LINGVISTIK CHEKLOVLAR – KOREFERENTLIKNI AVTOMATIK HAL ETISHNING ASOSI

Abdisalomova Shahlo
ToshDO‘TAU tayanch doktoranti.

Annotatsiya: Tabiiy tilni qayta ishlash (NLP) sun’iy intellektning kichik sohasi bo‘lib, uning maqsadi kompyuterlar va odamlar o‘rtasida o‘zaro aloqalarni o‘rnatishdan iborat. Ma’lumki, tilning o‘ziga xos jihatlari matndan avtomatik ravishda ma’no chiqarish jarayonida qiyinchiliklarni yuzaga keltiradi. Inson ongi bevosita matn mazmunini tushunishga qodir. Ammo sun’iy intellekt uni to‘g‘ri talqin qilishi uchun koreferentlik yuqori aniqlikda hal etilgan bo‘lishi lozim. Bundan tashqari, mashina tarjimasi, savollarga javob berish, matnni umumlashtirish, hissiyotlarni tahlil qilish, matnni tasniflash, nutqni tanish, nomga ega obyektni aniqlash, chatbot kabi sohalar rivoji ham aynan koreferentlikning hal etilishi bilan bog‘liqdir. Ushbu maqolada NLP ning koreferentlikni hal etish vazifasi, uning ahamiyati yoritiladi, lingvistik xususiyatlarning bu jarayondagi ishtiroki masalasi o‘rganiladi.

Kalit so‘zlar: NLP, koreferentlik, antisident, anafor, lingvistik xususiyat, vektor, olmosh, matn.

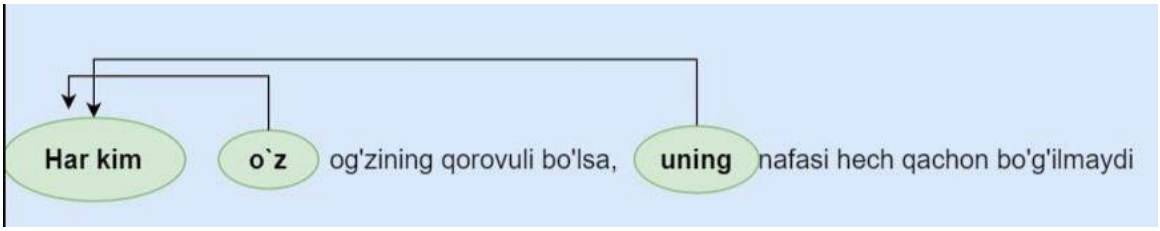
Annotation: Natural Language Processing (NLP) is a subfield of Artificial Intelligence, that aims to facilitate interactions between computers and humans. It is known, that the specific features of the language create difficulties in the process of automatically extracting meaning from the text. The human mind is able to directly understand the content of the text. But in order for the Artificial Intelligence to interpret it correctly, the coreference must be resolved with high accuracy. In addition, the development of fields such as Machine Translation, Question Answering, Text Summarization, Sentiment Analysis, Text Classification, Speech recognition, Named Entity Recognition, Chatbot is also related to the Coreference Resolution. In this article, the Coreference Resolution is a subfield of NLP, its importance is highlighted, problem of the participation of linguistic features in this process is investigated.

Keywords: NLP, coreference, antecedent, anaphor, linguistic feature, vector, pronoun, text

Аннотация: Обработка естественного языка (НЛП) — это область искусственного интеллекта, целью которой является облегчение взаимодействия между компьютерами и людьми. Известно, что особенности языка создают трудности в процессе автоматического извлечения смысла из текста. Человеческий разум способен непосредственно понимать содержание текста. Но для того, чтобы искусственный интеллект правильно интерпретировал ее, кореференция должна быть решена с высокой точностью. Кроме того, с решением кореференции связано развитие таких областей, как машинный перевод, ответы на вопросы, обобщение текста, анализ настроений, классификация текста, распознавание речи, идентификация объекта по имени, чат-бот. В данной статье подчеркивается роль НЛП в разрешении кореференции, ее значение, изучается проблем участия языковых особенностей в этом процессе.

