

UZBEKISTAN

O'ZBEKISTON

LANGUAGE & CULTURE

TIL VA MADANIYAT

**KOMPYUTER
LINGVISTIKASI**

ISSN 2181-922X

2023 Vol. 1 (6)

www.compling.tsuull.uz

MUNDARIJA

Eşref Adalı

Corpus for what.....6

Victor Zakharov

Functionality of the russian national corpus.....18

Botir Elov, Dilrabo Elova

NLPda koreferens masalasi.....27

Botir Elov, Shahlo Hamroyeva, Oqila Abdullayeva,

Zilola Xusainova, Nizomaddin Xudayberganov

O'zbek, turk va uyg'ur tillarida pos

teglash va stemming.....40

Dilrabo Elova, Sabohat Allanazarova

O'zbek tili matnlarida sentiment tahlil usullari.....65

Oqila Abdullayeva, Sabura Xudayarova

O'zbek tilida so'z birikmalarining lisoniy sintaktik qoliplari va

ularni modellashtirish masalasi77

Xolisa Axmedova

Statistik usullar yordamida turli so'z turkumlari orasidagi

omonimiyani aniqlash.....91

NLPDA KOREFERENS MASALASI

Botir Elov¹,
Dilrabo Elova²

Annotatsiya

Qidiruv tizimlarida amalga oshirilgan so'rovni bajarish uchun sun'iy intellekt metodlari ma'lumotlar bazasidagi matnlarni avtomatik qayta ishlash NLP (tabiiy til) vazifasini samarali hal etishni talab qiladi. Mashinali o'qitishni amalga oshirishning muhim vazifalaridan biri qoidalar tizimini ishlab chiqishdir. Ushbu maqolada o'zbek tili matnlaridagi gaplarda ifodalangan mazmuniy masofalarni avtomatik aniqlashga qaratilgan dastlabki tadqiqotlardan hisoblangan koreferensiya muammosi haqida fikr yuritiladi. Shuningdek, koreferensiya muammosini hal etishning asosiy qoidalari, bosqichlari tahlil qilinadi. Koreferensiya muammosini yuzaga keltiradigan anaforok munosabatlar va ularning asosiy turlari, asosiy qo'llanadigan anafora, antisedent, katafora, klasterlash, singleton kabi bir qator terminlarning izohlari keltiriladi.

Kalit so'zlar: *NLP, Coreference Resolution, Mention Detection, NER, pipeline konveyeri, tokenizatsiya, koreferensiya, atoqli otlar, olmoshlarni aniqlash.*

Kirish

Ushbu maqolada o'zbek tili matnlaridagi koreferensiya masalasini ko'rib chiqamiz. Bu NLPning nutq tahlili qismiga tegishli bo'lgan vazifalardan biri hisoblanib, matndagi gaplar yaxlit bir tadqiqot obyekti tashkil etadi. Matndagi koreferensiya masalasini hal qilish – NLPning klasterlash vazifasi bo'lib, u gap/hujjat/

¹ *Elov Botir Boltayevich* – texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti.

E-pochta: elov@navoiy-uni.uz

ORCID: 0000-0001-5032-6648

² *Elova Dilrabo Qudratilloevna* – filologiya fanlari falsafa doktori. Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti.

E-pochta: elova@navoiy-uni.uz

ORCID: 0000-0002-2329-1811

korpusdagi bir obyekt/hodisaga ishora qiluvchi barcha nomlar (bir referentni nomlovchi barcha birliklar) orasidagi mazmun masofasini aniqlashga xizmat qiladi [McShane, Nirenburg, 2021; Toshniwal, Xia, Wiseman, 2021].

Obyektlarga tashkilot, shaxs, joy, predmet nomi kabilarni ifodalovchi til birliklarini misol sifatida keltirish mumkin. *NLPdagi savol-javob tizimlari, hujjatlarni avtomatik umumlashtirish* yoki *mashina tarjimasi* kabi ko'plab tabiiy til ilovalari boshlang'ich qadamida matndagi obyektlarga havolalarni aniqlash lozim bo'ladi [Urbizu, Soraluze, Arregi, 2020; Sukthanker, Poria, Cambria, 2020; Stylianou, Vlahavas, 2021]. Koreferensiya masalasini tushunib olish uchun quyidagi misolni ko'rib chiqamiz:

Nilufarning aytishicha, Azimjon juda iltifotli yigit, chunki u har doim Azizani ishga o'zi olib keladi, mashinaning eshigini ham o'zi ochib, ayolini ishxonasiga kuzatib qo'yadi.

