ISSN 2181-371X
SJIF: 5.305 VOLUME 4 ISSUE 34
INNOVATIONS IN TECHNOLOGY AND SCIENCE EDUCATION

GAP O'XSHASHLIGI ASOSIDA PEREFRAZLARNI ANIQLASH JARAYONLARINI MODELLASHTIRISH

Axmedova Xusniya Xusanovna
"Muhammad al-Xorazmiy nomidagi TATU" katta o'qituvchisi
khusniya150586@gmail.com

Annotatsiya: *Tabiiy tilga ishlov berish (NLP) sohasidagi turli vazifalarda, jumladan, plagiatni aniqlash, mualliflikni aniqlash, savol-javob tizimlarida qo'llash, matnni qisqartirish, mashina tarjimasi, umumiy matn tahlili va boshqalarda muhim rol o'ynaydi. Matnlar o'rtasidagi semantik o'xshashlikni o'lchash bo'yicha dastlabki yondashuvlardan biri axborot izlash sohasida vektorli fazo modeli (Vector Support Model) hisoblanadi. Katta hajmdagi teglanmagan korpuslardan olingan matnlar bilan neyron tarmoqni (NN-training) o'qitish tasodifiy parametrlarga asoslangan VSMni yaratishga olib keladi. Perefrazlarni aniqlash vazifasida natija 0 dan 1 gacha bo'lgan ehtimollik ko'rinishida ifodalanadi, bu yerda 1 ga yaqin qiymat gaplar juftligini bir-birining perefrazi sifatida ko'rsatadi, 0 esa ularni semantik jihatdan farqlanishni bildiradi.*

Kalit so'zlar. *Perefrazni aniqlash, semantic o'xshashlik, Mashinali o'qitish (ML), Text Preprocessing Pipline (TPP)*

Perefrazlash jarayoni matnni qayta yozishni, so'zlar va ularning tartibini o'zgartirishni, shu bilan birga asl ma'noni saqlab qolishni o'z ichiga oladi. Qayta yozishni aniqlash Tabiiy tilga ishlov berish (NLP) sohasidagi turli vazifalarda, jumladan, plagiatni aniqlash, mualliflikni aniqlash, savol-javob tizimlarida qo'llash, matnni qisqartirish, mashina tarjimasi, umumiy matn tahlili va boshqalarda muhim rol o'ynaydi. Matnlarning semantik o'xshashlik darajasini aniqlash kabi umumiyroq vazifa NLP sohasida katta ahamiyatga ega.

Perefraz qilingan matn plagiatni aniqlash tizimlari duch keladigan eng qiyin muammolardan biridir. Ko'pgina plagiatni aniqlash tizimlari odatda umumiy so'zlar va kichik o'zgarishlarni aniqlash uchun yaratilgan, ammo keng semantik tuzilma va jumladagi o'zgarishlarni aniqlashga qodir emas. Shu sababli, ko'plab qayta yozish holatlari sezilmay qoladi.

Perefrazlarni aniqlashning asosiy usullari quyidagilardan iborat:

- Vektorli fazo modeli – uning asosi vektorizatsiya yoki matnni ko‘chirish va ikki matn orasidagi masofa/o‘xshashlikni hisoblash (*Jakkard koeffitsienti, Evklid masofasi, kosinus masofasi, Word Mover masofasi va boshqalar*);
- Sun’iy neyronli chuqur o‘rganish modellaridan foydalanish (*konvolyutsion, rekurrent, Siyam, encoder-decoder, e’tibor algoritmiga ega transformatorlar va boshqalar)*;
- Masofalarni hisoblash va neyron tarmoq tasnifi natijalarini yakuniy tasniflagichni o‘rgatish uchun alohida xususiyat sifatida qo‘llash.

Matnlar o‘rtasidagi semantik o‘xshashlikni o‘lchash bo‘yicha dastlabki yondashuvlardan biri axborot izlash sohasida vektorli fazo modeli (Vector Support Model) hisoblanadi [1]. VSMning maqsadi to‘plamdagi har bir obyektni (so‘zlardagi harflar, gaplardagi so‘zlar, mazmundagi gaplar, korpusdagi mazmun) n-o‘lchovli fazoda nuqta, ya’ni VSMda vektor sifatida ifodalashdir [2]. Ushbu fazoda nuqtalar qanchalik yaqin joylashgan bo‘lsa, ular semantik jihatdan shunchalik o‘xshash bo‘ladi va aksincha.

Berilgan **k** ta matnlar to‘plami $D = \{D_1, D_2, \dots, D_k\}$ uchun, D_i matni $D_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$ vektor shaklida ifodalanadi. Klassik so‘zga asoslangan VSMda har bir o‘lcham matnlar to‘plamidagi bitta atama/so‘zga mos keladi. Vazn turli vazn o‘lchash sxemalariga asoslanib aniqlanadi; *Bag-of-Words va TF-IDF* usullari so‘zga asoslangan VSMlarda keng qo‘llaniladi. Ikki D_i *va* D_j matnlar o‘rtasidagi o‘xshashlik *Jakkard koeffitsienti, Evklid masofasi, kosinus o‘xshashlik* va boshqa usullar yordamida hisoblanadi.

