O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2025, [1/8]
ISSN 2181-7324

FILOLOGIYA
http://journals.nuu.uz
Social sciences

UO‘K:811.512.133’32

Shahlo ABDISALOMOVA,
Toshkent davlat o‘zbek tili va adabiyoti universiteti tayanch doktoranti
E-mail: abdisalomovashahlo@gmail.com

THE COREFERENCE RESOLUTION USING A DECISION TREE

Annotation
In today’s world, where computers managed perform almost all tasks, Coreference Resolution issues for every language has become an urgent and unavoidable task. This article is dedicated to phenomenon of Coreference and discusses the classification phase of Coreference Resolution and the importance of the Decision Tree algorithm in this process. It also includes the results and analysis of tests conducted on a small dataset of texts in Uzbek.
Key words: Coreference, automatic, decision tree, classification, algorithm, indicator, evaluation.

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ КОРЕФЕРЕНЦИИ С ИСПОЛЬЗОВАНИЕМ ДЕРЕВА РЕШЕНИЙ

Аннотация
В современном мире, когда компьютеры выполняют функции человеческого сознания, решение проблемы кореференции для каждого языка стало актуальной задачей, не терпящей отлагательства. Данная статья посвящена феномену кореференции, в которой рассматривается этап классификации автоматического определения кореференции и роль алгоритма дерева решений в этом процессе. Также представлены результаты теста, проведенного на база данных на узбекском языке, и его анализ.
Ключевые слова: Кореферентность, автоматический, дерево решений, классификация, алгоритм, индикатор, оценка.

KOREFERENSIYANI QARORLAR DARAXTI YORDAMIDA AVTOMATIK ANIQLASH

Annotatsiya
Inson ongi vazifalarini kompyuter bajarayotgan bugungi kunda har bir til uchun koreferensiya muammosini bartaraf etish kechiktirib bo‘lmas dolzarb vazifaga aylandi. Ushbu maqola koreferensiya hodisasiga bag‘ishlangan bo‘lib, unda koreferensiyani avtomatik aniqlashning klassifikatsiya bosqichi va bu jarayonda qarorlar daraxti algoritmining ahamiyati haqida so‘z yuritiladi. Shuningdek, o‘zbek tilidagi kichik hajmli matnlar to‘plamida o‘tkazilgan sinov natijalari va uning tahlili ham havola qilinadi.
Kalit so‘zlar: Konferensiya, avtomatik, qarorlar daraxti, tasniflash, algoritm, ko‘rsatkich, baholash.

Kirish. Tabiiy tilga ishlov berish jarayonida mashinaning matn qismlari o‘rtasidagi semantik munosabatni to‘g‘ri talqin qilishi muhimdir. Bu vazifa NLPda koreferensiyani avtomatik aniqlash tizimlariga yuklatilgan. Koreferensiyani avtomatik aniqlash matndagi bir xil obyektga ishora qiluvchi barcha havola bo‘laklarni avtomatik topish jarayonidir. 1-rasmda ushbu jarayonga namuna berilgan:

1-rasm. Koreferensiya hodisasi va uni avtomatik aniqlash

Shuni ta’kidlash kerakki, matndagi havolalarni aniqlash va hal qilishning umumiy muammosi sifatida koreferensiyani avtomatik aniqlash vazifasiga murojaat qilamiz. Biroq texnik jihatdan havolalarning bir nechta turlari mavjud bo‘lib, ularning ta’riflari bahsli masaladir.

Koreferensiyani avtomatik aniqlashdan farqli holat hisoblangan masala anaforani avtomatik aniqlashdir. Anaforik munosabat matnda bir termin boshqasiga ishora qilganda, ikkinchisining talqinini belgilab berganda yuzaga keladi. Quyidagi misolda (1) va (2) to‘g‘ridan to‘g‘ri turli xil real dunyo obyektlariga tegishli ekanligini ko‘rish mumkin. Ammo ular bir xil kontekstda ishlatiladi va (2) talqin (1) ga tayanadi. Bu eslatmalar qo‘shma gap emas, balki anafora bilan bog‘liq.

