TIL KORPUSI ASOSIDA UNIGRAM, BIGRAM, VA TRIGRAM MODELLARINI ISHLAB CHIQISH

Ahmedova Maftuna Qahramon qizi
Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti
Kompyuter lingvistikasi mutaxasisligi magistranti
E-mail: maftuna_ahmedova97@mail.ru

Annotatsiya. Ushbu maqolada tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sohasida keng qo‘llaniladigan unigram, bigram va trigram modellarining nazariy va amaliy jihatlari tahlil qilinadi. Til korpusi asosida tuzilgan bu statistik modellar so‘zlar ketma-ketligini aniqlash, matnni tahlil qilish muhim rol o‘ynaydi. N-gram modellarining ehtimollik asosidagi ishlash mexanizmi, bayon etiladi. Har bir modelning afzalliklari va cheklovlari ko‘rib chiqilib, real til korpusi asosida olingan natijalar orqali ularning samaradorligi baholanadi.

Kalit so‘zlar: N-gram model, unigram, bigram, trigram, til korpusi, tabiiy tilni qayta ishlash (NLP), statistik til modeli, smoothing usullari, ehtimollik modeli.

Annotation. This article analyzes the theoretical and practical aspects of the unigram, bigram, and trigram models, which are widely used in the field of Natural Language Processing (NLP). These statistical models, constructed on the basis of a language corpus, play an important role in identifying word sequences and analyzing text. The probability-based working mechanism of N-gram models is described. The advantages and limitations of each model are examined, and their effectiveness is evaluated through results obtained from a real language corpus.

Keywords: N-gram model, unigram, bigram, trigram, language corpus, Natural Language Processing (NLP), statistical language model, smoothing techniques, probabilistic model.

Аннотация. В данной статье анализируются теоретические и практические аспекты моделей униграмм, биграмм и триграмм, которые широко применяются в области обработки естественного языка (Natural Language Processing – NLP). Эти статистические модели, построенные на основе языкового корпуса, играют важную роль в определении последовательностей слов и анализе текста. Описывается вероятностный механизм работы N-граммных моделей. Рассматриваются преимущества и ограничения каждой модели, а также оценивается их эффективность на основе результатов, полученных из реального языкового корпуса.

Ключевые слова: N-граммная модель, униграмма, биграмма, триграмма, языковой корпус, обработка естественного языка (NLP), статистическая языковая модель, методы сглаживания, вероятностная модель.

KIRISH

ISSN 3030-3907

So‘nggi yillarda sun’iy intellekt va axborot texnologiyalarining jadal rivojlanishi tabiiy tilni qayta ishlash sohasida yangi imkoniyatlar yaratmoqda. Ushbu sohada til modellari matnlar ichidagi bog‘liqliklarni tushunish, so‘zlarning ehtimoliy ketma-ketligini aniqlash hamda sun’iy intellekt tizimlariga tilni “o‘rganish” imkonini berishda asosiy rol o‘ynaydi. Matnni avtomatik tarzda tahlil qilish, tozalash va strukturaviy ko‘rinishga keltirish ushbu yo‘nalishning asosiy tadqiqot sohalaridan biri hisoblanadi.

Matndandan so‘z ehtimolini aniqlash uchun zarur bo‘lgan tabiiy tilning mavhum tushunchasi bir qator vazifalarni bajarish uchun ishlatilishi mumkin. Yaxshi til modeli orqali matnlarni ekstraktiv yoki mavhum umumlashtirishni amalga oshirish mumkin. Agar bizda turli tillar uchun modellar mavjud bo‘lsa, mashina tarjimasi tizimini osongina ishlab chiqish mumkin. Shuningdek, til modellari savol-javob tizimlarining muhim komponenti hisoblanadi [1].

Ayniqsa, kam resursli tillar, jumladan o‘zbek tili uchun bu boradagi dasturiy yechimlar va nazariy modellar hozircha yetarli darajada shakllanmagan.O‘zbek tilidagi matnlarni avtomatik tozalash va ularni semantik jihatdan tahlil qilishda an’anaviy yondashuvlar samarali natijalar bermayapdi. Shuning uchun ham statistik til modellari, xususan N-gram modellariga asoslangan yondashuvlar orqali matn ichidagi strukturaviy bog‘liqliklarni aniqlash dolzarb ahamiyat kasb etadi.