Ushbu modelning asosiy kamchiliklari yuqori o‘lchovlilik, siyraklik va lug‘at bilan bog‘liq muammolarni o‘z ichiga oladi. Shu sababli, VSMning turli modifikatsiyalari va umumlashmalari ishlab chiqilgan. Chuqur o‘rganishdagi oldingi usullar singari, hujjatlar yoki matnlar Doc2Vec usuli yordamida vektor sifatida shakllantiriladi. Bundan tashqari, so‘zlar ham Word2Vec usuli asosida vektorlar shaklida tasvirlanadi [3].

So‘zlarning vektorli tasvirini o‘rganish uchun matritsani dekompozitsiya qilish usuli, masalan, LSA asosida variantlar mavjud. Boshqa algoritmlar kontekstga asoslangan usullardan foydalanadi, masalan, Skip-gramm, Continuous Bag of Words. Ushbu vektorlar kosinus masofasi yoki boshqa o‘xshashlik o‘lchovlari yordamida solishtiriladi.

Word2Vec modelining paydo bo‘lishi tadqiqotchilarni Doc2Vec, FastText, GloVe, USE va ELMO kabi boshqa vektor modellarini yaratishga undadi. Ushbu

modellarni umumiy qilib *2Vec modellar deb atashadi, chunki ular matnni (so‘zlar, iboralar, gaplar, bo‘limlar va to‘liq hujjatlar shaklida) vektor shakliga o‘zgartiradi va n-o‘lchovli vektorli fazolarni yaratadi.

Katta hajmdagi teglanmagan korpuslardan olingan matnlar bilan neyron tarmoqni (NN-training) o‘rgatish tasodifiy parametrlarga asoslangan VSMni yaratishga olib keladi. Bu parametrlar orasida eng muhimlari quyidagilar: vektor fazosining o‘lchami, so‘zning minimal chastotasi, o‘rganish tezligi va har bir so‘zning kuzatuv oynasi/kontekst hajmi.

Word2Vec ikki kichik modeldan iborat:

- **CBOW (continuous bag of words)** – joriy so‘z ishtirokidagi kontekst berilganda, shu so‘zni bashorat qiladi.
- **Skip-gram** – berilgan so‘zdan foydalangan holda uning kontekstini bashorat qiladi.

Perefrazlash vazifasi ikki kichik vazifaga bo‘linadi:

perefrazlarni aniqlash

perefrazni yaratish.

Perefrazlarni aniqlash vazifasida natija 0 dan 1 gacha bo‘lgan ehtimollik ko‘rinishida ifodalanadi, bu yerda 1 ga yaqin qiymat gaplar juftligini bir-birining perefrazi sifatida ko‘rsatadi, 0 esa ularni semantik jihatdan farqlanishni bildiradi.

So‘nggi yillarda NLP uchun chuqur neyron tarmoqlardan foydalanish sezilarli darajada o‘sdi. Tadqiqotlarda perefrazlarni aniqlash uchun Siyam, konvolyutsion, rekurrent, konvolyutsion va rekurrent neyron tarmoqlarning kombinatsiyasiga asoslangan murakkab model arxitekturasi, transformatorlar va boshqa yondashuvlar qo‘llaniladi. Modelning aniqligi ma’lumotlarning taqsimlanadigan sinflari soni, an’anaviy yoki chuqur o‘rganish usullarini tanlash va ularning so‘z vektorizatsiyasi (nesting) va Text Processing Pipeline (TPP –Matnni qayta ishlash jarayoni) usullari bilan kombinatsiyasiga bog‘liq. Yangi ma’lumotlar to‘plami uchun matnda perefrazlarni aniqlovchi axborot tizimini (AT) yaratishda faqat ML modellarining turli usullari va arxitekturalarini emas, balki TPP algoritmlarini, matnni raqamli formatda ifodalash usullarini va ularning mumkin bo‘lgan kombinatsiyalarini ham baholash zarur. Bu esa tipik ATda qayta yozishni aniqlash sifatining (aniqligining) eng yuqori darajasiga erishish imkonini beradi.

Ba’zi xorijiy ilmiy manbalarda mualliflar qisqa gaplarning semantik bog‘liqligini o‘lchash uchun yangi algoritmni taklif qilganlar. Ushbu algoritm korpusga asoslangan usullar (masalan, nuqta bo‘yicha o‘zaro ma’lumot va latent semantik tahlil) yoki WordNet ma’lumotlariga asoslangan yuqorida tilga olingan masofalarni qo‘llaydi.