Zidan(1) Realda debyut qilganida, chiptalar(2) bir necha daqiqada sotilib ketdi.

Anaforik munosabatlar, koreferensiyadan farq qilsa ham, aksariyat hollarda, biri boshqasiga teng. Bunday tafovutlar va boshqa turli xil havolalar haqida ko‘plab misollar mavjud. Biroq koreferensiya hodisasi bilan bog‘liq tadqiqotlar keng qamrovga ega va mavjud ishlarning katta qismini qamrab oladi. Umuman olganda, anafora va koreferensiyani avtomatik aniqlash vazifalari matnni semantik tahlil qilish bosqichlaridir. Ular yordamida mashinaning inson kabi “fikrlash” salohiyatini rivojlantirish mumkin.

Mavzuga oid adabiyotlarning tahlili. Qarorlar daraxti algoritmi, uning umumiy tavsifi va amaliy qo‘llanish sohalari M.Onarqulov, M.Yusupov va N.Xudoyberdiyev hammuallifligidagi tadqiqotda bayon

qilingan [1]. M.Onarqulov va E.Omonaliyevaning maqolasida sun’iy intellekt sohasida qarorlar daraxti va uni kiritish algoritmi mavzusiga aloqador muhim ma’lumotlar ta’kidlangan [2]. Koreferensiyani avtomatik aniqlash jarayonida tasniflash uchun qo‘llaniladigan turli usullar mavjud bo‘lsa-da, ko‘pgina tadqiqotlarda qarorlar daraxti algoritmidan foydalanilganiga guvoh bo‘lamiz. Bunday ishlar sirasiga Aone va Bennett, Soon, Ng va Cardia, V.S.Ram va S.L.Devi, K.Šteflović va J.Kapusta tomonidan amalga oshirilgan tadqiqotlarni kiritish mumkin. Aone va Bennettning maqolasida koreferensiyani avtomatik aniqlash uchun qarorlar daraxti algoritmidan foydalanish jarayoni va MUC-6 korpusidagi sinov natijalari qayd etilgan. Ishda koreferent birliklarni aniqlashda 12 ta xususiyat to‘plamiga tayanilib, qoidaga asoslangan usullarga nisbatan yuqori aniqlikka erishilgan [3]. Soon va bir guruh olimlar otli birikmalar uchun koreferensiyani avtomatik aniqlash masalasini o‘rgandilar. Ular bu jarayonda mashinali o‘qitish usullariga murojaat qilgan va C5 qarorlar daraxti algoritmi yordamida koreferensiya klasterlarini shakllantirgan [4]. Ng va Cardia maqolasida keltirilishicha, ular tomonidan qo‘lda yig‘ilgan xususiyatlar to‘plami qarorlar daraxti algoritmi asosida trening qilingan va F1 ko‘rsatkichi MUC-6 korpusida 64.3 dan 70.4, MUC-7 korpusida 61.2 dan 63.4 oralig‘idagi aniqlik darajasiga ega [5]. V.S.Ram va S.L.Devi tasodifiy o‘rmon usuli asosida koreferensiyani avtomatik aniqlash jarayonini bayon etgan [6]. K.Šteflović va J.Kapusta yangiliklardan iborat matnlarda koreferensiya hodisasini avtomatik aniqlash masalasiga e’tibor qaratgan va turli yondashuvlarni, xususan, qarorlar daraxti algoritmini ham bu jarayonga tatbiq etgan va natijalarni qiyosiy tahlil qilgan [7].

