TABIIY TIL QOLIPLARINI N-GRAM METODI VOSITASIDA ANIQLASH

Mastura Primova

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti
E-mail: primovamastura@navoiy-uni.uz

KEYWORDS

Til modellari, n-gram metodi, til korpusi, til qoliplari, unigram, bigram, trigram, mashinali o'qitish.

ABSTRACT

N-gram metodi – matn ma'lumotlaridagi qolip va munosabatlarni aniqlash uchun tabiiy tilni qayta ishlash (Natural Language Processing, NLP)da qo'llaniladigan matnni tahlil qilish usuli. Ushbu metod matnni n-gram deb ataladigan kichikroq birliklarga bo'lish va matn ma'lumotlari haqida tushunchaga ega bo'lish uchun ushbu n-grammlarning chastotasi va ularning korpusda tarqalishini tahlil qilishni o'z ichiga oladi. N-gramlar so'zlar, belgilar yoki boshqa har qanday mazmunli matn birliklaridan iborat bo'lishi mumkin. Til korpusidahi N-gramlar tahlili muhim ahamiyatga ega bo`lib, u matn ma'lumotlarini tahlil qilish va ma'lumotlar ichidagi qolip va munosabatlarni aniqlashning sodda, ammo samarali usulini taklif qiladi. N-gram metodi tilni modellashtirish, matnni tasniflash va his-tuyg`ularni tahlil qilish kabi turli xil NLP ilovalarini ishlab chiqish uchun foydali bo'lishi mumkin. N-gram tahlili tilni modellashtirishda matn ma'lumotlaridagi qolip va munosabatlarni aniqlash hamda tabiiy tilni qayta ishlash vazifalari uchun bashoratli modellarni yaratish uchun qo`llaniladi. Shungdek, N-gram tahlili matnni tasniflashda matnning asosiy xususiyatlarini aniqlash va matnni oldindan belgilangan toifalarga ajratish uchun ishlatiladi.

Kirish.

N-gram – tabiiy tilni qayta ishlashda matnda birga paydo bo'ladigan N so'z yoki belgilar ketma-ketligidir [1]. Korpus lingvistikasida n-gramlar tokenlar (so'zlar) hisoblanadi. N-gramlar tilshunoslikda **MWE (multiword expressions)** yoki **ko'p so'zli birikmalar** deb ham ataladi [2]. Til korpusida eng ko'p uchraydigan n-gramlar ro'yxatini yaratish bizga lingvistik tahlilning boshqa usullari bilan unchalik aniq bo'lmagan tildan foydalanish qoliplarini aniqlashga yordam beradi.

NLPdagi N-gramalar kontekstual ma'lumotni va matndagi so'zlar orasidagi munosabatlarni tushunishga yordam berish uchun matndan olingan "**n**" ta so'zlarning ketma-ketligidir. Element bitta so'z (unigram) yoki bir nechta so'z, belgilar yoki bo'g'inlar (bigram, trigram, quadrigram va boshqalar) bo'lishi mumkin. Tabiiy tilni qayta ishlashda matn tahlili uchun N-gram tushunchasi keng qo'llaniladi. 1 o'lchamdagi N-gram - "**unigram**", 2 o'lchamli - "**bigram**", 3 o'lcham - "**trigrama**" deb nomlanadi [4].

1-jadval.

N-gramlarga namuna

Matn	N-gram
Ajoyib	1-gram (unigram)
Men keldim	2-gram (bigram)
Bu kitobni o`qidim	3-gram (trigram)
Biz birgalikda universitetga boramiz	4-gram (quadrigram)

Berilgan matn yoki til korpusda n ta so'zdan iborat oynani siljitish orqali n-gramlar ro`yxatini shakllantirish mumkin. Ushbu n-gramlarni ajratib olgandan so'ng, ma'lum so'z ketma-ketliklarining paydo bo'lishini tahlil

qilish, ko`p uchraydigan so'zlarni aniqlash va matndagi til qoliplarini tushunish mumkin.