Statistik til modelida ehtimollik til korpusidagi barcha gaplar bo'yicha taqsimlanadi. Statistik til modellari mavjud matnlar asosida so'zlarning paydo bo'lish ehtimolini o'rganadi. Oddiyroq statistik til modellarida so'zlarning qisqa ketma-ketligi kontekstini ko'rib chiqishi mumkin bo'lsa, kattaroq statistik til modellari gaplar yoki paragraflar darajasida ishlashi mumkin. [2].

N-gram modellari – ya’ni ketma-ket kelgan n ta so‘z yoki belgilar asosida qurilgan statistik modellarning qo‘llanilishi ushbu jarayonda muhim ahamiyatga ega. Bu modellar matndagi ketma-ket so‘zlar orasidagi ehtimoliy bog‘liqliklarni aniqlovchi statistik yondashuvlar hisoblanib, ularning soddaligi va ishlash tezligi bu modelni NLP da keng qo‘llanilishini ta’minlaydi.

Unigram modeli – bu eng oddiy n-gram modeli bo‘lib, har bir so‘z mustaqil holda paydo bo‘ladi, ya’ni ehtimollik faqat so‘zning o‘ziga bog‘liq bo‘ladi:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i)$$

Unigram modellar asosan matnni ehtimollik asosida ajratish, matnni indekslash yoki ehtimolliklar bo‘yicha tartiblashda ishlatiladi. [3]

Bigram modeli esa har bir so‘zdan oldingi bitta so‘zga asoslanib, navbatdagi so‘z ehtimolini aniqlaydi:

$$P(w_i \mid w_{i-1})$$

ISSN 3030-3907

Bu model real hayotdagi matnlar uchun yanada aniqroq natijalar beradi. Masalan, “ilmiy maqola” so‘z birikmasining ehtimoli:

$$P(\text{ilmiy, maqola}) = P(\text{ilmiy}) \cdot P(\text{maqola} \mid \text{ilmiy})$$

Ko‘rinishida hisoblanadi ya’ni birinchi so‘zning mustaqil ehtimoli va undan keyingi so‘zning unga nisbatan shartli ehtimoli ko‘paytiriladi.

O‘zbek tilining sintaktik xususiyatlari – ergash gaplarning ko‘pligi, so‘z tartibining dapdagi erkinligi kabi omillar – modelning yanada chuqurroq ishlashini taqozo etadi. N-gram modellari esa ushbu murakkabliklarni hisobga olib, gap tarkibini mantiqan tahlil qilish va imloviy xatolarni kontekst asosida to‘g‘ri aniqlash imkonini beradi. Bu esa imlo tuzatishda nafaqat leksik darajadagi, balki sintaktik va semantik darajadagi aniqlikni ham oshirishga yordam beradi.

Trigram modeli esa har bir so‘zdan oldingi ikkita so‘zga qarab ehtimollikni baholaydi:

$$P(w_i \mid w_{i-2}, w_{i-1})$$

Masalan, “mazkur ilmiy maqola” uchligi uchun:

$$P(\text{mazkur, ilmiy, maqola}) = P(\text{mazkur}) \cdot P(\text{ilmiy} \mid \text{mazkur}) \cdot P(\text{maqola} \mid \text{mazkur, ilmiy})$$

Bu yondashuv til modeli uchun ko‘proq kontekst beradi, lekin korpusda kam uchraydigan kombinatsiyalar uchun ma’lumot siyrakligi (data sparsity) muammosi kuchayadi.

N-gram modellar, ayniqsa, resurslari cheklangan tillarda (masalan, o‘zbek tilida) soddaligi va ishlash tezligi bilan ajralib turadi. Garchi zamonaviy chuqur o‘rganishga asoslangan transformer modellar (BERT, GPT) katta imkoniyatlarga ega bo‘lsa-da, n-gramlar hali ham kichik korpuslarda samarali natijalar beradi [4].

Ushbu modellarning afzalliklari – tez ishlash, resurs talab qilmasligi va o‘zbek tili kabi morfologik boy tillarda moslashuvchanligi tufayli hali ham dolzarb hisoblanadi.

