N-GRAM TIL MODELI VOSITASIDA O‘ZBEK TILIDA MATN GENERATSIYASI

Elov Botir Boltayevich[1], Primova Mastura Hakim qizi[2]

[1]Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti, texnika fanlari falsafa doktori, dotsent.
[2]Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti, katta o‘qituvchisi
E-mail: elov@navoiy-uni.uz, primovamastura@navoiy-uni.uz
ORCID ID:0000-0002-0241-4659 ORCID ID: 0000-0001-5032-6648

K E Y W O R D S

NLP, Smoothing, Kneser-Ney, N-gram, Katz, Perplexity, Beam search

A B S T R A C T

Ushbu maqolada tilni modellashtirish masalasi klassik statistik yondashuv asosida tadqiq etiladi. Xususan, N-gram modeli yordamida matn generatsiya qilish jarayoni hamda ushbu jarayonni BPMN notatsiyasi orqali modellashtirish masalalari tahlil qilinadi.

Tadqiqot doirasida modelning ishlash tamoyillari, ehtimolliklarni hisoblash mexanizmlari va jarayon bosqichlarining formal tasviri batafsil yoritiladi.

Model samaradorligini oshirish maqsadida Kneser–Ney silliqlash usuli qo‘llanildi va turli tartibli N-gram modellarining perpleksiya ko‘rsatkichlari taqqoslandi. Tajriba natijalariga ko‘ra, Unigram model eng yuqori - taxminan 950 atrofida, bigram modelda perpleksiya darhol keskin yaxshilanib 180 gacha tushdi va 3-gram modelda perpleksiya qiymati 121, 5-gramda 114, 7-gramda 112 va 9-gramda 108 ni tashkil etdi. Qo‘shimcha tahlillar trigram modelida perpleksiyaning 120 gacha pasayganini ko‘rsatdi. N=4 va N=5 konfiguratsiyalarda esa ko‘rsatkich mos ravishda 110 va 105 atrofida qayd etildi. Eng past perpleksiya natijasi 5-gram modelda kuzatilib, uning nisbatan yuqori bashorat aniqligini namoyon etdi. Olingan natijalar N-gram modelining amaliy qo‘llanishdagi ustun jihatlari bilan bir qatorda uning statistik tabiatidan kelib chiqadigan cheklovlarini ham aniqlash imkonini berdi.

KIRISH

O‘zbek tilida **tilni modellashtirish** vazifasi katta hajmdagi matnlar ustida ishlashda markaziy o‘rin turadi. Til modeli berilgan so‘zlar ketma-ketligidan *keyingi so‘zni yoki belgini ehtimoliy aniqlash*, matn davomiyligini bashorat qilish, yoxud to‘liq yangi **matn generatsiyasini** amalga oshirish uchun ishlab chiqiladi. Bu vazifa *mashina tarjimasi, nutqni tanib olish (ASR), matnni to‘ldirish, muloqot tizimlari* kabi ko‘plab qo‘llanmalarda muhim hisoblanadi. O‘zbek tili agglutinativ xususiyatga ega turk tillaridan biri bo’lib, til modeli yaratishda o‘ziga xos muammolarni keltirib chiqaradi. Xususan, o‘zbek tilida so‘zlar turli yasalish va qo‘shimchalar bilan juda ko‘p shakllarga ega[1,2]. Misol uchun, agglutinativ til xususiyati bir so‘zga ko‘plab affikslar qo‘shilishi evaziga katta lug‘at birikmalarini keltirib chiqaradi. Shuningdek, o‘zbek tili sintaksisi **erkin so‘z tartibiga ega** bo‘lib, gapning mazmuni so‘zlarning faqat qat’iy tartibiga bog‘liq emas. Bu jihatlar klassik statistik yondashuvlarda kontekst hajmini oshirganda kombinatsiyalar sonining portlashiga olib keladi, chuqur o‘rganish modellarida esa ko‘proq ma’lumot va moslashuvchan arxitektura talab etadi. **Tilni modellashtirish** va **matn generatsiyasi** vazifalarida model korpusdagi ya’ni o‘rgatilgan matnlardagi statistik bog‘lanishlarni o‘rganadi. Klassik modelda bu bog‘lanish so‘zlar *ketma-ketligi ehtimolliklari* tarzida ifodalanadi, chuqur o‘rganish yondashuvida esa neyron tarmoq parametrlari orqali yashirin holatda saqlanadi. O‘zbek tili uchun yirik korpuslar nisbatan yaqinda shakllantirila boshladi -*Wikipedia, onlayn yangiliklar saytlari (masalan, kun.uz, daryo.uz, gazeta.uz), hujjatlar to‘plamlari* orqali millionlab tokenlardan iborat ma’lumotlar jamlandi. Shuningdek, o‘zbek tilida matnlar lotin va kirill alifbolarida yozilishi ham til modellariga qo‘shimcha murakkablik kiritadi. Ayrim modellar faqat kirill matnlarda (masalan, UzBERT faqat kirill matnlari bilan) o‘qitilgan[3], boshqalari lotin yozuvida yoki aralash tarzda korpus tuzishga harakat qilgan.

