O‘ZBEK TILI MATNLARI UCHUN UNIGRAM TIL MODELINI ISHLAB CHIQISH: MUAMMO VA YECHINLAR

Botir Elov[1], Nizomaddin Xudayberganov[2], Mastura Primova[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]Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: nizomaddin@navoiy-uni.uz
[2]Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti [3]
E-pochta: primovamastura@navoiy-uni.uz

K E Y W O R D S

Til modellari, n-gram til modeli, unigram, modelni baholash, Laplas silliqlash, mashinali o‘qitish.

A B S T R A C T

Tabiiy tilni qayta ishlashda kontekstdagi keyingi qaysi so‘zni bashorat qilish vazifasi tilni modellashtirish deb ataladi. Usbu maqolada avvalo unigram modeli, ya’ni yagona so‘zlarga asoslangan model tavsiflanadi. So‘ngra, “noma’lum” so‘zlar deb nomlanuvchi modelda o‘rgatilmagan so‘zlar ehtimolligini bashorat qilish jarayoning murakkabligi va bu muammoni bartaraf qilish uchun Laplasni silliqlash usuli keltirladi. So‘ngra ushbu silliqlash usuli turli modellarning kombinatsiyasi bo‘lgan model interpolyatsiyasi sifatida ko‘rish mumkinligini namoyish qilinadi. Shuningdek, yuqoriroq n-gramm modellari va ushbu n-gramm modellarining model interpolyatsiyasining ta’sirini ko‘rsatiladi. Ma’lumotlar to‘plami (dataset) sifatida Alisher Navoiyning “Navodir un-nihoya” devoni – o‘quv ma’lumotlar to‘plami (train data), “Badoyi-ul-bidoya” devoni va Pirimqul Qodirovning “Avlodlar dovoni” asari n-gram (n=1) modelida baholash to‘plami (evaluation text) sifatida tanlangan. Maqolada matnlar uchun o‘rtacha logarifmik ehtimollik baholash ko‘rsatkichini hisoblash usullari va algortimlari keltirilgan bo‘lib, unigram til modelini amalda qo‘llanilish ketma-ketligi tavsiflangan. Yakuniy natija shuni ko‘rsatadiki, berilgan baholash matnlarining unigram modeli mos ravishda -11.98 va -14.86 o‘rtacha logarifmik ehtimolga ega bo‘lgan. O‘quv matni va 1-kitobning o‘rtacha logarifmik ehtimoli natijalarning o‘xshashligi ularning bir xil seriyaning birinchi va ikkinchi kitoblari ekanligida hisoblanadi. Biroq, Pirimqul Qodirovning “Avlodlar_dovoni” matnining unigram taqsimoti o‘quv taqsimotidan tubdan farq qiladi. Chunki ular turli vaqtlar, janrlar va turli mualliflarning ikki xil kitobidir.

Kirish

So‘nggi o‘n yil ichida strukturlanmagan matnli ma’lumotlardan zarur tushunchalarni ajratib olishga qaratilgan ko‘plab NLP ilovalarini ishlab chiqish uchun asos sifatida til modellari shakllantirildi. Tilni modellashtirish – gapdagi so‘zning ehtimolini bashorat qilishni amalga oshirishga asoslangan tabiiy tilni qayta ishlashning asosiy vazifalaridan biri. Tilni modellashtirish amaliyotda avtoto‘ldirish (autocomplete), imloni tuzatish (spelling correction) yoki matnni generatsiya qilish (text generation) kabi ko‘plab NLP ilovalarida qo‘llaniladi. Til modeli oldingi yozuvlar asosida gapdagi keyingi so‘zni bashorat qilish uchun ishlatiladigan so‘zlar bo‘yicha ehtimollik taqsimoti vositasida mashinali o‘qitishdan foydalanadi. Til modellari katta hajmdagi matnlarni o‘rganadi va matnni yaratish, matndagi keyingi so‘zni bashorat qilish, nutqni aniqlash, belgilarni optik aniqlash va qo‘l yozuvini aniqlash uchun ishlatilishi mumkin. Amalda, til modeli ma’lum bir so‘z ketma-ketligining "haqiqiy" bo‘lish ehtimolini beradi. Ushbu kontekstdagi haqiqiylik grammatik haqiqiylikka ishora qilmaydi.

Kontekstdan 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.

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

Shuningdek, til modellari savol-javob tizimlarining muhim komponenti hisoblanadi.

Hozirgi vaqtda neyron tarmoqlariga asoslangan til modellari eng zamonaviy hisoblanadi: ular atrofdagi/qurshovdagi so‘zlar asosida gapdagi joriy so‘zni samarali bashorat qiladi. Biroq, ushbu maqolada neyron tarmoqlariga asoslangan til modellari asosi hisoblangan klassik n-gram modellari haqida fikr yuritiladi.

Maqolada avvalo unigram modeli, ya'ni yagona so‘zlarga asoslangan model tavsiflanadi. So‘ngra, "noma'lum" so‘zlar deb nomlanuvchi modelda o‘rgatilmagan so‘zlar ehtimolligini bashorat qilish jarayoning murakkabligi va bu muammoni bartaraf qilish uchun *Laplasni silliqlash usuli* keltirladi. Keyingi qadamda ushbu silliqlash usuli turli modellarning kombinatsiyasi bo‘lgan model interpolyatsiyasi sifatida ko‘rish mumkinligini namoyish qilinadi. Shuningdek, yuqoriroq n-gramm modellari va ushbu n-gramm modellarining model interpolyatsiyasining ta'sirini ko‘rsatiladi. So‘ngra, n-gramm modellarni birlashtirish uchun eng yaxshi vazn/og‘irliklarni aniqlash uchun kutish-maksimallashtirish algoritmidan foydalaniladi.