N-gram tahlili tilni qayta ishlashning muhim metodi (usuli) bo‘lib, u til tuzilishini tushunishga va gapda keyin nima kelishini bashorat qilishga yordam beradi. N-gram tahlilini amalga oshirish matnni n-gramga ajratish orqali, har bir n-gram chastotasini hisoblash va matn ma'lumotlari haqida tushunchaga ega bo'lish uchun n-gram chastotasi va ularning tarqalishini tahlil qilishni o'z ichiga oladi. Bu jarayon NLP dasturi yordamida avtomatlashtirilishi yoki matndagi n-gram chastotasini hisoblash orqali qo‘lda bajarilishi mumkin. N-gramlarni tahlil qilishdan *matnni tasniflash, matn yaratish, imloni tuzatish* va *his-tuyg`ularni (sentiment)* tahlil qilish kabi murakkab NLP vazifalarni mashinali o'qitish modellarini o'rgatishda muhim rol o'ynaydi. NLP modellari til qoliplarini yaxshiroq tushunishi va qaysi so'zlar birgalikda paydo bo'lish ehtimolini o'rganish orqali joriy so`zlar ketma-ketligini aniq bashorat qilishlari mumkin. Bu *mashina tarjimasi*, *chatbotlar* va *qidiruv tizimlari* kabi turli NLP ilovalarini yaxshilashga yordam beradi.

Adabiyotlar tahlili

B.William va John M. Trenkle (1994) o`z tadqiqotlarida N-gram metodi asosida matnni tasniflash NLP vazifasini hal qilishi natijasida, elektron pochtada xabarlaridagi imlo va grammatik xatolar hamda OCR orqali kelgan hujjatlardagi belgilarni aniqlash xatolarini bartaraf qilgan va turli tillarda yozilgan Usenet yangiliklar guruhidagi maqolalarida 99.8% to‘g‘ri tasniflash ko‘rsatkichiga erishdgan [5]. Shuningdek, Peter Náther o`zining “N-gram based Text Categorization” nomli ilmiy tadqiqotida N-gram metodiga asoslangan matnni tasniflash algoritmini taklif qilgan va katta bo`lmagan ma`lumotlar to`plamiga qo`llagan [6].

Bashir Ahmad, Sung-Hyuk Cha va Charlz Tappertlar N-gramlarning ad-hoc kumulyativ chastota qo'shilishidan foydalangan holda samarali til tasniflagichini ishlab chiqishgan [7]. N-gramm asosidagi tartibli statistik ma'lumotlardan foydalangan holda til tasnifi juda aniq va bosma xatolarga sezgir emasligi ko'rsatilgan va natijada bu usul tilni qayta ishlash adabiyotlarida keng tadqiq qilingan.

Richard Zens va Hermann Ney statistik mashina tarjimasi uchun N-Gram ehtimolliklari bo`yicha tadqiqot olib borishgan va nutqni avtomatik aniqlash hamda mashina tarjimasida ishonchni baholash uchun keng tarqalgan yondashuvni ishlab chiqishgan [8]. Ular tomonidan ishlab chiqilgan metodning xitoycha-inglizcha NIST vazifasida qo`llanilishi BLEU ballining mutlaq yaxshilanishi 1,1% dan 1,6% gacha sezilarli yaxshilanishlarini qayd etishgan.

Bugungi kunda kompyuterlar va Internetdan tobora keng foydalanilishi natijasida tabiiy tillaridagi katta hajmdagi ma'lumotlar hosil bo`lmoqda. Ushbu avtomatik ma'lumotlarni qayta ishlash va qidirish qizimlarini (information retrieval, IR) optimallashtirish dolzarb vazifa hisoblanadi. Jumladan, P.Majumder, M.Mitra va B.B.Chaudhuri ko'p tilli mamlakat hisoblangan Hindiston kontekstidagi IRda N-grammlardan muvaffaqiyatli foydalanilgan [9].

Atanu Dey, Mamata Jenamani va Jitesh J. Thakkarlar N-grammlar asosida his-tuyg`ularni tahlil qilishga mo`ljallangan Sentiment-Na-Gram tizimini ishlab chiqishgan [10]. O'quv ma'lumotlari yetarli bo'lmaganda, his-tuyg`ularni lu`gatga asoslangan tahlil qilish yondashuvi MLga asoslangan usullardan afzalroqdir. Mavjud lu`gatlarga faqat unigram va hissiyot ballari mavjud. Unigramlarni kuchaytiruvchilar yoki inkorlar bilan birlashtirish natijasida hosil bo'lgan hissiyot n-grammlari yaxshilangan natijalarni ko'rsatishi kuzatilgan. Tavsiya etilgan qoidaga asoslangan yondashuv n-gram hissiyot ballarini mahsulot sharhlari va besh balllik shkalada mos keladigan raqamli reytingni o'z ichiga olgan tasodifiy korpusdan chiqarilgan. Senti-N-Gram lug`atdan foydalanilganda, taklif qilingan usul VADER va SO-CAL n-gram hissiyot tahlili yondashuvidan foydalangan holda taniqli unigram-lug`atga asoslangan yondashuvdan ustunligi isbotlangan.