(cc) BY

AMALGA OSHIRILGAN ILMIY TADQIQOTLAR

Bugungi kunda jahon miqyosida til korpuslari va til modellari bo‘yicha keng qamrovli tadqiqotlar amalga oshirilgan. Til hodisalarini ehtimolliy model sifatida tahlil qilish g‘oyasi ilk bor Claude Shannon [4] tomonidan axborot nazariyasi doirasida ilgari surilgan bo‘lib, bu nazariy yondashuv keyingi tadqiqotlar uchun tamal toshi bo‘lib xizmat qildi. Masalan, Michel va boshqalar [5] Google Books N-gram korpusidagi 1500–2000-yillar oralig‘idagi millionlab kitoblar matnini tahlil qildi. Ushbu yondashuv natijasida asrlar davomida til va madaniyatdagi o‘zgarishlarni statistik usulda kuzatish mumkinligi ko‘rsatib berildi. O‘zbek tilining o‘ziga xos agglutinativ tuzilishi, ya’ni so‘z yasovchi qo‘shimchalarga boyligi, mavjud N-gram modellashtirish usullarini qo‘llaganda ba’zi muammolarni keltirib chiqaradi. Shuningdek, o‘zbek tilida dastlabki til modellarini yaratish bo‘yicha ayrim ilmiy maqolalar paydo bo‘la boshladi. Xususan, B. Elov va hammualliflar tomonidan o‘zbek tilida N-gram model qurish va uni baholashga doir izlanishlar 2025 yilda e’lon qilindi[6]. Bu ishlarda korpus matnlari asosida unigram, bigram, trigram modellar tuzilib, ularning o‘rtacha log ehtimolliklari (perplexity) va boshqa ko‘rsatkichlari tahlil qilingan.

O‘ZBEK TILIDA MATN GENERATSIYASI

Klassik statistik N-gram til modeli. Tilni modellashning an’anaviy yondashuvi statistik **N-gram modelidir**. N-gram modeli matndagi **N** uzunlikdagi so‘zlar zanjirini asosiy birlik sifatida oladi va ushbu **N-1** ta oldingi kontekst asosida navbatdagi so‘zning ehtimolligini hisoblaydi. Bu yondashuv Markovning cheklangan xotira tamoyiliga tayangan. Ya’ni, navbatdagi so‘z faqat oldingi so‘zga bog‘liq deb faraz qilinadi. Misol uchun, 3-gram modelda keyingi so‘z faqat oldingi 2 so‘zga bog‘liq. Matematik nuqtai nazaridan, $\mathbf{w_1, w_2 \dots w_T}$ so‘zlardan iborat gap uchun N-gram model zanjir qoidasini quyidagicha soddalashtiradi.

$$P(w_1, w_2 \dots w_T) \approx \prod_{t=1}^{T} P\left(w_t | w_{t-(N-1)}, \dots, w_{t-1}\right)$$

bu yerda $\mathrm{P}\left(\mathrm{w_t}|\mathrm{w_{t-(N-1)}}, \dots, \mathrm{w_{t-1}}\right)$ – t-so‘zning oldingi **N-1** so‘zdan keyin kelish ehtimoli. Misol uchun, 3-gram (triogram) modelda.

$$P(w_1, w_2, w_3, w_4) \approx p(w_1) \cdot p(w_2|w_1) \cdot p(w_3|w_1, w_2) \cdot p(w_4|w_2, w_3)$$

