O‘ZBEK TILI MILLIY KORPUSI UCHUN N-GRAM MODELLINING KONKORDANS QIDIRUV TIZIMI TAHLILI

Z.Uzoqov,
Samarqand davlat chet tillar instituti magistranti,
mktiu93@gmail.com
X.A.Primova,
Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti Samarqand filiali professori,
primova@samtuit.uz

Annotatsiya: Milliy korpus muayyan tilning ma’lum davr (yoki davrlar) dagi maqomi, janrlari, uslublari, hududiy hamda ijtimoiy ko’rinishlari va boshqalarni o‘z ichiga oladi. Milliy korpus tilshunoslikning korpus lingvistikasi mutaxassislari tomonidan tuziladiki, bu ilmiy tadqiqot va tilni o‘rganish uchun xizmat qiladi. Ushbu tezisda O‘zbek tili milliy korpusi uchun n-gram modellining konkordans qidiruv tizimi asosida tahlil etishning boshlang‘ich qadamlari ko‘rib chiqilgan.

Kalit so‘zlar: *korpus, chastota, o‘zbek tili milliy korpusi, n-gramm, unigramma, orqaga qaytish algoritmi.*

O‘zbek tili korpusini ishlab chiqish bugungu kunda e’tirof etilgan standartlar asosida amalga oshiriladi. So‘z birikmalarining bog‘liqligini va ularning xatoliklarini aniqlovchi algoritmlar tahlil etilishi muhim masala hisoblanadi. Korpusning qidiruv tizimi so‘zlar, so‘z birikmalari, lemma, token hamda *n*-gram modellining konkordans qidiruv tizimi va dasturini ishlab chiqish bilan baholanadi. Milliy korpusining lemma, token hamda *n*-gram modellining konkordans qidiruv tizimi algoritm va dasturini ishlab chiqish dolzarb masala hisoblanadi [1].

Milliy til korpusining paydo bo‘lishi uning insoniyat hali korpus lingvistikasi fani paydo bo‘lmasdan oldingi davrda, ya‘ni XVIII asrdan boshlab tadqiq etila boshlashgan. Bibliyani tadqiq etish hamda lug‘atlar yaratish [2], tillarni o‘qitish [3], deskriptiv grammatika [4] va boshqalar tahlil etilgan. Kvirk korpusi bir

million so‘z birikmalarini o‘zida jamlagan bo‘lib, har biri o‘n yetti qator matndan iborat million kartotekadan iborat. Mazkur korpusning yaratilishiga 25 yil vaqt sarflangan va Kvirk korpusining ishlari 1989-yilda oxirgida yakunlagan. Bu paytda texnologiyalar yuqori sur‘atda rivojlanganligi bois korpus tezlikda elektron shaklga o‘tkazilgan. Ma’lumotlarda aniqlilik muammosi tufayli tilni ravonlashtirish *n*-gramm til modeli muhim texnik algoritmdir. Biroq, ba’zi hollarda u ko‘rinmas *n*-grammlarga noto‘g‘ri ehtimollik miqdorini belgilaydi. Shuning uchun ushbu maqolada tilning agglyutinativ xususiyatlaridan foydalangan holda ko‘rinmas grammlarning noto‘g‘ri tayinlangan ehtimollarini moslashtiradigan yangi usuli taqdim etiladi. Tilda a morfemaning grammatik toifasini oldingi morfemani bilish orqali oldindan aytish mumkin. Ushbu belgidan foydalanib, grammatik jihatdan noto‘g‘ri bo‘lgan *n*-grammalarning nisbatan yuqori ehtimollikka erishishiga yo‘l qo‘ymaslikka harakat qilinadi. Asisoy eksperimental natijalar shuni ko‘rsatmoqdaki, ushbu taklif qilingan usul Katz algoritmida 8,6% — 12,5% va Kneser-Ney Smoothing algoritmlarida 4,9% — 7,0% gacha tildagi chalkashlikni kamaytirishga imkon bergan [4, 5].

Milliy korpus qanday rivojlanadi? Rus tilining milliy korpusi, avvalo, o‘z ichiga XVIII asrning o‘rtalaridan to XXI asrning boshlarigacha bo‘lgan davrni qamrab oldi. Bu davr xoh o‘tgan, xoh yangi bo‘lishidan qat’iy davr u sotsiolingvistik ko‘rinishdagi badiiy so‘zlashuv, jonli so‘zlashuv, qisman matnlarni tashkil qiladi. Korpusga badiiy qimmatga ega bo‘lgan va til o‘rgatishga qiziqish qimmatga ega bo‘lgan va til o‘rgatishga qiziqish uyg‘otadigan badiiy adabiyot namunalari kiritiladi. Badiiy she’rlardan tashqari, yozma adabiyotning boshqa, yozma adabiyotning namunalaridan publisistika, ilmiy ommabop va ilmiy adabiyotlar, shaxsiy chiqishlar (ma’ruzalar), shaxsiy yozishmalar, kundaliklar, hujjatlar va kiritiladi.

Milliy korpusning o‘ziga xos muhim xususiyati me’yorlashtirilgan muayyan tarkibga ega ekanligi bilan xarakterlanadi. Bu korpus ma’lum tilda berilgan (turli

badiiy janrlar: publitsistik, o‘quv, ilmiy, ish yuritish, so‘zlashuv, publitsistik, o‘quv, ilmiy, ish yuritish, so‘zlashuv, shevaviy kabi), ularning barchasi imkon shevaviy kabi), ularning barchasi imkon darajasida ma’lum doiraga oid ma’lumotlarning proportsional matnlar hisoblanadigan og‘zaki va yozma ko‘rinishlarining barchasini o‘z ichiga oladi. Korpusning qoniqarli darajada bo‘lishi uchun uning ko‘lamiga e’tibor qaratish kerakligini nazardan chetda qoldirmaslik kerak (masalan, o‘n va yuz milliongacha so‘z qo‘llash).