Adabiyotlar sharhi

Til modellarining evolyutsiyasi 1990-yillarda mashhurlikka erishgan statistik til modellari (Statistical Language Models, SLM)) bilan boshlandi. Bigram va trigram til modellari kabi bu modellar Markov zanjiriga asoslangan holda, kontekstni hisobga olgan holda keyingi so‘zni bashorat qilishda qo`llanilgan va belgilangan kontekst uzunligi n-gramm til modellari sifatida belgilangan. SLMlar axborotni qidirish (IR) va tabiiy tilni qayta ishlash (NLP) kabi sohalarda vazifalar samaradorligini yaxshilashda qo‘llanildi [1]. Biroq, SLMlarda ko‘pincha *"o‘lchovlilik la'nati (curse of dimensionality)"* muammosi yuzaga keladi, bu esa o‘tish ehtimolining eksponensial o‘sishi tufayli samarasizlikka olib keladi. Birinchi SLM paydo bo‘lganidan beri texnologik taraqqiyot va MLning paydo bo‘lishi, katta hajmdagi ma'lumotlar bilan birga, LMlarni sezilarli darajada oshirdi va ko‘pincha insonning ishlashidan oshib ketdi. SLMlardan so‘ng neyron til modellari (Neural Language Model, NLM) joriy etildi. Ushbu modellar so‘z ketma-ketligi ehtimolini modellashtirish uchun Recurrent Neural Networks (RNN) va Long Short Term Memory (LSTM) kabi turli xil neyron tarmoqlardan foydalangan. NLMlar n-gramm modellari duch keladigan o‘lchamlilik muammosini hal qildi.

N-gram til modeli asosida matnni tasniflash masalasini hal uchun eng mos keladi. Unigram modeli (n=1) tilning alifbosini shakllantirishda va turkumlashtirishda juda samarali natijalarni taqdim etadi. N-gram usuli orqali so`zlarni stemlash masalasini Mayfield va McNamee (2003) hal qilishgan [2]. Cavnar va Trenkle (1994) n-gram asosidagi matn tasniflash usulini taklif qilgan va amalga oshirgan [3]. Ushbu usulda tilni tasniflash uchun Term Frequency (TF)dan foydalanadigan bo`lib, bugungi kunda Cavnar usuli sifatida tanilgan. Amalga oshirish qulayligi va qayta ishlash talablari soddaligi tufayli bu usuldan hali ham tillarni tasniflashda keng foydalaniladi. Biroq, ularni amalga oshirish bugungi kundagi ulkan hujjatlar to‘plamining toifalash katta hajmdagi xotirani talab qiladi. G.Sidorov va F.Velasquez (2014) tomonidan n-gramlar (sn-gramlar) gapdagi sintaktik daraxtlardagi munosabatlar asosida aniqlangan. Sn-gramlar sintaktik bilimlardan support vector machines (SVM) va naive Bayes (NB) kabi mashinali o‘qitish usullarida foydalanish imkonini bergan [4].

I.Zitouni (2007) tomonidan yashirin n-gram hodisalari ehtimolini yaxshiroq baholash uchun n-gramm tilining ierarxik sinf modellari ishlab chiqilgan [5]. Ushbu ko‘p darajali sinf ierarxiyasi tilini modellashtirish yondashuvi taniqli backoff n-gram tilini modellashtirish usulini umumlashtiradi. U so‘z kontekstlarini aniqlash uchun sinf ierarxiyasidan foydalanadi. Tadqiqiot natijalari Wall Street Journal (WSJ) korpusida ikkita 5000 so‘z va 20 000 so‘zdan iborat lug‘atlar to‘plamidan foydalangan holda taqdim etilgan. Test to‘plamida oz sonli ko‘rinmas hodisalarni o‘z ichiga olgan 5000 so‘zli lug‘at bilan o‘tkazilgan tajribalar ierarxik sinf n-gramm tili modellaridan foydalanganda ko‘rinmas hodisalarning chalkashligini 10% gacha yaxshilashni ko‘rsatgan.

S.Aouragh va A.Yousfi (2021) tomonidan n-gram til modelining yangi bahosi ishlab chiqilgan va n-gram til modelidagi kamchiliklarini bartaraf

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

etish imkonini bergan, hamda yangi model tez va aniq natijalar berishini ta`kidlagan [6].

Matn tilini aniqlash (Identifying the language) ko`plab tabiiy tillarni qayta ishlash dasturlari uchun muhim qadamdir. Bugungi kundagi tilni identifikatsiyalash (language identification, LID) tizimlari standart ma'lumotlar to‘plamlarida o‘zaro bog‘liq bo‘lmagan tillarni ajratishda juda yaxshi ishlaydi. Biroq, LID vazifasida o‘xshash tillar yoki til turlarini farqlashda qiyinchiliklar mavjud. Shuningdek, LID usuli twitter, telegeam va youtube sharhlaridagi qisqa matnlar bilan ishlashda yaxshi natija bermaydi. A.Fanani va S.Suyanto (2021) tvitlarni Braziliya va Yevropa portugal variantlarida tasniflash uchun silliqlashtirilgan n-gramm tili modellaridan foydalanishni taklif qilishgan [7]. So‘z va belgilar n-gram til modellari besh xil tasniflagich orqali birlashtirilgan va baholangan. Eng yaxshi konfiguratsiya 6 grammli Lidstone (0.1) belgili, Gud-Tyuring so‘zining unigrami va Witten-Bell so‘zining bigram modellarini Logarifmik ehtimol nisbati baholash usuli bilan birlashtirgan gibrid modeli yordamida 92,71% aniqlikka erishilgan.