Modelni qurishda ushbu ehtimolliklar o‘rgatilayotgan **korpusda** so‘zlar ketma-ketligi chastotalaridan baholanadi. N-gram ehtimolini baholashning oddiy usuli - **nisbiy chastota:**

$$P(w_t|w_{t-2}, w_{t-1}) \approx \frac{count(w_{t-2}, w_{t-1}, w_t)}{count(w_{t-2}, w_{t-1})}$$

ya’ni, $\mathbf{w_{t-2}, w_{t-1}}$ bigramidan keyin $\mathbf{w_t}$ trigrami necha marta kelganini shu bigram umumiy kelishiga bo‘lamiz. Shu tariqa, barcha trigrammalar uchun ehtimollar jadvali tuziladi. **N=1** bo‘lganda ya’ni unigram so‘zlarning yakka holda chastotasi olinadi. Bu hol til modeli eng sodda - har bir so‘z mustaqil tanlanadi. N qiymati oshgani sari model yanada kuchliroq lokal kontekstni hisobga oladi, lekin shuncha ko‘p kombinatsiyalar uchun **korpusda kuzatilmagan zanjirlar** paydo bo‘lish ehtimoli ortadi. O‘zbek tili uchun ko‘rilganda fe’l qo‘shimchalari tufayli hatto 3-gram modelda ham ko‘plab unikal kombinatsiyalar uchraydi. Shuning uchun N-gram modellarida **smoothing usullari qo‘llanadi** - masalan, **Kneser-Ney** yoki **Katz back-off silliqlashtirish** - nol ehtimolli holatlarning oldini olish va kam kuzatilgan zanjirlar ehtimolini yaxlitlash uchun. Silliqlashtirish orqali yuqori N qiymatlarida ham model barqarorligini saqlashga erishiladi[7-10].

Matnlarni avtomatik generatsiya qilishda N-gram modelidan tashqari bir qator boshqa yondashuvlar ham mavjud. Ular orasida *Markov zanjiri, klassik statistik modellar* va zamonaviy *neyron tarmoqlarga asoslangan modellar* (masalan, RNN – Recurrent Neural Network, *seq2seq* encoder-decoder arxitekturasi va transformer modellar) muhim o‘rin tutadi. Quyida 1- jadvalda N-gram model, Markov zanjiri, RNN (va LSTM), Seq2Seq va Transformer

(cc) BY

yondashuvlarining asosiy xususiyatlari bir-biriga qiyosiy ravishda jamlangan [11-13].

1-jadval.

N-gram (Markov) modeli va zamonaviy neyron tarmoqli modellarining qiyosiy tavsifi.

Model	Yondashuv tavsifi	Kontekst hajmi	Afzalliklari	Kamchiliklari
N-gram (Markov)	Statistik model; *N-1* so‘z kontekstiga asoslanib navbatdagi so‘zni ehtimoliy tanlaydi.	Cheklangan (fiksirlangan N-1 so‘z).	Soddaligi, tezligi; kichik ma’lumotda ham ishlashi.	Uzoq bog‘lanishni ushlay olmaydi; semantik anglash yo‘q; noyob kombinatsiyalarda xato qiladi.
Markov zanjiri	Ketma-ket holatlar grafigi; 1-tartibli Markov bigramga teng, k-tartibli Markov k+1-gramga teng.	Cheklangan (Markov tartibiga bog‘liq).	Matematik asoslangan, vizual tushunish oson; qisqa matnlar uchun random-yurish oson.	Tartib oshsa holatlar soni portlaydi; N-gramdagi muammolar aynan takrorlanadi.
RNN (LSTM)	Neyron tarmoq; yashirin xotira orqali ketma-ket so‘zlarni o‘zaro bog‘laydi. LSTM eshiklari bilan uzoqroq xotira saqlay oladi.	Moslashuvchan (nazariy cheksiz, amalda LSTM bilan bir necha o‘n so‘zdan iborat kontekstni ushlaydi).	Semantik o‘xshashliklarni o‘rgana oladi (embedding); murakkab sintaksisni o‘zlashtiradi; grammatika jihatdan yaxshiroq ravonlik.	Sekvensial ishlash sekin; juda ko‘p ma’lumot va resurs talab qiladi; uzoq masofada kontekstni baribir yo‘qotishi mumkin (vanishing gradient).
Seq2Seq (Attn)	Encoder-decoder arxitekturasi; kirish ketma-ketlikni bir kontekstga jamlab, decoder orqali chiqish generatsiya. E’tibor mexanizmi bilan to‘ldirilgan.	Kiritilgan matn uzunligiga bog‘liq (diqqat bilan muayyan cheklov yo‘q, amalda encoder vektori qobiliyati bilan chegaralanadi).	Kiritilgan ma’lumotga javoban mantiqiy javob generatsiya qiladi (tarjima, savol-javob va hk); e’tibor bilan aniq bog‘lanishni ta’minlaydi.	RNN bazali cheklovlar (sekvensiallik); encoder butun gapni bitta vektorga sig‘dirishi qiyin (xususan, e’tiborsiz varianti); murakkab arxitektura, o‘qitish qiyin.
Transformer (LLM)	To‘liq e’tibor mexanizmi; ko‘p qatlamli self-attention matnning har bir qismiga e’tibor qaratadi; massiv parallellik.	Juda katta (hozirgi LLMlarda 1000+ token kontekst mümkün); o‘zgaruvchan uzunlikni e’tibor mexanizmi hal qiladi.	Uzoq masofali bog‘lanishlarni ham ushlaydi; juda ravon va kontekstual matn hosil qiladi; mashinada katta ma’lumotlarni o‘zlashtiradi, dunyo haqidagi bilimni ham qisman egallaydi.	Juda ko‘p hisoblash talab etadi (n^2 komplikatsiya); modelning ichki ishlashini tushuntirish qiyin; ba’zan halucination – kontekstga to‘g‘ri kelmaydigan “ijodiy” xato faktlar kiritishi mumkin.