Tadqiqot ishlari va maqolalar tahlil etilganda Sinclair shunday yozadi: "Matnni o'rganayotgan har bir kishi unda har xil so'z shakllari qanchalik tez-tez uchraydiganligini bilishi kerak". Tribble va Jones til sinfida matnlardan foydalanishning pedagogik metodologiyasini belgilab, matnni tushunishning eng samarali boshlang'ich nuqtasi chastota bo'yicha saralangan so'zlar ro'yxati ekanligini taklif qiladi. Juilland ispan, rumin va frantsuz tillari uchun bir qator chastotali lug'atlarni yaratdi.

Kneser-Neining ta’kidlashicha silliqlash an’anaviy tekislash algoritmlariga nisbatan o'quv tanlanmalar bo'yicha yaxshiroq ishlash imkonini beradi. [1, 2] da taklif qilingan takomillashtirilgan Kneser-Ney silliqlash algoritmi boshqa tekislash algoritmlaridan yaxshiroq samaradorlikni berdi. Biroq, bu o‘zbek tilida hali ham muammoli hisoblanadi. Bunda asosan funksional morfemalardan keyin kelgan kontekstlar soni odatdagi kontent morfemalari bilan birga keladigan kontekstlar sonidan ko'p.

Ushbu maqolada mavjud usullar va til korpusi tahlil etildi. Ushbu maqolada quyi tartibli n-gramm modeli bilan tekislash algoritmlarining solishtirma tahlili o‘tkazilganda n-grammlarga noto‘g‘ri ehtimollik tayinlash muammosiga ega ekanligi kelib chiqdi.

Katz orqaga qaytish algoritmidan foydalangan holda tekislash algoritmlarida quyi tartibli n-gramm modellari masalasini hal qilgan. Tekislash algoritmi [6] va eng

yaxshi ma'lum bo'lgan tekislash algoritmlaridan biri bo'lgan Katz orqaga qaytish algoritmi [7] da batafsil yoritilgan.

Shuni ta’kidlash kerakki, kutubxonalar til xususiyatlaridan ko‘ra matnning mazmun-mohiyati bilan qiziquvchilar tomonidan yaratiladi. Milliy korpus kutubxonalardan farqli o‘laroq “qoniqarli” va “foydali” matnlar to‘plami hisoblanib qolmasdan, asosan, til o‘rganish uchun xizmat qiluvchi hisoblanadi. Bunda oddiy o‘rtamiyona yozuvchilarning romanlari ham, oddiy so‘zlashuv matnlari, mumtoz badiiy adabiyot namunalari qatorida mumtoz badiiy adabiyot namunalari qatorida foydalanilaveradi.

Milliy korpuslar milliy va chet tillarni o‘qitish uchun ham muhim ahamiyat kasb etadiki, ko‘pgina darslik va o‘quv rejalari hozirda milliy korpusga moslangan bo‘lishi kerak. Korpusdan foydalanish orqali mualliflari notanish so‘zning grammatik shakllarni ajnabiy ham, o‘qituvchi ham, jurnalist ham, redaktor yoki yozuvchi ham tez va oson tekshirib olishi mumkin bo‘ladi. Demak, milliy korpus oddiy qiziquvchimi, undan foydalanuvchilar millatidan qat’i nazar o‘rganuvchilar uchun birdek foydalanish imkonini beradi.

Korpus tilshunosligi, uning milliy shakllarini yaratish, shuningdek, o‘zbek milliy tajribalardan kelib chiqib o‘zbek milliy korpusining nazariy va amaliy asoslarini ishlab chiqish, uni keng jamoatchilikka keng tadbiq qilish mutaxassislarning galdagi dolzarb vazifalaridan biridir. Natijada o‘zbek tili mumtoz va zamonaviy matnlari, turli janrlarda yaratilgan durdona asarlar “ikkinchi hayot”ga yo‘llanma oladi va kelgusida ko‘p ming (ikki yuz, besh yuz ming yoki million) so‘zli lug‘atlarini yaratish imkoni tug‘iladi.

Adabiyotlar:

1. S. Chen. Building Probabilistic Models for Natural Lan-guage. PhD thesis, Harvard University, 1996.

2. S. Chen, D. Beeferman, and R. Rosenfeld. Evaluation metrics for language models. Proceedings of the DARPA Broadcast News Transcription and Understanding Work-shop, pages 275–280, 1998.

3. H. Chung. Statistical Korean Dependency Parsing Model based on the Surface Context Information. PhD thesis, Ko-rea University, 2004.

4. K. W. Church and W. A. Gale. A comparison of the en-hanced good-turing and deleted estimation methods for es-timating probabilities of english bigrams. Computer Speech and Language, 5(1):19–54, Jan. 1991.

5. F. Jelinek. Statistical Methods for Speech Recognition. The MIT Press, London, 1997.

6. R. Rosenfeld and X. Huang. Improvements in stochastic language modeling. In *HLT ’91: Proceedings of the workshop on Speech and Natural Language*, pages 107–111, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

7. S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 35(3):400–401, 1987.