N-gramlar

N-gram modeli his-tuyg`ularni tahlil qilishda muhim ahamiyatga ega bo`lib, *neytral*, *ijobiy* yoki *salbiy* his-tuyg'ularni ifodalovchi so'zlarning qoliplarini tahlil qila oladi. Masalan, "**juda yaxshi film**" so`z birikmasidagi "**juda yaxshi**" va "**yaxshi film**" kabi bigramlar mavjud bo'lib, ular ijobiy his-tuyg'ularni bildiradi.

Matnni sentiment tahlil qilish algoritmlari katta hajmli matnli ma'lumotlar to'plamlarida ushbu N-gramlarni tahlil qilish orqali turli xil hissiyotlarga tegishli umumiy kombinatsiyalarni aniqlashi mumkin. Shuningdek, xususiyatlarni ajratib olishda N-gramlar matndagi ma'noni bildiruvchi *muhim so'zlarni* yoki *so`z birikmalarini* aniqlashga yordam beradi.

Ijttimoiy tarmoq foydalanuvchilarining kompaniya mahsulotini haqidagi quyidagi sharhini ko'rib chiqamiz: **"ajoyib mahsulot"** yoki **"past sifatli"** kabi bigramlar sharhlar *ijobiy* yoki *salbiy* bo'lishidan qat'i nazar, sharh hissini aniq tasniflash uchun xususiyat sifatida xizmat qilishi mumkin.

N-Gram til modellari

N-gram tili modellari *og`zaki nutqni matnga aylantirish modellarini takomillashtirish* va *gapdagi so'z ehtimolini bashorat qilish* imkonini beradi. Ushbu modellar til qoliplarini yaxshiroq tushunish uchun N ta so'z yoki belgilar ketma-ketligini tahlil qilish usulidan foydalanadi.

2-jadval.

Berilgan matnning N-gramlari

Qayerdan keldi bu yoqimli ovoz?	
Unigram	qayerdan; keldi; bu; yoqimli; ovoz
Bigram	qayerdan keldi; keldi bu; bu yoqimli; yoqimli ovoz
Trigram	qayerdan keldi bu; keldi bu yoqimli; bu yoqimli ovoz
Quadrigram	qayerdan keldi bu yoqimli; keldi bu yoqimli ovoz

N-gram tili modellari og`zaki nutqni matnga aylantirish modellarini takomillashtirish va gapdagi so'z(lar) ehtimolini bashorat qilish imkonini beradi. Ushbu modellar til qoliplarini yaxshiroq tushunish uchun N ta so'z yoki belgilar ketma-ketligini tahlil qilish orqali ishlaydi. N-gram modellarining eng keng tarqalgan turlari *unigramlar, bigramlar* va *trigramlardir*. Ularning har biri turli darajadagi kontekstni qamrab oladi.

Unigramlar

Unigrammalar matndagi unikal so'zlardir. Ular kontekstga bog`liq bo`lmasdan, alohida so'zlar haqida asosiy ma'lumotlarni beradi. Masalan, "**Uni kim chorlayapti?**" gapidagi unigrammalar: **"uni", "kim"** va **"chorlayapti"**.

Bigrammalar

Har qanday gapda bigrammalar ikkita qo'shni so'zdan iborat. Ular cheklangan kontekstdagi so'z munosabatlari haqida tushuncha beradi. Misol uchun, **"qora mushuq"** so`z birikmasida bigramma mushukning rangini tavsiflaydi; **"tug'ilgan kuningiz bilan"** so`z birikmasi esa bayram tadbirini aks ettiradi. Bu juftliklar til tuzilishi va ma'nosini tushunishga yordam beradi.

Trigrammalar

Matndagi yonma-yon kelgan uchta so'zlar ketma-ketligi trigrammalar deb ataladi. Bigramlar bilan solishtirganda, trigrammalar til kontekstini chuqurroq tushunish imkonini beradi. Masalan, **"qimizi olma daraxti"**da trigramma mevaning rangini ham, daraxtga munosabatini ham ochib beradi. Trigrammalar til qoliplari va ma'nosini tahlil qilish uchun muhim ahamiyat kasb etadi.

