

KOMPYUTER LINGVISTIKASIDA QO'LLANILADIGAN METODLAR

Mohichehra Abramatoва Abdurazzoq qizi

Termiz davlat universiteti o'zbek filologiyasi fakulteti

abramatovamohichehra@gmail.com

Ilmiy rahbar. Tolipov Izatullabek Abdulhoy ogli

<https://doi.org/10.5281/zenodo.8084752>

Annotatsiya: Fan-texnika taraqqiyotining hozirgi bosqichida ilmiy bilimlar ko'p tarmoqli, fanlararo va transdisipliner loyihalarga kiritilib, o'z imkoniyatlarini to'liq ochib bermoqda. Shu bilan birga, ilmiy ontologiya tabiiy tilda yozilgan katta hajmdagi ma'lumotlarga asoslanadi. Ushbu tendentsiyalar amaliyotdan kelib chiqadigan murakkab muammolarga javob berishga qodir bo'lgan tarmoq sifatida hisoblash va eksperimental tilshunoslikning faol rivojlanishiga yordam beradi.

Ko'rib chiqilgan ishlarda ma'ruzachining ovozi va yozma nutqi bo'yicha hissiy holatlarni aniqlash, yozma matnlarda muallifning tajribasi va huquqiy hujjatlarni avtomatlashtirilgan tasdiqlash, shuningdek, hisoblash tilshunosligi sohasida tizimli terminologiya va qurilish metodologiyasi kabi vazifalar.

Kalit so'zlar: avtomatlashtirilgan identifikatsiya va tasniflash; tabiiy tilni qayta ishlash; semantik grafik; ontologiya; mantiqiy qoidalar; ovoz; yozma matn; yuridik matn; ekspertiza; tasdiqlash; sun'iy neyron tarmoqlari; hissiy holat; tilshunoslikda statistik usullar; stilmetriya; avtografik tekshiruv; ironiyaning grafik belgilari; terminologiya; maxsus lug'at; matematik tilshunoslik; korpus tilshunosligi; hisoblash tilshunosligi; terminologiya.

Fan-texnika taraqqiyotining hozirgi bosqichida ilmiy bilimlar ko'p tarmoqli, fanlararo va transdisipliner loyihalarga kiritilib, o'z imkoniyatlarini to'liq ochib bermoqda. Shu bilan birga, ilmiy ontologiya tabiiy tilda yozilgan katta hajmdagi ma'lumotlarga asoslanadi. Ushbu tendentsiyalar amaliyotdan kelib chiqadigan murakkab muammolarga javob berishga qodir bo'lgan tarmoq sifatida hisoblash va eksperimental tilshunoslikning faol rivojlanishiga yordam beradi.

. Ko'rib chiqilgan ishlarda ma'ruzachining ovozi va yozma nutqi bo'yicha hissiy holatlarni aniqlash, yozma matnlarda muallifning tajribasi va huquqiy hujjatlarni avtomatlashtirilgan tasdiqlash, shuningdek, hisoblash tilshunosligi sohasida tizimli terminologiya va qurilish metodologiyasi kabi vazifalar. A.Isxakova, D.Volf va R.Meshcheryakovlar ijtimoiy internet tarmog'ida joylashtirilgan ma'ruzachi ovozining audioyozuvidan emotsional holatni avtomatlashtirilgan tarzda aniqlash algoritmini ishlab chiqmoqda. Tadqiqotda har biri 1 dan 4 soniyagacha davom etadigan 5255 ta emotsional nutqning audioyozuvlaridan iborat korpusdan foydalaniladi. Korpus SAVEE, RAVDESS, TESS, Emo-DB ma'lumotlar bazalaridan yig'ilgan. Rivojlangan neyron tarmog'ini sinab ko'rish uchun korpus birliklari o'rtacha hisoblangan. Algoritm chuqur o'rganish printsipi bo'yicha o'qitildi (o'quv namunasi 3941 audio yozuv edi). Tajriba uchun bir guruh erkak va ayol ovozlari uchun 8 ta hissiy holat (g'azab, charchoq, jirkanish, qo'rquv, quvonch, qayg'u, ajablanish, neytral hissiy holat) matritsasi ishlatilgan.