Ushbu namunadagi bir referentga ega subyektni nomlashi mumkin bo'lgan barcha eslatmalarni alohida rangda ajratib ko'rsatamiz. Bular:

Azimjon, u, o'zi, Aziza, ayol

Demak, ushbu ko'rsatilgan eslatmalarning aynan qaysilari bir referentga egaligini avtomatik aniqlash koreferensiya muammosini hal qilishning mohiyatini ifodalaydi.

Matndagi koreferensiya muammosini avtomatik aniqlash vazifasi quyidagi ikki bosqichda amalga oshirilishi mumkin: (kichik vazifa (bosqich)ga ajratish mumkin [Kantor, Globerson, 2020; Dobrovolskii, 2021]:

1. *obyektga havola nomlarni aniqlash; (obyektni eslatuvchi nomlar-eslatma)*

2. *klasterlash.*

1. Obyektga havola nomlar (eslatmalar)ni aniqlash

Ushbu bosqichning asosiy maqsadi obyektlarga ishora qilishi mumkin bo'lgan barcha **nomzodlarni aniqlash**dir. Masalan, quyidagi misolda nomzod ilovalarni ajratib ko'rsatamiz:

Nilufarning aytishicha, Azimjon juda iltifotli yigit, chunki u har doim Azizani ishga o'zi olib keladi, mashinaning eshigini ham o'zi ochib, ayolini ishxonasiga kuzatib qo'yadi.

Ushbu bosqichda uch xil havola aniqlanadi [Lata, Singh, Dutta, 2022; Cruz, Rocha, Cardoso, 2020; Ji, Gao, Fei, Teng, Ren, 2020]:

- *olmoshlar aniqlanadi;*

- *atoqli otlar (NER) aniqlanadi;*
- *otli birikma (noun-phrases, NP)lar aniqlanadi.*

Olmoshlarni aniqlash

Olmosh mustaqil ma'noli so'zlar – *ot, sifat, son, ravish* kabi *shaxs, predmet, belgi* yoki miqdorni bildirmay, *so'zlar, so'z birikmasi* va *gap* o'rnida qo'llanuvchi, ularga ishora qiluvchi so'z turkumidir. Olmoshlar noaniq va umumiy ma'noga ega bo'lganligi sababli, ularning asosiy ma'nosi va qaysi so'z turkumi o'rnida qo'llanishi matn kontentiga ko'ra aniqlanadi. Olmoshlarning ushbu ma'noviy xususiyatini quyidagi matn orqali ko'rish mumkin:

Zamira bugun uyda. **U** hovli-joylarni supurib, kir yuvishga tutindi. **Guldor** kiyimlarni alohida ajratdi, chunki **bunday** kiyimlarni boshqalariga aralashtirib bo'lmaydi. **Bitta** sovunni maydalab kir mashinasiga solar ekan, "Oyim ham **shuncha** sovun ishlatarmidilar", - deb o'yladi.

Ushbu matndagi *u, bunday, shuncha* olmoshlari *Zamira (ot), guldor (sifat), bitta (son)* o'rnida qo'llanib, gapda ham ular bajargan vazifani bajarmoqda. Biroq ushbu gaplar matndan ajratib olinsa, ularning ma'nosi mavhum bo'lib qoladi. Olmoshlar anglatadigan nisbiy ma'nolari va grammatik belgilariga ko'ra 7 xil turga bo'linadi: [Sayfullayeva, Mengliyev va b., 2005]

1. *Kishilik olmoshlari* - shaxs ma'nosini bildiruvchi otlar o'rnida almashinib keladi va ularga ishora qiladi.
2. *Ko'rsatish olmoshlari* - Mustaqil so'zlar orqali ifodalangan shaxs, predmet belgi, harakatlarni ko'rsatib keladi.
3. *So'roq olmoshlari* –shaxs, narsa, belgi-xususiyat, harakat-hoatlarni so'rash uchun ishlatiladigan olmoshlar.
4. *Birgalik-ajratish olmoshlari* - mustaqil so'zlar o'rnida kelib, uning ma'nosini jamlab, umum-*lashtirib* keluvchi olmoshlar.
5. *Bo'lishsizlik olmoshlari* - Shaxs-narsa, belgi-xususiyat bildiruvchi mustaqil so'zlar o'rnida kelib, uning inkorini bildiradi.
6. *O'zlik olmoshi* - uchala shaxsdagi kishilik olmoshlari o'rnida qo'llanib, ma'nosini ta'kidlab keladi.
7. *Gumon olmoshi* - Shaxs-narsa, belgi-xususiyat bildiruvchi mustaqil so'zlar o'rnida kelib, uning ma'nosini taxminlab, gumonlab keluvchi olmoshlar.

