavjudligi ehtimoli. Tavsiflar quyidagi formula bo‘yicha hisoblanadi:

$$F(w,t) = \frac{C_t(w,t)}{C(t)}$$

– $P(t_i|t_{i-1})$ - bu t_{i-1} tegidan keyin t_i tegining kelish ehtimollik holati. Bunda $i = 1$ bo‘lganda t_i tegi gapda birinchi teg hisoblanadi. Hisoblash formulasi quyidagicha:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1},t_i)}{C(t_{i-1})}$$

Yashirin Markov modeliga asoslanib, omonimlikni aniqlashning statistik metodi yordamida ishlash natijasida n uzunlikdagi jumlada $T_i \in D(w_i)$ bo‘lganda $T=\{T_1,T_2,\ldots,T_n\}$ teglarning ehtimoliy ketma-ketligi topiladi [7].

XULOSA

Xulosa qilib aytganda, dunyo kompyuter lingvistikasida omonimlikni bartaraf etish usullari o‘rganilganida, bu xususdagi tajribadan foydalanib o‘zbekcha matnlardagi so‘zshakllarning tegishli tekshirish formulasi yaratildi. Omonimlikni bartaraf etish uchun har bir so‘zni “tasniflash” kerak, ya’ni uni lemma – gap bo‘lagi va morfologik xususiyatlar majmui bilan taqqoslash mumkin, ular qulaylik uchun bir tegga qo‘shiladi. Barcha mumkin bo‘lgan teglarni o‘rganish uchun morfologik lug‘atdagi so‘zlarga tegishli havolalarni topish yoki MyStem kabi morfologik analizatorni ishlatish yetarli bo‘lib, u so‘z teglarini topishda yordam beradi. Shundan so‘ng bir nechta teglar orasidan faqat tegishli tegni tanlash kerak bo‘ladi.

Omonim so‘zshakllarni tahlil qilishda qo‘llanilgan optimal lingvistik usul matnlarni tahrir va tahlil qilish, mashina tarjimasi, matnlarni qayta ishlash jarayonlarida muhim omil bo‘ladi.

REFERENCES

1. Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. Учебное пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – Москва: МИЭМ, 2011. – 272 с.
2. Порохнин А.А. Анализ статистических методов снятия омонимии в текстах на русском языке. Вестник АГТУ. Сер.: Управление, вычеслительная техника и информатика. – 2013. № 2. – С. 168-174.
3. Rahmatullayev Sh. O‘zbek tili omonimlarining izohli lug‘ati. – Toshkent: O‘qituvchi, 1984. – B.5.
4. Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari: Monografiya. – Toshkent, 2020. – B. 25-27.
5. Rizayev S. O‘zbek tilshunosligida lingvostatistika asoslari. – Toshkent: Fan, 2006. – B . 18.
6. Baum, L. E.; Sell, G. R. Growth transformations for functions on manifolds. Pacific Journal of Mathematics. 27 (2) 1968. – P. 211–227.; https://en.wikipedia.org/wiki/Hidden_Markov_model.
7. http://www.academia.edu/15517740/Анализ_статистических_алгоритмов_снятия_морфологической_омонимии_в_русском_языке.