Tadqiqot metodologiyasi. Klassifikatsiya – berilgan ma’lumot nuqtalari sinfini bashorat qilishga mo‘ljallangan nazoratli o‘qitish usuli. Klassifikatsiyaning algoritmlari xilma-xil. Asosiy klassifikatsiya algoritmlari qarorlar daraxti, soddalashtirilgan Bayes klassifikatori, sun’iy neyron tarmog‘i (ANN) va K-eng yaqin qo‘shni usuli (KNN) hisoblanadi. Qarorlar daraxti toifali va uzluksiz atributlarni boshqarishda qulayligi sababli tasniflashda eng ko‘p qo‘llaniladi. U ildiz tugun, shoxlar, ichki tugunlar va yaproq tugunlaridan tashkil topgan ierarxik daraxt tuzilishiga ega. Bu algoritmda ma’lumotlar xususiyatlar qiymatlariga asoslangan “agar...., u holda” qoidalari to‘plami sifatida ko‘rib chiqiladi. Tasniflashda qarorlar daraxti bir qancha ustunlikka ega. Ushbu afzalliklar quyida keltiriladi:

Xususiyatlarni tanlash imkoniyati;

Diskret va uzluksiz atributlar bilan ishlash qobiliyati;

Noma’lum qiymatlarni qayta ishlay olishi;

Qoidalar to‘plamini oson o‘zgartirish mumkinligi;

Tushunish va izohlash uchun oddiyligi;

Katta hajmli ma’lumotlarni qisqa vaqt ichida tahlil qila olishi.

Qarorlar daraxti algoritmi “yuqoridan pastga” tamoyili asosiga qurilgan, ya’ni:

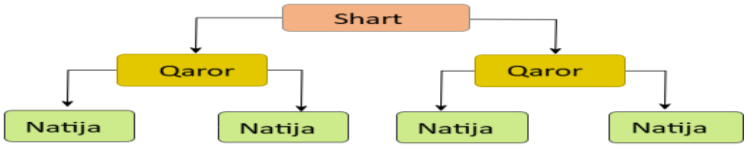
– Algoritm butun ma’lumotlar to‘plamini ifodalovchi “ildiz tugun” dan boshlanadi;

– Ma’lumotlarni aniq guruhlarga ajratish maqsadida “Qaysi atribut keltirilgan tugun uchun eng yaxshi tanlov?” savoliga javob beradigan asosiy xususiyatni qidiradi;

– Savolga berilgan javobga tayanib, ma’lumotlarni kichikroq to‘plamlarga ajratadi va yangi tarmoqlarni yaratadi;

– Algoritm savollar berishda davom etadi va ma’lumotlarni har bir tarmoqda bashorat qilingan natijalar yoki tasniflarni ifodalovchi yakuniy “yaproq tugunlari” ga yetguncha taqsimlaydi.

Qarorlar daraxtida ma’lumot to‘plamiga qarab turli algoritmlardan foydalanish mumkin. Eng ommabop qarorlar daraxti algoritmlariga ID3, C4.5 [8], CART va tasodifiy o‘rmon usuli kabilarni kiritish mumkin. Qarorlar daraxti algoritmining umumiy tuzilishi va ish faoliyatini 2-rasmda tasvirlaymiz:



2-rasm. Qarorlar daraxti arxitekturasi

Ko‘rinadiki, qarorlar daraxti kirish o‘zgaruvchilari va ularning natijalari o‘rtasidagi munosabatlarni tushunish va yakuniy qarorga yordam beradigan eng muhim xususiyatlarni aniqlash uchun zaruriy vositadir.

Tahlil va natijalar. Ma’lumki, modelning ishlashi uchun koreferensiya korpusi zarur. Biz maqolada tasniflashni o‘zbek tili matnlaridan tashkil topgan kichik ma’lumot to‘plami (6,185 token) ustida amalga oshiramiz. Ushbu ma’lumotlar to‘plami asosida eslatmalar – nomzod antesedent va anaforani tasniflash uchun 11 ta xususiyatlardan foydalanamiz:

HEAD_MATCH: eslatmalar bir xil hokim so‘zga egami? Ehtimoliy qiymat: {ha, yo‘q}

STR_MATCH: eslatmalar aynan bir xilmi? Ehtimoliy qiymat: {ha, yo‘q}

NUMBER: eslatmalar son (birlik/ko‘plik) da o‘zaro muvofiqmi? Ehtimoliy qiymat: {ha, yo‘q}