Emotsional holatni aniqlash bo'yicha tasniflash topshirig'ini hal qilishda mutlaq aniqlik 47,56% ni tashkil etdi (eng yaxshi aniqlangan hissiy holatlar ajablanib - 63%, jirkanish - 62, g'azab - 56, qo'rquv - 58, quvonch - 54, bosib chiqarish - 47%) , ma'ruzachining jinsini aniqlashda - 90%. Audio yozuvlarni emotsional holat turiga ko'ra uch guruhga birlashtirish:

ijobiy (quvonch, ajablanib), salbiy (g'azab, jirkanish, qo'rquv) va aralash (qayg'u, zerikish, neytral hissiy holat) ularni aniqlashning aniqligini oshirishga imkon berdi. 66,6% gacha. Yumshoqlashtirish (dasturiylashtirish) muammosining bir qismi sifatida - mutaxassislarni odamlar bilan bir xil natijalarni beradigan raqamli xizmatlar bilan almashtirish - assotsiativ-ontologik yondashuvga asoslangan huquqiy hujjatlarni tasdiqlash usuli ko'rib chiqiladi.

Mutaxassisning xatti-harakatlarini modellash, mualliflar yuridik matnni tekshirishda matematik model bajarishi kerak bo'lgan quyidagi harakatlar ketma-ketligini aniqlaydi: 1) matnni til me'yorlariga muvofiqligini, tushunchalarning mumkin bo'lgan noaniqligini, bayonotlarning to'liqligi va to'liqligini, korrelyatsiyani tekshiring. matn mazmunining huquqiy va axloqiy me'yorlarga ega bo'lishi; 2) hujjatda har bir tomon tomonidan bir ma'noda o'qiladigan majburiy va kerakli yuridik til mavjudligini tekshirish; 3) hujjatning majburiy bo'limlari - tarkibiy elementlarning mavjudligini tekshirish. Mualliflarning fikriga ko'ra, bugungi kunda hujjatlarni avtomatik tahlil qilishda dominant yondashuv past unumdorlikni ko'rsatadigan so'zlarning paydo bo'lishiga asoslangan statistik tahlil bo'lib qolmoqda.

Mualliflar tomonidan taklif qilingan yondashuv, tekshirish uchun taqdim etilgan yuridik matndan avtomatik ravishda olingan tizimli elementlarni ushbu turdagi matnning zarur va o'zgaruvchan elementlarining tuzilishini aniqlaydigan shablon bilan taqqoslashga asoslangan. Ushbu yondashuv quyidagi tamoyillarga asoslanadi: - qoidalar majmuasi strukturaviy elementlarni belgilaydi; - qoidalar haqiqiy huquqiy hujjatlar elementlarining tuzilishi tavsiflari va advokatlar tomonidan qabul qilinadigan va nomaqbul foydalanish misollari asosida shakllantiriladi; - natijada olingan qoidalar izohlovchi xususiyatlarga ega va foydalanuvchi tomonidan tahrirlanishi mumkin. Shablon tuzilmasi federal qonunlar va amaldagi huquqiy me'yorlar asosida qo'lda yaratiladi. Qoidalar qo'lda aniqlangan struktura shabloni asosida qurilgan va semantik grafik sifatida taqdim etilgan. Qurilish strukturaviy elementning to'g'ri matnini kiritish yoki tarkibiy elementni tavsiflovchi semantik grafikni tahrirlash, shuningdek, assotsiativ-ontologik yondashuv asosida belgilangan haqiqiy hujjatlarning bo'laklarini avtomatik o'rganishdan foydalanish orqali qo'lda amalga oshirilishi mumkin.

