TABIIY TILNING STATISTIK MODELLARI

Botir Elov[1], Ruhillo Alayev[2], Abdulla Abdullayev[3]

[1]texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti.
E-pochta: elov@navoiy-uni.uz
[2]texnika fanlari falsafa doktori. Mirzo Ulug'bek nomidagi O'zbekiston Milliy universiteti
E-pochta: mr.ruhillo@gmail.com
[3]"Urganch innovatsion university" NTM ta'limni kredit tizimini boshqarish bo'lim boshlig'i
E-pochta: abdulla_abdullayev9270@mail.ru

KEYWORDS

Til modellari, Statistical Language Model, SLM, tabiiy tilning statistik modeli, Laplas silliqlash, mashinali o'qitish.

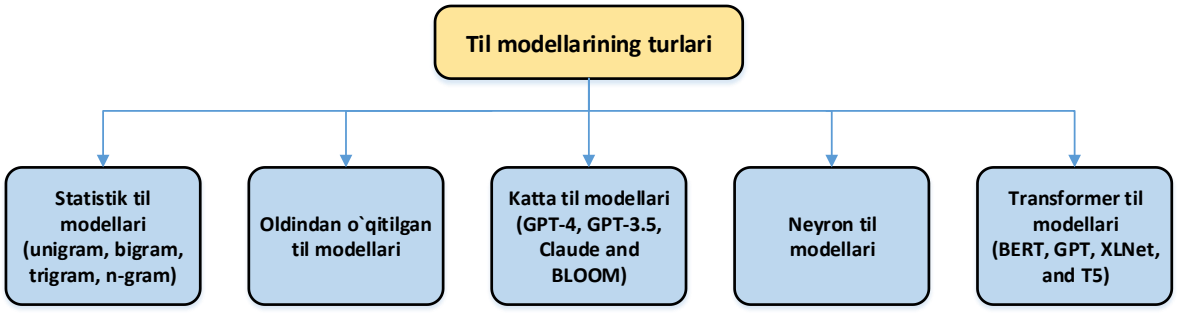
ABSTRACT

Tabiiy tilning statistik modeli (Statistical Language Model, SLM) – tabiiy tilni qayta ishlashda qo'llaniladigan zamonaviy vosita bo'lib, u ma'lum tildagi so'zlar ketma-ketligi ehtimolini bashorat qilishga qaratilgan. SLM asosida gapdagi muayyan ketma-ketlikdan keyingi so'z bashorat qilinadi. SLM so'zlarning tabiiy til ma'lumotlari korpusida paydo bo'lishiga asoslangan ketma-ketlik ehtimolini hisobga oladi. Katta hajmdagi matn ma'lumotlarini tahlil qilish orqali model so'zlarning tilda qanday qo'llanilishi qoliplarini o'rganishi va ushbu qoliplar asosida keyingi ehtimoli yuqori so'zni bashorat qilishi mumkin. NLP sohasi rivojlanishda davom etar ekan, statistik til modellari tilni tushunish va qayta ishlash uchun asosiy vosita bo'lib hisoblanadi. SLMlar yordamida tabiiy til texnologiyasida mumkin bo'lgan chegaralarni kengaytirishni davom ettirishimiz va yanada innovatsion va kuchli NLP ilovalarni yaratishimiz mumkin. Ushbu maqolada tabiiy tilning statistik modellaridan hiosblangan N-gram modelini o'zbek tili korpusi asosida ishlab chiqish usullari keltiriladi. Shuningdek, N-gram modellarining matematik tavsifi va baholash usullari hamda umumlashtirish, sezgirlik, OOV (noma'lum so'zlar), maxsus kontekst muammolari va ularni bartaraf qilish yo'llari keltiriladi.

Kirish

Tilni modellashtirish tabiiy tilni qayta ishlash (Natural Language Processing, NLP)ning asosiy vazifasi bo'lib, u so'zlar ketma-ketligidan keyingi so'zni bashorat qilishni amalga oshiradi. Odatda, til modeli ma'lum bir tabiiy tildagi so'z ketma-ketligi bo'yicha ehtimollik taqsimotini o'rganadi. Til modeli ma'lum bir kontekstda paydo bo'lish ehtimoli asosida so'z ketma-ketligiga ehtimolini belgilaydi.

Til modeli (Language Model, LM) – bu belgilar (yoki so'zlar) ketma-ketligiga ehtimolliklarni belgilaydigan model. Quyidagi 1-rasmda bugungi kundagi til modellarining turlari keltirilgan [1,2].



1-rasm. *Til modellarining turlari*

1. Statistik til modellari (Statistical Language Model, SLM). Statistik til modellari sifatida N-gram modellari va ularning variantlari ishlatiladi. Statistik til modellari so'zlarning ketma-ketligi ehtimoli haqida bashorat qilish uchun ma'lumotlardagi statistik qoliplardan foydalanadigan model turidir. Statistik til modelini yaratishning asosiy yondashuvi n-gram ehtimolliklarni hisoblashdir.

2. Oldindan o‘qitilgan til modellari (Pre-Trained Language Models, PLM). Hozirgi kunda NLP sohasida oldindan o‘qitilgan til modellaridan keng miqyosida foydalanilmoqda. Chuqur o‘qitish usullarining rivojlanishi bilan PLMlarni o‘qitish va ulardan foydalanish an’anaviy statistik LMlarga qaraganda ancha samarali natijalarni taqdim etdi [3]. Til modellari muayyan NLP vazifalariga qo‘llanilishidan oldin, sintaktik va semantik bilimlarni o‘z ichiga olgan universal ko‘rinishlarni o‘rganishlari uchun katta til korpuslari to‘plamida (gaplar to‘plami) oldindan o‘qitiladi. Oldindan o‘qitilgan so‘ng, PLMlar NLP vazifalari uchun sozlanadi va olingan bilimlar turli vazifalarga qo`llaniladi.

3. Katta til modellari (Large Language Models, LLM). Katta til modellari odatda katta hajmdagi *kitoblar, maqolalar* va boshqa *yozma materiallar* to‘plamlari kabi katta hajmdagi matn ma’lumotlarida o‘qitiladi [4]. Ular tilning **qoliplari** va **strukturasini** o‘rganish uchun murakkab algoritmlar va neyron tarmoqlardan foydalanadi. Bu esa ularga turli xil so‘rovlarga mos izchil va mazmunli javoblarni yaratishga imkon beradi. Ushbu modellar NLPda **Generativ AI modellari** sifatida ham tanilgan bo`lib, matnni turli usullar bilan yaratishi mumkin. Jumladan, *gapdagi keyingi so‘zni bashorat qilish, gap yoki paragrafni yakunlash* yoki *berilgan mavzu asosida yangi matn yaratish* kabi vazifalarni bajaradi. Katta til modellariga misol sifatida GPT-4, GPT-3.5, BERT, RoBERTa, Klod va BLOOMlarni keltirish mumkin.