Og`zaki nutqni matnga o'tkazish modellarida N-gramlar oldingi so'zlar asosida keyingi so'zni bashorat qilishga yordam beradi. Misol uchun, agar til modeli **"men kitobni"** degan so'z birikmasini qabul qilsa, u mashg'ulot ma'lumotlaridagi N-gramlarni tahlil qilish natijasida o'rgangan kontekstga qarab **"o`qiyman"** yoki **"oldim"** so`zlarini taxmin qilishi mumkin.

Shuningdek, N-gram metodi berilgan so'zlar ketma-ketligidan keyin so'zning paydo bo'lish ehtimolini hisoblashga yordam beradi. Tabiiy tilni qayta ishlashda turli NLP vazifalari uchun ehtimollikni aniqlash/hisoblash juda muhimdir.

Masalan, "kitobni" so'zidan keyin ehtimolligi katta matnli korpusdagi N-gramlarni tahlil qilish asosida **"o`qimoq"** leksemasining grammarik shakllari **"olmoq"** so`ziga qaraganda yuqori bo'lishi mumkin. Umuman olganda, N-gram til modellari tabiiy tilni *samarali tushunish* va *qayta ishlashda* muhim rol o'ynaydi.

N-gram metodining qo'llanilishi

Tabiiy tilni tushunish uchun N-gram metodi tahlilini turli sohalarda qanday qo'llanilishini

ko'rib chiqamiz.

His-tuyg'ularni tahlil qilish va matnni tasniflash

N-gram analizatori his-tuyg'ularni tahlil qilish va matnlarni tasniflashda keng qo'llaniladi. Bunda, his-tuyg'ular yoki mavzularni aniqlash uchun matndagi qoliplar aniqlanadi. Masalan, "**Menga meva juda yoqdi**" gapidagi "**juda yoqdi**" bigrammasi ijobiy his-tuyg'ularni bildiradi. Bu sharhlar yoki ijtimoiy media xabarlarini to'g'ri tasniflashga yordam beradi.

3-jadval.

Berilgan matnning N-gramlari

Unigram	**Bigram**	**Trigram**
Gi`jduvon	Gi`jduvon shashligining	Gi`jduvon shashligining mazasi
shashligining	shashligining mazasi	shashligining mazasi yomon
mazasi	mazasi yomon	mazasi yomon emas
yomon	yomon emas	
emas		

Yuqoridagi 3-jadvalda **"Gi`jduvon shashligining mazasi yomon emas"** gapining unigram, bigram va trigram modellari keltirilgan bo`lib, matnni tahlil qilish uchun unigram yoki bitta so'zni ko'rib chiqsak, salbiy **"yomon"** so'zi matnni noto'g'ri bashorat qilishga olib keladi. Ammo agar biz bigramdan foydalansak, **"yomon emas"** bigrammasi matnni ijobiy his-tuyg'u sifatida taxmin qilishga yordam beradi.

Nutqni tanib olish

Nutqni tanib olish va uni matnga o'tkazish ilovalarida N-gram tahlili keyingi so'zni kontekstga qarab bashorat qilishga yordam beradi. Misol uchun, kimdir **"Iltimos, menga olmani"** desa, model umumiy so'zlar ketma-ketligiga qarab **"bering"** yoki **"uzating"** so`zlarini taxmin qilishi mumkin. Bu transkripsiya qilingan nutqning aniqligini oshiradi.

Matn tasnifi

Mashinali o'qitish modellarida N-gramlar matnni turli toifalar bilan bog'liq qolip va xususiyatlarni tanib olish uchun muayyan toifalarga ajratishga yordam beradi. Masalan, kino sharhlarini ikki toifaga ajratish mumkin deb faraz qilsak: *ijobiy* va *salbiy*. Quyidagi sharhni tahlil qilamiz: **"Film mutlaqo fantastik va ajoyib edi"**. Ushbu gapdagi bigrammalar quyidagicha bo'ladi:

- "film mutlaqo";
- "mutlaqo fantastik";
- "fantastik va";
- "va ajoyib";
- "ajoyib edi".

Mashinali o'qitish modeli ushbu bigrammalardan **"mutlaqo fantastik"** va **"ajoyib edi"** kabi bigrammalar odatda ijobiy sharhlar bilan bog'liqligini bilish uchun foydalanishi mumkin. Model ushbu n-gramlarni oldingi sharhlar bilan solishtirib, yangi sharhlar ijobiy yoki salbiy ekanligini taxmin qilishi mumkin.