GENDER: eslatmalar jinsda o‘zaro mosmi? Ehtimoliy qiymat: {ha, yo‘q, noma’lum}

ANIMACY: eslatmalar jonli/jonsizlik nuqtayi nazaridan mosmi? Ehtimoliy qiymat: {ha, yo‘q, noma’lum}

BOTH_PRONOUNS: ikki eslatma ham olmoshmi? Ehtimoliy qiymat: {ha, yo‘q, noma’lum}

BOTH_PROPER_NOUNS: ikkala eslatma ham atoqli otmi? Ehtimoliy qiymat: {ha, yo‘q, noma’lum}

APPOSITIVE: eslatmalardan biri ikkinchisining qisqartmasimi? Ehtimoliy qiymat: {ha, yo‘q}

SEMCLASS: eslatmalar bitta semantik sinfdami? Ehtimoliy qiymat: {ha, yo‘q, noma’lum}

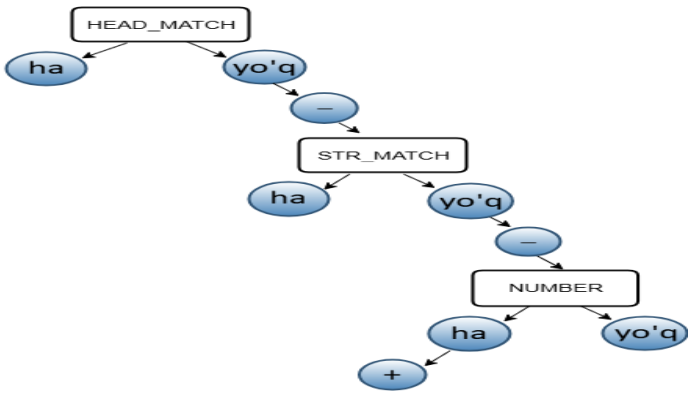
DISTANCE: eslatmalar orasida nechta so‘z yoki gap mavjud? Ehtimoliy qiymat: {0, 1, 2, 3...}

ALIAS: bir eslatma boshqasining taxallusi yoki laqabimi? Ehtimoliy qiymat: {ha, yo‘q}

Bizning ma’lumotlar to‘plamimizda obyekt-obyekt tipidagi bog‘lanishlar ko‘p uchraydi. Ularni bir sinfga birlashtirish yoki birlashtirmaslik masalasini qarorlar daraxti misolida tahlil qilamiz:

Barvasta kishi engashib, Anvarning yelkasiga qoqdi: “Mard bo‘l, o‘g‘lim, mard bo‘l!”

Keltirilgan gapda “Anvar” ni nomzod antesedent, “o‘g‘lim” so‘zini esa nomzod anafora sifatida belgilaylik. Ularning kategoriyasini aniqlash strukturasi quyidagicha:



3-rasm. Qarorlar daraxtida tasniflash jarayoni

Biz namuna sifatida 3 ta xususiyat misolida qarorlar daraxtining ish jarayonini tavsifladik. Unda negativ qiymatlar pozitiv qiymatlarga nisbatan ko‘proq. 1-jadvalda 11 ta xususiyatlar to‘plamining qarorlar daraxti algoritmiga tatbiqi batafsil keltirildi:

1-jadval. Qarorlar daraxti algoritmida tasniflash jarayoni tahlili

№	Gap	Nomzod antesedent	Nomzod anafora	Xususiyatlar qiymati	Koreferent
1.	Barvasta kishi engashib, Anvarning yelkasiga qoqdi: “Mard bo‘l, o‘g‘lim, mard bo‘l!”	Anvar	o‘g‘lim	–, –, +, +, +, –, –, –, +, 0, +	ha
2.	“Nima deding, Fitna? Nega dilgir bo‘libdilar?” – so‘radi so‘fi Qurvonbibidan.	Qurvonbibi	Fitna	–, –, +, +, +, –, +, –, +, 0, +	ha
3.	Keksa otaxon uyali telefon ustasiga telefoni o‘chib qolganini aytib, uni tuzatib berishini so‘radi.	Uyali telefon	telefoni	+, –, +, +, +, –, –, –, +, 0, –	ha
4.	Toshkentda Mamat akani tanimagan odam yo‘q edi. Xalq hofizi Muhammadjon Karimov borgan to‘yga ayricha fayz qo‘nardi.	Mamat aka	Xalq hofizi Muhammadjon Karimov	–, –, +, +, +, –, +, –, +, 1, +	ha