Gaplarning semantik bog‘liqligini o‘lchash formulasida so‘zlar juftligi o‘xshashligi va ularning o‘ziga xosligini o‘lchovlari birlashtiriladi. Ushbu formula kirish matnlarining o‘xshashligini aniqlash uchun yaxshi ko‘rsatkich bo‘lishi mumkin, chunki aniqlik, to‘g‘rilik, to‘liqlik va F-o‘lchovining olingan qiymatlari TF-IDF asosida vektor formatiga o‘zgartirilgan ikkita kirish gapi orasidagi kosinus masofasini o‘lchashning asosiy yondashuvi bilan taqqoslaganda yaxshiroq natijalar beradi [4]. Undan tashqari ko’plab xorijiy ilmiy maqolalarda semantik bog‘liqlikni aniqlash uchun usul taklif qilingan, bu usul so‘zlar uchun semantik profil yaratadi va u asosiy konseptual xususiyatlarga asoslanadi, ular esa ensiklopedik bilimlardan to‘plangan. Modelning asosiy g‘oyasi shundan iboratki, so‘zning ma’nosi bevosita kontekstda bo‘lgan kontseptsiyalar orqali aniqlanadi. Ushbu usul ikki asosiy qadamdan iborat. Birinchi qadamda, Wikipedia asosida asosiy konseptual so‘zlar bilan tegishli havolalar bilan annotatsiyalangan korpus yaratiladi. Keyin, ushbu korpus so‘zlar va matnlar o‘rtasidagi semantik o‘xshashlikni o‘lchash uchun ishlatiladi [5].

Vektorlar o‘rtasidagi kosinus masofasining formulasi matnlarning semantik o‘xshashligini hisoblash uchun moslashtirish maqsadga muvofiq ekanligini ko’rishimiz mumkin [6]. Taklif qilingan algoritm WordNet lug‘atlaridan olingan so‘z o‘xshashligi haqidagi ma’lumotlardan foydalanadi. Gaplar juftligi o‘rtasidagi o‘xshashlikni hisoblash uchun gaplarni ifodalovchi vektorlar o‘rtasidagi kosinus masofasi va so‘z juftliklarining o‘xshashligi haqidagi ma’lumotlarni o‘z ichiga olgan semantik o‘xshashlik matritsalari, shuningdek, WordNetning olti xil iyerarxiya asosidagi semantik masofalaridan olingan ma’lumotlar ishlatiladi. Har bir gap ikkilik vektor sifatida ifodalanadi (agar so‘z gapda mavjud bo‘lsa, element 1 ga teng, aks holda 0 ga teng).

NN yordamida olingan so‘zlarning vektorli joylashuvlari va so‘zlarning an’anaviy vektorli tasvirlari (so‘zlarning birga uchrashuvlarini hisoblashga asoslangan) o‘rtasida taqqoslash tadqiqoti o‘tkazilgan [7]. Mualliflar ikkita kichik hajmdagi vazifani (so‘zlarning ma’nolarini tahlil qilish va gaplarning o‘xshashligini aniqlash) va ikkita katta hajmdagi vazifani (perefrazlarni aniqlash va dialog harakatlarini belgilash) bajarishgan. Perefrazlarni aniqlash vazifasi uchun tadqiqotchilar vektorlar o‘rtasidagi kosinus masofasidan foydalanishgan. NN yordamida olingan so‘zlarning vektorli joylashuvlarini qo‘llash perefrazlarni aniqlash va dialog harakatlarini belgilash uchun model sifatini sezilarli darajada yaxshilaydi, ammo kichik hajmdagi vazifalarni hal qilishda aniqlikda sezilarli o‘sish bermaydi. Boshqa manbalarda esa, mualliflar so‘zlarning semantik o‘xshashlilik darajasini o‘lchash uchun korpusga asoslangan o‘lchovlardan va eng uzun umumiy ketma-ketlik

(LCS) satrlarni moslashtirish algoritmining normallashtirilgan va o‘zgartirilgan versiyasidan foydalanishni taklif qilishgan [8].

Matnlarni plagiatga tekshirish jarayoni turli NLP algoritmlari yordamida yaxshilangan: bir xil trigramlarning soni, til modeli ehtimolligi, eng uzun umumiy ketma-ketlik, bog‘lanishlar o‘rtasidagi moslik [9]. Olingan xususiyatlar Naive Bayes klassifikatorini plagiat mavjudligi va perefrazlash darajasi va turiga qarab ikkita yoki to‘rtta kategoriyada o‘rgatishda ishlatilgan. Tajriba natijalari korpusdagi hujjatlarni ikki kategoriya (qayta yozilgan va noyob) bo‘yicha tasniflashda qoniqarli bo‘ldi: 38 ta plagiatsiz hujjatdan 37 tasi to‘g‘ri tasniflangan, va faqat bir necha plagiat hujjatlari plagiatsiz deb tasniflangan (57 dan 5 tasi). Plagiatning uchta turli darajasini ajratish ancha murakkab vazifa ekanligi isbotlandi. Plagiatni tasniflash modelining qoniqarli bo‘lmagan aniqligi vazifaning murakkabligi va annotatorlar tomonidan yuzaga kelgan xatoliklar bilan bog‘liq [10].

Dunyoning ko’plab tillarida gap perefrazini yaratish bo’yicha qator tadqiqotlar olib borilgan. Perefraz yaratish uchun mustahkamlash bilan chuqur o‘qitish yondashuvi taqdim etilgan [11]. Ushbu muammoni hal qilish uchun yangi tuzilma tasvirlangan bo‘lib, u generator va baholovchidan iborat. Generator, ketma-ketlikdan ketma-ketlik (sequence to sequence) arxitekturasiga asoslangan takroriy neyron tarmog‘i sifatida qurilgan bo‘lib, kiruvchi gapni perefraz qiladi. Baholovchi modellar esa ikki gap bir-birining perefraz ekanligini tasniflaydi. Generator avval chuqur o‘qitish orqali o‘qitiladi va keyin mustahkamlash o‘qitish yordamida qayta sozlanadi.