Matnni bashoratlash (Predictive Text) va avtoto'ldirish (autocomplete)

Bashoratli matn va avtomatik to'ldirish funksiyalari N-gram modeliga tayanib, foydalanuvchi yozganda so'z yoki so`z birikmalarni taklif qiladi. Misol uchun, **"Men … boryapman"** kontekstida, tizim tez-tez uchraydigan trigrammalarga asoslangan holda **"maktabga"** yoki **"ishga"** so`zini taklif qilishi mumkin. Bu yozishni tezlashtiradi va xatolarni kamaytiradi.

Mobil qurilmalar va kalit so'zlardagi bashoratli matn kiritishlari mavjud so'zlar kontekstiga asoslangan ketma-ketlikda keyingi so'zni taklif qilish uchun n-gramdan foydalanadi. Misol uchun, smartfonda quyidagi so`zlar yozilgan bo`lsin: **"Men sut sotib …"**. Bashoratli matn tizimi kiritilgan kontekstga qarab keyingi so'zni taklif qilishi mumkin. Ba'zi takliflar orasida "**olmoqchiman**", "**oldim**", "**olmadim**" va shunga o`xshash "**olmoq**" fe`lining grammatik variantlari bo'lishi mumkin.

Nomlangan obyektni tan olish (NER)

N-gramlar NER tizimlariga *nomlar, manzillar, tashkilotlar, sanalar* va shu kabi nomlangan obyektlarni aniqlash va tasniflashda yordam beradi.

Tematik (mavzuni) modellashtirish

N-gramlar hujjatlar to'plamidagi asosiy mavzularni yoki mavzularni klasterlash va kontent asosida tasniflashga yordam beradi. Misol uchun, katta hajmdagi maqolalar to'plami va maqolalarda muhokama qilinadigan asosiy mavzularni aniqlash lozim bo`lsin. Bunday holda, n-gram modellashtirish algoritmiga salomatlik, sanoat, sun'iy intellekt va boshqalar kabi umumiy so'z ketma-ketligini aniqlashga yordam beradi. Ushbu n-gramlarning chastotasiga asoslanib, algoritm maqolalarni guruhlashi mumkin.

Mashina tarjimasi

N-gramlar mashina tarjimasining umumiy sifatini yaxshilash uchun kengroq kontekstdagi so`z birikmalarni tushunish va tarjima qilishga yordam beradi. Masalan, quyidagi gapni tahlil qilamiz: **"Tezda tepadan tush."** N-gram modeli tizimga ushbu gapdagi **"tush"** inson ko`radigan tush (ot so`z turkimi) emas, balki harakat (fe`l so`z turkimi) ekanligi anglatishi mumkinligini tushunishga yordam berishi mumkin. Agar tizim boshqa kontekstda **"tush"** so'ziga duch kelsa, n-gram modeli **"tush ta`biri"** kabi so`z birikmalar yordamida to'g'ri ma'noni taxmin qilish uchun atrofdagi so'zlardan foydalanishi mumkin.

N-Gram tahlilini SEOga qo`llash

SEO (search engine optimization) – qidiruv tizimlariga kontentni tushunishga yordam berish va foydalanuvchilarga saytlarni topishga va qidiruv tizimi orqali saytlarga tashrif buyurishi kerakligi haqida qaror qabul qilishga yordam berishdir. SEOda N-gram modelini qo`llash natijasida veb-sahifalarda kalit so'zlardan foydalanishni tushunish uchun so'zlar ketma-ketligini (N-gram) o'rganishni mumkin. Misol uchun, agar sog'lom retseptlar haqida veb-sahifani tahlil qilsangiz, **"sog'lom ovqatlanish"** yoki **"retsept qoidalari"** kabi umumiy bigrammalarni o'z ichiga olishi mumkin. Ushbu qolip va chastotalarni o'rganish orqali SEO mutaxassislari tegishli kalit so'zlar yordamida sayt tarkibini optimallashtirishlari mumkin. Shuningdek, **"sog'lom retseptlar"** va **"taomni rejalashtirish"** kabi birgalikdagi so`z birikmalarini tahlil qilish tegishli atamalarni tushunishga yordam beradi.