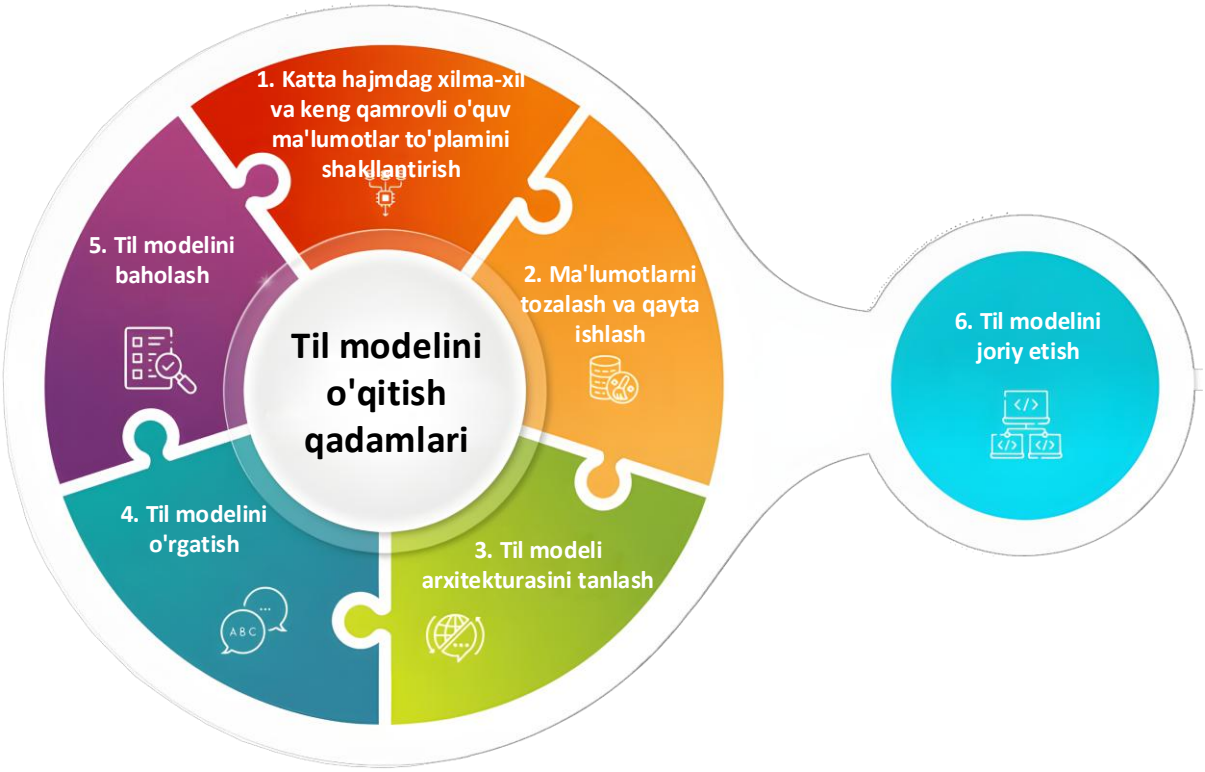
4. Neyron til modellari (Neural Language Models, NLM). Neyron til modellari, so‘zlar ketma-ketligi ehtimolini bashorat qilish uchun neyron tarmoqlardan foydalanadi [5]. Ushbu modellar matn ma’lumotlarining katta korpusida o‘qitilgan va tilning asosiy tuzilishini o‘rganishga qodir. Neyron til modellari kontekstni an'anaviy statistik modellarga qaraganda yaxshiroq aniqlaydi. Shuningdek, ular murakkabroq til strukturalarini va so‘zlar orasidagi uzoqroq bog‘liqlikni aniqlashi mumkin.

5. Transformer til modellari (Transformer Language Models, TLM). Transformer til modellari matnlardagi uzoq masofali o‘zaro bog‘liqlikni samarali aniqlaydigan **self-attention** mexanizmini joriy etish orqali til modellashtirishda inqilob qildi [6]. TLMlar kirish ketma-ketligini parallel ravishda qayta ishlashga imkon beruvchi ko‘p sonli self-attention va uzatish neyron tarmoqlaridan iborat. Transformerga asoslangan mashhur modellarga BERT, GPT, XLNet va T5 larni misol sifatoda keltirish mumkin. Ushbu modellar tabiiy tilni qayta ishlash sohasini sezilarli darajada kengaytirdi.

Bugungi kunda til modellaridan ko`plab NLP ilovalarini ishlab chiqishda foydalanilmoqda:

- **Matn yaratish:** Til modellari izchil, kontekstga mos matn yaratishga xizmat qiladi. Til modellari vositasida matn yaratish metodlaridan kontent yaratish, kopirayterlik va ijodiy yozish kabi turli sohalaridagi ilovalarda foydalaniladi.
- **Til tarjimasi:** Til modellari mashina tarjimasi tizimlari samaradorligini sezilarli darajada yaxshiladi. Til modellari orqali turli tabiiy tillarning konteksti va nuanslarini tahlil qilinib, aniqroq va kontekstga mos tarjimalar olindi. **Language Learning** va **Google Translate** kabi ikkita mashhur tarjima ilovalari til modellari asosida ishlab chiqilgan.
- **Virtual yordamchilar:** Siri, Alexa, Google Assistant va Cortana kabi virtual yordamchilar foydalanuvchi so‘rovlarini qayta ishlash uchun va ularga aniq javoblar berish maqsadida til modellariga tayanadi. Ushbu virtual yordamchilar murakkab algoritmlar va katta ma’lumotlar to‘plamidan foydalangan holda og‘zaki nutqni tanib oladi va unga javob beradi.
- **His-tuyg‘ularni tahlil qilish va Opinion Mining:** Matn ma’lumotlarida bildirilgan his-tuyg‘ular va fikrlarni tushunish *biznes, marketologlar* va *siyosatchilar* uchun foydalidir. Til modellari asosida jamoatchilik fikrini tahlil qilish, brendni baholash va rivojlanayotgan tendentsiyalarni aniqlash uchun ijtimoiy media postlari, mijozlar sharhlari va maqolalardan foydalanishi mumkin.

Til modelini ishlab chiqish bosqichlari



2-rasm. *Til modelini ishlab chiqish bosqichlari*

1. **Katta, xilma-xil va keng qamrovli o‘quv ma’lumotlar to‘plami (dataset)ni shakllantirish:** LM modelini yaratish uchun birinchi qadamda bu modelni o‘qitish uchun foydalanadigan matnli hujjatlar ma’lumotlar to‘plamini shakllantirish lozim. Ushbu ma’lumotlar to‘plami model ishlatiladigan til va sohani ifodalashi kerak. Keng qamrovli ma’lumotlar to‘plamini shakllantirish uchun turli yondashuvlardan foydalanish mumkin. Ushbu yondashuvlarning ba'zilari *umumiy ma’lumotlar to‘plamlari (public datasets), Internet saytlarini skanerlash (crawling the web), ma’lumotlarni qo‘lda yig‘ishdir (manually collecting data).*

2. **Ma’lumotlarni tozalash va boshlangi`ch qayta ishlash:** Ma’lumotlar to‘plamini to‘plaganidan so‘ng, uni tozalash va boshlang`ich qayta ishlas kerak. Bu bosqichda *shovqinni olib tashlash, xatolarni tuzatish* va *ma’lumotlarni til modeli tomonidan ishlatilishi mumkin bo‘lgan formatga aylantirish* amalga oshiriladi.

3. **Til modeli arxitekturasini tanlash:** Til modelini ishlab chiqishda tanlash mumkin bo‘lgan juda ko‘p turli til modeli arxitekturalari mavjud. Eng mashhur arxitekturalardan ba'zilari **seq2seq, transformer, gpt-3** hisoblanadi.