Arab va ingliz tillaridagi matnlarni perefraz qiluvchi SimAll tizimi ishlab chiqilgan bo‘lib, bu tizim tasma taqqoslash, korpus va bilimlarga asoslangan usullarni birlashtiradi va ingliz va arab tillarini qo‘llab-quvvatlaydi. Umuman olganda, tizim 61 ta gaplarning semantik o‘xshashligini baholash algoritmlarini qo‘llab-quvvatlaydi, har bir algoritmning baholash natijalari normallashtirilgandan so‘ng [0, 1] oraliqqa tegishli. Foydalanuvchi natijalarni yig‘ish usulini tanlashi kerak: o‘rtacha, yig‘indi yoki maksimal qiymat. Bilimlarga asoslangan metrikalar 6 ta semantik o‘xshashlik masofasini o‘z ichiga oladi: Path (yo‘l), Leacock & Chodorow (lch), Wu & Palmer, Resnik (res), Lin (lin), Jiang & Conrath (jcn). Metrikalar faqat ingliz tilida mavjud [10-11].

Tadqiqot davomida olib borilgan izlanishlar shuni ko’rsatdiki, o’zbek tilidagi gaplarni perefraz qilish, matn tarkibidagi perefraz gaplarni aniqlash gibrid yondashuvlarga ehtiyoj sezadi.

Qoidaga asoslangan yondashuv asosida perefrazni aniqlash.

Berilgan matndagi so‘zlarni tezaurus yordamida sinonimlarga almashtiriladi.

Masalan, UzbekWordNet yoki boshqa leksik ma’lumotlar bazasidan foydalanish demakdir. O’zbek tili WordNet axborot tizimi ishlab chiqilmaganligi sababli bu ishda boshqa usulni qo’lladik. O’zbek tilidagi so’zlarning sinset to’plamlari mavjud bo’lib, ular yordamida gaplarni perefraz qilish mumkin. Buning uchun o’zbek tilidagi o’zak so’zlarni sinonimlar to’plami ajratilgan bo’lishi lozim. Sinset to’plami quyidagicha bo’ladi.

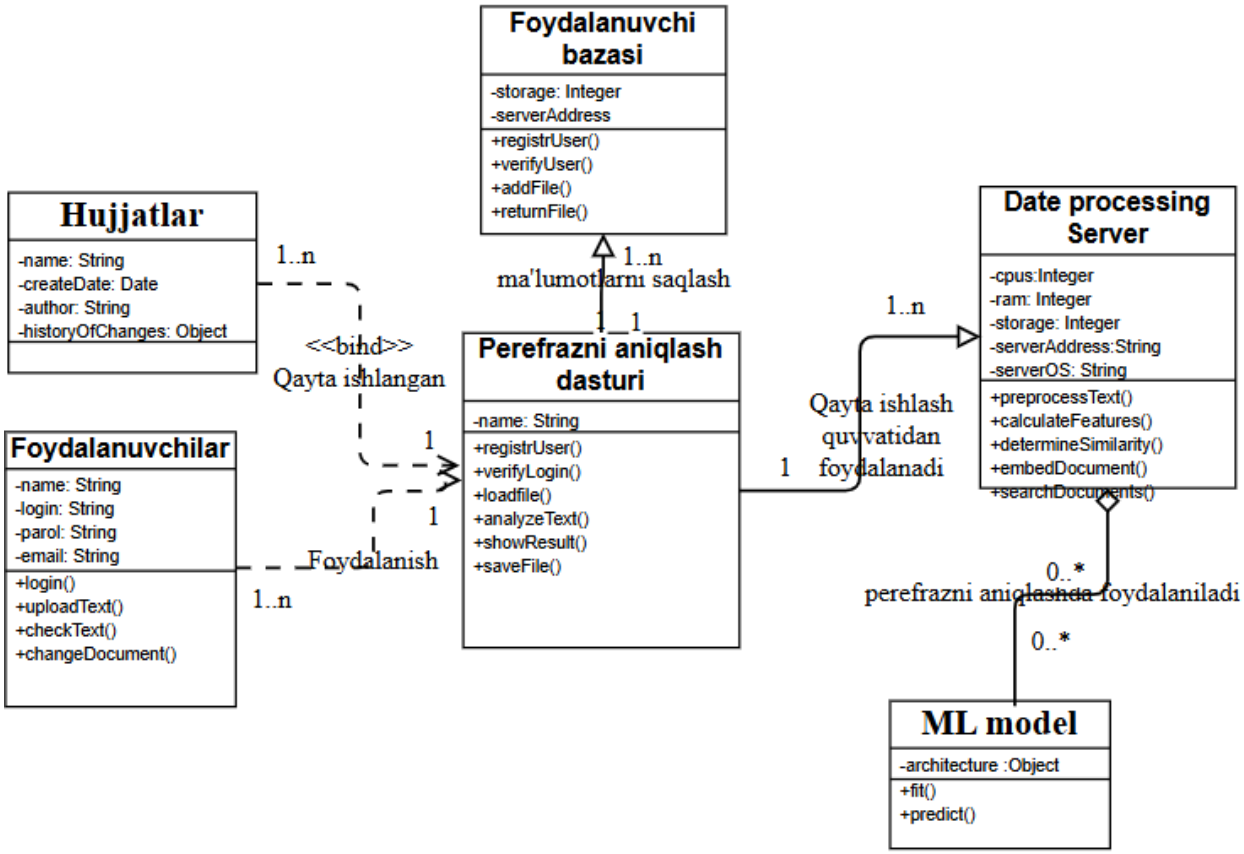
talaba → o‘quvchi
ilm → bilim
universitet → oliygoh
bordi → yo‘l oldi