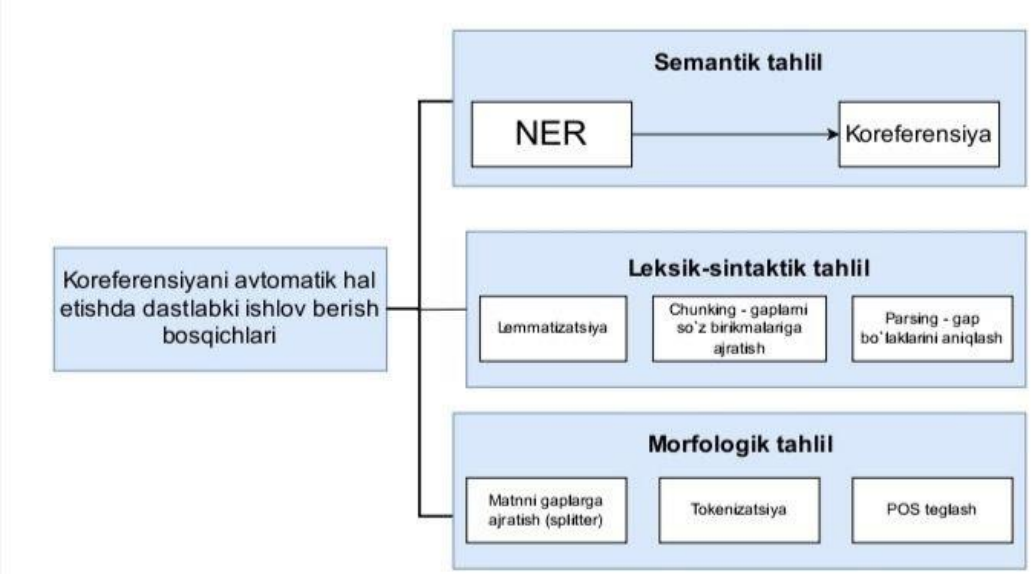
Ключевые слова: НЛП, кореференция, антисидент, анафора, лингвистический признак, вектор, местоимение, текст

Nutqdagi har qanday jumlaning mohiyati kontekstda fikr yuritilayotgan shaxs, predmet yoki voqea-hodisa yordamida anglashiladi. Inson ongi bevosita matn mazmunini tushunishga qodir. Ammo sun'iy intellekt uni to'g'ri talqin qilishi uchun matn qismlarining o'zaro sinonimiyasini, koreferentlikni hosil qiluvchi bir yadroviy zanjirga birikkan nutqiy birliklarni aniqlash muhimdir. Bu vazifa NLPda **Koreferensiyani hal etish (CR – Coreference Resolution)** tizimlariga yuklatilgan. Koreferensiyani hal etish obyekt haqidagi matn ma'lumotlarini birlashtirishning asosiy texnologiyasi [Vincent Ng, 2017; 4877], matndagi bir xil obyektga ishora qiluvchi barcha havola bo'laklarni avtomatik topish jarayoni [Elango, 2005; 1] bo'lib (1-rasm), bunda axborot uzatish mavqeyi jihatidan mustaqil referent **"antisident"** va ma'nosi antisident vositasida namoyon bo'ladigan referent **"koreferent/anafor"** deb yuritiladi.



1-rasm. Koreferensiya (koreferentlik) hodisasi

Koreferentlik dastavval qoidaga asoslangan usullar orqali hal etilgan, keyinchalik uning o'rnini mashinaviy o'rganish usullari egalladi. Koreferentlikni avtomatik hal etish imkoniyati dastlabki ishlov berish bosqichlarining sifat darajasi bilan belgilanadi. Koreferentlikni hal etishda dastlabki ishlov berish bosqichi deyarli barcha tillar uchun umumiy. Bu jarayon *morfologik ishlov berish, gapni otli birikmalarga ajratish (NP chunker), POS teglash, nomlangan obyektni aniqlash (NER), avtomatik sintaktik teglash (parsing)* ni o'z ichiga oladi [Shi Chunqi va b., 2014; 54]. Ushbu bosqichlar esa nutqiy birliklarning lingvistik tahlili asosiga quriladi (1-chizma):



1-chizma

Bugungi kunda koreferentlikni bartaraf etuvchi mashinaviy o‘rganish usullarining barchasi [Soon va b., 2001; 524] tomonidan taklif etilgan til birliklarining inobatga olish lozim bo‘lgan 12 ta xususiyatlar to‘plamidan foydalanib yaratilgan. Ushbu xususiyatlar vektori ikkita ajratib olingan ***i* (nomzod antisident) va *j* (anafor)** belgilariga asoslanadi (Ingliz tilidagi qisqartmalari bilan berildi; 1-jadval):