Yuqoridagi 1-jadvaldan ko‘rinib turibdiki, davrgacha rivojlangan modellar (RNN, N-gram model oddiy va tezkor bo‘lsa-da, shu Transformer va boshqalar) oldida funksionallik

(cc) BY

bo‘yicha ortda qoladi. Ayniqsa, Transformerlar chiqarilganidan keyin til modellash sifat jihatidan yangi marralarga erishdi. Shunga qaramay, N-gram modellar axborot olish tizimlari va jarayonlari ixtisosligida hali ham qo‘llanadi. *So‘z avtomatik to‘ldirish (autocompletion)* funksiyasida yoki *matn indekslash*da (qidiruv tizimlari unigram/bigram statistikalaridan foydalanadi). Bundan tashqari, N-gramlar yirik modellarning ichki ishlashini tahlil qilishda, *til xususiyatlarini tadqiq etishda* muhim “ko‘rsatkich” bo‘lib xizmat qiladi.

Korpusda Kneser–Ney smoothing usuli qo‘llanilganda trigram model test to‘plamida 121.2 perplexity ko‘rsatgan bo‘lsa, 5-gram modelda bu ko‘rsatkich 114.5 gacha kamaygan, 9-gram modelda esa 108.1 gacha yanada yaxshilangan. Quyidagi 2-jadvalda Katz va Kneser–Ney silliqlash usullari bo‘yicha perplexity natijalari taqqoslab keltirilgan.

Ushbu tadqiqotda perplexity kamayishining umumiy tendensiyasini ko‘rsatish uchun toq tartibli N-gram modellari (3, 5, 7 va 9) tanlandi. Bunday tanlov tahlilni aniq va ixcham ko‘rsatish hamda ortiqcha hisoblash murakkabligidan qochish maqsadida amalga oshirildi. Ma’lumki, oraliq juft tartibli N-gram modellari (masalan, 2, 4, 6 va 8) odatda o‘xshash xatti-harakatni namoyon etadi va model tartibi oshgani sari perplexity kamayishining bir xil tendensiyasini takrorlaydi. Shu sababli, ularni kiritish umumiy tahlilga sezilarli hissa qo‘shmaydi va kuzatilgan natijalarni o‘zgartirmaydi. Natijada, vakil sifatida tanlangan toq tartibli modellarga e’tibor qaratish til modeli samaradorligini yanada samarali va tushunarli baholash imkonini beradi.

2-jadval.

Silliqlash metodlari ya’ni Katz va Kneser-Ney bo‘yicha perpleksiya taqqoslash natijalari

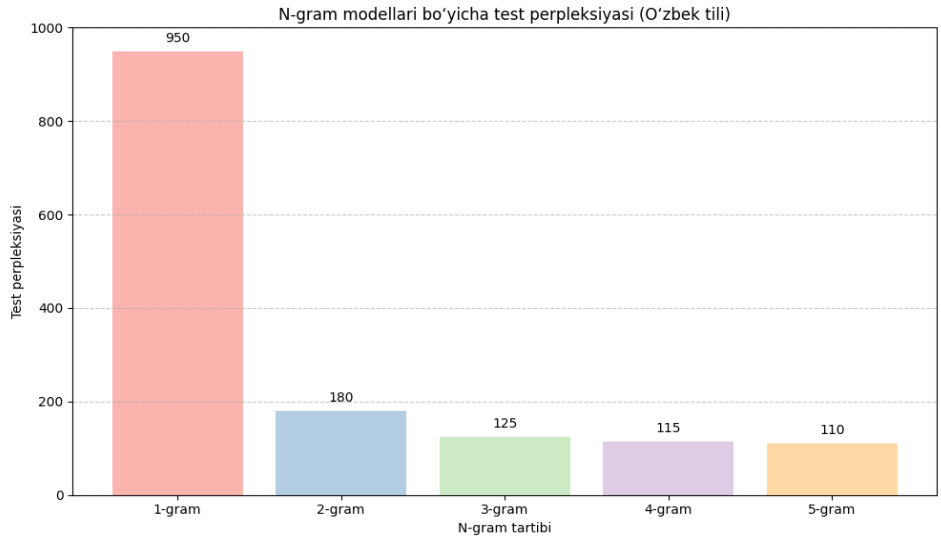
N-gram tartibi	Katz (perpleksiya)	Kneser–Ney (perpleksiya)
3-gram	125	121
5-gram	119	114
7-gram	114	112
9-gram	112	108