Axborot manbasi (shu jumladan ijtimoiy tarmoqlar) bir xil mavzuga tegishli nashrlar yoki savollarni birlashtirish orqali tarkibni siyrakligini kamaytirish uchun tizimdan foydalanishi mumkin. Shuningdek, ushbu resurslar matn muallifini aniqlash uchun perefrazni aniqlash tizimidan foydalanishi mumkin. Ikkala mashina tarjimasi dasturlari ham, QA (savol-javob) tizimlari ham taklif etilayotgan ATning funksiyalaridan kirish xabarini soddalashtirish uchun foydalanishi mumkin. Bundan tashqari, QA tizimlari savollarga javoblarni ilg‘orroq izlash uchun perefrazni aniqlashni qo‘llashi mumkin.

- Class diagrammasini yaratish uchun yettita asosiy class aniqlangan:
- Foydalanuvchilar-axborot tizimining individual foydalanuvchilari,
- Hujjatlar-axborot tizimiga tekshirish uchun yuklangan hujjatlar,
- Perefrazni aniqlash dasturi-hujjat tarkibidagi perefraz jumlalarni aniqlash,
- Foydalanuvchilar bazasi -foydalanuvchilarning tizimdan foydalanishi haqidagi ma’lumotlar,
- Data Processing Server-hujjatlar ustida bajariladigan jarayonlar,
- ML Model- Mashinali o’qitish modellari.

3-rasmdagi sinf diagrammasi tizimdan faqat individual foydalanuvchi o‘z matnlarining noyobligini tekshirish uchun foydalanishni ko‘rsatadi. Tashqi dasturlar tomonidan foydalaniladigan perefrazani aniqlash axborot tizimining dastur sinf diagrammasi mezbon dastur bilan so‘rovlar yordamida muloqot qilish uchun qo‘shimcha usullarni belgilashi kerak. Foydalanuvchi hujjatni yuklab olish, o‘zgartirish, tekshirish va natijalarni olish imkoniyatiga ega.

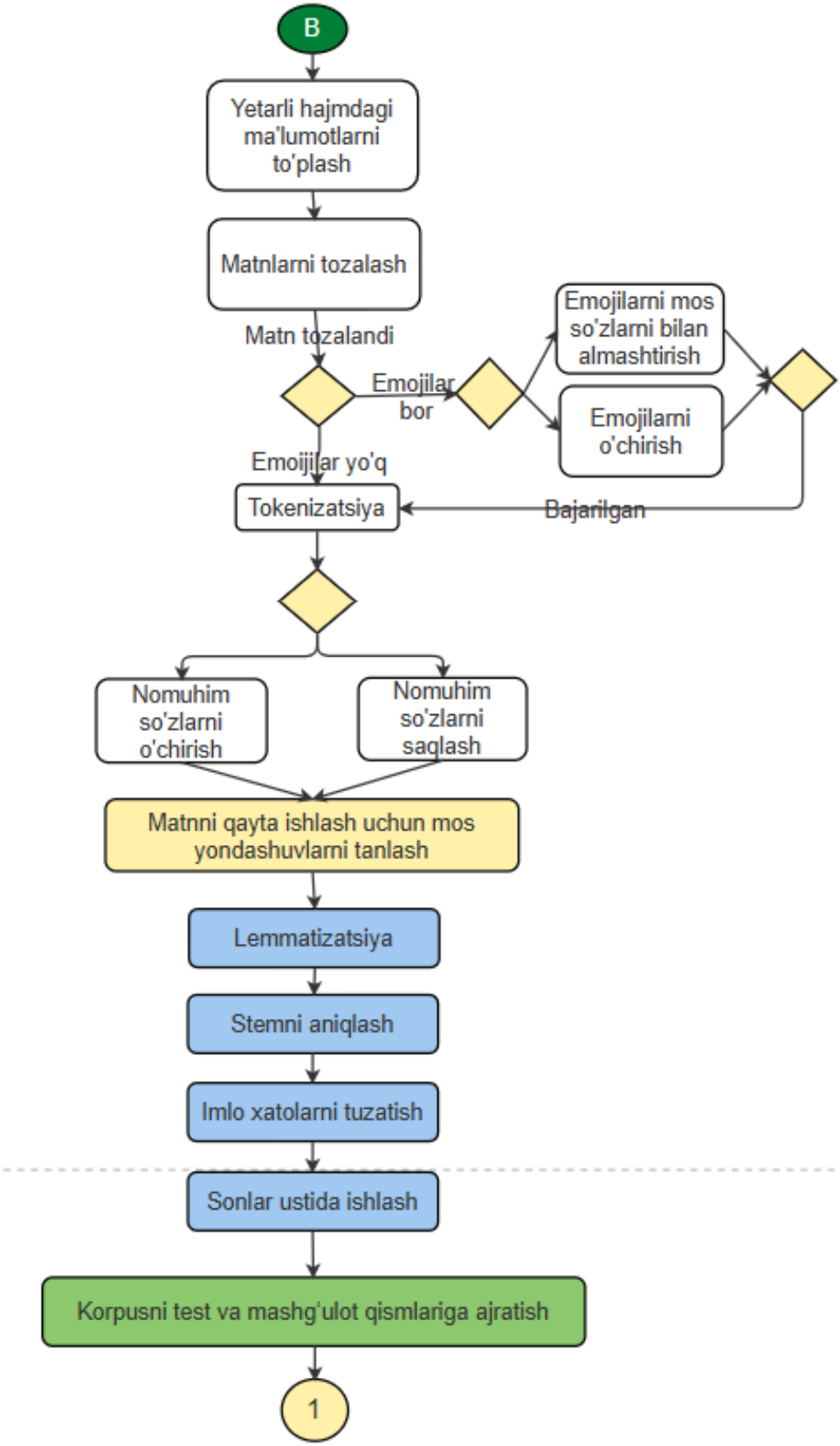
ISSN 2181-371X
SJIF: 5.305 VOLUME 4 ISSUE 34
INNOVATIONS IN TECHNOLOGY AND SCIENCE EDUCATION