1. Masofa (DIST)	anafora va antisident orasida joylashgan gaplar sonini ifodalaydi. Agar 326anafora va uning antisidenti bir gap ichida qo‘llanilsa, qiymat 0 ga teng; agar ular orasida bitta gap mavjud bo‘lsa, qiymat 1 ga, ikkita gap mavjud bo‘lsa, qiymat 2 ga teng va h.k.
2. i-olmosh (i_PRONOUN)	agar antisident olmosh bo‘lsa *rost* qiymatini, aks holda, *yolg‘on* qiymatini beradi.
3. j-olmosh (j_PRONOUN)	agar 326anafora olmosh bo‘lsa *rost* qiymatini, aks holda, *yolg‘on* qiymatini beradi.
4. Matn (string) mosligi (STR_MATCH)	artikl va ko‘rsatish olmoshlari olib tashlangandan so‘ng, antisident matn anafora matniga mos kelsa, ya’ni biri ikkinchisining aynan o‘zi bo‘lsa, rost qiymatini beradi. Masalan, *dasturchi bu dasturchi* birligiga mos keladi.
5. Aniq ot birikmasi (DEF_NP)	agar antisident *“the”* artiklini olgan otli birikma bo‘lsa, *rost* qiymatini , aks holda, *yolg‘on* qiymatini beradi.
6. Ko‘rsatish olmoshiga ega ot birikmasi (DEM_NP)	agar antisident ko‘rsatish olmoshi bilan birikkan otli birikma bo‘lsa, *rost* qiymatini, aks holda, *yolg‘on* qiymatini beradi.
7. Son (birlik/ko‘plik) muvofiqligi (NUMBER)	agar antisident va anafora son bo‘yicha o‘zaro muvofiq bo‘lsa (ikkalasi ham birlikda yoki ikkalasi ham ko‘plikda bo‘lsa) *rost* qiymatini, aks holda, *yolg‘on* qiymatini beradi.
8. Semantik guruh muvofiqligi (SEMCLASS)	Har bir tizim o‘zining 326anafora guruhlarini belgilashi mumkin. Soon va b. “Erkak”, “Ayol”, “Shaxs”, “Vaqt” kabi anafora guruhlarni belgilaydi. Semantik guruhlar ISA iyerarxiyasida[27] joylashtirilgan. Anafora va uning antisidenti, agar ulardan biri ierarxiyada ikkinchisining hosila birligi bo‘lsa yoki ular o‘zaro bir xil maqomda bo‘lsa muvofiq keladi. *Ros t(1), yolg‘on (0)* yoki *noma’lum (2)* qiymatlaridan biri beriladi.
9. Jins muvofiqligi (GENDER)	agar 326anafora va uning antisidenti jins jihatidan o‘zaro muvofiq kelsa *rost (1)*, muvofiq kelmasa *yolg‘on (0)*, agar anafor yoki antisidentning jinsi noma’lum bo‘lsa, *noma’lum (2)* qiymatini beradi.
10. Atoqli ot (PROPER_NAME)	agar 326anafora ham, uning antisidenti ham atoqli ot bo‘lsa *rost*, aks holda, *yolg‘on* qiymatini beradi.

[27] ISA iyerarxiyasi narsa-predmetlar joylashuvi bo‘lib, unda asosiy va hosila birliklarning munosabati ifodalanadi. Asosiy birlik hosila birliklarning o‘zagi hisoblanadi. A ISA B tarzida keltirsak, har bir A obyekti B obyekti ham bo‘la oladi. Masalan, bizda A guruhi=shaxs, B guruhi=ayol, erkak. Agar B A bo‘lsa, B dagi har bir birlik A bo‘ladi, ya’ni har bir erkak va ayol shaxsdir.

11. Taxallus (ALIAS)	agar 327anafora antisidentning taxallusi bo‘lsa *rost*, aks holda, *yolg‘on* qiymatini beradi.
12. Izohlovchi (APPOSITIVE)	anafora antisidentning izohlovchisi bo‘lsa *rost*, aks holda, *yolg‘on* qiymatini beradi.