Bundan tashqari, maqolada ishlab chiqilgan algoritmning har bir moduli va ularning funksionalligi tavsifi berilgan. Algoritmni sinab ko'rish maqsadida Rossiya kompaniyalarida qo'llaniladigan "shaxsiy ma'lumotlarni qayta ishlashga rozilik" turidagi 62 ta hujjat materiali bo'yicha tajriba o'tkazildi. Yuklab olingan hujjatlarning algoritm qo'llagan shablon tuzilishiga (16 ta bo'limni o'z ichiga oladi) muvofiqligini tekshirish 76% aniqlikni ko'rsatdi. Yaratilgan qoidalarni qo'lda tuzatish bilan mashinani o'rganishdan foydalangan holda tekshirishning aniqligi 68% ni, avtomatik ravishda yaratilgan qoidalardan foydalanganda - 82% ni tashkil etdi. T. Litvinova matn uslubida "o'zgarishlar" ni keltirib chiqaradigan turli janrdagi yozma matnlar orqali muallifni aniqlash mumkinmi, degan savolga javob topishga harakat qilmoqda. Stilmetriya metodologiyasidan foydalangan holda, muallif RusIdiolect2 ma'lumotlar bazasidan olingan rus tilida so'zlashuvchi talabalarning 206 yozma matnlaridan foydalanadi. Har bir ma'lumot beruvchi ikkita matnni taqdim etdi: birinchisi - bir qator rasmlarning tavsifi, ikkinchisi - o'nta mavzudan biriga oid insho. Shunday qilib, T. Litvinova bahsli matnga mualliflikni aniqlash uchun taqqoslash amalga oshiriladigan, tasdiqlangan matn bilan birga keladigan vaziyatni modellash tiradi.

Tadqiqotda "kontekstdan mustaqil" deb ataladigan quyidagi stilistik xususiyatlar ishlatilgan

[o'sha erda, p. 281–283]:

- 1) eng tez-tez uchraydigan tokenlar (so'z shakllari);
- 2) leksik murakkablik ko'rsatkichi;
- 3) leksik xilma-xillik indeksi;
- 4) o'qish qobiliyati indeksi;
- 5) morfologik xilma-xillik indeksi;
- 6) morfologik ko'rsatkichlar;
- 7) n-gramm morfologik belgi va tinish belgilari.

Lematizatsiya va morfologik belgilar Treetagger¹ yordamida amalga oshirildi. Ma'lumotlarni tahlil qilish jarayonida quyidagi statistik usullar qo'llanildi: Pearson korrelyatsiya koeffitsienti, dispersiyaning bir tomonlama tahlili (ANOVA) va eng yaqin markazlashtirilgan siqish usuli (eng yaqin qisqargan markaz - NSC). Olingan ma'lumotlar tahlili natijasida ma'lum bo'ldiki, matnning janri va predmeti o'zgarganda qo'llanilgan aksariyat uslubiy xususiyatlar to'liq barqaror bo'lmagan. Faqat leksik murakkablik, leksik xilma-xillik va morfologik xilma-xillik ko'rsatkichi bilan bog'liq xususiyatlar matn uslubining o'zgarishiga qarshi immunitetga ega edi. T. Litvinova leksik va morfologik xilma-xillik bilan bog'liq xususiyatlarni ushbu xususiyatlarning ayrimlarining matn uzunligi bilan yuqori bog'liqligini hisobga olgan holda tanlash kerakligini aniqlaydi.

Leksik murakkablik bilan bog'liq xususiyatlarning janr, mavzu va matn uzunligining mustaqilligi ular til qobiliyatining etarlicha barqaror xususiyatlarini aks ettirganligi bilan bog'liq bo'lishi mumkin. Vaziyat (matn janri) va barqaror interlesional o'zgaruvchanlik (jins, yosh - maktab o'quvchilari / talabalarning yosh guruhi, mualliflarning psixologik xususiyatlari - Frayburg shaxsiyati so'rovi) omillarining o'zaro ta'sirini tahlil qilishni davom ettirib, T.A. Litvinova va uning hamkasblari matnni tasniflash algoritmlarini ishlab chiqish uchun korpus va hisoblash lingvistikasi usullaridan foydalanadilar. Tajribalarda do'stga xat, rasm tavsifi (jami 1500 ta murojaatchi), shuningdek LiveJournal platformasidagi blog matnlari (2364 ta blog) kabi janrlar tahlil qilinadi. Matn xususiyatlari sifatida quyidagi markerlar ishlatilgan: eng tez-tez uchraydigan belgilar, shu jumladan tinish belgilari; eng tez-tez uchraydigan lemmalar; belgilarning trigrammalari; morfologik xususiyatlar; sintaktik munosabatlar; leksik xilma-xillik o'lchovlari. Tahlil usullari sifatida - qisman eng kichik kvadratlar (PLS-DA) yordamida asosiy komponentlar va diskriminant tahlillari usuli. Tajribalar turli janrdagi matnlar uchun o'tkazildi.