N-gramlar vositasida ma`lumot izlash tushunchasi qidiruv tizimining natijalari sahifalarida (SERP) veb-sahifaning ko'rinishi va reytingini yaxshilash bilan birga foydalanuvchi qidiruv maqsadiga mos keladigan natijalarni shakllantirishga yordam beradi. Bu oxir-oqibat saytga ko'proq trafikni olib keladi.

N-gram modelidagi cheklovlar

Boshqa har qanday til modeliga o'xshab, N-gram modelida ham bir nechta kamchilik va cheklovlar mavjud:

1. Katta hajmdagi ma'lumotlar to'plamlari va hisoblash murakkabligi.

N-gram modelini shakllantirishda katta hajmdagi xotira va qayta ishlash quvvatiga ehtiyoj tufayli katta hajmdagi ma'lumotlar to'plami bilan bog'liq muammolar yuzaga keladi. Katta hajmdagi matnni tahlil qilish hisoblash resurslarini yuqori bosimda ishlashi, tahlilni sekinlashtirishi yoki tizimning ishdan chiqishiga olib kelishi mumkin.

Bu murakkablik, N-gram modellari N so'zlarning barcha mumkin bo'lgan ketma-ketligini ko'rib chiqishi kerakligidan kelib chiqadi. Chunki ma'lumotlar to'plami hajmi o'sishi bilan ularga qo`yilgan talab kuchayadi. Ushbu hisoblash talablarini samarali boshqarish uchun samarali algoritmlar va apparat vositalari muhim ahamiyatga ega.

2. Ma'lumotlarning kamligi (Data Sparsity) va haddan tashqari moslash (Overfitting Issues) muammolari.

Ba`zida N-gram modelidagi o`quv ma'lumotlarning kamligi va haddan tashqari o'qitilishi bilan bog'liq muammolarga duch kelinadi. Ma'lumotlarning siyrakligi ma'lumotlar to'plamida kamdan-kam uchraydigan so'z birikmalari paydo bo'lib, noto'g'ri prognozlarga olib keladi.

Model o'quv ma'lumotlarini juda yaxshi eslab qolsa, yangi, ko'rinmas ma'lumotlarda yomon ishlaganda haddan tashqari moslashish sodir bo'ladi. Foydali qoliplarni qo'lga kiritish va xotiraga saqlashdan qochish o'rtasidagi muvozanat juda muhimdir. To'g'rilash va tartibga solish kabi usullar modelni sozlash orqali ushbu muammolarni hal qilishga yordam beradi.

N-gram modelining turli NLP ilovalar samaradorligini optimallashtirish uchun ba'zi eng yaxshi tavsiyalarni keltiramiz:

**1. N-gram modeli uchun ma'lumotlarni boshlang`ich qayta ishlash usullari.** Ma'lumotlarni boshlang`ich qayta ishlash N-gram modeli uchun muhim bo`lib, quyida keltirilgan usullar bilan amalga oshirilishi mumkin:

- **Tokenlash (Tokenization):** matnni alohida so‘z yoki belgilarga ajratish;
- **Kichik harflarga o`tkazish (Lowercasing):** so'zlar tahlilini samarali amalga oshirish uchun matnni to`liq kichik harflarga aylantirish;
- **Nomuhim so'zlarni olib tashlash (Removing Stopwords):** muhim ma'noga ega bo'lmagan **"va","uchun",** yoki **"bilan"** kabi umumiy so'zlarni chiqarib tashlash;
- **Tinish belgilarini tahrirlash (Handling Punctuation):** Matndagi tinish belgilarini saqlash yoki olib tashlashni hal qilish.

2. **Aniq NLP vazifalari uchun N-gram modellarini optimallashtirish.** N-gram modelini optimallashtirish uchun turli NLP vazifalarni to'g'ri tushunish va amalga oshirish muhimdir. Bu vazifalarga quyidagilar kiradi:

- **Moslashtirish (Alignment):** N-gram modeli parametrlarini, masalan, N qiymatini, muayyan vazifa talablariga moslashtirish;
- **Xususiyatlarni tanlash (Feature Selection):** NLP vazifasi maqsadlariga ko'proq hissa qo'shadigan tegishli N-gramlarni tanlash/ajratish;
- **Algoritmni tanlash:** N-gram modeli uchun tegishli algoritmlar yoki usullarni NLP vazifasining xususiyatlari va maqsadlaridan kelib chiqqan holda tanlash;
- **Baholash ko'rsatkichlari (Evaluation Metrics):** NLP vazifasi maqsadlariga erishishda modelning ishlashini baholash uchun aniq baholash ko'rsatkichlarini belgilash.