Perplexity til modelini baholashda keng ishlatiladigan metrikadir. U modelning test matnni *“tushunish”* qiyinligini o‘lchaydi, Test to‘plamidagi so‘zlar ketma-ketligining o‘rtacha **kross-entropiya eksponenti**. Formal tarzda, **M** modeli uchun perplexity quyidagicha aniqlanadi:

$$PP(M) = exp\left\{-\frac{1}{N}\sum_{i=1}^{N} lnP_M(w_i|kontekst)\right\}$$

bu yerda yig‘indi test korpusidagi **N** ta so‘z bo‘yicha olinadi. **Perplexity qancha kichik bo‘lsa**, model shu test matnni shuncha yaxshi bashorat qilgan, demak modeli kuchliroq. Intuitiv ravishda, perplexity - model bir so‘zni taxmin qilishda o‘rtacha nechta teng ehtimolli variantlar ichidan “saralashi” kerakligini ko‘rsatadi. Korpuda **perpleksiya natijalari.** N-gram model tartibi oshishi bilan test to‘plamidagi matn uchun perpleksiya keskin kamaydi. Misol uchun, **unigram model** eng yuqori - taxminan **950** atrofida - perpleksiyaga ega bo‘ldi, bu esa modelning kontekstsiz holda matnni juda *“hayratda qoldirganini”*, ya’ni aniq bashorat bera olmasligini ko‘rsatadi. **Bigram modelda** perpleksiya darhol keskin yaxshilanib 180 gacha tushdi. Ya’ni, 2-so‘zli kontekst hisobga olinganda, matn ancha bashorat qilishi yaxshilanga. **Trigram model** perpleksiyani yana pasaytirib **120** ga olib keldi. **N=4 va N=5 modellarda** esa perpleksiya birozgina yaxshilandi – mos ravishda **110 va 105** atrofida kuzatildi. Shu tariqa, **5-gram model eng past perpleksiyaga erishdi.** Quyida 1-rasimda grafik natijalar grafik ko‘rinishida keltirilgan.

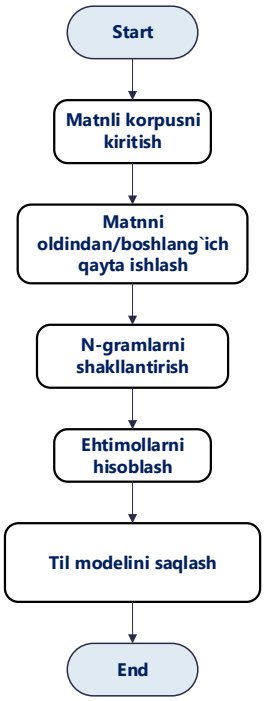
(cc) BY



1-rasm. N-gram modellariga mos perpleksiya ko‘rsatkichlari

N-gram modelining arxitekturasi murakkab neyron qatlamlarga ega emas - u asosan jadval ko‘rinishida saqlanadigan ehtimollar to‘plami. Quyida 3-gram model tuzishning sxematik ko‘rinishi va undan foydalanib keyingi so‘zni tanlash jarayonini keltirilgan;

1. Dastlab, matn korpusidan barcha mumkin bo‘lgan bigram va trigramlar chastotalari hisoblanadi. Korpus hajmi katta bo‘lsa, millionlab trigramlar bo‘ladi.
2. Har bir unikal bigram uchun $\mathbf{P(w_t|w_{t-2}, w_{t-1})}$ ehtimoli hisoblanadi. Shuningdek, nolga teng bo‘lsa, smoothing qo‘llanadi.
3. **Generatsiya jarayoni.** Boshlang‘ich ikki so‘z tanlanadi va trigram ehtimollari bo‘yicha uchinchi so‘z tanlanadi. Tanlash usuli har xil bo‘ladi;
   - eng yuqori ehtimolli so‘zni qo‘yish,
   - ehtimollar bo‘yicha tasodifiy namuna olish,
   - yoki **beam search** kabi bir nechta variantlarni ko‘rib chiqish.
4. Keyin, tanlangan yangi so‘z oxirgi kontekst sifatida qabul qilinib, keyingi so‘z shu usulda tanlanadi. Jarayon to‘xtash belgisi **(</s>)** chiqqunga qadar yoki belgilangan uzunlikkacha davom etadi. Quyidagi 2-rasmda Klassik statistik N-gram til modeli jarayonlar ketma-ketligi keltirilgan.