M.Costa-Jussa va J.Fonollosa (2009) statistik mashina tarjimasi (statistical machine translation, SMT)da n-gramga asoslangan qayta tartiblash (Ngram-based reordering, NbR) yondashuvidan foydalangan. Ushbu tadqiqotda statistik mezonlarni qayta tartiblash cheklovlari SMT tizimiga taqdim etilgan va bu SMT dekodlash qidiruvini kengaytirish imkonini bergan. Tarjima samaradorligini oshirish EPPS topshirig‘i (ispan va nemis tilidan ingliz tiliga) va BTEC topshirig‘i (arabchadan inglizchaga) bilan ko‘rsatilgan.

Yuqoridagi ilmiy tadqiqotlardan til modellaridan turli NLP vazifalarini hal qilishda foydalanilganligini qayd etish mumkin. O`zbek tili matnlari asosida n-gram til modelini shakllantirish maqsadida ma'lumotlar to‘plami va 2 ta baholash matni tanlandi va tahlil qilindi.

Ma’lumotlar to‘plami (dataset)

Ushbu maqoladagi **o‘quv ma’lumotlar to‘plami (train data)** – A.Navoiyning "Navodir un-nihoya" nomdagi mashhur devon. So‘ngra ushbu kitob asosida shakllantirilgan til modeli uchun ikkita baholash matnidan foydalanilgan:

- **uzb_text1** deb nomlangan birinchi baholash matni A.Navoiyning "Badoyi-ul-bidoya" devoni bo‘lib, o‘quv matniga o`xshash. Keyinchalik bu matnlardan til modelini sozlash uchun foydalaniladi.
- **uzb_text2** deb nomlangan ikkinchi baholash matni Pirimqul Qodirovning tomonidan yozilgan Boburning sadoqatli, jasur farzandi Humoyun va nabirasi haqida hikoya qiluvchi tarixiy "Avlodlar dovoni" kitobidir. Ushbu kitob o‘quv matnidan juda farq qiladigan matnga baholanganda bizning til modelimiz qanday bo‘lishini ko‘rsatish uchun tanlangan.

Unigram tili modeli

Tabiiy tilni qayta ishlashda n-gramm n ta so‘zdan iborat ketma-ketlikdir [9]. Masalan, "NLP" - unigram (n = 1), "kompyuter lingvistikasi" - bigram (n = 2), "sun`iy intellekt metodlari" - trigram (n = 3).

Modelni o‘qitish

Til modeli gapdagi so‘zning ehtimolligini baholaydi, odatda undan oldin kelgan so‘zlarga asoslanadi. Masalan, “Men yangi kitobni” so‘zlar birikmasi uchun bizning maqsadimiz shu jumladagi oldingi so‘zlar asosida gapdagi har bir so‘zning ehtimolini baholashdir:

Berilgan gap:	"**Men yangi kitobni o‘qidim** </s>"	
Bashorat qilish:	P("men")	
	P("yangi"	| "men")
	P("kitobni"	| "men yangi")
	P("o‘qidim")	| "men yangi kitobni")
	P("</s>")	| "men yangi kitobni o‘qidim")

Odatda, til modelida barcha so‘zlar kichik harflar bilan yoziladi va tinish belgilari e'tiborga olinmaydi. **</s>** belgisi gapning oxirini bildiradi.

Unigram tili modeli quyidagi taxminlarga asoslanadi [10,11]:

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

1. Har bir so‘zning ehtimoli o‘zidan oldingi so‘zlarga bog‘liq emas.
2. Har bir so‘z ehtimoli mashg‘ulot matnidagi barcha so‘zlar orasidagi uchrash qiymatiga bog‘liq. Ya`ni, modelni o‘rgatish o‘quv matnidagi barcha unigramlar uchun ushbu qiymatlarni hisoblashdan iborat.

$$P_{o`quv}(\text{"o'qidim"} \mid \text{"men yangi kitobni"}) = P_{o`quv}(\text{"o'qidim"}) = \frac{n_{o`quv}(\text{"o'qidim"})}{N_{o`quv}}$$

o'quv matnidagi so'zlarning umumiy soni

1-rasm. *O‘quv matnidan "o‘qidim" unigramsining taxminiy ehtimoli*

Modelni baholash

Barcha unigram ehtimolliklari aniqlanganidan so‘ng, matndagi har bir gapning ehtimolini hisoblash uchun ushbu ehtimollik qiymatlaridan foydalanamiz: *har bir gapning ehtimoli undagi so‘zlar ehtimolliklarining yig‘indisidan iborat* [12].

$$P_{umumiy}(\text{"Men yangi kitobni o‘qidim </s>"}) =$$

$$P_{o`quv}(\text{"men"})P_{o`quv}(\text{"yangi"}\mid\text{"men"})P_{o`quv}$$

$$(\text{"kitobni"}\mid\text{"men yangi"})P_{o`quv}$$

$$(\text{"o‘qidim"}\mid\text{"men yangi kitobni"})$$

$$P_{o`quv}(\text{"[END]"}\mid\text{"men yangi kitobni o‘qidim"}) =$$

$$P_{o`quv}(\text{"men"})P_{o`quv}(\text{"yangi"})P_{o`quv}(\text{"kitobni"})$$

$$P_{o`quv}(\text{"o‘qidim"})P_{o`quv}(\text{"</s>"})$$

Yuqorida keltirilgan qoidani *uzb_text1* yoki *uzb_text2* kabi matnlarning umumiy ehtimolini hisoblash mumkin. Matndagi har bir gapning boshqa gaplardan mustaqil, degan sodda taxmin asosida, bu ehtimolni matndagi gaplar ehtimollarining yig‘indisi sifatida aniqlanadi.