Olmoshlar turlari 1-jadvalda keltirilgan:

1-jadval. O'zbek tilida olmoshlar

Kishilik olmosh-lari	Ko'r-satish olmosh-lari	So'roq olmosh-lari	Birgalik olmosh-lari	Bo'lish-sizlik olmosh-lari	O'zlik olmoshi	Gumon olmoshi
men	U	kim?	hamma	hech kim	o'z	kimdir
sen	bu	nima?	barcha	hech nima		nimadir
u	shu	qaysi?	bari	hech qan- cha		qaysidir
biz	o'sha	qancha?	har kim	har qan- day		allakim
siz	mana bu (manavi)	necha?	har nima	har qaysi		allanarsa
ular	mana shu	nimaga?	har qancha	hech bir		allaqanday
	ushbu	qani?	har bir			

Olmosh – ot, otli birikmalarning o'rnini bosadigan va odatda anaforani o'z ichiga olgan so'z bo'lib, unda olmoshning ma'nosi oldingi gap (**antecedent**)ga bog'liq bo'ladi. Ammo yuqorida ko'rsatilgan olmoshlarning barcha turlari anaforik munosabatlarni hosil qilavermaydi. Ba'zi olmoshlar (3-shaxs *kishilik olmoshi*, *ko'rsatish olmoshlari*, *o'zlik olmoshi*) tilshunoslikda **anaforani** ifodalaydi va antesedentga bog'liq bo'ladi. Masalan:

A'zam bilan Toshkentda ko'rishgan edim, **bu** xabarni **o'sha** aytib berdi.

Bu jumladagi **“o'sha”** – ko'rsatish olmoshi va **“A'zam”** so'zi bitta referentga ega bo'lib, olmosh jumlaning avvalgi qismidagi atoqli otga ishora qilyapti.

O'zbek tilidagi anaforani ifodalaydigan olmoshlar quyidagi jadvalda keltirilgan:

2-jadval. O'zbek tilida olmoshlar

Kishilik olmoshlari	Ko'rsatish olmoshlari	O'zlik olmoshi
	<i>u</i>	<i>o'z</i>
	<i>o'sha</i>	
<i>u, ular</i>	<i>bu, ushbu</i>	
	<i>shu</i>	

Atoqli otlar (NER)ni aniqlash

Ayrim shaxs yoki narsa-buyumga atab qo'yilgan nomlar atoqli otlar deyiladi va ular bosh harflar bilan yoziladi. Ularga quyidagilar kiradi:

- **Kishilarning ismi-sharifi va taxalluslari:** Dilmurod, Ravshan Mahmudov, Zarifa Hakimjon qizi, Cho'lpon, Julqunboy kabi.
- **Uy hayvonlariga atab qo'yiladigan nomlar:** Olapar, To'rtko'z, Mosh, Boychibor, Yo'lbars kabi.
- **Geografik va astronomik nomlar:** O'sh, Chirchiq, Oqtosh, Kavkaz, Dnepr, Mars, Oy, Quyosh kabi.
- **Respublikalar va yuqori tashkilotlarning nomlari:** O'zbekiston Respublikasi, Vazirlar Mahkamasi, Yevropa Taraqqiyot va Tiklanish Banki, Oliy Kengash kabi.
- **Ilmiy muassasalar, oliy o'quv yurtlari, vazirlik va korxonalar nomining birinchi qismi:** Til va adabiyot instituti, Moliya vazirligi, Bank-moliya akademiyasi kabi.
- **Yuqori lavozim, oliy faxriy unvon nomlari:** O'zbekiston Prezidenti, Bosh Vazir, O'zbekiston Qahramoni kabi.
- **Fabrika, zavod, turli tashkilotlar, jamoa xo'jaliklari, kino, teatr, kitob, gazeta, jurnal nomlari qo'shtirnoqqa olinib, bosh harf bilan yoziladi:** "Qizil tong" fabrikasi, "Sharq yulduzi" jurnali, "Umid" jamg'armasi, "Muqimiy" teatri kabi.
- **Tarixiy voqealar, sanalar, ilmiy anjuman, hujjatlar nomlarining birinchi qismi:** Mustaqillik kuni, Toshkent deklaratsiyasi, Qurbon hayiti kabi.
- **Bosh harflari qisqartirilgan qisqartma otlarning birinchi so'zi:** JIDU Jahon iqtisodi va diplomatiya universiteti, BMA Bank-moliya akademiyasi kabi.