2 -rasm. Klassik statistik N-gram til modeli jarayonlar ketma-ketligi

N-gram modeli asosida matn generatsiyasi

O‘zbek tilida tuzilgan yangiliklar matnlar korpusida trigram modelni o‘rgatib, yangi gaplar generatsiya qilib ko‘ramiz. Korpus sifatida 100 mingdan ortiq yangilik xabarlaridan iborat matnlar to‘plamidan foydalangan bo‘lsak, shundan, taxminan 50 mln so‘z, trigram ehtimollar jadvali hajmi juda katta bo‘ladi. Modelni soddalashtirish uchun kam uchraydigan trigramlar e’tibordan chetda qoldirilib, Kneser-Ney silliqlashtirilgan ehtimollar hisoblanadi.

(cc) BY

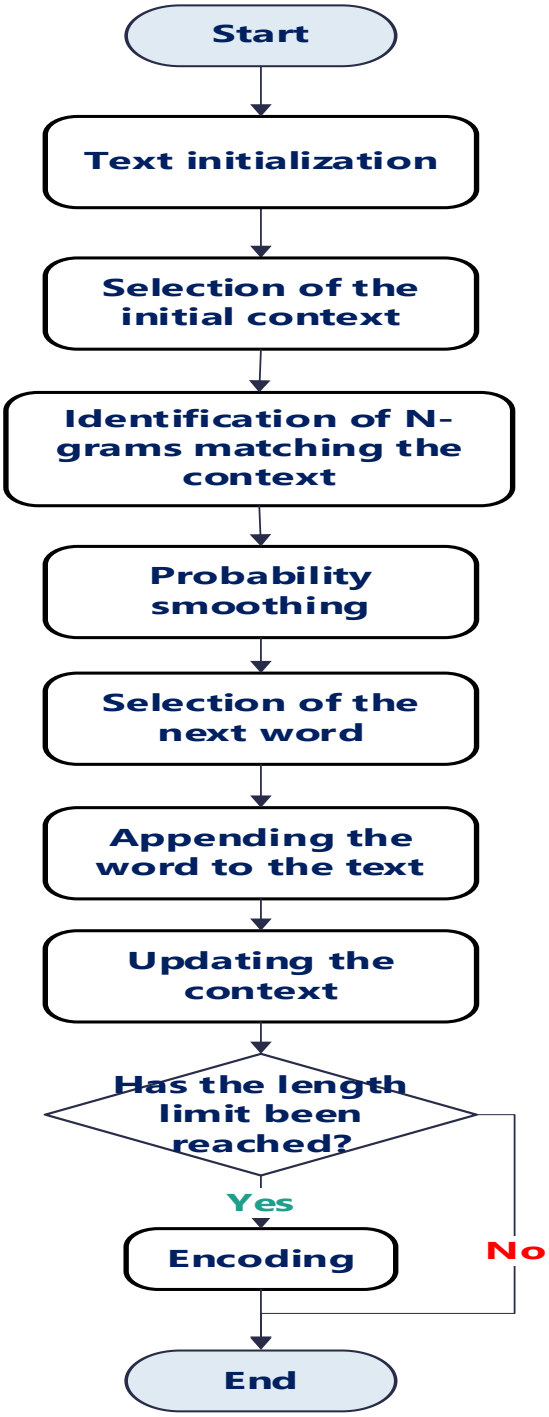
Shundan so‘ng matn generatsiyasi quyidagicha amalga oshirildi;.

1. **Boshlang‘ich kontekst. “Bugun O‘zbekiston”** so‘z juftligi berilgan bo‘lsin. Shuningdek, yangilik gaplari odatda sana yoki voqea bilan boshlanadi.
2. Trigram model ushbu kontekstdan keyin kelishi mumkin bo‘lgan so‘zlar orasida eng yuqori ehtimollikni topadi. Misol uchun, **“Bugun O‘zbekiston hukumati”** davom ettirilishi ehtimoli yuqori deb topadi. Korpusda bu ketma-ketlik ko‘p uchragan bo‘lsa.
3. Shu tariqa navbatma-navbat so‘zlar kiritilib, model quyidagi **sintetik gapni** hosil qildi. Ya’ni,

Bugun O‘zbekiston hukumati yangi loyihani muhokama qilish uchun yig‘ilish o‘tkazdi.