4. **Til modelini o‘qitish:** Til modeli arxitekturasi tanlangandan so‘ng, modelni o‘qitish kerak. Bu bosqichda modelga o‘quv ma’lumotlarini uzatiladi va ma’lumotlardagi qoliplar aniqlanadi. O‘qitish jarayoni *ma’lumotlar to‘plamining hajmiga* va *til modeli arxitekturasining murakkabligiga* qarab uzoq vaqt talab qilishi mumkin. Shuningdek, til modelini o`qitishda ma`lumotlar to`plami 3 ta qimga ajratilad: **oq`uv (training), basholash (validate)** va **test**.

5. **Til modelini baholash:** Til modeli o‘qitilgandan so‘ng, uni baholash kerak. Bu bosqich modelni baholash to‘plamida sinovdan o`tkazish va uning samaradorligini aniqlashni o‘z ichiga oladi. Til modelini baholash uchun foydalanish mumkin bo‘lgan ba'zi ko‘rsatkichlar **aniqlik (accuracy)**, **bleu baho (bleu score)**, **rouge baho (rouge score)**, **F1 baho (F1-score)**ni o‘z ichiga oladi.

6. **Til modelini joriy (deploy) qilish:** Agar til modelining ishlashi barcha mezonlarga javob bersa, uni joriy qilish mumkin. Bu bosqichda shakllantirilgan til modeli foydalanuvchilarga ishlatish uchun taqdim qilinadi. Til modelini qo‘llashning ba'zi usullari orasida *modelni API bilan integratsiyalash, veb-ilovani yaratish, modelni chatbotga joylashtirish* kiradi.

Ushbu maqolada o`zbek tili korpusidan[1] tanlab olingan **1120458** ta gaplar asosida N-gram statistik til modellari ishlab chiqildi, baholandi va model samaradorligini oshirishga oid tavsiyalar berildi [7,8,9].

Statistik til modellari

Tabiiy tillar gramatikasiga ega bo‘lsa-da, ular juda katta lug‘atga ega. Shuningdek, so‘z birikmalarini hosil qilishda o‘ziga xos jihatlar mavjud. Insonlarning o‘zaro muloqotida turli noaniqliklar tabiiy ravishda yuzaga keladi. Vaqt o‘tishi bilan tabiiy tildagi qoidalar ham o‘zgaradi. Xulosa sifatida gramatik qoidalar va sintaktik tuzilmalarni barcha foydalanish holatlari uchun aniqlab bo‘lmasligini qayd etish lozim.

Shuning uchun til modellari katta hajmdagi matnni tahlil qilish orqali tilning "tuzilmasini" o‘rganishga harakat qiladi. Ushnu yondashuv gramatik qoidalarga asoslangan emas, balki statistikdir. Qaysidir ma’noda, til modellari til sintaksisi va semantikasini qamrab oladi.

Odatda, tilni modellashtirish uchun so‘zlar ketma-ketligi haqida gapiradigan bo‘lsak-da, amalda tokenlar ketma-ketligini tahlil qilishga to‘g`ri keladi. Token **gap, so‘z birikmasi, so‘z, n-gram, morfema** yoki **harf** bo‘lishi mumkin [9]. Lemmatizatsiya qo‘llanilganda, bir nechta so‘z shakllari bitta tokenga qisqartiriladi. Qo‘yilgan NLP vazifasiga ko‘ra LMga mos tokenizatsiya usuli tanlanadi.

Tilni modellashtirish – bu tabiiy tildagi gaplarga ehtimolliklarni belgilash vazifasi. Shuningdek, u berilgan so‘zlar ketma-ketligiga ehtimolini belgilaydi va N-1 ta so‘zlar ketma-ketligidan keyingi yangi so‘zning ehtimolini baholaydi. Statistik til modellari – oldingi so‘zlarni hisobga olgan holda keyingi so‘zni bashorat qilish uchun ishlatiladigan ehtimollik modellari.

Og‘zaki nutqni tanib olish, imloni yoki gramatik xatolarni tuzatish va mashina tarjimasi kabi NLP vazifalarida ehtimolliklarni aniqlash muhim ahamiyatga ega. Ushbu maqolada tabiiy tildagi matnlarga mos n-gram statistik til modelini ishlab chiqish bosqichlari keltiriladi.

N-gram modellari

Ehtimoliy til modelining maqsadi gap yoki so‘zlar ketma-ketligi ehtimolini hisoblashdan iborat bo‘lib, quyidagi ikkita vazifani hal qilish nazarda tutiladi [10,11]:

1) $\boldsymbol{P(W) = P(w_1, w_2 \ldots w_N)}$ ketma-ketlikning hosil bo‘lish ehtimolini aniqlash;
2) Ketma-ketlikdan so‘ng keladigan so‘z ehtimolini $\boldsymbol{P(w_N \mid w_1, w_2 \ldots w_{N-1})}$ aniqlash.

Ngram modelining asosi shundan iboratki, butun ketma-ketlik asosida berilgan so‘zning ehtimolligini hisoblash o‘rniga, faqat oxirgi bir necha so‘zlar orqali taxminiy hisoblashlar amalga oshiriladi. Demak, oddiy Ngram modelida **h** yoki $\boldsymbol{P(w|h)}$ qiymatdan foydalanib, **w** so‘zining hosil bo‘lish ehtimoli bashorat qilinadi.

Quyidagi keltirilgan gap misolida ushbu vazifani korib chiqamiz:

Talabalar darsdan so‘ng uyga _____. (1)

Bo‘sh joyni qanday so‘zlar bilan to‘ldirish mumkin? Quyida ba`zi variantlar keltirilgan:

qaytish(a)di;
borish(a)di;
kelish(a)di.

Ushbu so‘zlar ro‘yxatini davom ettirish mumkin. Biz (1) gapda keying so‘zning “*qaytishdi*” bo‘lish ehtimolini quyidagi formula orqali aniqlaymiz:

$$P\left(\begin{matrix}"qaytishdi" \mid \\ "Talabalar\ darsdan\ so`ng\ uyga"\end{matrix}\right) \quad (2)$$

Bu yerda,

w = "qaytishdi"

h = "Talabalar darsdan so‘ng uyga"

Ushbu qo‘shma ehtimollikni zanjir qoidasi yordamida hisoblash mumkin:

[1] https://uznatcorpara.uz/

$$P(x_1, x_2 \dots x_N) = P(x_1)P(x_2|x_1)P(x_3|x_{1:2}) \dots P(x_n|x_{1:n-1}) = \prod_{k=1}^{N} P(x_k|x_{1:k-1}) \quad (3)$$

Shunday qilib, (1) gap uchun,

$$P(\text{"Talabalar darsdan so'ng uyga"}) = P(\text{"Talabalar"})\, P(\text{"darsdan"}|\text{"Talabalar"})\, P(\text{"so'ng"}|\text{"Talabalar darsdan"})\, P(\text{"uyga"}|\text{"Talabalar darsdan so'ng"})$$

Ehtimollik qiymatlarini aniqlaymiz:

$$P(\text{"qaytishdi"} \mid \text{"Talabalar darsdan so'ng uyga"}) == \frac{\square(\text{"Talabalar darsdan so'ng uyga }\square\text{"})}{\square(\text{"Talabalar darsdan so'ng uyga"})} = \frac{\square(\text{"Talabalar darsdan so'ng uyga qaytishdi"})}{\square(\text{"Talabalar darsdan so'ng uyga"})} = (4)$$

Quyidagi jadvalda o'zbek tili korpusidan ajratib olingan 15 mln. gapdagi (1) namunaga mos unigramalar soni keltirilgan.