Berilgan matn:

"**Men yangi kitobni o‘qidim. U juda qiziqarli ekan**"

$$P_{baholash}(T) = = P_{umumiy}(\text{"men yangi kitobni o‘qidim </s>"})$$
$$P_{umumiy}(\text{"u juda qiziqarli ekan </s>"})$$
$$P_{baholash}(\text{"men yangi kitobni o‘qidim </s>"}) = P_{o`quv}(\text{"men"})P_{o`quv}(\text{"yangi"})P_{o`quv}(\text{"kitobni"})$$
$$P_{o`quv}(\text{"o‘qidim"})P_{o`quv}(\text{"</s>"})$$
$$P_{baholash}(\text{"u juda qiziqarli ekan </s>"}) = P_{o`quv}(\text{"u"})P_{o`quv}(\text{"juda"})P_{o`quv}(\text{"qiziqarli"})$$
$$P_{o`quv}(\text{"ekan"})P_{o`quv}(\text{"</s>"})$$
$$\rightarrow P_{baholash}(T) = P_{o`quv}(\text{"men"})P_{o`quv}(\text{"yangi"})P_{o`quv}(\text{"kitobni"})$$
$$P_{o`quv}(\text{"o‘qidim"})P_{o`quv}(\text{"</s>"})$$
$$P_{o`quv}(\text{"u"})P_{o`quv}(\text{"juda"})P_{o`quv}(\text{"qiziqarli"})$$
$$P_{o`quv}(\text{"ekan"})P_{o`quv}(\text{"</s>"})$$

Gapni tugash belgilarining roli

Yuqorida ta’kidlanganidek, til modeli nafaqat so‘zlarning, balki matndagi barcha gaplarning ham ehtimolliklarni aniqlashga xizmat qiladi. Natijada, barcha mumkin bo‘lgan gaplarning ehtimollari 1 ga to‘g‘ri kelishini ta’minlash uchun har bir gapning oxiriga **</s>** belgisini qo‘shishimiz va uning ehtimolini xuddi haqiqiy so‘z kabi baholashimiz kerak.

Baholash ko‘rsatkichi: o‘rtacha logarifmik ehtimollik (average log likelihood)

Matnining ehtimolligi baholash uchun yuqoridagi tenglamaning har ikki tomonidagi logarifini oladigan bo‘lsak, **matnning log(arifmik) ehtimoli** har bir so‘z uchun log ehtimolliklarining yig‘indisiga teng bo‘ladi [13]. Nihoyat, o‘lchovimiz matndagi so‘zlar soniga bog‘liq bo‘lmasligini ta'minlash uchun ushbu log ehtimolini baholash matnidagi so‘zlar soniga bo‘lamiz.

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

$$P_{baholash}(T) = \prod_{so`z} P_{o`quv}(\text{so'z})$$

$$\log\left(P_{baholash}(T)\right) = \prod_{so`z} \log\left(P_{o`quv}(\text{so'z})\right)$$

$$O'rtacha\ logarifmik\ ehtimollik_{baholash} = \frac{\prod_{so`z} \log\left(P_{o`quv}(\text{so'z})\right)}{N_{baholash}}$$

↑
baholash matnidagi so'zlarning umumiy soni

Odatda, n-gram modellar uchun 2 asosli logarifm funksiyasidan foydalaniladi.

Yuqoridagi hisoblashlar natijasida o'rtacha logarifmik ehtimollik ko'rsatkichiga erishamiz, bu esa baholash matnidagi har bir so'zning o'qitilgan logarifmik ehtimolliklarining o'rtacha ko'rsatkichidir. Ya`ni, bizning til modelimiz qanchalik yaxshi bo'lsa, uning baholash matnidagi har bir so'zga tayinlagan ehtimoli yuqori bo'ladi.

Til modellari uchun umumiy baholash ko'rsatkichlari sifatida **kross-entropiya (cross-entropy)** va **chalkashlik (perplexity)ni** keltirish mumkin [13,14]. Biroq, ushbu baholash ko'rsatkichlari ham lografimlarga asoslanadi: kross-entropiya o'rtacha logarifmik ehtimolining manfiy ko'rsatkichidir, chalkashlik esa kross-entropiyaning eksponentidir.