1-jadval. Koreferentlikni bartaraf etish uchun lingvistik cheklovlar

Keltirilgan lingvistik cheklovlarni o‘zbek tilidagi matnlarda koreferensiyani hal etish jarayoniga tatbiq etamiz. Quyida berilgan parchadagi *Qaffol Shoshiy* antisidenti va *u* anafori xususiyat vektorini o‘zaro bog‘laymiz:

Qaffol Shoshiy *Hirotdagi Nizomiya madrasasiga mudarrislik qilgan.* ***U*** *fiqh, hadis, usul, tilshunoslik fanlari va she’r san’ati bo‘yicha Movarounnahrda olimlar orasida tengsiz edi.* ***Qaffol ash-Shoshiy*** *qonunshunoslik, mantiq kabi sohalarga oid asar yozgan.*

Belgilarning xususiyat vektori (i=Qaffol Shoshiy, j=u)

Xususiyat	Qiymat	Izoh
DIST	1	i va j orasida bitta gap mavjud
i_PRONOUN	-	i olmosh emas
j_PRONOUN	+	j olmosh
STR_MATCH	-	matn mosligi mavjud emas
DEF_NP	-	j aniq ot emas
DEM_NP	-	j ko‘rsatish olmoshi emas
NUMBER	+	i va j ning ikkisi ham birlik sonda
SEMCLASS	1	i va j ning ikkisi ham shaxs
GENDER	1	i va j bir jinsga mansub
PROPER_NAME	-	Faqat i atoqli ot
ALIAS	-	j i ning taxallusi yoki laqabi emas
APPOSITIVE	-	j i ning izohlovchisi emas

Bu yerda o‘zbek tilida koreferentlikni hal etishda *u olmoshi* bilan bog‘liq noaniqliklar yuzaga keladi, ya’ni u olmoshi ko‘rsatish yoki kishilik olmoshi ekanligini aniqlash, barcha jinsni u olmoshining o‘zi ifodalashi bizdan yangi cheklov kiritishni talab etadi.

Yuqoridagi matn asosida *Qaffol Shoshiy va Qaffol ash-Shoshiy* birliklarning xususiyat vektorini ko‘rib chiqamiz:

Belgilarning xususiyat vektori (i=Qaffol Shoshiy, j=Qaffol ash-Shoshiy)

Xususiyat	Qiymat	Izoh
DIST	2	i va j orasida ikkita gap mavjud
i_PRONOUN	-	i olmosh emas
j_PRONOUN	-	j olmosh emas
STR_MATCH	+	matn mosligi mavjud
DEF_NP	+	j aniq ot
DEM_NP	-	j ko‘rsatish olmoshi emas
NUMBER	+	i va j ning ikkisi ham birlik sonda
SEMCLASS	1	i va j ning ikkisi ham shaxs
GENDER	1	i va j bir jinsga mansub
PROPER_NAME	+	i va j atoqli ot

ALIAS	-	j i ning taxallusi yoki laqabi emas
APPOSITIVE	-	j i ning izohlovchisi emas

Shu tariqa, nomzod antisident va anafor birliklar xususiyat vektorlarining qiymati aniqlangach, keyingi jarayonda mashinani o‘rganish algoritmlariga murojaat etiladi va lingvistik xususiyatlar vektori natijalariga ko‘ra antisident-anafor juftliklari klasterlanadi. O‘zbek tili matnlarida koreferentlikni bartaraf etishda [Soon va b., 2001; 524] yoki [Zhang va b., 3] tadqiqotini o‘zgarishlarsiz qo‘llash orqali yuqori natijaga erishib bo‘lmaydi. Ushbu xususiyat vektorlari ingliz tili uchun ishlab chiqilgan va ular orasida o‘zbek tili tabiatiga mos kelmaydigan cheklovlar ham mavjud. Masalan, artikllar orqali aniq va noaniq otni ajratish cheklovi nafaqat o‘zbek tili, balki barcha turkiy tillar oilasi uchun amal qilmaydi. Shu sababli bu cheklovlarni o‘zbek tilining lingvistik tabiatiga moslab o‘zgartirish, kengaytirish maqsadga muvofiqdir.

Foydalanilgan adabiyotlar:

1. Vincent Ng. Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research. / Proceedings of the Thirty-First AAAI, Conference on Artificial Intelligence (AAAI-17), pp.4877-4884.
2. Elango P. Coreference Resolution: A Survey. – University of Wisconsin, Madison, WI, pp.1-8, 2005.
3. Shi Chunqi, Verhagen Marc, Pustejovsky James. A Conceptual Framework of Natural Language Processing Pipeline Application. / Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT, pp.53-59, Dublin, Ireland, 2014.
4. Soon, W., H. Ng, and D. Lim (2001): A Machine Learning Approach to Coreference Resolution of Noun Phrases. / Computational Linguistics 27 (4), pp. 521–544.
5. Yimeng Zhang, Yangbo Zhu. Machine Learning for Coreference Resolution: Recent Developments, pp.1-23
6. https://www.cs.cmu.edu/~yimengz/papers/Coreference_survey.pdf