NER modeli orqali strukturlanmagan matndan atoqli otlar aniqlanadi va toifalarga ajratiladi [Elov, Samatboyeva, 2023]. Masalan shaxs nomlari, tashkilot nomlari, joy nomlari, mahsulot nomlari kabi. NER obyektlarini aniqlash jarayoni quyidagi bosqichda amalga oshiriladi:

1. *Ma'lumotni ajratib olish* – NERni aniqlashda birinchi qadam gapda, paragrafda, matnda ko'rsatilgan obyektlarni ajratib olishdir. Bu bosqichda butun matn belgilanib, matnlar chegarasi aniqlanadi. Bu o'rinda bosh harflar (Capital letter) tizimiga ko'ra matn gaplarga ajratiladi.
2. *Tokenizatsiya jarayoni* – matndan bo'laklarga ajratilgan gaplar endi o'z ichida tokenlarga ajratiladi.
3. *"IOB" yoki "BILUO" sxemasiga ko'ra tokenlar chegarasini aniqlab, ularni qayta "yig'ish"* – bu o'rinda model asosida bir necha tarkibli NERlar tokenlari "birlashtiriladi".
4. *Obyektlarni qidirish* – NERdagi navbatdagi jarayon tokenlar

tarkibidagi NER obyektlarni qidirish.

5. Aniqlangan NER obyektlariga to'g'ri toifa berish.

Yuqoridagi jarayonni boshqacha tushuntirganda, aniqlangan otlar nafaqat morfologik, balki semantik tahlil ham qilinadi. Matn tarkibidan ajratilgan otlar "atoqli ot" (proper noun)lik xususiyatiga ko'ra aniqlanadi. Atoqli otlar "NER xususiyatlari" (matn tarkibida bosh harfda kelishi, lug'atdan tashqari leksika ekanligi, qo'shimchalarning undan so'ng qo'shilishi, obyektning qayta "nom" bilan atalishi va h.)ga ko'ra tahlil qilinadi. Lug'atlar bazasi ham NER obyektlarini aniqlashda yordamchi vosita sifatida xizmat qiladi. Ammo bu baza mukammal samarali yechim bo'la olmaydi.

NER – matndagi asosiy obyektlarni ajratib olish va ularni oldindan belgilangan toifalarga ajratishi mumkin bo'lgan NLP usuli. Matndan *shaxs ismlari, joylashuv nomlari, kompaniya nomlari* va shu kabi lug'atda mavjud bo'lmagan nomlangan obyektlarni aniqlash jarayoni ko'plab NLP vazifalarini hal qilishda muhim bosqich sanaladi. NLPda nomlangan obyektни tanib olish, odatda, *obyektни identifikatsiyalash, obyektни ajratib olish yoki obyektни bo'laklarga ajratish* deb ham yuritiladi. NER obyektlarini aniqlash algoritmlari quyidagi modellar hisoblanadi [www.analyticsvidhya].

- *qoidalarga asoslanib tahlil qilish;*
- *lug'at qidirish;*
- *POS tagging (morfologik teglash);*
- *Parsing (sintaktik teglash).*

Matn tarkibidan NER obyektlarini aniqlash jarayonini tushunish uchun quyidagi gapni ko'rib chiqamiz:

Toshkent LOC – O'zbekistonning LOC poytaxti va eng yirik shahri, aholisi bo'yicha Markaziy Osiyodagi LOC eng katta shahar.

1-rasm. NER obyektlarini aniqlash

Bu yerda ko'k ranglilar **ot** so'z turkumidagi so'zlar. Ushbu otlarning ba'zilar dunyoda mavjud bo'lgan real obyektlarni ifodalaydi. Masalan, yuqoridagilardan quyidagi otlar xaritada mavjud joylarni ifodalaydi: "Toshkent"; "O'zbekistonning"; "Markaziy Osiyodagi".

