BIGRAM MODELI ASOSIDA GAP GENERATSIYASIDA LAPLAS VA ADD-K SILLIQLASH USULLARINI TAQQOSLASH

Abdullayeva Mohinur Dilmurod qizi
Kompyuter lingvistikasi yo‘nalishi talabasi
ongorovamohinur@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbek tilidagi psixologiya sohasiga oid 200 ming so‘zdan tashkil topgan korpus asosida bigram modeli yordamida matn generatsiyasi masalasi o‘rganildi. Nol ehtimollik muammosini bartaraf etish maqsadida Laplace va Add-k silliqlash algoritmlaridan foydalanildi hamda ularning natijalari o‘zaro taqqoslandi. Tajribalar natijasida Add-k usuli yuqori chastotali so‘z birikmalarining ehtimolliklarini yaxshiroq saqlab qolishi va tabiiyroq matn generatsiyasini ta’minlashi aniqlandi.

Kalit so‘zlar: *Laplace silliqlash, Add-K silliqlash, Data Sparsity, n-gram, matn generatsiyasi.*

Abstract. This article investigates the problem of text generation using a bigram model based on a corpus consisting of 200,000 words related to the field of psychology in the Uzbek language. To address the zero-probability problem (data sparsity), Laplace and Add-k smoothing algorithms were applied, and their results were comparatively analyzed. The experimental results demonstrate that the Add-k method better preserves the probabilities of high-frequency word combinations and ensures more natural text generation.

Keywords. *Laplace smoothing, Add-k smoothing, Data Sparsity, n gram, text generation.*

Kirish

Zamonaviy kompyuter lingvistikasi va sun’iy intellekt tizimlarida matnlarni avtomatik tahlil qilish, keyingi so‘zni bashorat qilish va til modellarini qurish asosan ehtimollik statistikasiga tayanadi. Matematik nuqtai nazardan eng sodda va samarali hisoblangan N-gramm modellarida matndagi so‘zlarning ketma-ket kelish zanjirlari

o‘rganiladi [1]. Biroq, har qanday chekli o‘lchamli matnlar to‘plamida tilning barcha so‘z birikmalari qamrab olinishi imkonsizdir. Statistik modellar amaliyotda uchramagan va korpusda mavjud bo‘lmagan yangi birikmalarga duch kelganda ularning shartli ehtimolligini nol deb baholaydi. “Nollik ehtimollik” muammosi kompyuter lingvistikasida juda muhim o‘rin tutadi, chunki bu holat umumiy model zanjirining ko‘paytma ehtimollik qiymatini butunlay nolga tushirib qo‘yadi. Ushbu muammoni tizimli hal etish uchun korpus lingvistikasida turli silliqlash usullari qo‘llaniladi.

Ushbu maqolada psixologiya sohasiga oid 200 mingta so’zdan tashkil topgan korpus asosida matn generatsiyasi amalga oshirildi. Matn yaratishda N-gram modellari ichida eng sodda va samarali usullardan biri hisoblangan Bigram modelidan foydalanildi [2].

Bigram modelida har bir keyingi so‘zning paydo bo‘lish ehtimoli faqat undan oldingi bitta so‘zga bog‘liq deb qaraladi. Shuning uchun gapning umumiy ehtimoli quyidagi formula orqali hisoblanadi:

$$P(\omega_{1,}\, \omega_{2,} \ldots, \omega_n) = (\omega_1 \prod_{i=2}^{n} P(\omega_i \,|\omega_{i-1})$$

Bu yerda:

$\omega_{I,}$ - gapdagi i-so‘z;

$P(\omega_i|\, \omega_{i-1})$ - oldingi so‘z berilganda keyingi so‘zning shartli ehtimoli.

Bigram ehtimoli korpusdagi so‘zlar chastotasidan quyidagicha aniqlanadi:

$$\mathrm{P}(\omega_i|\omega_{i-1}) = \frac{C(\omega_{i-1}, \omega_i)}{C(\omega_i - 1)}$$

Bu yerda:

$C(\omega_{i-1,}\omega_i)$ - bigramning uchrash soni;

$C(\omega_i - 1)$ - oldingi so‘zning umumiy uchrash soni.

Misol uchun, “*shaxs ijtimoiy*” bigram juftligi korpusda 21 marta uchragan bo‘lsa va **“**shaxs“ so‘zi jami 150 marta ishlatilgan bo‘lsa,

$$P(ijtimoiy|shaxs) = \frac{21}{150} = 0{,}14$$

ga teng bo’ladi.

Quyidagi jadvalda korpusda eng ko‘p uchragan bigramlar chastotasi keltirilgan.