Ushbu statistika asosidagi yondashuvlar tilni aniqlash, imlo tekshiruvi, matn klassifikatsiyasi, avtomatik tarjima va ovozni matnga aylantirish tizimlarida qo‘llaniladi. Ammo n-gram modellarning asosiy cheklovi – smoothing texnikalariga ehtiyoj, chunki ba'zi kombinatsiyalar korpusda mavjud bo‘lmasligi va yangi kombinatsiyalar uchun nol ehtimollik kabi muammolar mavjud bo‘lib, ular turli smoothing texnikalari orqali yumshatiladi. [5]

ASOSIY QISM

Zamonaviy tabiiy tilni qayta ishlash sohasida statistik modellar til birliklarini matematik modellashtirishda muhim vosita hisoblanadi. Ayniqsa, unigram, bigram va trigram modellariga asoslangan yondashuvlar matnlarni tahlil

ISSN 3030-3907

qilish, tilni avtomatik aniqlash, avtomatik tarjima, avtomatik matn generatsiyasi va boshqa ko‘plab amaliyotlarda keng qo‘llaniladi.[6]

Statistik til modellarining negizida **so‘zlar ketma-ketligini ehtimollik asosida ifodalash** yotadi. Bu modelga ko‘ra, berilgan biror matnning yuzaga kelish ehtimoli undagi so‘zlarning yuzaga kelish ehtimollarining ko‘paytmasi sifatida baholanadi. Til modellarining eng sodda shakli – **unigram model** bo‘lib, unda har bir so‘z mustaqil deb qaraladi. Masalan, matnning ehtimoli quyidagi formula orqali aniqlanadi:

$$P(w_1, w_2, ..., w_n) \approx \Pi\, P(w_i)$$

Bunda, har bir w_i – bu mustaqil tarzda tanlangan so‘zdir.

Unigram model juda sodda bo‘lishiga qaramasdan, real tilning o‘ziga xosliklarini to‘liq aks ettira olmaydi. Chunki til birliklari orasidagi **kontekstual bog‘liqlik** hisobga olinmaydi. Shu sababli undan keyingi yondashuv – **bigram model** keng qo‘llaniladi. Bu modelda har bir so‘z o‘zidan oldingi so‘zga bog‘liq deb qaraladi:

$$P(w_1, w_2, ..., w_n) \approx \Pi\, P(w_i \mid w_{i-1})$$

Bigram modeli tabiiy tilni qayta ishlash jarayonida so‘zlar orasidagi lokal bog‘liqlikni hisobga olish imkonini beradi. Ushbu model har bir so‘zning ehtimolini bevosita undan oldingi so‘zga nisbatan baholaydi. Masalan, “men ketdim” degan gapda “ketdim” so‘zining “men” so‘zidan keyin kelish ehtimoli yuqoriroq bo‘ladi. Bunday yondashuv matnning mantiqiy tuzilishini va so‘zlar ketma-ketligidagi tabiiy qonuniyatlarni modellashtirish imkonini yaratadi.

Bigram modelining samarali ishlashi va aniqligini ta’minlash uchun katta hajmdagi til korpusi zarur bo‘ladi. Chunki ehtimollarni ishonchli baholash va so‘zlar o‘rtasidagi haqiqiy statistik bog‘lanishlarni aks ettirish uchun real matnlar asosida shakllangan keng qamrovli ma’lumotlar bazasi talab etiladi. Til korpusida turli so‘z birikmalarining yetarli darajada namoyon bo‘lishi modelga kam uchraydigan kombinatsiyalar uchun ham aniq ehtimollarni hisoblash imkonini beradi va ma’lumotlar siyrakligi (data sparsity) muammosini yumshatadi. Shuningdek, bunday modellardan foydalanish imloviy xatolarni aniqlash va tuzatishda ham yuqori natijalar beradi, chunki ular kontekstga mos keladigan so‘zlar ketma-ketligini aniqlashda samarali vosita bo‘lib xizmat qiladi.

Yana-da kuchli model – bu **trigram model** bo‘lib, u har bir so‘zni o‘zidan oldingi ikkita so‘z bilan bog‘laydi:

$$P(w_1, w_2, ..., w_n) \approx \Pi\, P(w_i \mid w_{i-2}, w_{i-1})$$

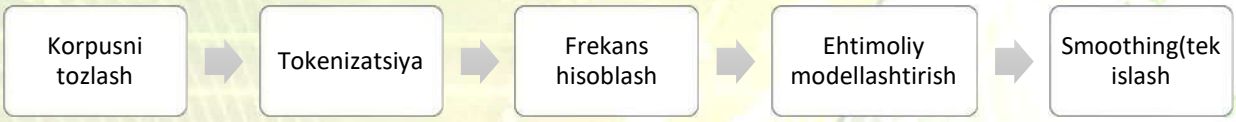
Bu model yordamida tilning kontekstual xususiyatlari ancha chuqurroq tahlil qilinadi. Trigram model, ayniqsa, **murakkab sintaktik tuzilmalarni aniqlash**da foydali hisoblanadi.