Pozitiv va negativ qiymatlar teng kelib qolganda, ularni sinfga birlashtirish jarayoni qiyinlashadi. Shu sababli bunday holatlarda entropiyaga murojaat qilinadi. Entropiya “betartiblik darajasi”, “noaniqlik o‘lchovi” degan ma’nolarni bildirib, qarorlar daraxti algoritmida eng yaxshi bo‘linishni tanlashda qo‘llaniladi. Quyida entropiya formulasi keltirilgan [9]:

$$E(S) = -p(+)\log p(+) - p(-)\log p(-)$$

Bu yerda:

p+ – pozitiv sinf taxmini;

p– – negativ sinf taxmini;

S – trening misollari to‘plami.

Axborot olish (Information Gain) – bu qaysidir belgi-xususiyat berilganda noaniqlikning kamayishini o‘lchaydigan ko‘rsatkichdir va u qaysi atribut qaror tuguni yoki ildiz tugun sifatida tanlanishi kerakligini aniqlashda muhim omil hisoblanadi:

Axborot olish = E (Y) – E (Y/X)

Ushbu formulalar yordamida qaror qabul qilish uchun eng to‘g‘ri atributlarni tanlashga erishiladi. Natijada model aniqligi pozitiv qiymatni aks ettiradi.

Mashinali o‘qitishning klassifikatsiyalash turida modellarning qanchalik to‘g‘ri hisoblashini tekshirishning bir nechta o‘lchovlari bor. Biz maqolada tasniflashni o‘zbek tili matnlaridan tashkil topgan kichik ma’lumot to‘plami (6,185 token) ustida amalga oshiramiz. Ushbu jarayonni baholashda MUC ko‘rsatkichidan foydalanamiz. MUC ko‘rsatkichi model to‘g‘ri bog‘lanishlarning nechtasini topganini hisoblaydi. Ushbu jarayonni tekshirishda F1 o‘lchovini tanlaymiz. Chunki u pozitiv va negativ qiymatlar nisbati teng bo‘lmagan holatda aniqroq natijani qayd etadi. F1 ko‘rsatkichini o‘lchashda dastlab Precision (aniqlik) va Recall (qamrov) parametrlari aniqlab olinadi [10]:

Precision – model tomonidan pozitiv deb belgilangan natijalar ichidan nechtasi haqiqatan ham to‘g‘ri ekanligini ko‘rsatadi. Formulasi:

$$Precision_{MUC} = \frac{Tizim\ topgan\ to'g'ri\ bog'lanishlar\ soni}{Tizimdagi\ umumiy\ bog'lanishlar\ soni}$$

Recall – etalon pozitiv bog‘lanishlar ichidan model nechta holatni to‘g‘ri topganini ifodalaydi:

$$Recall_{MUC} = \frac{Gold\ (etalon) klasterdagi\ to'g'ri\ bog'lanishlar\ soni}{Etalon\ umumiy\ bog'lanishlar\ soni}$$

$F1\ score = \frac{2*(Recall*Precision)}{Recall+Precision}$F1 Precision va Recall orqali ifodalanadigan o‘lchov bo‘lib, uning formulasi quyidagicha:

Model aniqligini hisoblash algoritmi quyida keltirildi:

```
from sklearn.metrics.cluster import precision_score, recall_score
# Qayta bog‘lanishlar (Gold standarti) va model natijalari
gold_clusters = [[1, 2, 3], [4, 5], [6]]
pred_clusters = [[1, 2], [3, 4, 5], [6]]
def muc_score(gold, pred):
correct_links = 0
total_links = 0
for g in gold:
correct_links += len(g) - 1 # Gold standarti bo'yicha to‘g‘ri bog‘lanishlar
for p in pred:
total_links += len(p) - 1 # Model tomonidan topilgan bog‘lanishlar
muc_recall = correct_links / total_links
muc_precision = correct_links / total_links
return muc_recall, muc_precision
```

