

O'ZBEK TILI MATNLARIDA KOREFERENSIYANI HAL QILISH BOSQICHLARI

Elova Dilrabo Qudratullayevna

Filologiya fanlari falsafa doktori PhD,
ToshDO'TAU katta o'qituvchisi

elova@navoiy-uni.uz

Annotatsiya. Tabiiy tilni qayta ishlash (Natural language processing, NLP) metodlari orqali odamlar va mashinalar o'rtasida aloqa o'rnatiladi. NLP – sun'iy intellektning murakkab sohalaridan biri bo'lib, inson tabiiy tili kompyuterlar o'rganishi qiyin bo'lgan semantik noaniqliklarga boydir. Bu muammoni bartaraf qilish usullaridan biri kontent mazmunini tushunishga xalal beradigan noaniq va umumiy mazmunni ifodalaydigan ifodalarni aniqlashtirishdan iborat. Kompyuterlar matndagi noaniq ifodalar qatnashgan jummalarni tushunishi osonroq bo'lgan aniq ifodalarga almashtirish NLPda koreferensiya muammosini hal qilish orqali amalga oshiriladi. Ushbu maqolada o'zbek tili matnlarida koreferensiya muammosi hal etish bosqichlari keltiriladi.

Kalit so'zlar: *koreferensiya, anaphora, anaforik munosabat, anaforik so'z, koreferensiyani hal qilish, katafora, antecedent, postsedent, eslatma.*

Abstract. Communication between people and machines is established through the methods of natural language processing (NLP). NLP is one of the complex areas of artificial intelligence, and human natural language is rich in semantic ambiguities that are difficult for computers to learn. One way to overcome this problem is to clarify expressions that express vague and general content that hinders the understanding of content content. In order to replace a sentence with ambiguous expressions in the text into an easier-to-understand expression, it is necessary to solve the problem of coreference in NLP. This article presents the stages of solving coreference problems in Uzbek language texts.

Key words: *coreference, anaphora, anaphoric relationship, anaphoric word, resolution of coreference, cataphora, antecedent, postsedent, note.*

Абстракт. Связь между людьми и машинами устанавливается с помощью методов обработки естественного языка (Natural language processing, NLP). NLP — одна из сложных областей искусственного интеллекта, а естественный человеческий язык богат семантическими неоднозначностями, которые трудно изучить компьютерам. Одним из способов преодоления этой проблемы является уточнение выражений, выражающих расплывчатое и общее содержание, затрудняющее понимание содержания. Для того чтобы заменить предложение с неоднозначными выражениями в тексте на более простое для понимания выражение, необходимо решить проблему кореферентности в NLP.

В данной статье представлены этапы решения проблем кореференции в текстах на узбекском языке.

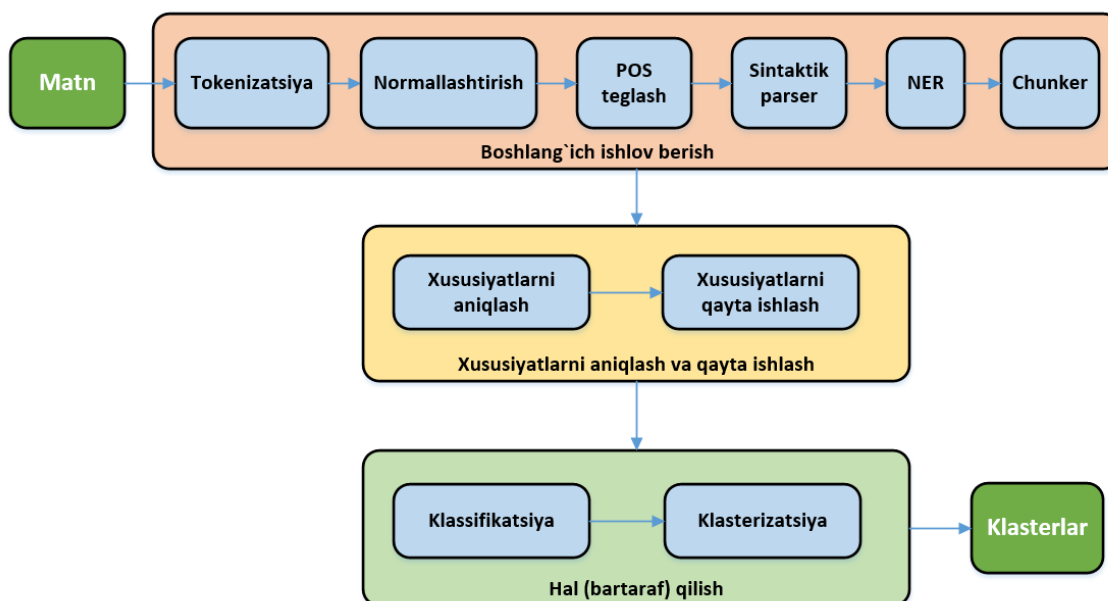
Ключевые слова: *кореференция, анафора, анафорическое отношение, анафорическое слово, разрешение кореферентности, катафора, антецедент, апостериор, нота.*