Agar biz matn tarkibidan otlarni – xususan, **nomlangan otlarni** mana shunday tog'ri natija bilan topa olsak, ushbu ma'lumotlar bilan NLPda matndagi nomlangan obyektlar ro'yxatini avtomatik ravishda aniqlay olishimiz mumkin. Demak, NERning maqsadi bu otlarni o'ziga mos real-dunyo tushunchalari bilan

aniqlash va belgilashdir. (Masalan: Unicef – tashkilot (ORG); Alisher – odamning ismi (PER))

NER tizimlari shunchaki oddiy lug'atdan qidirishni amalga oshirmaydi. Balki, ular matn tarkibidagi gapda so'zning qanday paydo bo'lishi va bu so'z qaysi turdagi otni ifodalashini aniqlash uchun statistik modeldan foydalanadi [www.analyticsvidhya]. O'zbek tilidagi NER obyektlarini aniqlash masalasi B.Elov va M.Samatboyevalar tomonidan tahlil qilingan [Elov, Samatboyeva, 2023].

Otli birikmalar (Noun-phraseslarni) aniqlash

O'zbek tilida tobe hokim munosabatda erkin bog'langan, bosh so'zi ism bilan ifodalangan so'z birikmalari otli birikmalardir. Masalan: shamoldan tez, qiziqarli kitob, aqliy bilish, tezkor xotira kabi. Ko'rinib turibdiki, o'zbek tilida otli birikmalar deganda, bosh so'z ot so'z turkumidan tashqari sifat, son, ba'zi ravishlar, harakat nomi kabi so'zlar ifodalanishi mumkin. Koreferensiya muammosini hal qilish aniq nomlarga qaratilgani sababli, faqat shu nomlarning modifikatsion belgilarini ifodalaydigan otli birikmalarga e'tibor qaratamiz. Ya'ni nom va uni tasvirlashga xizmat qilayotgan so'zlar bilan ifodalangan atributiv munosabatli birikmalarni aniqlaymiz.

Masalan: yosh, iste'dodli, mas'uliyatli xodim.

2. Klasterlash

Birinchi bosqichdagi vazifalarni amalga oshirgandan keyin bir obyektga ishora qiluvchi nomzodlar/eslatmalarni aniqlashga harakat qilamiz. Kontentda ifodalangan ishora qiluvchi nomzodlar/eslatmalarni matnda ko'rsatilgan obyektlarga mos keladigan klasterga birlashtirish jarayonini amalga oshiramiz.

Masalan: *Nilufarning aytishicha, **Azimjon** juda iltifotli yigit, chunki **u** har doim **Azizani** ishga **o'zi** olib keladi, mashinaning eshigini ham **o'zi** ochib, **ayolini** ishxonasiga kuzatib qo'yadi.*

Yuqoridagi namunada koreferensiyani yuzaga keltirgan nomzod/eslatmalar guruhini alohida ranglarda belgiladik. Ushbu gapda **Azimjon** va **u, o'zi** hamda **Aziza** va **ayol** bir xil obyektga tegishli bo'lib, har bir obyektни ifodalayotgan nomzod/eslatmalar alohida klaster (guruh)ni ifodalaydi. Kontentda havolaga ega bo'lmagan *otlar* yoki *obyektlar* nomi singleton (faqat bitta birlikni o'z ichiga olgan klaster) deb ataladi.

Asosiy terminlar

Korpusdakoreferensiya masalasini hal qilishda qo'llaniladigan atamalarni quyida keltirilgan gaplar misolida izohlaymiz:

Qizaloq bog'dan chiqishni istamadi, chunki **u** arg'imchoqlarning barchasida uchmoqchi edi.

- **Antesedent – jumlaning/matnning/kontentning oldingi qismida qatnashgan obyekt** – matndagi keyingi olmoshning ma'nosini aniqlashtiruvchi birlik. Masalan, yuqoridagi misolda [**u**] olmoshining ma'nosi [**qizaloq**]da aniqlashadi. Shuning uchun [qizaloq] [u]ning antesedenti hisoblanadi.