Yuqoridagi hisoblash natijalariga ko‘ra, o‘zbekcha matnlarning kichik ma’lumotlar to‘plamidagi aniqlik 2-jadvalda ifodalandi:

2-jadval. Qarorlar daraxti algoritmining sinov natijalari

	MUC
Precision	73.53%
Recall	29.07%
F1	41.64%

Natijalar tahlilidan ko‘rish mumkinki, model aniqligi yuqori ko‘rsatkichga ega emas. Shunday bo‘lsa-da, Precision uchun qarorlar daraxti algoritmi samarali ishladi. Model aniqligining pastligi bizda ma’lumotlar bazasining kam miqdorda ekanligi bilan bog‘liqdir. Ma’lumotlar to‘plami hajmi qanchalik ko‘p bo‘lsa, model aniqligi ham shunchalik yuqori bo‘ladi. Bu kichik bir tajriba natijasi va keyingi tadqiqotlarimizda ushbu xatoliklarni bartaraf etishga harakat qilamiz.

Xulosa va takliflar. Ushbu maqolada koreferensiya hodisasini bartaraf etishda qarorlar daraxtining ahamiyati va tatbiq natijalari haqida fikr yuritildi. Tajriba o‘zbek tilidagi kam hajmli matnlar doirasida o‘tkazilgani bois model yuqori aniqlikka ega bo‘la olmadi. Bu esa sinovni boshqa baholash o‘lchovlarida ham amalga oshirishga va natijalarni qiyosiy tahlil qilishga yo‘l ochadi. Ko‘pgina ishlarda koreferensiya masalasi bitta emas, balki bir nechta zamonaviy usullar yordamida hal etilgan. Ularning safida qarorlar daraxti hamisha mavjud. Chunki qarorlar daraxti algoritmi qo‘llanilgan tadqiqotlarni tahlil qiladigan bo‘lsak, ularning barchasida model aniqligi ijobiy ko‘rsatkichga ega ekanligini ko‘ramiz. Demak, davrlar o‘tsa ham samaradorlik mavqeyi jihatidan qarorlar daraxti tasniflashning keng tarqalgan usullaridan bo‘lib qolaveradi.

ADABIYOTLAR

1. Onarqulov M.K., Yusupov M.A., Xudoyberdiyev N.Z. Qarorlar daraxtini qurish algoritmlari. / “Izlanuvchi” ilmiy-metodik jurnali, 1-son, B. 32-38. / www.phoenixpublictaion.net
2. Onarqulov M.K., Omonaliyeva E.U. Qarorlar daraxti va uni kiritish algoritmi. / Science and innovation in the education system international scientific-online conference, B. 66-73.
3. Aone C., Bennett S.W. Evaluating automated and manual acquisition of anaphora resolution strategies. / Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1995, pp. 122–129.
4. Soon W., Ng H. and Lim D. A Machine Learning Approach to Coreference Resolution of Noun Phrases, Computational Linguistics 27 (4), 2001, pp. 521–544.
5. Ng V., Cardie C. Improving machine learning approaches to coreference resolution, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 104-111.
6. Ram V.S., Devi S.L. Clause Boundary Identification Using Conditional Random Fields. / A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, 2008, pp.140-150.
7. Šteflović K., Kapusta J. Coreference Resolution for Improving Performance Measures of Classification Tasks. / Applied Sciences, 13, 9272, 2023, pp. 1-20. / https://doi.org/10.3390/app13169272
8. Quinlan J.R. C 4.5: Programm for Machine Learning. / Morgan Kaufmann publishers, Vol. 16, 1993, pp. 235-240.
9. www.geeksforgeeks.org/decision-tree
10. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/#