Koreferensiya muammosi hal qilish (Coreference resolution, CR) – bu matndagi bir obyektga tegishli bo‘lgan barcha eslatmalar (lingvistik ifodalar)ni aniqlashga asoslangan NLP vazifasi (1-rasm). Ushbu *eslatmalarni aniqlash* va *guruhlash (klasterlash)* orqali **noaniq ifodalarni obyektning aniq nomlari bilan almashtirib** matndagi koreferensiya muammosini bartaraf qilish mumkin. Obyekt nomlari **bir so‘zdan iborat token** bilan ham, yaxlit sintaktik butunlik - **otli birikmalar** bilan ham ifodalanishi mumkin.

“Men Tohir Malikni tanlayman, chunki u mening qiziqishlarimga mos asarlar yozgan” – dedi Akbar

1-rasm. Gapdagi bir obyektga tegishli bo‘lgan barcha eslatmalar

Koreferensiya muammosi hal etish – bu *matnni tushunish, ma’lumot olish, mashinali tarjima qilish, kayfiyatni tahlil qilish* yoki *hujjatlarni umumlashtirish* kabi ko‘plab NLP ilovalarida qo‘llanilishi mumkin bo‘lgan muhim NLP vazifasi [1]. Koreferensiya muammosini hal etish kompyuterlar matndagi noaniq ifodalar qatnashgan jummalarni tushunishi osonroq bo‘lgan aniq ifodalarga almashtirishning zamonaviy usuli hisoblanadi. Quyidagi 2-rasmda o‘zbek tili matnlarida koreferensiya muammosi hal etish bosqichlari keltirilgan.



2-rasm. Koreferensiya muammosi hal etish bosqichlari

Koreferensiya va anaforik munosabat

Shuni ta'kidlash kerakki, koreferensiya muammosini hal qilishda *matndagi havola (eslatma)lar* va *ular orasidagi koreferent bog'liqlikni aniqlash* lozim. Biroq, matnda bir nechta turdagi havolalar mavjud bo'lib, ularning ta'riflari munozarali masaladir.

Koreferensiya muammosini hal qilishda ajralib turadigan bir holat **anafora muammosini hal qilish (anaphora resolution, AR)** hisoblanadi. Matnda bir atama ikkinchisiga ishora qilganda, ikkinchisining yagona talqinini belgilaganda anafora koreferentlikni (anaforada har doim ham koreferentlik xususiyati namoyon bo'lmaydi) yuzaga keltiradi. [2].

Quyida keltirilgan misolda (1) va (2) ifodalar to'g'ridan-to'g'ri turli xil real obyektlarni nomlaganini ko'rish mumkin. Ammo bu ifodalar bir jumlada qo'llanganligi sababli, (2)ning talqini (1)ga tayanadi. Ushbu eslatmalar o'zaro koreferent bo'lmasa-da, anaforik munosabatni hosil qiladi.

Aziza ¹ tug'ilgan kuniga yaqin qolganda, **uyini** ² chinniday tozaladi.

Anaforani hal etish masalasi koreferensiyani bartaraf qilishdan farq qilsa-da, aksariyat hollarda biri boshqasiga teng keladi.