1-rasm: Matndagi perefrazni aniqlash axborot tizimining ER diagrammasi

Perefraza mavjudligini batafsilroq tekshirish uchun hujjat avval tasniflanishi yoki vektorli ko‘rinishga aylantirilishi kerak. Bu yondashuv hisoblash resurslarini tejash va so‘rovni qayta ishlash vaqtini qisqartirish uchun qo‘llaniladi.

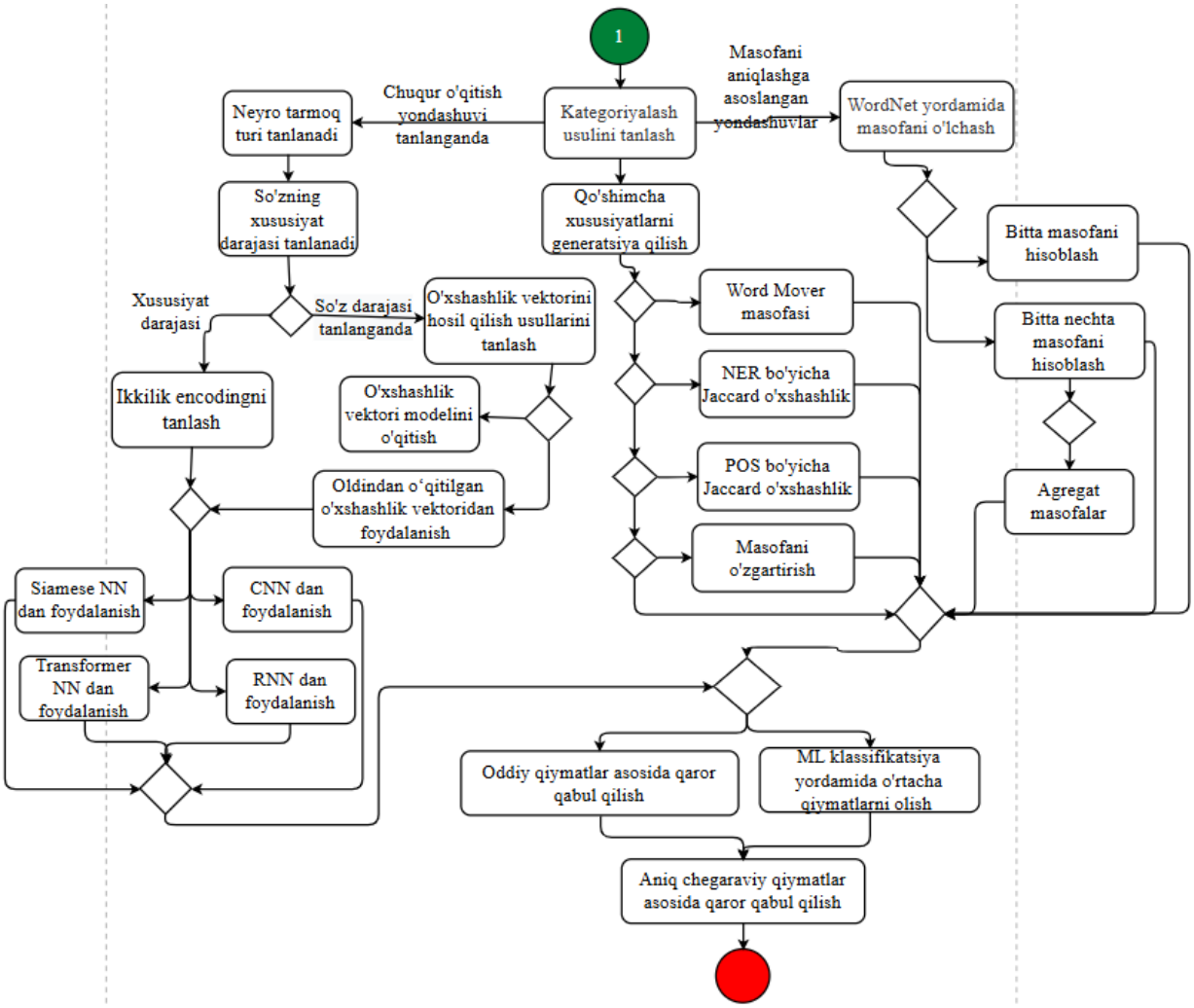
Muhim ichki tizim sozlamalari foydalanuvchi bilan taqqoslanadigan hujjatlar sonini o‘z ichiga oladi. Ko‘p hajmli hujjatlar tekshirish jarayonini sekinlashtiradi, kam hajmli hujjatlar esa uning aniqligi va to‘liqligini kamaytiradi. Semantik o‘xshashlik darajasini aniqlash yoki ikkilik tasniflash natijalari qaytariladi va oldingi oddiy tekshirish natijalari bilan birga vizual tarzda ko‘rsatilishi kerak.



