TABIIY VA SUN’IY YARATILGAN MATNLARNI FARQLASHDA LINGVISTIK-STATISTIK BELGILAR TAHLILI

Muallif: Fozilov Murodjon Marifjon o‘g‘li
Tashkilot: Andijon davlat universiteti

Annotatsiya. Maqolada tabiiy matnlar hamda sinonimlash va Markov zanjirlari yordamida yaratilgan sun’iy matnlarni farqlash masalasi lingvistik-statistik belgilar nuqtayi nazaridan ko‘rib chiqiladi. Manbada keltirilgan 21 ta miqdoriy xususiyat, jumladan so‘z uzunligi, tinish belgilari, xizmat so‘zlari, gap tuzilishi, grammatik xatolar va semantik yadro ko‘rsatkichlari qiyosiy tahlil qilinadi. Natijalar ayrim belgilar sun’iy matn sinflarini ajratishda foydali bo‘lishini, biroq bitta belgi asosida ishonchli xulosa chiqarish yetarli emasligini ko‘rsatadi. Shu sababli belgilar majmuasidan foydalanish va matnning turli lingvistik darajalarini birgalikda baholash zarurligi asoslanadi.

Kalit so‘zlar: sun’iy matn, tabiiy matn, lingvistik belgi, statistik tahlil, invariant, Markov zanjiri, sinonimlash.

Kirish

Matnlarning avtomatik yaratilishi internet kontenti, elektron hujjatlar va axborot almashinuvi ko‘lamining ortishi bilan dolzarb tadqiqot yo‘nalishiga aylandi. Matnning tabiiy yoki avtomatik yaratilganligini aniqlashda faqat mazmunni o‘qish bilan cheklanib bo‘lmaydi: so‘zlar, belgilar, gaplar va matn tuzilishiga oid o‘lchanadigan xususiyatlarni ham tahlil qilish talab etiladi.

Ushbu maqolaning maqsadi — tabiiy matnlar bilan sinonimlash hamda Markov zanjiri usullarida yaratilgan matnlar o‘rtasidagi farqlarni manbada berilgan lingvistik-statistik ko‘rsatkichlar asosida tavsiflash. Maqola yangi tajriba o‘tkazilganini da’vo qilmaydi; tahlil dissertatsiya materialida keltirilgan jadval qiymatlariga tayanadi.

Tadqiqot materiali va usuli

Tadqiqot materiali sifatida manbada tabiiy matn, sinonimlash yo‘li bilan o‘zgartirilgan matn va Markov zanjiri asosida yaratilgan matn uchun berilgan 21 ta xususiyat qiymatlari olindi. Ko‘rsatkichlar orasida o‘rtacha so‘z uzunligi, 1000 belgiga to‘g‘ri keladigan tinish belgilari, xizmat so‘zlari va olmoshlar chastotasi, qisqa va uzun so‘zlar ulushi, gapdagi o‘rtacha so‘zlar soni, grammatik xatolar, murakkab gaplar hamda semantik yadroga oid ko‘rsatkichlar bor.

Taqqoslash tavsifiy-statistik yondashuvda bajarildi: har bir belgi bo‘yicha uchta matn sinfining qiymatlari yonma-yon ko‘rib chiqildi, sinflar orasidagi farq ko‘proq ko‘rinadigan ko‘rsatkichlar ajratildi. Ko‘rsatkichlarning o‘lchov birliklari va hisoblash usullari manbadagi ta’rifga muvofiq qabul qilindi.

Natijalar va muhokama

Manbadagi jadvalga ko‘ra, tabiiy matnlar uchun o‘rtacha so‘z uzunligi 5,83; sinonimlangan matnlar uchun 6,213; Markov zanjiri matnlari uchun 5,851 qiymatida berilgan. Bu ko‘rsatkichning o‘zi uch sinfni to‘liq ajratmaydi: tabiiy va Markov matnlari qiymatlari bir-biriga yaqin, sinonimlangan matn qiymati esa nisbatan yuqori.

Tinish belgilarining 1000 belgiga nisbatan o‘rtacha soni tabiiy matnda 31,742, sinonimlangan matnda 29,035, Markov matnida 21,545 deb qayd etilgan. Demak, ayni tanlanmada Markov usulida yaratilgan matnlar tinish belgilari chastotasi bo‘yicha tabiiy matnlardan sezilarli farq qilgan. Biroq bu kuzatuvni barcha generatorlarga umumlashtirish mumkin emas.