NLPda koreferensiyani hal qilishdagi asosiy terminlar ta'rifini keltiramiz:

– **Anafora** – o'zidan avval keltirilgan shaxs, narsa, holat kabilarni ko'rsatuvchi, ishora qiluvchi, antesedentga bog'liq so'z[3]. Anaforik so'z real borliqdagi bir obyektga yoki turli obyektlarga ishora qilishi mumkin. O'zbek tilshunoslida Azim Hojiyev anaforik so'z sifatida olmoshlarning ba'zi turlarini ko'rsatib o'tgan [4]. Lekin anaforik munosabatlar nafaqat olmoshlar vositasida, balki qarindoshlikni ifodalovchi nomlar, kasb-hunar nomlari, lavozim nomlari kabi so'zlar vositasida ham yuzaga keladi.

– **Katafora** – bu anaforaning bir ko'rinishi bo'lib, olmosh u nazarda tutgan otdan oldinroq ifodalanishi bilan farqlanadi [3].

– **Koreferent** – real borliqdagi bir obyektga ishora qiluvchi nomlar [3].

Anafora va Katafora

Anafora va kataforaning asosiy farqi shundaki, *anafora gapda u nazarda tutgan so'zdan keyin, katafora esa undan oldin joylashadi*. Anaforadan oldin joylashgan nom **antesedent**, kataforadan keyingi so'z esa **postsedent** deb ataladi.

O'zbekiston terma komandasi JChda g'alabani qo'lga kiritdi. Bu ularning ilk musobaqasi edi.
antecedent *anafora*

Maktabda uning baholari juda yomon bo'lishiga qaramay, Tomas Edison buyuk ixtirochi bo'ldi.
katafora *postcedent*

Antesedentlarni ajratish

Bunda (2) olmoshi bir nechta oldingi (1) ni anglatadigan anaforik ifoda.

Hamid Ne'matov va Asqar G'ulomov ¹ - buyuk tilshunos olimlar edi.

Ular ² ta'limga sezilarli hissa qo'shishda hamkorlik ham qilishgan.

Noaniq olmoshli havolalar

Ba'zi matnlarda ifodalangan olmosh va boshqa nomlar o'rtasida hech qanday bog'liqlik bo'lmasa-da, koreferensiya muammosini hal etish algoritmini yaratishda ushbu turdagi havolalarni e'tibordan chetda qoldirmaslik kerak. Bunday havolalarga gapdan chiqarilganda ham mazmunga putur yetkazmaydigan olmoshlarni, badiiy matnda, so'zlashuv uslubida ta'sirchanlikni oshirish, ta'kidni kuchaytirish maqsadida o'zi ko'rsatayotgan obyekt nomi bilan yonma-yon qo'llangan olmoshlarni, pleonastik qo'llangan olmoshlarni, rasmiy ish uslubida buyruq va talabnomalarda qat'iylikni oshirish uchun ishlatilgan olmoshlarni, tarjimai hol, tushuntirish xati yozilganda atoqli ot bilan yonma-yon kelgan olmoshlarni kiritishimiz mumkin.

Koreferensiyani hal qilish bosqichlari

Yuqorida keltirilgan fikr mulohazalardan kelib chiqqan holda, o'zbek tili matnlarida koreferensiyani hal qilish uchun quyidagi 3-bosqichdan iborat amallarni bajarish lozim:

1-bosqich. *Potentsial so'z, birikma(nomzod)larni aniqlash.*

2-bosqich. *Nomzodlarni guruhga ajratish.*

3-bosqich. *Olmoshlarni haqiqiy obyekt nomlari bilan almashtirish.*

1-bosqich. *Potensial nomzodlarni aniqlash.*

Koreferensiya muammosini hal qilishning dastlabki bosqichida alohida **so'zlar/tokenlar** yoki **spanlarni** qayta ishlash lozim [5].