2-rasm: Perefrazni aniqlash modulini yaratishning dastlabki qadamlari

Perefrazani avtomatik aniqlash modulini rivojlantirishning muhim qismi bu

matnlarni tasniflash vazifasi uchun ML modelini ishlab chiqish va tayyorlashdir. Tekshiruv natijalarining maksimal sifatlisini olish uchun masofalarni hisoblashning turli usullarini, ML-algoritmlarini va TPP-usullarini o‘rganish kerak. Faoliyat diagrammasi ikki qismga bo‘linadi (2-3-rasm). TPP va tizim qurilishida barcha mumkin bo‘lgan tarmoqlarni ko‘rsatadi. Shunday qilib, tabiiy tilga ishlov berish axborot tizimi ishlab chiqish jarayonida mumkin bo‘lgan usullar va ularning kombinatsiyalarining tarmoqlanish tizimiga duch kelish tabiiy. Ikkilik tasniflashda korpus hajmining yetarlicha katta va muvozanatli ekanligiga ishonch hosil qilish kerak, chunki ma’lum bir kategoriya bo’yicha jamlangan matnlar sonining ustunligi sun’iy Neyron tarmoqning noto‘g‘ri yechimiga olib kelishi mumkin. Ya’ni har bir kategoriya bo’yicha teng miqdorda matnlar teglanishi lozim.



3-rasm:Matn tarkibidagi perefraz jumlalarni aniqlashda foydalaniladigan usullar ketma-ketligi

Agar dataset yetarli bo‘lmasa, sun’iy ma’lumotlar yaratish yoki mavjud yozuvlarni o‘zgartirish usullari qo‘llaniladi. Qisqa matnlar, ayniqsa ijtimoiy media

postlari tarkibida ko‘proq emojilar bo‘lishi mumkin. Emoji matnning tonal yoki emotsional rang-barangligini ko‘rsatishi mumkin, shuning uchun axborot tizimi ishlab chiquvchilari ularni tegishli so‘zlar bilan almashtirishi yoki olib tashlashi mumkin. Keyingi qadam gaplarni tokenlarga bo‘lishdir. Tokenizatsiya jarayoni ham murakkab bo’lib, uni amalga oshirishda bir qator algoritmlar va til qoidalaridan foydalaniladi. Ya’ni o’zbek tili grammatikasiga ko’ra nafaqat gap, so’z tokenlarga ajratish mumkin, balki so’z birikmalari, iboralar ham token sifatida qabul qilinadi. Navbatdagi qadam matn tarkibidagi nomuhim so’zlar (StopWords) ajratishdir. Nomuhim so’zlar semantik rang-baranglikka ega emas, shuning uchun ularni olib tashlash ixtiyoriydir. TPPning keyingi bosqichlari lemmatizatsiya (so‘zning dastlabki shakli), stemming (so‘z ildizi), imlo xatolarini tuzatish va raqamlarni olib tashlashni o‘z ichiga oladi. Nomuhim so‘zlarni olib tashlash bilan bir qatorda, yuqoridagi bosqichlarni amalga oshirish ixtiyoriy, ammo tasniflash aniqligiga turlicha ta’sir ko‘rsatishi mumkin. TPPdan so‘ng, ma’lumotlar to‘plami o‘quv va sinov namunalariga bo‘linadi. 3-rasm jumlalar juftligini ikkilik tasniflash uchun asosiy usullar va ularning perefraz bo‘lishiga qarab kombinatsiyalarini ko‘rsatadi. Yakuniy qarorga kelishda kerakli metrikani qo‘llash yoki logistik regressiyasi kabi ML usullaridan foydalanib bir nechta qiymatlarni yig‘ish mumkin. Perefraz juftliklarni aniqlash uchun o’xshashliklarni aniqlash eng asosiy bosqich bo’lib, unda so’zlar hamda jumlalar orasidagi masofalar aniqlanadi. Xorij tajribasiga ko’ra masofalarni aniqlashda WordNet yordamida hisoblash usulini tanlash, masofalarni birlashtirish usuli, so‘zlarni vektor ko’rinishiga olib kelish kabi jarayonlarni amalga oshirish lozim. Ammo, o’zbek tilining Wordnet tizimini ishlab chiqish yakunlanmaganligini hisobga olsak, vazifani amalga oshirishda boshqa usullarni qo’llash maqsadga muvofiq. Perefrazni avtomatik aniqlash tizimining joylashuv diagrammasi ma’lumotlarni qayta ishlash serveri protsessorlari va ikkita ma’lumotlar bazasini ko‘rsatadi. Foydalanuvchi axborot tizimi funksiyalaridan yuklab olingan dasturning grafik interfeysi yoki onlayn platforma orqali foydalanishi mumkin. Ma’lumotlar bazasi foydalanuvchilar va hujjatlar haqidagi ma’lumotlarni saqlaydi.