- **Anafora (Anaphora)** – kontekstdagi shaxs yoki narsaga (obyekt), oldin aytilganga (ifodalangan obyektga) ishora qiluvchi birlik. Misol uchun,

Tort yesang, muzlatkichda **bir bo'lak** bor. Jumlaning keyingi qismidagi [**bir bo'lak**] so'zi [**tort**] so'ziga ishora qilayotgan anaforik birlik.

- **Katafora (Cataphora)** – bu anaforaning bir ko'rinishi bo'lib, olmosh u nazarda tutilgan otdan oldinroq ifodalanishi bilan farqlanadi. Misol uchun,

*Negadir **u** musobaqada kutilgan natijani ko'rsata olmadi, **Anvar** odatda bunday tadbirlarga puxta tayyorlanar edi.*

Bu jumladagi [**u**] ishora qiluvchi olmosh, ishora qilayotgan [**Anvar**] so'zidan (obyektdan) oldin joylashgan.

Umumiy xususiyatlar

Tildagi *grammatik qoidalardan* foydalangan holda eslatmalar o'rtasida asosiy aloqalarni aniqlashga yordam beradigan ba'zi umumiy xususiyatlar mavjud bo'lib, ularni quyidagi misollar orqali ko'rib chiqamiz.

Olmoshga yaqin masofada ifodalangan obyektlarga havola qilinish holati. Masalan, *Jasmin ingliz tili kurslariga qatnaydi, endi **Malika** ham unga qo'shildi, chunki **u** ham ingliz tilini o'rganmoqchi.*

Ushbu gapda [**u**] havolasi (eslatmasi) [**Malika**]ga ishora qiladi, chunki gapda ushbu so'z [**Jasmin**]ga nisbatan yaqinroq joylashgan.

Gapda subyekt pozitsiyasi obyekt pozitsiyasiga qaraganda ko'proq havola qilinadi. Masalan, *Jasmin ingliz tili kurslariga qatnaydi, endi Malika ham **unga** qo'shildi, chunki u ham ingliz tilini o'rganmoqchi.*

Bu holda [**unga**] havolasi [**Jasmin**]ga ishora qiladi, chunki [**Malika**] obyekt pozitsiyasida, [**Jasmin**] esa subyekt pozitsiyasidir.

Parallellik

Nazokat Feruza bilan kinoga bordi, **Aziza** esa **u** bilan savdo markaziga bordi.

Ushbu gapda **bilan** so'zi parallel tuzilmani yaratganligi sababli, **u Nazokat**ga ishora qiladi. Quyidagi gapda "**uning**" so'zi "**talabalarga**" so'ziga emas, balki "**professor**" so'ziga tegishli ekanligini qayd etish kerak:

Professor navbatdagi ma'ruzani o'qigach, **talabalarga uning** mavzulari yakunlanganini, mashg'ulotlar nihoyalaganini, endi **ular** imtihonga puxta tayyorgarlik ko'rishlari kerak ekanligini e'lon qildi.

1-rasm. Gapdagi so'zlarning o'zaro semantik bog'lanishi

Fe'l semantikasi

Quyidagi ikkita gapni solishtiramiz:

- *Qizim qunt bilan **tort** pishirdi va endi **u** san'at darajasiga yetdi.*

- ***Qizim** qunt bilan **tort** pishirdi, endi **u** juda charchadi.*

Strukturasiga ko'ra yuqoridagi gaplar bir xil qurilishga ega. Ammo bu gaplarning har birida "**u**" olmoshi nimani (qaysi obyektni) anglatishini to'g'ri aniqlash lozim.

Semantika bizga gapdagi eslatmalar o'rtasidagi bog'lanish haqida ham ma'lumot berishi mumkin.

"The animal didn't cross the street because it was too tired."

"The animal didn't cross the street because it was too wide."

Birinchi gapda **it** so'zi hayvonga, ikkinchi gapda esa ko'chaga tegishli ekanligini **tired** va **wide** so'zlari orqali bilamiz.