Xulosa

N-gramlar til modellarining qurilish bloklari bo'lib, so'zlar ketma-ketligini bashorat qilishga yordam beradi. Ular qo‘shni so‘z yoki belgilar guruhlarini tahlil qilish orqali kontekstni qamrab oladi. Yuqori tartibli n-gramlar ko‘proq kontekstni ta’minlaydi, lekin ko‘proq ma’lumot talab qiladi.

N-gram til modeli rivojlanishda davom etar ekan, rivojlanayotgan tendentsiyalar murakkab til tuzilmalarini boshqarishga qodir bo'lgan yanada murakkab modellarni ishlab chiqishni taqozo qiladi. Turli NLP vazifalarini ishlab chiqish, ularning aniqlik va samaradorlikni oshirish uchun ularni N-gram modeli bilan integratsiyalash orqali yuqori samaradorlikka ega NLP ilovalarini yaratish mumkin. Aniq o'lchamdagi N-gramlardan tashqari kontekstual ma'lumotlarni o'z ichiga olgan holda til modellarini ishlab chiqish orqali tabiiy tildagi qoliplarni yanada aniqroq tushunish imkonini beradi. Ushbu maqolada keltirilgan fikr-mulohaza va yondashuvlar asosida tabiiy tillarni qayta ishlash bo'yicha yangi imkoniyatlar tavsiflandi. NLPda n-gramlardan foydalanishning asosiy afzalliklari quyidagilar:

- **Tilni modellashtirish:** N-gramlar ma'lum bir tabiiy tildagi so'zlarni tarqalish ehtimolini aniqlashga yordam berib, nutqni aniqlash, mashina tarjimasi va avtoto'ldirish kabi NLP ilovalari samaradorligini yaxshilaydi.
- **Matnni bashorat qilishni yaxshilash:** Til korpusida eng ko`p uchraydigan n-gramlar asosida keyingi so'zni ketma-ketlikda bashorat qilishga yordam beradi. Bu yangi matn yaratish va avtoto'ldirish NLP ilovalarida foydalidir.
- **Axborot qidirish:** N-gramlar ma'lumot qidirish vazifalarida tegishli natijalarni olish uchun hujjatlarni moslashtirish va tartiblashda yordam beradi.

Matn konteksti va semantikasini aniqlash: N-gramlar so'zlar ketma-ketligidagi kontekst va ma'nosini aniqlashga yordam beradi, bu esa tabiiy tilni tushunishni osonlashtiradi.

Foydalanilgan adabiyotlar

1. Pauls, A., & Klein, D. (2011, June). Faster and smaller n-gram language models. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 258-267).
2. Takahashi, S., & Morimoto, T. (2012, November). N-gram language model based on multi-word expressions in web documents for speech recognition and closed-captioning. In *2012 International Conference on Asian Language Processing* (pp. 225-228). IEEE.

3. Chakraborty, R., Deka, M., & Sa\rma, S. K. (2024). Syntactic Category based Assamese Question Pattern Extraction using N-grams. *Procedia Computer Science*, *235*, 214-230.
4. B. Elov, A. Abdullayev, A., N.Xudoyberganov. (2024). O'zbek tili korpusi matnlari asosida til modellarini yaratish. *Contemporary technologies of computational linguistics*, 2(22.04), 344-353.
5. Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (Vol. 161175, p. 14).
6. Náther, P. (2005). N-gram based Text Categorization. *Comenius University, Bratislava, Slovakia*.
7. Ahmed, B., Cha, S. H., & Tappert, C. (2004). Language identification from text using n-gram based cumulative frequency addition. *Proceedings of Student/Faculty Research Day, CSIS, Pace University*, *12*(1).
8. Zens, R., & Ney, H. (2006, June). N-gram posterior probabilities for statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation* (pp. 72-77).
9. Majumder, P., Mitra, M., & Chaudhuri, B. B. (2002, November). N-gram: a language independent approach to IR and NLP. In *International conference on universal knowledge and language* (Vol. 2).
10. Dey, A., Jenamani, M., & Thakkar, J. J. (2018). Senti-N-Gram: An n-gram lexicon for sentiment analysis. *Expert Systems with Applications*, *103*, 92-105.