ISSN 3030-3907

Ushbu modellarni qurishda til korpusi asosiy rol o‘ynaydi. Til korpusi – bu tizimli tarzda to‘plangan, belgilangan, tozalangan va tahlilga tayyorlangan matnlar majmuasidir. O‘zbek tilida bu borada muayyan yutuqlarga erishilmoqda.

Unigram model asosida o‘zbek tilidagi eng ko‘p ishlatiladigan so‘zlar: “va”, “bu”, “bir”, “uchun”, “ham” kabi funksional birliklar bo‘lishi mumkin. Bigram model esa “bu yerda”, “shu sababli”, “ular bilan” kabi bog‘liqliklarni aniqlaydi. Trigram esa “o‘z navbatida bu”, “u yerda ular” kabi murakkab bog‘lanmalarni aniqlash imkonini beradi.

Ushbu modellarni yaratish jarayoni bir necha bosqichda amalga oshiriladi:



1. **Korpus tozalash:** HTML belgilar, maxsus belgilar va kirill-lotin aralashuvlari yo‘q qilinadi.
2. **Tokenizatsiya:** matn so‘zlarga bo‘linadi.
3. **Frekans hisoblash:** Har bir unigram, bigram va trigramning chastotasi aniqlanadi.
4. **Ehtimoliy modellashtirish:** Frekanslar asosida ehtimollar hisoblanadi.
5. **Smoothing (tekislash):** Nol ehtimollarni bartaraf etish uchun Laplace yoki smoothing usullari qo‘llaniladi [7].

Unigram modelning afzalligi – soddalik va tezlikdir. Biroq, unigram modeli kontekstni inobatga olmaydi. Bigram va trigram modellar kontekstni hisobga oladi, lekin bu modellarda **data sparsity** (ma’lumot yetishmasligi) muammosi yuzaga keladi. Ayniqsa, kichik korpuslarda ishltganda trigram model deyarli har bir kombinatsiyani topishda qiynaladi. Bu holatda, smoothing algoritmlari yordamida ehtimollarni tekislash zarur bo‘ladi.

O‘zbek tilining morfologik jihatdan boyligi, ya’ni so‘zlar turli qo‘shimchalar bilan kelishi, bu modellar uchun yana bir muammoni yuzaga keltiradi. Shuning uchun, o‘zbek tilidagi model yaratishda **lemmatizatsiya** ya’ni so‘zlarni asosiy shaklga keltirish muhim hisoblanadi.

Amaliy jihatdan bu modellar quyidagi sohalarda qo‘llaniladi:

- **matnni avtomatik tozalash va klassifikatsiya qilish**
- **chatbotlar va virtual yordamchilarni yaratish**
- **avtomatik tarjima tizimlarida**
- **nutqni matnga aylantirish (ASR) tizimlarida**

ISSN 3030-3907

Masalan, trigram model yordamida chatbotlar foydalanuvchlarining oldingi ikki gapiga qarab keyingi jumlani taxminiy generatsiya qiladi. Yoki bigram model orqali imlo xatolari aniqlanib, to‘g‘ri variant tavsiya qilinadi.

Unigram, bigram va trigram modellarining har biri o‘zbek tilini statistik jihatdan tahlil qilishda muhim rol o‘ynaydi. Ular orqali til strukturasi, matn tuzilmasi, so‘zlar orasidagi bog‘liqlik va ehtimoliy tahlillar samarali amalga oshiriladi. Biroq, ularni samarali qo‘llash uchun til korpusining sifatli, to‘liq va keng qamrovli bo‘lishi zarur.