Tadqiqot shuni ko'rsatdiki, "idiolekt belgilarining turli guruhlarini qo'llashda "jins" va "yosh" signallari tasniflagichlar tomonidan tan olinadi, "psixologik" signal esa kamroq aniqlanadi. Tahlil qilingan matnlar orasida "gender" signali katta yoshdagi odamlarda ko'proq namoyon bo'ladi. "Muallifning "men"ining yuksak darajada ifodalanishini, shaxsning ongli qurilishini (do'stga maktub, blog) ifodalovchi janrlar matnlarini yaratuvchilarining jinsi va yoshiga qarab tasniflash mumkin, matnlar esa muallifning jinsi va yoshiga qarab tasniflanishi mumkin. "Rasm tavsifi" janrini bu asoslarga ko'ra tasniflash mumkin emas. Yozma muloqot jarayonida istehzoni aniqlashga imkon beradigan belgilarni izlashda P.Mikhailova ma'ruzachining intonatsiyasini etkazish uchun tinish belgilari va harflarni takrorlash, yuz ifodasini etkazish uchun qavs va emoji kabi grafik vositalarga murojaat qiladi. so'zlovchining ifodalari va imo-ishoralari. Tasniflash maqsadida muallif quyidagi belgilarning qarama-qarshiliklaridan foydalanadi: me'yoriy imlo/unli takrori, me'yoriy tinish belgilari/ortiqcha tinish belgilari, bir

nechta qavs (1–2)/qavslardan ortiqcha foydalanish (4 dan ortiq). Shu bilan birga, "normativ" adabiy me'yor sifatida emas, balki ataylab buzishdan farqli o'laroq, ko'proq yoki kamroq to'g'ri imlo sifatida tushuniladi. Kundalik muloqotning neytral mavzularida rus tilidagi 192 ta qisqa suhbatlar materiallari bo'yicha eksperimental tadqiqot o'tkazildi.

Rag'batlantiruvchi materialda ikkita satr bor edi: birinchi qator har doim o'zgarmagan, ikkinchi qatorga turli xil belgilar qo'shilgan, muallif ta'kidlaganidek, istehzo va kinoyaning turli xil variantlarini etkazishi mumkin. Ironiya va kinoyani aniqlash omillari sifatida muallif sakkizta marker kombinatsiyasidan foydalangan.

Tadqiqot quyidagi gipotezalarni tekshirishga qaratilgan edi: 1) barcha ko'rib chiqilgan omillarning o'rtacha qiymati A omilining o'rtacha qiymatidan oshadi; 2) S omilning o'rtacha qiymati B omilining o'rtacha qiymatidan oshadi; 3) bir nechta markerlarni - E, G va H omillarini birlashtirgan o'rtacha qiymatlar bitta markerni - B, C, D va D omillarini birlashtirgan o'rtacha qiymatlardan oshib ketadi. 350]. Tadqiqotda 143 respondent ishtirok etdi (o'rtacha yoshi 18-23 yosh). Ulardan dialoglarni o'qish va har birini o'lchov bo'yicha belgilash so'ralgan:

- 1) taklif neytral;
- 2) jumla kinoyadan ko'ra betarafroq,
- 3) gap neytraldan ko'ra istehzoli,
- 4) gap istehzoli / kinoyali.

Ma'lumotlar tavsiflovchi statistika, bir tomonlama dispersiya tahlili (ANOVA) va sezilarli farq testi (Tukey testi) yordamida tahlil qilindi. Natijada birinchi va ikkinchi farazlar tasdiqlandi, uchinchi qisman tasdiqlandi.