Qo'shimcha xususiyatlar

Boshqariladigan klassifikatorlarni yaratishga yordam beradigan yana bir nechta xususiyatlar (**PNG cheklovlari**) mavjud:

- *shaxs (1-shaxs, 2-shaxs, 3-shaxs);*
- *raqam (bir yoki ko'p);*
- *jins (erkak yoki ayol).*

O'zbek tili matnlaridagi koreferensiya

An'anaviy usul orqali berilgan matnda **pipeline** konveyeri jarayoni orqali birinchi navbatda barcha **olmoshlarni aniqlash** lozim. Buning uchun B.Elov, R.Alayev va Sh.Hamroyeva tomonidan ishlab chiqilgan **o'zbek tili morfoanalizatori** orqali matnda POS teglash jarayonini amalga oshirish kerak. Ikkinchi qadamda esa B.Elov va M.Samatboyeva tomonidan ishlab chiqilgan matndagi **NER obyektlarni aniqlash tizimidan** foydalanib matndagi bizga zarur NERlar ro'yxati aniqlanadi. Keyingi qadamda **sintaktik analizator** va **atoqli otlarga havolalarni aniqlash** va **koreferensiyalarni klasterlash** tizimlaridan foydalaniladi. Natijada 5 bosqichdan iborat ketma-ketlik orqali o'zbek tili matnlaridagi koreferensiyaga o'zaro bog'liq so'z va so'z birikmalari to'plami aniqlanadi.

Xulosa

Ushbu maqolada o'zbek tili matnlaridagi koreferensiya muammosini hal qilishning asosiy shartlari, bosqichlari tavsiflandi. Milliy tilni o'zida aks ettirish, uni elektron ko'rinish, raqamli formatda saqlash vazifasini bajaradigan elektron kutubxonalar, til korpuslari kabi mahsulotlar bugungi globallashuv davrining zaruriy vositalaridir. Tabiiy tilga ishlov berish, uni mashina tili darajasiga olib chiqish tilning yashovchanligini, mavqeyini oshirishga sezilarli darajada xizmat qiladi. Buni amalga oshirish uchun o'zbek tili matnlarini avtomatik tahlil qiladigan semantik analizatorlar, savol-javob tizimlari, mashina tarjimasini kabi NLP ilovalari amalda ishlab chiqilishi lozim. E'tibor qaratganimiz, koreferensiya muammosini avtomatik hal qilish ana shunday vazifalarni amalga oshirishda lozim bo'ladigan dolzarb masalalardandir.

Foydalanilgan adabiyotlar

- McShane, M., & Nirenburg, S. (2021). Basic Coreference Resolution. In *Linguistics for the Age of AI*. <https://doi.org/10.7551/mitpress/13618.003.0007>
- Toshniwal, S., Xia, P., Wiseman, S., Livescu, K., & Gimpel, K. (2021). On Generalization in Conference Resolution. 4th Workshop on Computational Models of Reference, Anaphora and Conference, CRAC 2021 - Proceedings of the Workshop. <https://doi.org/10.18653/v1/2021.crac-1.12>
- Urbizu, G., Soraluze, A., & Arregi, O. (2020). Sequence to Sequence Conference Resolution. Proceedings of the Third Workshop on Computational Models of Reference,

- Anaphora and Coreference, Crac.
- Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and conference resolution: A review. *Information Fusion*, 59, 139–162. <https://doi.org/10.1016/j.inffus.2020.01.010>
- Stylianou, N., & Vlahavas, I. (2021). A neural Entity Conference Resolution review. In *Expert Systems with Applications* (Vol. 168). <https://doi.org/10.1016/j.eswa.2020.114466>
- Kantor, B., & Globerson, A. (2020). Conference resolution with entity equalization. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. <https://doi.org/10.18653/v1/p19-1066>
- Dobrovolskii, V. (2021). Word-Level Conference Resolution. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*. <https://doi.org/10.18653/v1/2021.emnlp-main.605>
- Lata, K., Singh, P., & Dutta, K. (2022). Mention detection in conference resolution: survey. *Applied Intelligence*, 52(9). <https://doi.org/10.1007/s10489-021-02878-2>
- Cruz, A. F., Rocha, G., & Cardoso, H. L. (2020). Conference resolution: Toward end-to-end and cross-lingual systems. *Information (Switzerland)*, 11(2). <https://doi.org/10.3390/info11020074>
- Ji, D., Gao, J., Fei, H., Teng, C., & Ren, Y. (2020). A deep neural network model for speakers conference resolution in legal texts. *Information Processing and Management*, 57(6). <https://doi.org/10.1016/j.ipm.2020.102365>
- Сайфуллаева Р., Менглиев Б., Курбонова М., Боқиева Г., Юнусова З., Абузалова М. Ҳозирги ўзбек адабий тили. Ўқув қўлланмаси. Тошкент, 2005. 388 б.
- B.Elov, M.Samatboyeva. (2023). Identifying NER (Named Entity Recognition) objects in Uzbek language texts. *Science and Innovation International Scientific Journal*, Volume 2 Issue 4 April 2023 UIF-2022: 8.2 | ISSN: 2181-3337. <https://www.analyticsvidhya.com/blog/2021/06/part-1-step-by-step-guide-to-master-natural-language-processing-nlp-in-python/>