1-jadval.

O'zbek tili korpusidan olingan 15 mln. gapdagi 5 so'zga mos unigramalar.

	talabalar	**darsdan**	**so'ng**	**uyga**	**qaytishdi**
Unigram	19212	1259	133735	11525	435

Markov modeli

Markov modeli ketma-ketlikning muayyan qismi asosida keyingi elementni ehtimolini bashorat qilishga asoslangan modellar sinfidir [12]. Ushbu farazga asoslangan holda, Ngram til modellarini bigram (avvalgi bir so'z) trigram (avvalgi ikkita so'z) va shu tariqa N-gramga (avvalgi N-1 ta so'z) toifalarga ajratishimiz mumkin.

Bigram modeli

Bigram modelida so'zning ehtimolligi faqat oldingi so'zga bog'liq. Shunday qilib, zanjir ehtimoli bo'yicha quyidagi formulani hosil qilamiz:

$$P(x_1, x_2 \dots x_N) \approx P(x_1)P(x_2|x_1)P(x_3|x_2) \dots P(x_n|x_{n-1}) \approx \prod_{k=1}^{N} P(x_k|x_{k-1}) \quad (5)$$

Shunday qilib, (1) gap uchun, quyidagi tenglik o'rinli:

$$P(\textit{"qaytishdi"} \mid \textit{"Talabalar darsdan so'ng uyga"}) \approx P(\textit{"qaytishdi"} \mid \textit{"uyga"})$$

Yoki, umumiy holda:

$$P(x_n|x_{1:n-1}) \approx P(x_n|x_{n-1}) \quad (6)$$

$$P(x_n|x_{n-1}) = \frac{C(x_{n-1}x_n)}{C(x_{n-1})}$$

Keyingi so'zni bashorat qilish uchun kontekstdagi avvalgi ikkita so'z ko'rib chiqiladi. Shunday qilib, zanjirli ehtimollik formulasi quyidagicha soddalashtiriladi:

$$P(x_n|x_{1:n-1}) \approx \prod_{k=1}^{N} P(x_k|x_{k-1}, x_{k-2}) \quad (7)$$

Shunday qilib, (1) gap uchun, quyidagi tenglik o'rinli:

$$P(\text{"}qaytishdi\text{"} \mid \text{"}Talabalar\ darsdan\ so`ng\ uyga\text{"}) \approx P(\textit{"qaytishdi"} \mid \text{"}so`ng\ uyga\text{"})$$

Yoki,

Arzon narsa hech qachon qadrlanmaydi.

$$P(\textit{"arzon narsa hech qachon qadrlanmaydi"}) \approx P(\textit{"narsa"}|\text{"}<s>\text{"}, \textit{"arzon"})\, P(\textit{"hech"}|\textit{"arzon"}, \textit{"narsa"})\, P(\textit{"qachon"}|\textit{"narsa"}, \textit{"hech"})\, P(\textit{"qadrlanmaydi"}|\textit{"hech"}, \textit{"qachon"})$$

Xuddi shu usulda boshqa yuqori ngram modellarini umumlashtirish mumkin.

Til modellarini baholash

Til modellarining samaradorligini **ichki (intrinsic)** yoki **tashqi baholash (extrinsic evaluation)** yordamida amalga oshirish mumkin [13]. Tashqi baholash til modelini real vaqtda NLP ilovasiga qo'llsh va ishlashni tahlil qilish orqali

bajariladi. Biroq, ichki baholash qo‘llanilishidan qat'iy nazar samaradorlikni o‘lchash uchun turli metriklardan foydalaniladi. Til modelining ishlashini tahlil qilish uchun ishlatiladigan ichki ko‘rsatkich – **chalkashlik (perplexity)** hisoblanadi. Test to‘plamidagi til modelining chalkashligi (ba’zan qisqacha PP deb ataladi) test to‘plamining normallashtirilgan so‘zlar soni bo‘yicha teskari ehtimolligiga teng.

$\boldsymbol{W = w_1, w_2 \dots w_N}$ test to‘plami uchun:

$$PP(W) = P(w_1 w_2 w_3 \dots w_n)^{-\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{(w_1 w_2 w_3 \dots w_n)}} \quad (8)$$

Yoki **W** to‘plamning ehtimolini kengaytirish uchun zanjir qoidasidan foydalanishimiz mumkin:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{(w_i | w_1 \dots w_{i-1})}} \quad (9)$$

Bigram til modeli bilan hisoblangan W test to‘plamining chalkashlik ehtimoli qiymati bigram ehtimolliklari o‘rtacha geometrik qiymati orqali hisoblandi:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{(w_i | w_{i-1})}} \quad (10)$$

Chalkashlikni hisoblashning yana bir usuli bor: **tilning o‘rtacha vaznli tarmoqlanish (weighted average branching factor of a language)**.

Tilning tarmoqlanish omili (the branching factor) – har qanday so‘zdan keyin kelishi mumkin bo‘lgan keyingi so‘zlar soni.

2-jadval.

O‘zbek tili korpusiga mos perplexity qiymatlar.

	Unigram	**Bigram**	**Trigram**
Perplexity	12827	55835	243452

Demak, bu yerda unigram modelining chalkashligi 11544 ga teng, bu 11544 ta mumkin bo‘lgan keyingi so‘z borligini bildiradi va model ushbu 11544 ta mumkin bo‘lgan so‘zlardan ketma-ketlikning eng yaxshi keyingi so‘zini tanlashi kerak. Quyida Ngram modellari yondashuvi bilan bog‘liq muammolar keltirilgan.

1. **Umumlashtirish va sezgirlik (Generalization and Sensitivity):** Ngram modeli samaradorligi *o‘quv korpusiga (training corpus)* bog‘liq. Buning sababi shundaki, ba’zida ehtimolliklar o‘quv korpusiga oid aniq faktlarni kodlaydi va o‘quv korpusidagi ishlash samaradorligi **N** qiymatiga mutanosibdir.

Ushbu muammoni hal qilish uchun turli xil n-gram modellaridan tasodifiy gaplarni yaratish mumkin. Unigram modeli uchun bu usulni qanday ishlashini tasavvur qilish oson. O‘zbek tilidagi barcha so‘zlar 0 dan 1 gacha bo‘lgan ehtimollik maydonini qamrab olsin. Bu holda, har bir so‘z chastotasiga proportsional intervalni qamrab oladi. Biz 0 dan 1 gacha bo‘lgan tasodifiy qiymatni tanlaymiz va intervali ushbu tanlangan qiymatni o‘z ichiga olgan so‘zlarni aniqlaymiz. Tasodifiy raqamlarni tanlashda va gapning yakuniy tokenini tasodifiy yaratmagunimizcha so‘zlarni yaratishda davom etamiz.

Ushbu usulni bigram modeli uchun ham qo‘llash mumkin. Birinchi navbatda (uning bigram ehtimoli bo‘yicha) bilan boshlanadigan tasodifiy bigram hosil qilish lozim. Aytaylik, o‘sha bigramaning ikkinchi so‘zi w bo‘lsin. Keyingi qadamda w bilan boshlanadigan tasodifiy bigramani tanlash lozim va hokazo. Yuqori tartibli n-gramlar uchun modelning o‘quvlar to‘plamiga nisbatan sezgirligini pasaytiradi.