Span nima? Ko'p hollarda, matndagi olmosh bog'langan obyekt bir nechta qo'shni tokenlardan iborat bo'ladi. Shu sababli koreferensiyani hal qilish NLP vazifasida span sifatida gapdagi o'zaro bog'liq tokenlar ketma-ketligi (otli birikmalar) tushuniladi. NLP vazifalarida aksariyat hollarda matndagi **NER** obyektlar **span** sifatida talqin qilinadi [6]. Shunigdek ko'plab **eslatma/havola**

termini ham span o'rnida qo'llaniladi. Ko'p hollarda bir necha tokenlardan iborat eslatmalardagi element (so'z)lar yaxlit ma'noni anglatadi.

1- bosqich amalga oshirilishi natijasida **olmoshlar**, ular ishora qilayotgan **nomzod spanlar ro'yxati** aniqlanadi.

2-bosqich. *Nomzodlarni guruhga ajratish*

Ushbu bosqichda nomzod spanlar muayyan shartlar va qoidalar asosida guruhlarga birlashtiriladi [7,8]. Nomzod spanlarni birlashtirish **klasterlash** yoki **guruhlash** jarayoni deb ataladi. Umumiy guruhga mansub spanlar umumiy ma'noni ifodalashi lozim. NLPdagi so'zlarni klasterlash *kutubxonadagi kitoblarni toifalash* vazifasiga o'xshaydi. Bugungi kunda guruhlash (toifalash)ning ko'plab usullari mavjud bo'lib, guruhdagi element (span)lar o'xshash xususiyatlarga ega bo'lishi va boshqa klasterlardan farq qilishi kerak.

3-bosqich. *Olmoshlarni haqiqiy obyekt nomlari bilan almashtirish.*

Matndagi **[men, mening]** va **[u]** olmoshlari mos ravishda “**Akbar**” va “**Tohir Malik**” real referent nomlari bilan almashtirilishi natijasida, olmoshsiz semantik jihatdan o'zgarmagan matn hosil qilinadi.

“**Men** **Tohir Malikni** tanlayman, chunki **u** **mening** qiziqishlarimga mos asarlar yozgan” – dedi **Akbar**

Xulosa

O'zbek tili matnlarida koreferensiyani hal qilish NLP vazifasining maqsadi har qanday noaniq ifodalarni ular ishora qilayotgan obyektlarning real nomlarini *aniqlash, guruhlash* va *almashtirishdan* iborat. Ushbu maqolada koreferentlik va anaforik munosabatlarning o'xshash jihatlari hamda farqli tomonlari muhokama qilindi. Shuningdek, ularning o'zbek tili matnlaridagi koreferensiya va anafora muammolarini hal qilish usullari keltirildi.

O'zbek tili matnlarida koreferensiyani hal qilish uchun *potensial so'z, birikma(nomzod)larni aniqlash, nomzodlarni guruhga ajratish* va *olmoshlarni haqiqiy obyektlar bilan almashtirish* bosqichlari amaliy misollar asosida tushuntirildi. Keyingi ilmiy tadqiqot ishlarida o'zbek tili matnlarida koreferensiyani hal qilish modellari haqida batafsil ma'lumot keltiriladi.

Foydalanilgan adabiyotlar:

1. Marta Maslankowska, Paweł Mielniczuk. Intro to coreference resolution in NLP.
2. Christopher Manning. Stanford lecture (CS224n). 2019
<https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1162/handouts/cs224n-lecture10-coreference.pdf>
3. R. Mitkov. Anaphora Resolution. Pearson Education, 2002.
4. Хожиев А. Тилшунослик терминларининг изоҳли луғати. – Тошкент: Ўзбекистон миллий энциклопедияси, 2002. – 168 б.



5. B.B.Elov, Sh.M.Khamroeva, Z.Y.Xusainova. NLP (tabiiy tilga ishlov berish)ning pipeline konveyeri (2022).
6. B.Elov. Tabiiy tilni qayta ishlash (nlp)da spacy modulidan foydalanish. ИЛМ-фан ва инновацион ривожланиш, 2022 (4). 41-55
7. Urbizu, G., Soraluze, A., & Arregi, O. (2020). Sequence to Sequence Coreference Resolution. Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference, Crac.
8. Stylianou, N., & Vlahavas, I. (2021). A neural Entity Coreference Resolution review. In *Expert Systems with Applications* (Vol. 168). <https://doi.org/10.1016/j.eswa.2020.114466>