Shunday qilib, tadqiqot natijalariga ko'ra, yozma muloqotda istehzoni aniqlashda eng muhim belgi ko'p sonli qavslar yoki takrorlangan unlilar yoki haddan tashqari tinish belgilari bilan ko'p sonli qavslarning kombinatsiyasi hisoblanadi. Eksperimental va matematik tilshunoslik doirasida ko'rib chiqilgan tadqiqot materiali, olingan natijalarning amaliy ahamiyati bilan bir qatorda, nafaqat tegishli sohalaridagi usullar va yondashuvlar kombinatsiyasiga asoslangan maxsus tadqiqot metodologiyasini loyihalash zarurligiga e'tibor qaratadi. balki fanlararo terminologiyaning paydo bo'layotgan qatlamiga ham. O.N.ning so'zlariga ko'ra. Proxorova va uning hamkasblari tomonidan "hisoblash tilshunosligidagi vazifalar ko'lam, ularni hal qilish uchun ilmiy apparatlar va amaliy qo'llanmalar doimiy ravishda kengayib bormoqda, bu esa ushbu sohada ko'plab yangi atamalarining paydo bo'lishiga olib keladi".

Mualliflar "yadro-periferiya" modelidan foydalangan holda hisoblash tilshunosligi terminologiyasini tizimli ravishda tavsiflashga harakat qilishadi va yadro sifatida tabiiy tilni qayta ishlash (NLP) bilan bog'liq terminologiyani, periferiyaga - matematika, tilshunoslik terminologiyasini belgilaydilar. informatika, sun'iy intellekt va boshqa fanlar va nazariyalar. O'rganilayotgan materiallarda yadroning ulushi 28,2% ni, periferiya ulushi 71,8% ni tashkil etdi. Shu bilan birga, «periferik zonada tarmoqlararo atamalar soni yadroga qaraganda ancha ko'p. Periferiyadan o'zakgacha bo'lgan yo'nalishda hisoblash tilshunosligining maxsus nomlari takomillashtirilib, konkretlashtiriladi.

Tayanch atamalar yuqori ixtisoslashuv darajasiga ega bo'lib, periferik umumiy ilmiy va fanlararo atamalar asosida shakllanadi" [o'sha yerda, b. 34]. Maqolada atama shakllanishidagi tizimli ulanishlarning uchta asosiy turi tasvirlangan:

- 1) periferiyaning asl atamasiga aniqlik kiritish orqali yadro terminini olish (masalan,

matematik statistikaning "klasterlash" atamasidan hisoblash tilshunosligining asosiy terminologiyasining bir qismi bo'lgan "matnlarni klasterlash" iborasi shakllanadi.);

2) o'zak terminini atrof-muhitning o'xshash atamalari tarkibiga asoslanib olish (masalan, "avtomatik nutq sintezi" iborasini shakllantirishda umumiy ilmiy "sintez" atamasi va "nutq" lingvistik atamalarining ishtiroki. o'rganilayotgan terminologiyaning asosiy qismida);

3) periferiya atamalarini bosqichma-bosqich aniqlashtirish orqali yadro terminini olish (masalan, asl periferik "tahlil" atamasidan lingvistik terminologiyaning "sintaksis tahlili" iborasi olingan, bu esa o'z navbatida , kompyuter tilshunosligi terminologiyasi o'zagiga kiritilgan "avtomatik sintaktik tahlil" iborasi shakllandi).

Ma'lumotlar korpusiga asoslangan empirik tadqiqotning metodologik istiqboli maqolada tasvirlangan. Maqolada korpus lingvistikasining evolyutsiyasini ko'rib chiqishga harakat qilinadi va "bu ilmiy yo'nalishga yangi turtki bergan kompyuter texnologiyalarining paydo bo'lishi" deb ta'kidlaydi. Muallif ta'kidlaganidek, zamonaviy korpus lingvistikasi "elektron korpus shaklida saqlanadigan kuzatiladigan ma'lumotlarga empirik yondashuvni amalga oshirish, tillarni tavsiflash, tahlil qilish va o'rgatish uchun innovatsion imkoniyatlarni ta'minlaydigan til o'rganishning o'ziga xos uslubiy vositasidir". Maqolada tadqiqot korpusining umumlashtirilgan mezonlari keltirilgan, ba'zi taniqli lingvistik korpuslar (The Bank of English, British National Corpus, Corpus de Referencia del Español Contemporáneo, Corpus Diacrónico del Español) qisqacha tavsiflangan va lingvistik korpuslar uchun tasniflash mezonlari keltirilgan. Xulosa qilib aytganda, muallif "hozirda korpus har qanday lingvistik tadqiqotlar, xususan, hisoblash lingvistikasi uchun yagona manba hisoblanadi. Uning asosiy afzalliklari shundaki, u tilning haqiqiy namunalaridan iborat bo'lib, natijalar va xulosalarning ob'ektivligini ta'minlaydi, shuningdek, ma'lum bir nazariyaning ishonchliligini, uning afzalliklari va kamchiliklarini oddiygina tekshirish imkonini beradi.