2. **Noma’lum so‘zlar (Unknown Words):** Og‘zaki nutqni tanib olish yoki mashina tarjimasi kabi NLP vazifalarini hal qilishda oldindan tayyorlangan lug‘atlar yoki so‘z birikmalari jadvali mavjud bo‘ladi. Shuning uchun til modeli faqat shu lug‘at yoki so‘z birikmalari jadvalidagi so‘zlardan foydalanishi mumkin. Ba’zi NLP ilovalarida ilgari uchramagan so‘zlar bilan aniqlanadi. Bunday so‘zlar noma’lum so‘zlar yoki ba’zan OOV (out of vocabulary words) deb ham ataladi.

OOV muammosini hal qilishning ikkita usuli mavjud:

Birinchisi, oldindan tanlangan lug'at asosida muammoni yopiq lug'atga aylantirishdir. O'qitish jarayonida ushbu to'plamda bo'lmagan har qanday so'zni (har qanday OOV so'zini) matnni normallashtirish bosqichida noma'lum so'z tokeniga (unknown word token) aylantirish lozim. So'ngra o'quv to'plamidagi boshqa har qanday oddiy so'z kabi, uning soni bo'yicha ehtimolini hisoblash zarur.

Ikkinchi usulda, bizda oldindan lug'at mavjud bo'lmagan holatlarda, bunday lug'atni bilvosita yaratish, o'quv ma'lumotlaridagi so'zlarni ularning chastotasiga qarab almashtirish mumkin.

Misol uchun, o'quv to'plamida **n** martadan kam uchraydigan barcha so'zlarni almashtirish mumkin. Bundam **V** lug'at hajmini oldindan tanladi (aytaylik, 50 000) va chastota bo'yicha yuqori **V** so'zlarni tanlab, **UNK** tokeniga almashtiriladi. Har ikkala usulda ham OOV so'ziga biz odatdagi so'z kabi muomala qilib, til modelini o'rgatishda davom etamiz.

3. **Maxsus kontekst (Context Specific):** Qayta ishlanishni lozim bo'lgan muyyan so'zlar bizning lug'atimizda mavjud bo'lib, lekin test korpusida ko'rinmas kontekstda paydo bo'ladigan holatlar yuzaga kelishi mumkin. Til modeli ushbu ko'rinmas hodisalarga nol ehtimollik tayinlamasligi uchun, ko'proq sodir bo'ladigan hodisalardan bir oz ehtimollik massasini olib tashlashimiz va uni hech qachon ko'rmagan hodisalarga berishimiz kerak.

Ushbu modifikatsiya silliqlash (smoothing) yoki diskontlash (discounting) deb ataladi. Silliqlashning turli usullari mavjud: add-1 smoothing, add-k smoothing, stupid backoff, va Kneser-Ney smoothing.

Laplas silliqlash – silliqlashning eng oddiy usuli bo'lib, barcha bigrama qiymatlarini normallashtirishdan oldin ularga birni qo'shishga asoslangan. Ilgari nolga teng bo'lgan barcha qiymatlar endi 1 ga, 1 ga teng bo'lganlar esa 2 ga teng bo'ladi va hokazo.

$$P_{Laplace}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{\sum_w (C(w_{n-1}w)+1)} = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V} \quad (11)$$

3-jadval.

*O'zbek tili korpusidan olingan **1120458** ta gapdagi 5 ta so'zga mos normallashtirgan bigram ehtimoli (V = **347554**).*

	arzon	narsa	hech	qachon	qadrlanmaydi
arzon	9.38e-11	1.87e-10	9.38e-11	9.38e-11	9.38e-11
narsa	4.33e-10	4.33e-10	2.16e-09	4.33e-10	4.33e-10
hech	1.25e-09	1.35e-06	1.25e-09	2.36e-06	1.25e-09
qachon	3.11e-10	1.11e-10	1.25e-08	3.11e-10	6.22e-10
qadrlanmaydi	4.09e-13	4.09e-13	4.09e-13	4.09e-13	4.09e-13

3-jadvaldagi har bir katakchani o'z qatori uchun tegishli unigramaga bo'lish, unigram ehtimollarning quyidagi to'plamidan olingan):

arzon	narsa	hech	qachon	qadrlanmaydi
917	4270	12611	3057	3

Add-k silliqlash

Add-k silliqlsh: Modelni silliqlash usullaridan biri bu modelda mavjud hodisalardan ko'rinmas hodisalarga ehtimollik massasini kamroq hajmda o'tkazishdir. Laplas silliqlash usulidagi har bir ifodaga 1 qiymatni qo'shish o'rniga, **k** kasr sonini **(.5? .05? .01?)** qo'shiladi. Shuning uchun, bu algoritm **Add-k silliqlsh** deb ataladi [14].

$$P_{Add-k}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+k}{C(w_{n-1})+kV} \quad (12)$$

Kneser-Ney silliqlash

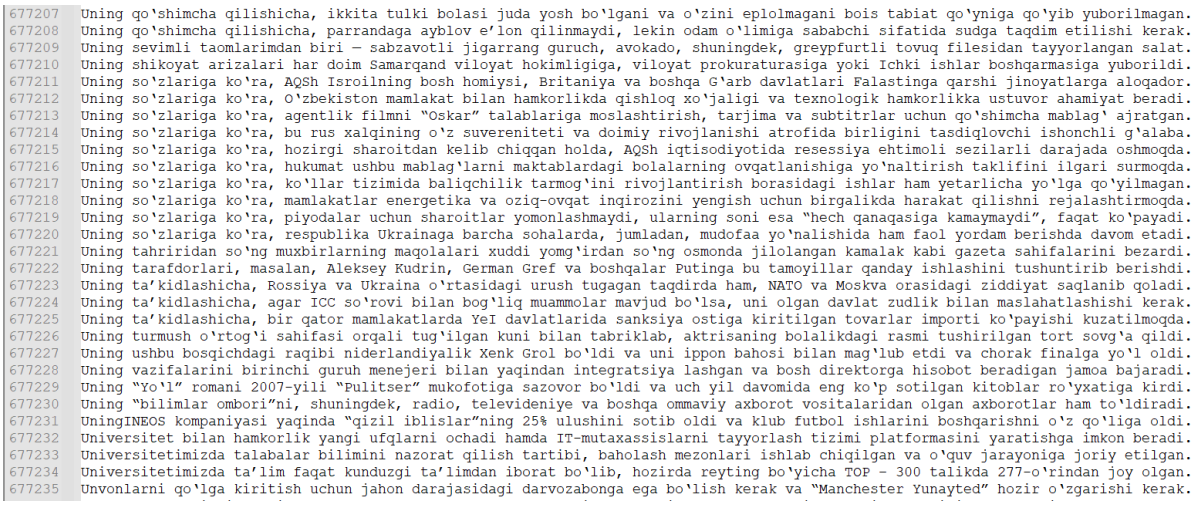
Kneser-Ney silliqlash: Kneser-Ney silliqlash usulida joriy so'z ketma-ketlikning

davomi bo‘lish ehtimolidan foydalaniladi. Interpolyatsiya qilingan Kneser-Ney silliqlash algoritmi diskontlangan ehtimolni quyi tartibdagi davom etish ehtimoli bilan aralashtirib yuboradi.

$$P_{AbsoluteDisconting}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) - d}{\sum_v C(w_{n-1}v)} + \lambda(w_{n-1})P(w_n) \quad (13)$$

Til korpusni va ma`lumotlarni boslang`ich qayta ishlash

Biz foydalanadigan korpus O‘zbek tili korpusi bo‘lib, u 2021-2024 yillarda ToshDO‘TAU kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi xodimlari tomonidan ishlab chiqilgan [9]. Amaliy tadqiqot uchun ushbu til korpusidan 1120458 ta gap tanlab olindi.