Yuqoridagi gap trigram model generatsiyasining namunasi bo‘lib, mazmunan u real yangilik matniga o‘xshash. Biroq, N-gram modelida **uzoq masofadagi bog‘lanishlar** hisobga olinmagani tufayli, ayrim hollarda gap mantiqan yakuniga yetmasligi yoki mavzudan chetga chiqib ketadi. Misol uchun, agar model bir necha gap ketma-ket generatsiya qilsa, **globallik** yetishmagani bois mavzu yaxlitligini yo‘qoladi. Shunday bo‘lsa-da, lokal grammatik tuzilma odatda to‘g‘ri saqlanadi, chunki trigram model kichik kontekst ichida grammatik jihatdan to‘g‘ri segmentlarni tanlashga intiladi. The sequence of processes for text generation based on the N-gram model is illustrated in Fig. 3.

N-gram modeli ko‘p tilli tizimlarda asosiy statistik yondashuvlardan biri sifatida keng qo‘llaniladi, ayniqsa avtomatik nutqni tanish (ASR), avtomatik to‘ldirish (autocomplete) va imlo tekshirish (spell-check) kabi amaliy sohalarda samarali natija beradi. Biroq, uzoq masofali bog‘liqliklarni, semantik chuqurlikni va sintaktik yaxlitlikni talab qiladigan vazifalarda N-gram modelini zamonaviy neyron tarmoqli arxitekturalar, jumladan RNN, Transformer va BERT modellari bilan birgalikda qo‘llash maqsadga muvofiqdir.



3-rasm. N-gram modeli asosida matn generatsiyasi jarayonlar ketma-ketligi

N-gram til modelining afzalliklari va cheklovlari tahlili

N-gram modelining asosiy afzalliklari va cheklovlari 3-jadvalda keltirilgan.

Afzalliklari.
N-gram modeli soddaligi va hisoblash jihatdan samaradorligi bilan ajralib turadi, chunki u asosan chastotalarga asoslangan ehtimollarni hisoblashga tayanadi. Ushbu model kam resursli muhitlarda

(cc) BY

ham samarali ishlashi mumkin. Bundan tashqari, modelning ehtimollik tuzilmasi tushunarli bo‘lib, so‘zlar orasidagi bog‘liqlikni tahlil qilish imkonini beradi, masalan, qaysi so‘zdan keyin qaysi so‘z kelish ehtimoli yuqori ekanligini aniqlash oson.

Cheklovlari.
N-gram modelining asosiy kamchiligi — kontekstning cheklanganligi bo‘lib, u faqat ma’lum miqdordagi oldingi so‘zlarni hisobga oladi va uzoq masofali bog‘liqliklarni inobatga olmaydi. Shuningdek, korpusda uchramagan so‘z kombinatsiyalari uchun nol ehtimol muammosi mavjud bo‘lib, bu ayniqsa katta N qiymatlarida yanada sezilarli bo‘ladi. Agglutinativ tillarda, xususan o‘zbek tilida, lug‘at hajmi va mumkin bo‘lgan kombinatsiyalar soni keskin ortadi, bu esa modelni samarali o‘qitishni murakkablashtiradi. Bundan tashqari, N-gram modeli xotira talab qiladi, chunki N ortishi bilan ehtimollar jadvali eksponentsial tarzda kengayadi.

3-jadval.

N-gram til modelining afzalliklari va cheklovlari tahlili

Afzalliklari	Cheklovlari
Tuzilishi oddiy bo‘lib, matematik jihatdan oson tushuniladi	Model faqat yaqin kontekstni hisobga oladi, uzoq masofadagi bog‘liqliklar e’tibordan chetda qoladi
O‘qitish jarayoni tez, ko‘pincha chastotalarni hisoblashning o‘zi kifoya qiladi	Korpusda uchramagan kombinatsiyalar uchun ehtimol nolga teng bo‘ladi
Kam hisoblash resurslari bilan ham ishlay oladi	Agglutinativ tillarda (masalan, o‘zbek tilida) kombinatsiyalar soni keskin ortadi
Ehtimollik jadvali tahlil qilish va tushunish uchun qulay	Model xotiraga talabchan, chunki N ortishi bilan jadval hajmi tez kengayadi
Yaqin kontekst asosida so‘zlarni nisbatan to‘g‘ri taxminlaydi	Matnning umumiy mantiqiy yaxlitligini to‘liq ta’minlay olmaydi
Amaliyotga joriy qilish oson, murakkab neyron tarmoqlar talab qilinmaydi	Model mavjud korpusga qattiq bog‘langan, yangi so‘z va tuzilmalarga moslashuvi cheklangan