Noma'lum unigramlar

Yuqoridagi keltirilgan unigram modelida katta muammo bor: o'quv matnida mavjud bo'lmagan, biroq baholash matnida mavjud unigram uchun uning o'quv matnidagi soni 0 ga tengligi sababli, ehtimoli ham 0 teng bo'ladi. Bunday unigram modelidan foydalanib bo'lmaydi: bu nol ehtimollik logarifmning qiymati manfiy cheksizga teng bo'lib, butun model uchun manfiy cheksiz o'rtacha logarifmik ehtimoliga olib keladi!

$$n_{o`quv}(unigram) = 0$$

$$\rightarrow P_{o`quv}(unigram) = \frac{0}{N_{o`quv}} = 0$$

$$\rightarrow log\left(P_{o`quv}(unigram)\right) = log(0) = -\infty$$

Laplas silliqlash

Til modelidagi nomalum unigramlar muammosini hal qilish uchun **Laplas silliqlash** deb ataladigan yondashuvdan foydalanish mumkin [15]. Laplas silliqlash usuli quyidagi qadamlardan iborat:

1. **O'quv matnidagi unikal unigramlar ro'yxatiga [UNK] deb nomlangan sun'iy unigram qo'shiladi.** Bu unigram til modelini baholash vaqtida duch kelishi mumkin bo'lgan barcha noma'lum belgilarni ifodalaydi. Albatta, ushbu unigramning soni o'quv to'plamida nolga teng bo'ladi va unigram lug'at hajmi (unikal

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

unigramlar soni) bu yangi unigram qo‘shilgandan keyin 1 ga ortadi.

2. **Lug‘atdagi barcha unigramlarga k qiymatning psevdo-soni qo‘shiladi.** k qiymatning eng keng tarqalgan qiymati 1 dir va bu usul "add-one smoothing" kabi nomlanadi.

$$P_{o`quv}(unigram) = \frac{n_{o`quv}(unigram) + k}{N_{o`quv} + kV}$$

unigram lug`ati hajmi
(o`quv matnidagi unikal unigramlar soni)

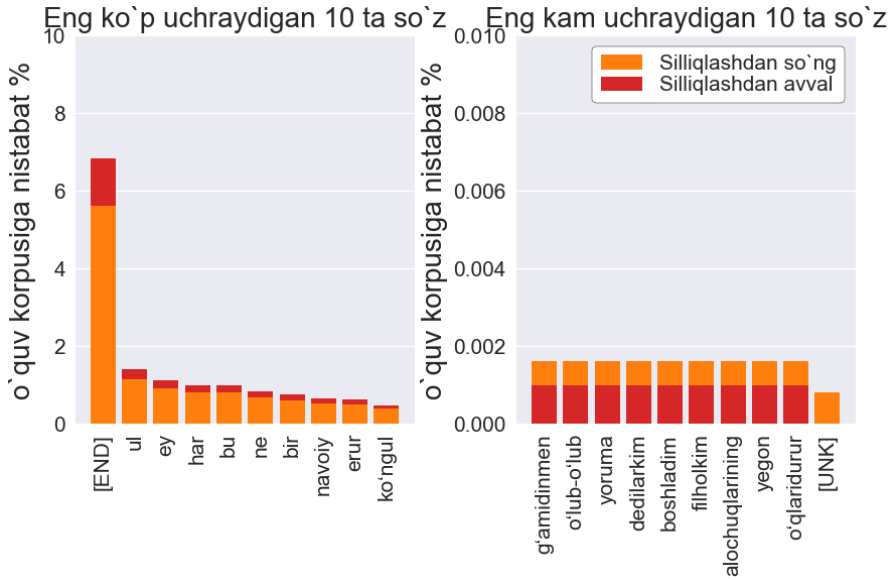
Natijada, har bir unigram uchun ehtimollik formulasining numeratori unigram va Laplas silliqlashdan olingan k psevdo-qiymat asosida aniqlanadi. Shuningdek, kasr mahrajida o‘quv matnidagi so‘zlarning umumiy soni va unigram lug‘at hajmining k qiymat bilan ko‘paytmasiga teng bo‘ladi. Chunki, bizning lug‘atimizdagi har bir unigramning soniga k qo‘shiladi, bu esa o‘quv matnidagi jami unigram soniga (k × lug‘at hajmi) qo‘shiladi.

Laplas silliqlashning til modeliga ta'siri

Har bir unigramga qo‘shimcha k psevdo-qiymat qo‘shilishi tufayli unigram modeli har safar baholash matnida noma'lum so‘zga duch kelganida, u joriy so‘zni [UNK] unigramga aylantiradi. Lug‘atdagi oxirgi unigram qiymati o‘quv matnida nolga teng, ammo k psevdo-qiymat tufayli endilikda manfiy bo‘lmagan ehtimolga ega:

$$P_{o`quv}([UNK]) = \frac{k}{N_{o`quv} + kV}$$

Shuningdek, Laplas silliqlash natijasida ba'zi ehtimollar oddiy tokenlardan unikal tokenlarga o‘tkaziladi. Ikkita unigramning soni 2 va 1 bo‘lib, ular bir Laslash sillishdan so‘ng mos ravishda 3 va 2 ga aylanadi. Ilgari keng tarqalgan unigramning ehtimoli kamroq tarqalgan unigramning ehtimoli ikki baravar ko‘p edi, ammo hozir boshqasiga nisbatan atigi 1,5 baravar ko‘p.



2-rasm. *Matnga Laslas sillqilashni qo‘llanilishi*

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

Yuqoridagi 2-rasmda matndagi 10 ta eng ko‘p va eng kam uchraydigan so‘zlarning foizlari keltirilgan. Buni o‘quv matnidagi 10 ta eng keng tarqalgan va 10 ta eng kam tarqalgan unigramning taxminiy ehtimollaridan ko‘rish mumkin. Bitta qo‘shimcha Laslas silliqlash amalga oshirilganidan so‘ng, birinchisi o‘z ehtimollarining bir qismini yo‘qotadi, ikkinchisining ehtimoli esa ularga nisbatan sezilarli darajada oshadi. Xulosa sifatida Laslas silliqlash natijasida unigramlarning ehtimollik taqsimotini tenglashtiradi, shuning uchun usul nomida "silliqlash" termini qo‘shilgan.