677207 Uning qo‘shimcha qilishicha, ikkita tulki bolasi juda yosh bo‘lgani va o‘zini eplolmagani bois tabiat qo‘yniga qo‘yib yuborilmagan.
677208 Uning qo‘shimcha qilishicha, parrandaga ayblov e’lon qilinmaydi, lekin odam o‘limiga sababchi sifatida sudga taqdim etilishi kerak.
677209 Uning sevimli taomlarimdan biri – sabzavotli jigarrang guruch, avokado, shuningdek, greypfurtli tovuq filesidan tayyorlangan salat.
677210 Uning shikoyat arizalari har doim Samarqand viloyat hokimligiga, viloyat prokuraturasiga yoki Ichki ishlar boshqarmasiga yuborildi.
677211 Uning so‘zlariga ko‘ra, AQSh Isroilning bosh homiysi, Britaniya va boshqa G‘arb davlatlari Falastinga qarshi jinoyatlarga aloqador.
677212 Uning so‘zlariga ko‘ra, O‘zbekiston mamlakat bilan hamkorlikda qishloq xo‘jaligi va texnologik hamkorlikka ustuvor ahamiyat beradi.
677213 Uning so‘zlariga ko‘ra, agentlik filmni “Oskar” talablariga moslashtirish, tarjima va subtitrlar uchun qo‘shimcha mablag‘ ajratgan.
677214 Uning so‘zlariga ko‘ra, bu rus xalqining o‘z suvereniteti va doimiy rivojlanishi atrofida birligini tasdiqlovchi ishonchli g‘alaba.
677215 Uning so‘zlariga ko‘ra, hozirgi sharoitdan kelib chiqqan holda, AQSh iqtisodiyotida resessiya ehtimoli sezilarli darajada oshmoqda.
677216 Uning so‘zlariga ko‘ra, hukumat ushbu mablag‘larni maktablardagi bolalarning ovqatlanishiga yo‘naltirish taklifini ilgari surmoqda.
677217 Uning so‘zlariga ko‘ra, ko‘llar tizimida baliqchilik tarmog‘ini rivojlantirish borasidagi ishlar ham yetarlicha yo‘lga qo‘yilmagan.
677218 Uning so‘zlariga ko‘ra, mamlakatlar energetika va oziq-ovqat inqirozini yengish uchun birgalikda harakat qilishni rejalashtirmoqda.
677219 Uning so‘zlariga ko‘ra, piyodalar uchun sharoitlar yomonlashmaydi, ularning soni esa “hech qanaqasiga kamaymaydi”, faqat ko‘payadi.
677220 Uning so‘zlariga ko‘ra, respublika Ukrainaga barcha sohalarda, jumladan, mudofaa yo‘nalishida ham faol yordam berishda davom etadi.
677221 Uning tahriridan so‘ng muxbirlarning maqolalari xuddi yomg‘irdan so‘ng osmonda jilolangan kamalak kabi gazeta sahifalarini bezardi.
677222 Uning tarafdorlari, masalan, Aleksey Kudrin, German Gref va boshqalar Putinga bu tamoyillar qanday ishlashini tushuntirib berishdi.
677223 Uning ta’kidlashicha, Rossiya va Ukraina o‘rtasidagi urush tugagan taqdirda ham, NATO va Moskva orasidagi ziddiyat saqlanib qoladi.
677224 Uning ta’kidlashicha, agar ICC so‘rovi bilan bog‘liq muammolar mavjud bo‘lsa, uni olgan davlat zudlik bilan maslahatlashishi kerak.
677225 Uning ta’kidlashicha, bir qator mamlakatlarda YeI davlatlarida sanksiya ostiga kiritilgan tovarlar importi ko‘payishi kuzatilmoqda.
677226 Uning turmush o‘rtog‘i sahifasi orqali tug‘ilgan kuni bilan tabriklab, aktrisaning bolalikdagi rasmi tushirilgan tort sovg‘a qildi.
677227 Uning ushbu bosqichdagi raqibi niderlandiyalik Xenk Grol bo‘ldi va uni ippon bahosi bilan mag‘lub etdi va chorak finalga yo‘l oldi.
677228 Uning vazifalarini birinchi guruh menejeri bilan yaqindan integratsiya lashgan va bosh direktorga hisobot beradigan jamoa bajaradi.
677229 Uning “Yo‘l” romani 2007-yili “Pulitser” mukofotiga sazovor bo‘ldi va uch yil davomida eng ko‘p sotilgan kitoblar ro‘yxatiga kirdi.
677230 Uning “bilimlar ombori”ni, shuningdek, radio, televideniye va boshqa ommaviy axborot vositalaridan olgan axborotlar ham to‘ldiradi.
677231 UningINEOS kompaniyasi yaqinda “qizil iblislar”ning 25% ulushini sotib oldi va klub futbol ishlarini boshqarishni o‘z qo‘liga oldi.
677232 Universitet bilan hamkorlik yangi ufqlarni ochadi hamda IT-mutaxassislarni tayyorlash tizimi platformasini yaratishga imkon beradi.
677233 Universitetimizda talabalar bilimini nazorat qilish tartibi, baholash mezonlari ishlab chiqilgan va o‘quv jarayoniga joriy etilgan.
677234 Universitetimizda ta’lim faqat kunduzgi ta’limdan iborat bo‘lib, hozirda reyting bo‘yicha TOP – 300 talikda 277-o‘rindan joy olgan.
677235 Unvonlarni qo‘lga kiritish uchun jahon darajasidagi darvozabonga ega bo‘lish kerak va “Manchester Yunayted” hozir o‘zgarishi kerak.

3-rasm. *O‘zbek tili korpusi matnlaridan namuna*

Birinchi qadamda, korpus ma’lumotlarni boshlang`ich qayta ishlash kerak. Bunda barcha gaplar kichik harflarga aylantiradi. Chunki katta-kichik registrlarning har xil harfli gramlarda bir xil bo‘lishini ta'minlanad. Masalan: (Men,bilan) bigrami (men,bilan) bilan bir xil bo‘lishi kerak. So‘ngra, gaplar ustida tokenlash amali bajariladi. Shuningdek, ushbu bosqichda tinish belgilari olib tashlanadi. Chunki ular modelga hech qanday ahamiyat bermaydi. Korpus matnlariga boshlang`ich ishlov berilgandan so‘ng quyidagi ko‘rinishga keladi:

[['[BEGIN]', 'toshkent', 'vaqti', 'bilan', '13', '12', 'da', '01', '01', '2024-yil', 'grinvich', 'bo‘yicha', '08', '12', 'da', 'Afg‘onistonda', 'zilzila', 'sodir', 'bo‘ldi', '[END]'], ['[BEGIN]', '0', '0', 'hisobida', 'yakunlangan', 'ana', 'shu', 'o‘yin', 'asosiy', 'jamoada', 'yagona', 'bo‘lib', 'qoldi', '[END]'], ['[BEGIN]', '1', 'rubl', '0', '58', 'so‘mga', 'qimmatladi', 'va', '136', '01', 'so‘mga', 'teng‘, 'bo‘ldi', '[END]']]

Korpus lug`atini shakllantirish

Ikkinchi qadamda, boshlang`ich qayta ishlangan ma’lumotlardan korpus lug`atini shakllantirish kerak. Lug‘at – bu korpus matnlarining unikal so‘zlari to‘plami. Lug‘atni yaratishning sababi shundaki, u OOV (Out of Vocabulary) so‘zlarini aniqlashga yordam beradi va OOV ko‘rsatkichini aniqlashga xizmat qiladi. Shuningdek, Add-k silliqlash usulidagi **|V|** qiymatni aniqlash uchun lug‘at hajmini bilish kerak. Korpus lug‘ati aniqlashgach, unga **[BEGIN]**, **[END]**, **[UNK]** kabi tokenlarni qo‘shish lozim.