Korporadan ma'lumotlar manbai sifatida foydalanishning alohida ahamiyati ona va chet tillarini o'rgatish yoki didaktik materiallar, lug'atlar, grammatika va mashina tarjimasi yoki nutq texnologiyalari bilan bog'liq boshqa mahsulotlarni ishlab chiqish jarayonida topiladi. Ushbu sharhda taqdim etilgan ishlar amaliy farazlarni sinab ko'rishga qaratilgan kompyuter, matematik va eksperimental tilshunoslik sohasidagi tadqiqotlarning rivojlanishidan dalolat beradi. Ushbu fan yo'nalishlarida olib borilgan tadqiqotlar natijalari katta ma'lumotlarga va ularni qayta ishlashning aniq usullariga tayangan holda o'rganilayotgan ob'ektlar va hodisalar to'g'risida bilimlarni to'plash imkonini beradi, bu esa ushbu natijalar asosida tuzilgan ontologiyalar o'rtasidagi muvofiqlik ehtimolini oshiradi. o'rganilayotgan haqiqat.

References:

1. Литвинова Т.А., Бирюкова Е.Д., Загоровская О.В. Диагностирование пола, возраста и психологических характеристик автора письменного текста с использованием методов корпусной и компьютерной лингвистики: возможности и ограничения // Вопросы криминологии, криминалистики и судебной экспертизы. – 2021. – 1 (49). – С. 105–111. – URL: https://elibrary.ru/download/elibrary_46223195_72711007.pdf
2. Системность терминологии компьютерной лингвистики / Прохорова О.Н., Польщикова О.Н., Польщикова А.К., Деев А.В. // Известия Юго-Западного государственного университета. Серия: Лингвистика и педагогика. – 2021. – № 11 (1). – С. 29–39. – URL: https://elibrary.ru/download/elibrary_45640867_50723245.pdf

3. Чилингарян К.П. Корпусная лингвистика: теория vs методология // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. – 2021. – Т. 12, № 1. – С. 196–218. – DOI: 10.22363/2313-2299-2021-12-1-196-218. – URL: <http://journals.rudn.ru/semiotics-semantics/article/view/25814/19099>
4. Iskhakova A., Wolf D, Meshcheryakov R. Automated destructive behavior state detection on the 1 D CNN-based voice analysis // Speech and Computer: 22 nd International Conference, SPECOM 2020 (St. Petersburg, Russia, October 7–9, 2020). – Cham : Springer, 2020. – LNAI 12335. – P. 184–193. – DOI: https://doi.org/10.1007/978-3-030-60276-5_19.
5. Kuleshov S., Zaytseva A., Nenausnikov K. Legal Tech: Documents' validation method based on the associative-ontological approach // Speech and Computer 22 nd International Conference, SPECOM 2020 (St. Petersburg, Russia, October 7–9, 2020). – Cham : Springer, 2020. – LNAI 12335. – P. 244–254. – DOI: https://doi.org/10.1007/978-3-030-60276-5_25.
6. Litvinova T. Stylometrics features under domain shift: Do they really «contextindependent»? // Speech and Computer: 22 nd International Conference, SPECOM 2020 (St. Petersburg, Russia, October 7–9, 2020). – Cham : Springer, 2020. – LNAI 12335. – P. 279–290.– DOI: https://doi.org/10.1007/978-3-030-60276-5_28..