1000 belgiga to‘g‘ri keladigan grammatik xatolar soni jadvalda tabiiy matn uchun 0,01, sinonimlangan matn uchun 6,215, Markov matni uchun 1,002 deb berilgan. Ushbu farq sinonimlash jarayoni ayrim hollarda grammatik moslikni buzishi mumkinligini ko‘rsatadi. Shu bilan birga, grammatik xato darajasi matn muallifining savodxonligi, tahrir sifati va matn mavzusiga ham bog‘liq bo‘lishi mumkin; shuning uchun u mustaqil yakuniy mezon sifatida qo‘llanmasligi lozim.

Gapdagi o‘rtacha so‘zlar soni tabiiy matnda 9,113, sinonimlangan matnda 9,987, Markov matnida 16,549. Bu qiymat Markov zanjiri bilan yaratilgan

matnlarda gaplarning uzunroq bo‘lishi mumkinligini ko‘rsatadi. Biroq boshqa belgilar — masalan, so‘z uzunligi — bir-biriga yaqin chiqishi matnni tasniflash uchun bir nechta belgi birgalikda ishlatilishi zarurligini tasdiqlaydi.

Semantik yadrodagi so‘zlar soni tabiiy matnda 66,025, sinonimlangan matnda 95,645, Markov matnida 75,268 sifatida berilgan. Bu natija sinonimlashdan keyin noyob yoki semantik yadroga kiritilgan so‘zlar soni ortishi mumkinligini ko‘rsatadi. Lekin ushbu ko‘rsatkichning mazmuni ishlatilgan lug‘at va semantik yadro qanday aniqlanganiga bog‘liq bo‘lgani uchun, uni alohida metodik izoh bilan qo‘llash kerak.

Umuman, 21 ta belgi bir xil darajada ajratuvchi emas. Ayrim belgilar sinflar orasidagi farqni yaqqol ko‘rsatsa, boshqalari tabiiy va sun’iy matnlar uchun o‘xshash qiymatga ega. Shuning uchun matn sinfining invariantini shakllantirishda belgilar to‘plamini tanlash, ularning o‘zaro bog‘liqligini hisobga olish va matnning leksik, sintaktik hamda grammatik qatlamlarini birgalikda ko‘rib chiqish muhim.

Xulosa

Manbada berilgan 21 ta lingvistik-statistik belgi qiyosi tabiiy, sinonimlash orqali o‘zgartirilgan va Markov zanjiri yordamida yaratilgan matnlar o‘rtasida ayrim muntazam farqlar borligini ko‘rsatadi. Ayniqsa, tinish belgilari chastotasi, grammatik xatolar va gap uzunligi ko‘rsatkichlari mazkur tanlanmada ajratuvchi axborot beradi.

Bitta xususiyatga tayanish ishonchli tasnif uchun yetarli emas. Amaliy aniqlash tizimida bir nechta miqdoriy va sifat belgilaridan tuzilgan invariantlar majmuasidan foydalanish, sinov tanlanmasini kengaytirish va har bir til hamda matn janri bo‘yicha alohida tekshirish tavsiya etiladi. Keltirilgan sonlar manba dissertatsiyasidagi jadvaldan olingan; ular ushbu maqola doirasida qayta o‘tkazilgan mustaqil tajriba natijasi sifatida talqin qilinmasligi kerak.

Jadval 1. Manbada keltirilgan ayrim xususiyatlar qiymati

Ko‘rsatkich	Tabiiy matn	Sinonimlangan matn	Markov matni
O‘rtacha so‘z uzunligi	5,830	6,213	5,851

1000 belgiga tinish belgilari	31,742	29,035	21,545
1000 belgiga grammatik xatolar	0,010	6,215	1,002
Gapdagi o‘rtacha so‘zlar soni	9,113	9,987	16,549
Semantik yadrodagi so‘zlar soni	66,025	95,645	75,268

Foydalanilgan adabiyotlar

1. Khmelev D. V. Распознавание автора текста с использованием цепей А. А. Маркова // Вестник Московского государственного университета. Серия 9: Филология. — 2000. — № 2. — С. 115–126.
2. Valgina N. S. Теория текста. — Москва: Логос, 2003. — 191 с.
3. Nikolina N. A. Филологический анализ текста: учебное пособие. — Москва: Академия, 2003. — 256 с.
4. Sun’iy yaratilgan matnlarni aniqlash usuli va dasturiy ta’minoti: foydalanuvchi taqdim etgan dissertatsiya materiali. — 2026. — 21 ta matn xususiyati keltirilgan jadval.