XULOSA VA TAKLIFLAR

Ushbu tadqiqotda o‘zbek til korpusi asosida unigram, bigram va trigram modellarini ishlab chiqish, ularning nazariy asoslari, imkoniyatlari, afzalliklari va amaliy qo‘llanilish holatlari har tomonlama tahlil qilindi. Tadqiqot davomida statistik til modellarining tabiiy tilni qayta ishlash sohasidagi o‘rni va ahamiyati chuqur o‘rganildi hamda ularning ishlash mexanizmlari tafsiliy yoritildi. Xususan, N-gram modellar yordamida matndagi so‘zlar o‘rtasidagi ehtimollik bog‘lanishlari aniqlanib, turli lingvistik vazifalarda qo‘llanilganda ularning samaradorlik darajalari amaliy misollar asosida baholandi. Shuningdek, o‘zbek tilining morfologik va sintaktik xususiyatlarini inobatga olgan holda, ushbu modellar tilning kontekstual tuzilishini modellashtirishda qanday imkoniyatlarga ega ekani tahlil qilindi. Tadqiqot natijalari n-gram modellarining nafaqat nazariy, balki amaliy jihatdan ham tabiiy tilni qayta ishlash tizimlarida muhim o‘rin tutishini ko‘rsatdi.

Statistik modellarning eng sodda turi – **unigram** modeli har bir so‘zning mustaqil ehtimolligiga asoslanadi. Bu yondashuv soddaligi tufayli ba’zi sathdagi TQQ vazifalarida foydali bo‘lishi mumkin, biroq kontekstni e’tiborga olmaydi. **Bigram** modeli esa har bir so‘zning undan oldingi so‘z bilan bog‘lanishini hisobga oladi, bu esa sintaktik strukturalarni tahlil qilishda ancha aniqlik beradi. **trigram** modelida esa uchta ketma-ket kelgan so‘zlar o‘rtasidagi bog‘liqlik aniqlanadi, bu esa tilning tabiiy oqimini va semantik mosligini yuqori darajada aks ettirish imkonini beradi.

Tadqiqot natijalari nafaqat nazariy, balki amaliy jihatdan ham muhim ahamiyatga ega. Xususan, bu modellar yordamida avtomatik tarjima, matnni tozalash, matndan kontekstli ma’lumot ajratish, so‘z tavsifi (autocomplete) kabi zamonaviy sun’iy intellekt tizimlarida samarali yechimlar ishlab chiqish mumkin. Ayniqsa, O‘zbek tili uchun moslashgan til modellari mavjud emasligi inobatga olinsa, ushbu tadqiqot o‘zbek tilida ishlov beruvchi texnologiyalarni rivojlantirishda muhim qadamlardan biri hisoblanadi.

ISSN 3030-3907

Shuningdek, aniqlangan natijalar asosida N-gram modellarning cheklovlari ham ko‘rsatildi. Jumladan, bu modellar uzoq kontekstni tushunishda sust ishlaydi, so‘zlar orasidagi semantik bog‘liqlikni chuqur anglay olmaydi. Shuning uchun ham kelgusidagi tadqiqotlarda ushbu modellarni chuqur o‘rgatuvchi neyron tarmoqlar asosidagi yondashuvlar – **Word2Vec**, **FastText**, **ELMo**, **BERT** kabi modellarga integratsiyalash orqali o‘zbek tilida ishlaydigan kuchli gibrid modellarni yaratish taklif qilinmoqda.

Mazkur tadqiqot ishi o‘zbek tilida statistik til modellarini ishlab chiqish bo‘yicha dastlabki muhim qadam bo‘lib xizmat qiladi. O‘zbek tilining grammatik va sintaktik xususiyatlariga moslangan bu modellar kelajakda tilni avtomatik tahlil qilish, tilshunoslik tadqiqotlari, avtomatik tarjima tizimlari, interaktiv chatbotlar va boshqa ko‘plab sun’iy intellekt yechimlarida keng qo‘llanilishi mumkin.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

1. Elov B., Xudayberganov N., Primova M., O‘zbek tili matnlari uchun unigramtil modelini ishlab chiqish: muammo va yechimlar. – Toshkent: Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali. 2024. – 130-132 b.
2. Elov B., Alayev R., Abdullayev A., Aloyev N., Yuqori n-gram modellarini o‘zbek tili matnlariga qo‘llash. – Toshkent: Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali. 2024
3. Christopher D. Manning, Hinrich Schütze. *Foundations of Statistical Natural Language Processing*.
4. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). *A neural probabilistic language model*. Journal of Machine Learning Research, 3(Feb), 1137–1155.
5. Chen, S. F., & Goodman, J. (1998). *An Empirical Study of Smoothing Techniques for Language Modeling*. Technical Report, Harvard University.
6. Jurafsky, D., Martin, J. H. *Speech and Language Processing*. Pearson Education. 2008
7. Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*.

ISSN 3030-3907