Modelni shakllantirish

Til modelini shakllantirishdan avval, quyidagi qo‘shimcha tokenlardan foydalanishga e'tibor bering: gap boshini bildirish uchun **[BEGIN]**, gap oxirini bildirish uchun **[END]** va lug‘atda mavjud bo‘lmagan so‘zlarni bildirish uchun **[UNK]**, ya'ni OOV so‘zlari. Quyidagi jadvalda berilgan til korpusi asosida shakllantirilgan bigram, trigram va quadrigramlarning eng ko‘p uchragan juftliklari keltirilgan

.

bigram	trigram	quadrigram
[(('[BEGIN]', 'Bu'), 27047), (('edi', '[END]'), 15190), (('mumkin', '[END]'), 14568), (('qildi', '[END]'), 12702), (('[BEGIN]', 'Shu'), 10691), (('bo‘ldi', '[END]'), 10016), (('[BEGIN]', 'Shuningdek'), 9996), (('[BEGIN]', 'U'), 9904), (('bo‘ladi', '[END]'), 9785), (('kerak', '[END]'), 8339)]	[(('[BEGIN]', 'Bundan', 'tashqari'), 4659), (('[BEGIN]', 'Bu', 'haqda'), 4225), (('[BEGIN]', 'Eslatib', 'o‘tamiz'), 4128), (('ma’lum', 'qildi', '[END]'), 3935), (('[BEGIN]', 'Shu', 'bilan'), 2979), (('[BEGIN]', 'Bu', 'haqida'), 2968), (('[BEGIN]', 'Qayd', 'etilishicha'), 2825), (('xabar', 'berdi', '[END]'), 2780), (('Shu', 'bilan', 'birga'), 2762), (('bo‘lib', 'o‘tdi', '[END]'), 2424)]	[(('[BEGIN]', 'Shu', 'bilan', 'birga'), 2730), (('[BEGIN]', 'Uning‘, 'so‘zlariga', 'ko‘ra'), 1187), (('[BEGIN]', 'Mazkur', 'holat', 'yuzasidan'), 802), (('harakatlari', 'olib', 'borilmoqda', '[END]'), 757), (('xizmati', 'xabar', 'berdi', '[END]'), 663), (('bilan', 'jinoyat', 'ishi', 'qo‘zg‘atilib'), 653), (('deb', 'xabar', 'bermoqda', 'Dunyo‘), 639), (('golli', 'uzatmani', 'amalga', 'oshirdi'), 629), (('uzatmani', 'amalga', 'oshirdi', '[END]'), 619), (('[BEGIN]', 'Eslatib', 'o‘tamiz', 'avvalroq'), 599)]

Yangi matnni generatsiya qilish

Yaratilgan til modellaridan foydalanib o‘zbek tilidagi yangi matnni generatsiya qilish masalasini ko‘rib chiqamiz. Ushbu amalni 3 xil usulda bajarish mumkin:

1) $w_1, w_2, \ldots$: ketma-ketlik asosida gapning keyingi qismini shakllantirish;

2) $w_1, w_2, \ldots, w_{n-1}, w_n$: gapning noma`lum qismini shakllantirish;

3) $\ldots, w_{n-1}, w_n$: ketma-ketlik asosida gapning boshlang`ich qismini shakllantirish.

Quyidagi jadvalda "bizning maktabimiz" bigramasiga mos generatsiya qilingan gaplarga namuna keltirilgan:

Bigrama	Generatsiya qilingan gap
Bizning maktabimiz …	[BEGIN] Bizning maktabimiz binosi referendum uchastkasi sifatida faoliyat yuritdi [END]
	[BEGIN] Bizning maktabimiz jamoasi nomidan samimiy minnatdorchilik bildirmoqchimiz [END]
… bizning maktabimiz …	[BEGIN] Odatda bizning maktabimiz bitiruvchilarining deyarli hammasi institut yo universitet talabasi bo‘lardi [END]
	[BEGIN] Odatda bizning maktabimiz kursantlari tomonidan amaliyot o‘tash uchun ishlatiladi [END]

Quyidagi jadvalda bigramalarga mos *gapning keyingi* va *boshlang`ich qismini* shakllantirish namunalari keltirilgan:+

	N	Nozmod N-gramlar
"bizning maktabimiz"	6	Nomzodlar davomidan: ('maktabimiz','binosi','referendum','uchastkasi','sifatida','faoliyat'):4.10e-06 ('maktabimiz','bitgan','bo‘lganida','balki','shu','hodisa') : 4.10e-06 ('maktabimiz','balki','respublikamizdagi','barcha','bitiruvchi','yoshlarimizning‘) :4.10e-06
		('maktabimiz','kuchli','ustozlarimiz','bor','edi','va'):4.10e-06
	6	('maktabimiz','binosini','ta’mirlashga','javobgar','mutasaddilar','aytilgan'): 4.10e-06 ('maktabimiz','binosi','referendum','uchastkasi','sifatida'):4.10e-06 ('maktabimiz','bitgan','bo‘lganida','balki','shu') : 4.10e-06 ('maktabimiz','balki','respublikamizdagi','barcha','bitiruvchi') : 4.10e-06 ('maktabimiz','kuchli','ustozlarimiz','bor','edi') : 4.10e-06 ('maktabimiz','binosini','ta’mirlashga','javobgar','mutasaddilar') : 4.10e-06 ('maktabimiz','zaligacha','issiq','turardi','[END]') : 4.10e-06 ('maktabimiz','uchun','g‘isht','quyardik','[END]') : 6.15e-06
	5	Nomzodlar boshidan: ('ko‘zlangan','asosiy','maqsad','va','bizning‘) : 4.10e-06 ('qanday','yakunlanishi','mumkin','va','bizning‘) : 4.10e-06 ('[BEGIN]','Xorijdan','kelgan','mutaxassis','bizning‘) : 4.10e-06 ('O‘zbekistonning‘,'2','million','fuqarosi','bizning‘) : 4.10e-06 ('uqtiradi','bizga','yaxshilab','yozing‘,'bizning‘) : 4.10e-06
"bizning maktab"	4	Nomzodlar davomidan: ('maktab', 'orzusi', 'barchani', 'lekin') : 4.10e-06 ('maktab', 'ekaniga', 'esa', 'shak-shubha') : 4.10e-06 ('maktab', 'haqida', 'salbiy', 'fikrlar') : 4.10e-06 ('maktab', 'tanlab', 'olinadi', '[END]') : 4.10e-06 ('maktab', 'bitiruvchilarini', 'zamonaviy', 'kasblar') : 8.20e-06 ('maktab', 'ta’mirtalab', 'hisoblanadi', '[END]') : 4.10e-06
	4	Nomzodlar boshidan: ('[BEGIN]', 'Xo‘sh', 'bu', 'bizning‘) : 4.10e-06 ('[BEGIN]', 'Xo‘sh', 'endi', 'bizning‘) : 4.10e-06 ('2', 'million', 'fuqarosi', 'bizning‘) : 4.10e-06 ('bizga', 'yaxshilab', 'yozing‘, 'bizning‘) : 4.10e-06
	5	Nomzodlar davomidan: ('maktab', 'o‘qituvchilari', 'va', 'bog‘cha', 'tarbiyachilari') : 1.84e-05 ('maktab', 'o‘quvchilari', 'o‘rtasida', 'so‘rovlar', 'o‘tkazib') : 4.10e-06 ('maktab', 'sumkasi', 'darsliklar', 'majmuasidan', 'iborat') : 1.02e-05 ('maktab', 'ta’limi', 'vazirligi', 'izoh', 'berdi') : 4.10e-06 ('maktab', 'binosi', 'rekonstruksiya', 'qilindi', '[END]') : 4.10e-06 ('maktab', 'masofaviy', 'ta’limga', 'o‘tdi', '[END]') : 4.10e-06
	5	Nomzodlar boshidan: ('qanday', 'yakunlanishi', 'mumkin', 'Va', 'bizning‘) : 4.10e-06 ('[BEGIN]', 'Xorijdan', 'kelgan', 'mutaxassis', 'bizning‘) : 4.10e-06 ('O‘zbekistonning‘, '2', 'million', 'fuqarosi', 'bizning‘) : 4.10e-06 ('uqtiradi', 'bizga', 'yaxshilab', 'yozing‘, 'bizning‘) : 4.10e-06