Unigram modelini qo‘llanilishi

Ma'lumotlarni boshlang‘ich qayta ishlash

Unigram modelini matnlarga qo‘llashdan oldin, matnlarni tokenlash orqali alohida so‘zlarga ajratish kerak.

```
def tokenize_text(text_path: str, token_text_path: str)
```

Matn faylidagi har bir satr paragraph/abzatsni ifodalaydi. Har bir satrni ketma-ket o‘qib olinib, kichik harflarga o‘tkazib tokenlash metodiga yuboriladi:

```
def generate_tokenized_sentences(paragraph: str)
```

Tokenlash metodi ichida paragraf gaplarga ajratiladi. Bu jarayon ancha murakkab hisoblanib, o‘zbek tili matnlarini tokenlashdagi muammolar va ularning yechimlari B.Elovning “O‘zbek tili korpusi matnlari asosida til modellarini yaratish”1 nomnli maqolasida keltirilgan. Bu almashtirishlar replace_characters metodi tomonidan oldindan amalga oshiriladi.

```
def replace_characters(text: str)
```

Keyingi qadamda har bir gap tokenlarga ajratiladi vahar bir tokenlashtirilgan gapning oxiriga maxsus </s> belgisi qo‘shiladi:

```
tokenized_sentence.append('</s>')
```

Nihoyat, har bir tokenlashtirilgan gap matnli fayliga yoziladi. Ushbu tokenlashtirilgan matn fayli keyinchalik til modellarini o‘qitish va baholash uchun ishlatiladi.

```
tokenize_raw_text('../data/train1_uz_raw.txt', '../data/train1_uz_tokenized.txt')

tokenize_raw_text('../data/text1_uz_raw.txt', '../data/text1_uz_raw_tokenized.txt')

tokenize_raw_text('../data/text2_uz_raw.txt', '../data/text2_uz_raw_tokenized.txt')
```

Tokenlashdan oldin o‘quv matni (train.txt) quyidagi ko‘rnishga ega:

Nafsning kelajagi — zulmdir, zulmning kelajagi esa — xorlikdir.

Biz ana shu oddiy haqiqatni bayon qilmoqchimiz.

Badan pokligi badanga, ruh pokligi ruhga hayot bag‘ishlar.

"Ana o‘shalar hidoyatga zalolatni sotib olgandirlar.

Va tijoratlari foyda keltirmadi hamda hidoyat topganlardan bo‘lmadilar.

Tokenlash jarayonidan keyin (train_tokenized.txt), undagi har bir satr tokenlardan iborat:

nafsning,kelajagi,zulmdir,zulmning,kelajagi,esa,xorlikdir,[END]

biz,ana,shu,oddiy,haqiqatni,bayon,qilmoqchimiz,[END]

badan,pokligi,badanga,ruh,pokligi,ruhga,hayot,bag‘ishlar,[END]

ana,o‘shalar,hidoyatga,zalolatni,sotib,olgandirlar,[END]

va,tijoratlari,foyda,keltirmadi,hamda,hidoyat,topganlardan,bo‘lmadilar,[END]

[1] B. Elov, A. Abdullayev, N. Xudoyberganov (2024). O‘ZBEK TILI KORPUSI MATNLARI ASOSIDA TIL MODELLARINI YARATISH. *«CONTEMPORARY TECHNOLOGIES OF COMPUTATIONAL LINGUISTICS»*, *2*(22.04), 344-353.

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

Modelni o‘qitish

Unigram ehtimollik formulalari juda soda bo‘lsa-da, ammo modelni tez ishlashini ta'minlash uchun quyidagi ishlar amalga oshirildi:

–Birinchidan, **UnigramCounter** klassi har bir tokenlashtirilgan o‘quv matn faylidan bir vaqtning o‘zida bitta satr/gapni o‘qiydi. Gapdagi har bir unigram sinfning **counts** atributi uning uchrash sonini oshirish maqsadida ishlatiladi. Bu har bir unigramni o‘quv matnidagi soniga moslashtiradigan qiymatdir. Nihoyat, bu sinf o‘zining **token_count** atributi orqali har bir matndagi so‘zlarning umumiy sonini ham saqlaydi. Bu sinfdan quyidagicha foydalanish mumkin:

```
train_counter = UnigramCounter('../data/train1_uz_tokenized.txt')
```

–Ikkinchidan, **Unigramodel** klassi o‘quv matni bilan bog‘langan **UnigramCounter** ob'ektini qabul qiladi va har bir unigramning mos keladigan ehtimolliklarini psevdo-qiymatli **k** bilan Laplas silliqlash formulasi asosida hisoblaydi. Hisoblangan ehtimolliklar sinfning **probs** atributida saqlanadi. Ushbu hisob-kitoblarning barchasini quyidagi tarzda qo‘llaniladigan modelning **train** usulida aniqlash mumkin (k=1):

```
train_model=Unigramodel(train_counter)
train_model.train(k=1)
```

Modelni baholash

Barcha unigram ehtimolliklar hisoblab chiqilganidan so‘ng, har bir matn uchun o‘rtacha logarifmik ehtimolini hisoblash uchun uni baholash matnlariga qo‘llash mumkin:

–Birinchidan, har bir baholash matnidagi unigramlar sonini hisoblash uchun **UnigramCounter** sinfidan foydalaniladi.

```
uzb_text1_counter = UnigramCounter('../data/text1_uz_raw_tokenized.txt')
uzb_text2_counter = UnigramCounter('../data/text2_uz_raw_tokenized.txt')
```

–O‘qitilgan **Unigramodel** dagi **evaluate** metodi baholash matni uchun o‘rtacha logarifmik ehtimolini hisoblash uchun ishlatiladi. U baholash matni uchun **UnigramCounter** qabul qiladi. Ushbu hisoblagichdagi har bir unigram uchun uning baholash matnidagi soni o‘quv matnidagi ehtimollik qiymatiga ko‘paytiriladi (ilgari **probs** atributida saqlangan). Agar unigram o‘quv lug‘atida mavjud bo‘lmasa, uni yuqorida aytib o‘tilganidek [UNK] unigramsiga aylantiriladi.

–Har bir unigram uchun yuqoridagi ko‘paytma qiymati baholash matnining log ehtimoliga qo‘shiladi va ushbu qadam matndagi barcha unigramlar uchun takrorlanadi. Oxirgi qadamda, matnning o‘rtacha logarifmik ehtimolini aniqlash uchun ushbu logarifmik ehtimolini baholash matnidagi so‘zlar soniga bo‘linadi.

uzb_text1 va *uzb_text2* matnlarida unigram modeli uchun baholash bosqichi natijasi quyidagicha:

```
uzb_text1_avg_log_likelihood = train_model.evaluate(uzb_text1_counter) # -11.98

uzb_text1_avg_log_likelihood = train_model.evaluate(uzb_text2_counter) # -14.86
```

Yakuniy natija shuni ko‘rsatadiki, *uzb_text1* unigram modeli -11.98 va *uzb_text2* unigram modeli 14.86 o‘rtacha logarifmik ehtimolga ega.

Natijalar tahlili

Unigram taqsimotining o‘rtacha logarifmik ehtimoli va o‘xshashligi

Yuqoridagi natijadan shuni ko‘ramizki, *uzb_text1* matni *uzb_text2* matniga nisbatan yuqori logarifmik qiymatga ega. Bu mantiqqa to‘g‘ri keladi, chunki bizda shunga o‘xshash matnda ushbu so‘zning ehtimolligi allaqachon mavjud bo‘lsa matndagi so‘zning ehtimolini aniq taxmin qilish osonroq bo‘ladi. Ya`ni, baholash matni uchun o‘rtacha logarifmik ehtimollik formulasini quyidagi tarzda ajratish mumkin:

$$O'rtacha\ log\ ehtimoli_{baholash} = \frac{\sum_{so`z} \log\left(P_{o`quv}(so`z)\right)}{N_{baholash}}$$

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

$$= \frac{\sum_{unigram} n_{o`quv}(unigram) \log\left(P_{o`quv}(unigram)\right)}{N_{baholash}}$$

$$= \sum_{unigram} \frac{n_{o`quv}(unigram)}{N_{baholash}} \log\left(P_{o`quv}(unigram)\right)$$

$$= \sum_{unigram} P_{baholash}(unigram) \log\left(P_{o`quv}(unigram)\right)$$

O‘rtacha logarifmik ehtimoli: baholash matnidagi har bir so‘zning o‘qitilgan logarifmik ehtimolliklarining o‘rtacha qiymatiga teng. Yuqoridagi mulohazalardan quyidagi xulosa/tamiyillarni qayd etish mumkin:

1. Baholash matnidagi har bir so‘z uchun (o‘quv matnidan hisoblangan) logarifmik ehtimolini qo‘shish o‘rniga, ularni unigram asosida qo‘shish mumkin.
2. O‘rtacha logarifmik ehtimolini maksimal darajada oshirish uchun o‘quv va baholash matnlari o‘rtasidagi unigram taqsimoti imkon qadar o‘xshash bo‘lishi kerak.

O‘quv va baholash matnlari o‘rtasidagi unigramlarning o‘xshash taqsimoti

	A	B
O’quv	0.2	0.8
Baholash	0.3	0.6

$$O'rtacha\ log\ ehtimoli_{baholash}$$
$$= P_{baholash}(A) \log\left(P_{o`quv}(A) + = P_{baholash}(B) \log\left(P_{o`quv}(B)\right.\right.$$
$$= 0.3 \log(0.2) + 0.6 \log(0.8)$$
$$= 0.3(-2.32) + 0.6(-0.32)$$
$$\approx -0.89$$

O‘quv va baholash matnlari o‘rtasidagi unigramlarning turli xil taqsimoti

	A	B
O’quv	0.8	0.2
Baholash	0.3	0.6

$$O'rtacha\ log\ ehtimoli_{baholash}$$
$$= P_{baholash}(A) \log\left(P_{o`quv}(A) + = P_{baholash}(B) \log\left(P_{o`quv}(B)\right.\right.$$
$$= 0.3 \log(0.8) + 0.6 \log(0.2)$$
$$= 0.3(-0.32) + 0.6(-2.32)$$
$$= -1.49$$