ID	Bigram	Chastota
1	yordam beradi	249
2	ahamiyatga ega	195
3	ta’sir qiladi	161
4	amalga oshiriladi	132
5	ichiga oladi	129
6	xizmat qiladi	128
7	yuzaga keladi	124
8	talab qiladi	123
9	keltirib chiqaradi	103
10	rol o‘ynaydi	99

Bigram modelining asosiy muammosi

Bigram modelining asosiy kamchiligi ma’lumot Data Sparsity muammosidir [3]. Agar korpusda ma’lum bir bigram uchramagan bo‘lsa, uning chastotasi nolga teng bo‘ladi:

$$C(\omega_{i-1}, \omega_i) = 0$$

Natijada ehtimollik ham nol bo‘ladi:

Gapning umumiy ehtimoli ko‘paytmadan iborat bo‘lgani sababli bitta nol ehtimollik butun gap ehtimolini nolga aylantiradi:

$$P(\omega_i|\omega_{i-1}) = \frac{0}{C(\omega_{i-1})} = 0$$

Natijada shartli ehtimollikning nol bo‘lishi:

$$P(\omega_1, \omega_2, \ldots, \omega_n) = 0$$

Bu esa yangi yoki kam uchraydigan so‘z birikmalarini generatsiya qilish imkoniyatini cheklaydi. Nol ehtimollik muammosini kamaytirish uchun Laplace silliqlash algoritmidan foydalaniladi. Ushbu usulda har bir bigram chastotasiga 1 qo‘shiladi:

$$P_{Laplas}(\omega_i|\omega_{i-1}) = \frac{C(\omega_{i-1}, \omega_i) + 1}{C(\omega_i - 1) + V}$$

Bu yerda:

V - lug‘atdagi noyob so‘zlar soni.

Natijada korpusda uchramagan bigramlar ham nol bo‘lmagan kichik ehtimollikka ega bo‘ladi.

Ikkinchi usul Laplace usulining umumlashtirilgan ko‘rinishi Add-k silliqlash hisoblanadi. Bunda 1 o‘rniga k qiymati qo‘shiladi:

$$P_{Add-k}(\omega_i|\omega_{i-1}) = \frac{C(\omega_i - 1, \omega_i) + k}{C(\omega_i - 1) + kV}$$

Bu yerda:

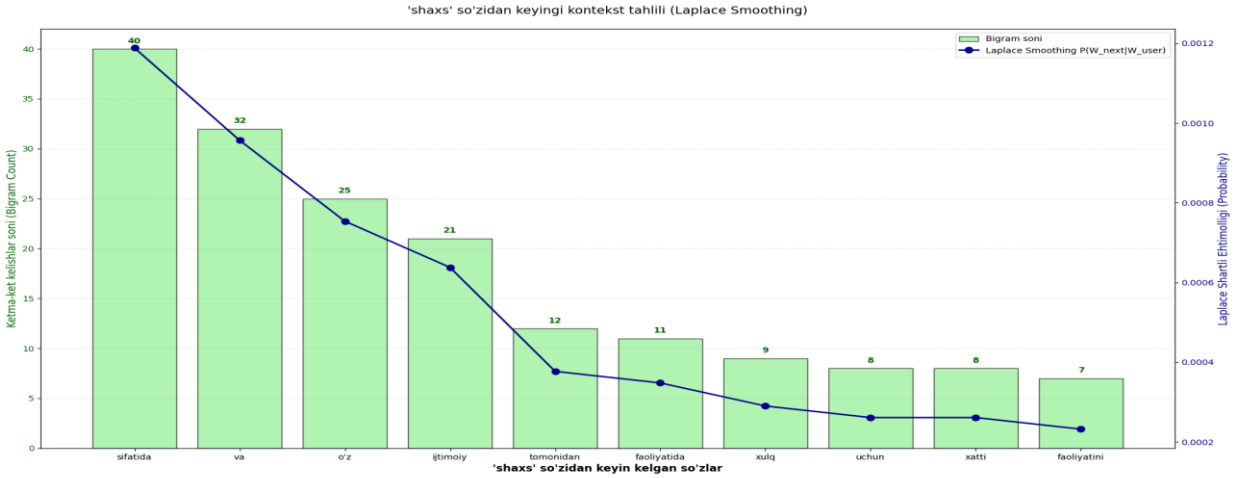
$k > 0$ - silliqlash koeffitsienti;

V - lug‘at hajmi.

Agar k=1 bo‘lsa, Add-k formulasi Laplace silliqlash formulasiga teng bo‘ladi. Kichik k qiymatlari masalan, 0.1 yoki 0.5 ehtimolliklarni haddan tashqari silliqlab yubormasdan, nol ehtimollik muammosini bartaraf etishga yordam beradi.[3]