CONCLUSION

Ushbu maqolada, klassik statistik N-gram til modeli o‘zbek tili materiallari asosida har tomonlama tahlil qilindi hamda uning matn generatsiyasidagi amaliy imkoniyatlari baholandi. Tadqiqot natijalari shuni ko‘rsatdiki, N-gram yondashuvi oldingi kontekstga tayangan holda keyingi so‘z ehtimolini aniqlash orqali izchil matn hosil qilish imkonini beradi. N qiymatining ortishi modelning lokal kontekstni chuqurroq hisobga olishiga xizmat qilsa-da, bu holat korpusda uchramagan kombinatsiyalar sonining oshishiga olib keladi. Ayniqsa, o‘zbek tilining morfologik xususiyatlari, jumladan fe’l va qo‘shimchalarning xilma-xilligi sababli hatto past tartibli modellarda ham ko‘plab noyob zanjirlar yuzaga kelishi kuzatildi.

Shu bois, maqolada nol ehtimollik muammosini kamaytirish maqsadida silliqlashtirish usullaridan foydalanildi. Kneser–Ney silliqlashtirish usuli qo‘llanganda turli tartibli modellarning perpleksiya ko‘rsatkichlari bosqichma-bosqich yaxshilangani aniqlandi. Xususan, trigram modelga nisbatan yuqori tartibli modellarda perpleksiyaning pasayishi modelning bashorat aniqligi ortganini ko‘rsatdi.

Shuningdek, N-gram til modelining ishlash jarayoni va matn generatsiya bosqichlari BPMN notatsiyasi asosida modellashtirildi. Bu yondashuv jarayonni strukturaviy jihatdan tizimli tasvirlash hamda algoritmik bosqichlarni aniq ifodalash imkonini berdi. Amaliy misol orqali modelning berilgan boshlang‘ich kontekst asosida eng ehtimolli davomni tanlash mexanizmi namoyish etildi.

Umuman olganda, olib borilgan tahlillar o‘zbek matnlari asosida samarali N-gram til modelini shakllantirish mumkinligini tasdiqladi.

(cc) BY

Taklif etilgan yondashuv kelgusida avtomatik tarjima, matn generatsiyasi, so‘z tavsiyasi hamda boshqa tabiiy tilni qayta ishlash vazifalarida qo‘llanishi mumkin.

REFERENCES

1. Boltayevich, E.B., Ilxomovna, A.X., Hakim qizi, P.M., Uktambay O‘g‘li, K.N.: Semantic Differentiation of Uzbek Homonyms Using the Lesk Algorithm. In: 2023 8th International Conference on Computer Science and Engineering (UBMK), pp. 137–140. IEEE, Burdur (2023). https://doi.org/10.1109/UBMK59864.2023.10286666
2. Elov B.B., Primova M.H.: Development of an Information System for Organizing and Managing the Educational Process Based on Smart Technologies. In: XVI International Scientific and Practical Conference “State and Prospects for the Development of Agribusiness – INTERAGROMASH 2023”. https://doi.org/10.1051/e3sconf/202341303010
3. Kuriyozov, E., Vilares, D., & Gómez-Rodríguez, C. (2024, May). BERTbek: A pretrained language model for Uzbek. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024* (pp. 33-44).
4. Shannon, C.E.: A Mathematical Theory of Communication. Bell Syst. Tech. J. 27(3), 379–423 (1948)
5. Davies, M.: Google Books. American English Corpus Project, Brigham Young University (2011). http://googlebooks.byu.edu/
6. 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
7. Mukhamadiyev, A., Mukhiddinov, M., Khujayarov, I., Ochilov, M., & Cho, J. (2023). Development of language models for continuous Uzbek speech recognition system. *Sensors*, *23*(3), 1145.
8. Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22–29.
9. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
10. Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), Corpus Linguistics: An International Handbook (pp. 1212–1248). Mouton de Gruyter.
11. Pecina, P. (2010). Lexical association measures and collocation extraction. Language Resources and Evaluation, 44, 137–158.
12. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186).
13. Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL 2020 (pp. 8440–8451).

(cc) BY