Modelni baholash

Yuqorida ko‘rsatilganidek, til modelining ishlashi ichki usul – **chalkashlik (perplexity)** yordamida baholanadi. Yaratilgan til modeli ko‘rinmas hodisalarga nol ehtimolliklarni belgilashiga yo‘l qo‘ymaslik uchun **silliqlash** yoki **diskontlash** jarayoni amalga oshiriladi. Ushbu maqolada shaklantirilgan N-gram modeli uchun Kneser-Ney silliqlashdan foydalaniladi. O‘quv korpusidan butunlay farq qiladigan test ma’lumotlari bo‘yicha til modelimizning ishlashini baholaymiz.

9-formula asosida W ehtimolini kengaytirish uchun zanjir qoidasidan foydalanishimiz mumkin. U holda, bigram til modeli uchun, W ehtimollik qiymati quyidagi 10-formula asosida aniqlanadi.

13-formuladan foydalanib, Kneser-Ney silliqlashni qo‘llaymiz.

_____

Perplexity of Unigram Model: 12827.902838562914

Perplexity of Bigram Model: 55835.11296314676

Perplexity of Trigram Model: 243452.72320897126

Xulosa

Til odamlar uchun asosiy aloqa usuli bo‘lib, u odamlarning o‘zaro ta’siri uchun asosdir. Shunday qilib, bugungi kunda tabiiy tilni qayta ishlash (natural language processing, NLP) tadqiqot va rivojlanishning hal qiluvchi sohasi sifatida shakllandi. Statistik til modellari (Statistical Language Model, SLM) kompyuterlarga tabiiy tilni tushunish va yaratishga imkon beruvchi NLPning asosiy usullaridan biridir. SLMlar til tasodifiy so‘zlarning to‘plami emas, balki qoidalar va qoliplar tizimi degan g‘oyaga asoslanadi. Ushbu qoliplarga statistik usullarni qo‘llash orqali SLMlar ma’lum bir kontekstda muayyan so‘zlar va birikmalarning ehtimolini aniqlashi va bashorat qilishi mumkin. Turli xil til qoliplari ehtimolini bashorat qilish qobiliyati SLMlarni tabiiy tilni qayta ishlashda muhim vositaga aylantiradi.

Statistik til modellarining asosan ikki turi mavjud: n-gram modellari va neyron tarmoqqa asoslangan modellar. N-gram modellari oldingi n-1 ta so‘zlarni hisobga olgan holda joriy so‘zning kontekstda paydo bo‘lish ehtimolini hisoblashga asoslanadi. Ehtimollik maksimal ehtimollikni baholash usuli yordamida hisoblanadi. Eng ko‘p ishlatiladigan n-gramli modellar bigram, trigram va 4 gramli modellardir. Ushbu maqolada o‘zbek tili korpusi asosida N-gram modellarini ishlab chiqish usullari va til modellarini baholash usullari keltirildi. Shuningdek, Ngram modellari yondashuvi bilan bog‘liq umumlashtirish, sezgirlik, OOV (noma’lum so‘zlar), maxsus kontekst muammolari va ularni bartaraf qilish usullari keltirildi.

Shunday qilib, til modellari gap yoki boshqa so‘zlar ketma-ketligiga ehtimollik belgilash va oldingi so‘zlardan so‘zni bashorat qilish usulini taklif qiladi. Odatda, N-gram til modellari chalkashlik perplexity bahosi foydalanib baholanadi. Test to‘plamining chalkashligi til modeliga ko‘ra, model tomonidan hisoblangan teskari ehtimolining geometrik o‘rtachasiga teng. Silliqlash algoritmlari n-gram ehtimolini baholashning murakkab usuli hisoblanadi. Odatda, N-gramlar uchun ko‘p ishlatiladigan silliqlash algoritmlari **orqaga qaytarish (backoff)** yoki **interpolyatsiya (interpolation)** orqali quyi tartibli n-gramlarga asoslanadi. Kneser-Ney silliqlash usuli yangi so‘zning davomi bo‘lish ehtimolidan foydalanadi. Interpolyatsiya qilingan Kneser-Ney silliqlash algoritmi diskontlangan ehtimolni quyi tartibdagi davom etish ehtimoli bilan aralashtiradi.

Foydalanilgan adabiyotlar

1. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases?. *arXiv preprint arXiv:1909.01066*.
2. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
3. Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., ... & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language

models: A survey. *ACM Computing Surveys*, *56*(2), 1-40.
4. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274.
5. De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., & Bruza, P. (2014, November). Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management* (pp. 1819-1822).
6. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... & Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
7. Elov, B., & Xudayberganov, N. (2024). O'zbek tili korpusi matnlarini pos teglash usullari. *Computer Linguistics: problems, solutions, prospects*, *1*(1).
8. Elov, B., & Abdullayeva, O. (2024). O'zbek tili korpusini sintaktik teglash masalasi. *Computer Linguistics: problems, solutions, prospects*, *1*(1).
9. Elov, B., & Xusainova, Z. (2024). Til korpuslarini lingvistik teglash bosqichlari. *Computer Linguistics: problems, solutions, prospects*, *1*(1).
10. Aouragh, S. L., Yousfi, A., Laaroussi, S., Gueddah, H., & Nejja, M. (2021). A new estimate of the n-gram language model. *Procedia Computer Science*, *189*, 211-215.
11. Suzuki, M., Itoh, N., Nagano, T., Kurata, G., & Thomas, S. (2019, May). Improvements to n-gram language model using text generated from neural language model. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7245-7249). IEEE.
12. Chiu, J. T., & Rush, A. M. (2020). Scaling hidden Markov language models. *arXiv preprint arXiv:2011.04640*.
13. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1-45.
14. Malagutti, L., Buinovskij, A., Svete, A., Meister, C., Amini, A., & Cotterell, R. (2024). The Role of $ n $-gram Smoothing in the Age of Neural Networks. *arXiv preprint arXiv:2403.17240*.