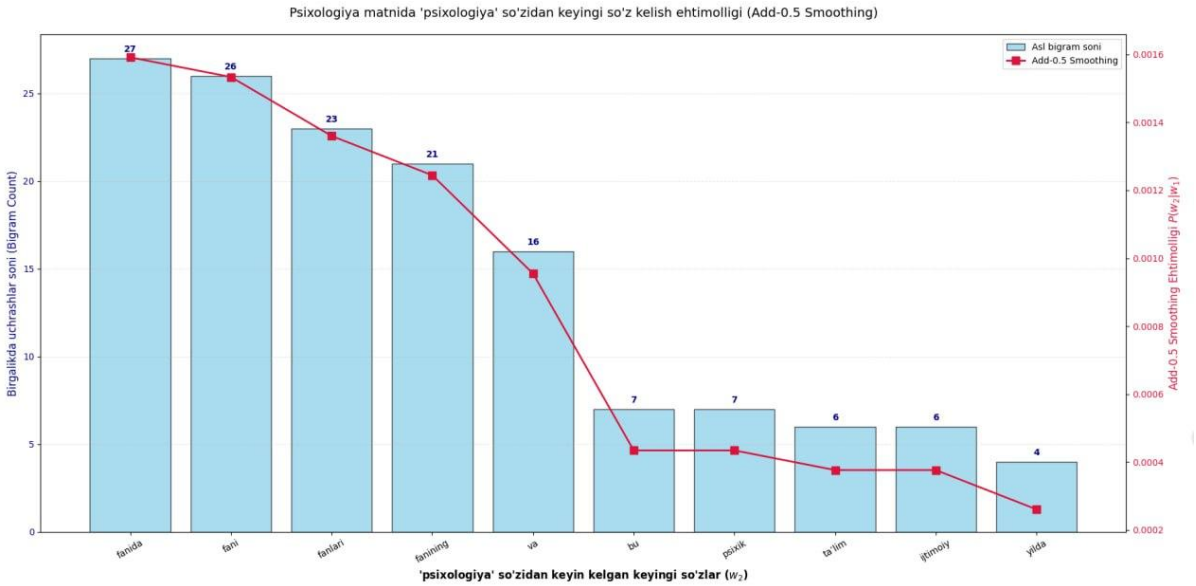
Ushbu maqolada bigram modeli yordamida gap generatsiyasi amalga oshirildi. Generatsiya jarayonida yuzaga keladigan nol ehtimollik muammosini bartaraf etish maqsadida Laplace va Add-k silliqlash algoritmlaridan foydalanildi hamda ularning natijalari o‘zaro taqqoslandi. Psixologiya sohasiga oid matnlar

korpusi asosida Laplace silliqlash usuli qo‘llanildi va olingan natijalar 1-rasmda keltirilgan. Natijada “**shaxs**” so‘zidan keyin kelishi mumkin bo‘lgan so‘zlarning ehtimollik taqsimoti ko‘rsatildi.



1-rasm. Bigram modelida Laplace silliqlash usuli qo‘llanilgandagi natijalar

2-rasmda add-k silliqlash usuli orqali hisoblangan natija keltirilgan. Unda, “**Psixologiya**” so‘zidan keyin eng ko‘p uchraydigan so‘zlarning Add-k silliqlash formulasi asosida hisoblangan. k = 0.5 qiymati tanlanganligi sababli yuqori chastotali so‘z birikmalari o‘zining nisbiy ustunligini saqlab qolgan. Shu bilan birga, korpusda uchramagan yoki kam uchraydigan so‘z birikmalariga ham nol bo‘lmagan kichik ehtimollik qiymatlari berilgan. Bu esa modelda nol ehtimollik muammosini kamaytirib, ehtimollik taqsimotining yanada barqaror shakllanishiga xizmat qiladi.



2-rasm. Add-k silliqlash usuli natijasi

Xulosa

Ushbu maqolada o‘zbek tilidagi psixologiya sohasiga oid 200 ming so‘zdan iborat korpus asosida bigram modeli yordamida matn generatsiyasi amalga oshirildi. Nol ehtimollik muammosini kamaytirish maqsadida Laplace va Add-k silliqlash algoritmlari qo‘llanib, ularning natijalari o‘zaro taqqoslandi. Tajriba natijalari Add-k usuli yuqori chastotali so‘z birikmalarining ehtimolliklarini yaxshiroq saqlab qolishi va tabiiyroq natijalar berishini ko‘rsatdi. Olingan natijalar bigram modellari uchun Add-k silliqlash usulining samaraliroq ekanligini tasdiqladi.

Foydalanilgan adabiyotlar ro‘yxati

1. Chew, Y. C., Mikami, Y., Marasinghe, C. A., & Nandasara, S. T. (2009). Optimizing n-gram order of an N-gram based language identification algorithm for 63 written languages. *The International Journal on Advances in ICT for Emerging Regions*, *2*(2)

2. Elov B.B., Tojieva G.N., Tokhtaeva M. X., Jurayeva N.J., Primova M.H.: N-gram language model for Uzbek texts. International Conference on Trends in Sustainable Computing and Machine Intelligence (ICTSM 2025). – Bangkok,

Thailand. – 19–20 September 2025. – P. 346-357. https://link.springer.com/chapter/10.1007/978-3-032-13177-5_27

3. Chen S.F., Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling // Computer Speech and Language. – 1999. – Vol. 13, No. 4. – P. 359–394.