ML uchun eng mashhur dasturlash tillari Python, R, Java, C++. Perefrazni aniqlash axborot tizimini ishlab chiqishda Pythonning boshqa tillardan ustunligi ko‘plab kutubxonalarni qo‘llab-quvvatlashida:

- ML usullari bilan ishlash: Scikit-Learn;
- Sun’iy NNlar va chuqur o‘rganish yaratish: TensorFlow, Keras, PyTorch;
- Tabiiy tilni qayta ishlash: NLTK, spaCy, WordNet;
- Massivlar va matritsalar bilan ishlash: NumPy;
- Jadvallar bilan ishlash: pandas;

- Ma’lumotlarni vizualizatsiya qilish (shu jumladan interaktiv): Matplotlib, seaborn, Plotly.

Scikit-Learn kutubxonasi ma’lumotlar o‘lchamlarini kamaytirish, regressiya, tasniflash yoki klaster tahlili uchun ML modellarini tanlashni qo‘llab-quvvatlaydi. Biroq, Scikit-Learn chuqur o‘rganish modellarini qurish uchun to‘liq qo‘llab-quvvatlashga ega emas. Sun’iy NN yaratish uchun Keras kutubxonasi tanlangan bo‘lib, u TensorFlow2 uchun yuqori darajali API sifatida xizmat qiladi. Keras graf shaklida ketma-ket modellarni qurishga imkon beradi, uning uchlari ma’lum turdagi va tugunlar soniga ega qatlamlardan iborat. Shuningdek, Keras yordamida NNning bir nechta alohida qismlarining natijalarini birlashtirib, ularni keyinchalik qayta ishlash mumkin. TensorFlow tayyorlangan ML modelini import qilish va uni boshqa dasturlarda foydalanish imkonini beradi. TensorFlow shuningdek, CPU, GPU va tensor protsessor birliklaridan foydalanib past darajadagi tensor operatsiyalarini qo‘llab-quvvatlaydi. Ushbu tadqiqotda NLTK kutubxonasi TPP uchun ishlatiladi:

- *tokenizatsiya;*
- *nomuhim so‘zlarni olib tashlash;*
- *stemming;*
- *lemmatizatsiya.*

Shuningdek, ushbu kutubxonadan olingan funktsiyalar yordamida eng mashhur n-grammalarni va individual tokenlar tilining qismlarini aniqlash, nomlangan obyektlarni tanib olish (NER-Name Entity Recognition) mumkin. Tabiiy til bilan ishlashni soddalashtiradigan qo‘shimcha kutubxonalar Regex va emojini o‘z ichiga oladi - mos ravishda regulyar ifodalar va emojilarni so‘zlarga almashtirish uchun foydalaniladi.

Foydalanilgan adabiyotlar

1. Bhagat, R., Hovy, E. What is a paraphrase? Computational Linguistics, 39 (3): 463–472 (2013).
2. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., LopezGazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G.; Uria, L., Wiebe, J. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. Proceedings of SemEval 2015 (2015)
3. Vila, M., Martí, M. A., Rodríguez, H. Is this a paraphrase? What kind? Paraphrase boundaries and typology. Open Journal of Modern Linguistics, 4 (01), 205 (2014)

4. Barrón-Cedeño, A., Vila, M., Martí, M. A., Rosso, P. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics, 39(4), 917–947 (2013).

5. Petrović, S., Osborne, M., Lavrenko, V. Using paraphrases for improving first story detection in news and Twitter. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 338–346. Association for Computational Linguistics (2012)

6. Pavlick, E., Nenkova, A. Inducing lexical style properties for paraphrase and genre differentiation. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 218– 224 (2015)

7. Wieting, J., Bansal, M., Gimpel, K., Livescu, K. Transactions of the Association for Computational Linguistics, vol. 3, pp. 345–358 (2015)

8. Hintz, G. Data-driven Paraphrasing and Stylistic Harmonization. In Proceedings of NAACL-HLT, pp. 37–44 (2016).

9. Eshkol-Taravella, I., Grabar, N. Paraphrastic Reformulations in Spoken Corpora. In International Conference on Natural Language Processing, pp. 425–437, Springer International Publishing (2014).

10. Li Z., Jiang X., Shang L., Li H., Paraphrase Generation with Deep Reinforcement Learning, ArXiv, 2017. DOI: 10.48550/arXiv.1711.00279.

11. Gomaa W., Fahmy A., SimAll: A flexible tool for text similarity, in: Language Engineering, 2017, https://www.academia.edu/35381793/SimAll_A_flexible_tool_for_text_similarity