THE PROBLEM OF COREFERENCE IN NLP

Botir Elov¹,
Dilrabo Elova²

Abstract

Artificial intelligence methods for performing a query in search engines require an effective NLP (natural language of processing) task of automatic processing of texts in the database. One of the important tasks of implementing machine learning is the development of a system of rules. This article discusses the problem of coreference calculated from preliminary studies aimed at automatic determination of meaningful distances expressed in sentences in Uzbek language texts. Also, the main rules and stages of solving the coreference problem are analyzed. Explanations of a number of terms such as anaphoric relations and their main types, the main used anaphora, antecedent, cataphora, clustering, singleton, etc., which create the problem of coreference, are given.

Keywords: *NLP, Coreference Resolution, Mention Detection, NER, pipeline pipeline, tokenization, coreference, nouns, pronoun detection.*

References:

- McShane, M., & Nirenburg, S. (2021). Basic Conference Resolution. In *Linguistics for the Age of AI*. <https://doi.org/10.7551/mitpress/13618.003.0007>
- Toshniwal, S., Xia, P., Wiseman, S., Livescu, K., & Gimpel, K. (2021). On Generalization in Conference Resolution. 4th Workshop on Computational Models of Reference, Anaphora and Conference, CRAC 2021 - Proceedings of the Workshop. <https://doi.org/10.18653/v1/2021.crac-1.12>

¹ Elov Botir Boltayevich – Elov Botir Boltayevich – doctor of philosophy of technical sciences (PhD), associate professor. Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.

E-mail: elov@navoiy-uni.uz

ORCID: 0000-0001-5032-6648

² Elova Dilrabo Qudratilloevna – doctor of philosophy in philology. Alisher Navoi Tashkent State University of Uzbek Language and Literature.

E-mail: elova@navoiy-uni.uz

ORCID: 0000-0002-2329-1811

- Urbizu, G., Soraluze, A., & Arregi, O. (2020). Sequence to Sequence Conference Resolution. Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference, Crac.
- Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and conference resolution: A review. *Information Fusion*, 59, 139–162. <https://doi.org/10.1016/j.inffus.2020.01.010>
- Stylianou, N., & Vlahavas, I. (2021). A neural Entity Conference Resolution review. In *Expert Systems with Applications* (Vol. 168). <https://doi.org/10.1016/j.eswa.2020.114466>
- Kantor, B., & Globerson, A. (2020). Conference resolution with entity equalization. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. <https://doi.org/10.18653/v1/p19-1066>
- Dobrovolskii, V. (2021). Word-Level Coreference Resolution. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*. <https://doi.org/10.18653/v1/2021.emnlp-main.605>
- Lata, K., Singh, P., & Dutta, K. (2022). Mention detection in coreference resolution: survey. *Applied Intelligence*, 52(9). <https://doi.org/10.1007/s10489-021-02878-2>
- Cruz, A. F., Rocha, G., & Cardoso, H. L. (2020). Conference resolution: Toward end-to-end and cross-lingual systems. *Information (Switzerland)*, 11(2). <https://doi.org/10.3390/info11020074>
- Ji, D., Gao, J., Fei, H., Teng, C., & Ren, Y. (2020). A deep neural network model for speakers conference resolution in legal texts. *Information Processing and Management*, 57(6). <https://doi.org/10.1016/j.ipm.2020.102365>
- Sayfullayeva R., Mengliyev B., Qurbonova M., Boqiyeva G., Yunusova Z. *Hozirgi o'zbek adabiy tili. – Toshkent, 2005. 388 b.*
- Elov B., Samatboyeva M. (2023). Identifying NER (Named Entity Recognition) objects in Uzbek language texts. *Science and Innovation International Scientific Journal*, Volume 2 Issue 4 April 2023 UIF-2022: 8.2 | ISSN: 2181-3337. <https://www.analyticsvidhya.com/blog/2021/06/part-1-step-by-step-guide-to-master-natural-language-processing-nlp-in-python/>