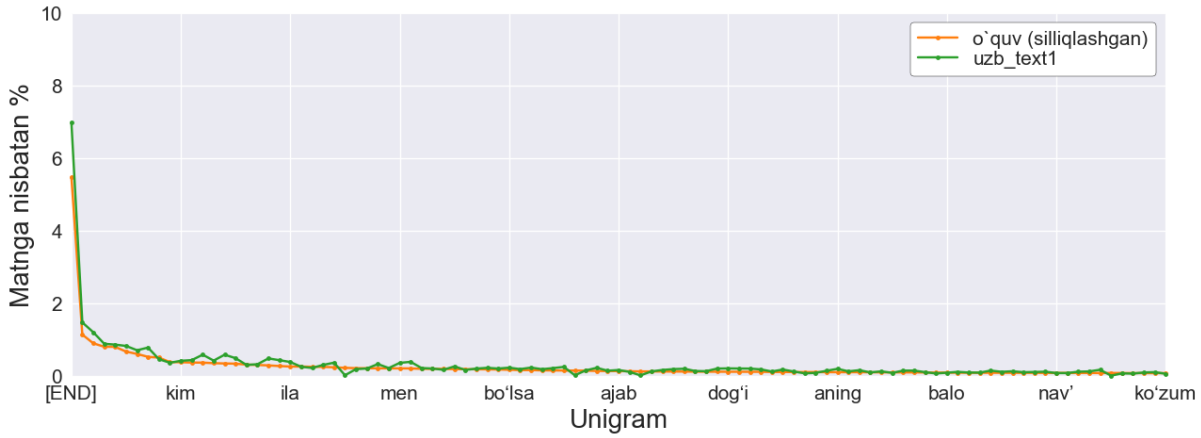
Quyidagi oddiy misolda lug‘at faqat ikkita A va B unigramdan iborat bo‘lib, yuqorida keltirilgan tamoyillarni tekshirish mumkin:

Yuqoridagi misoldan quyidagi tasdiqlarga ega bo‘lamiz:

1. O‘qitish ehtimoli yuqori bo‘lgan unigram (0.8) yuqori baholash ehtimoli (0.6) bilan birlashtirilishi kerak. O‘quv ehtimoli logarifmi ularning ko‘paytmasi (-0.19) kabi kichik manfiy qiymat bo‘ladi.
2. Shuningdek, kichik o‘qitish ehtimoli (0.2) ga teng bo‘lgan unigram past baholash ehtimoli (0.3) ga teng bo‘ladi. O‘quv ehtimoli logarfmik qiymati katta manfiy qiymat -2.32 ga teng bo‘ladi. Biroq, u 0.3 dan kichikroq baholash ehtimoli bilan moslashtiriladi va ularning manfiy ko‘paytmasi minimallashtiriladi.

Unigram taqsimotlarini solishtirish

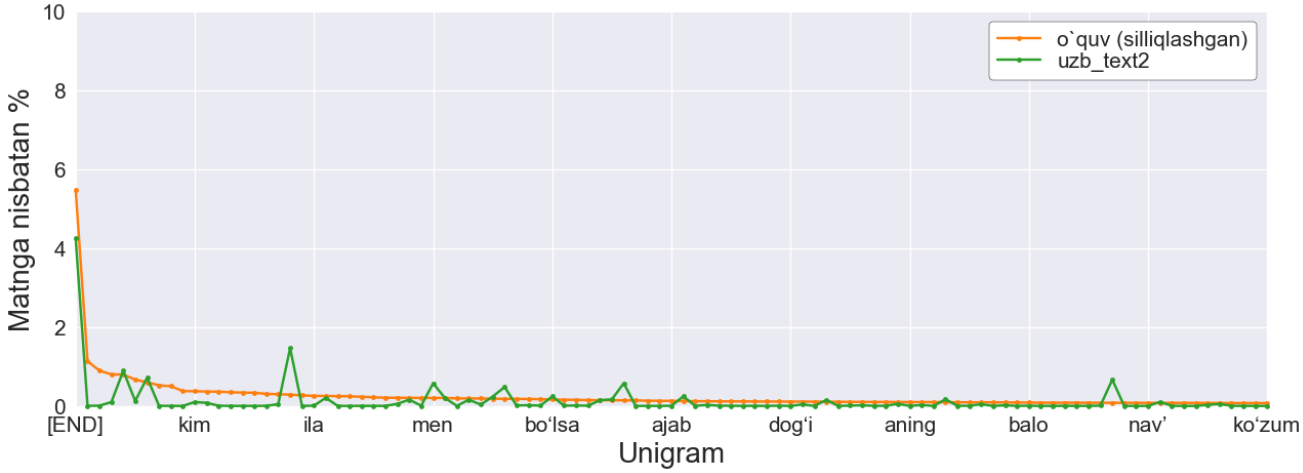
© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova



3-rasm. *O‘xshash matnlarning unigram taqsimotlarini solishtirish*

O‘quv matnining unigram taqsimoti *add-one smoothing* vositasida *uzb_text1* bilan solishtirilsa, ular unigramlarning juda o‘xshash taqsimotiga ega ekanligini qayd etish mumkin. Hech bo‘lmaganda o‘quv matnidagi 100 ta eng keng tarqalgan unigramlar uchun qiyidagi grafikni shakllantiramiz:

Natijalarning o‘xshashligi ularning bir xil seriyaning birinchi va ikkinchi kitoblari ekanligida hisoblanadi. E'tiborga molik istisno bu “ned” unigrami bo‘lib, u *uzb_text1* martnda sezilarli darajada pasayib ketgan. Chunki o‘quv asarning oxirida “ned” personaji qatl qilingan. Biroq, *uzb_text2* matnining unigram taqsimoti o‘quv taqsimotidan tubdan farq qiladi. Chunki ular turli vaqtlar, janrlar va turli mualliflarning ikki xil kitobidir.



4-rasm. *O‘xshash bo‘lmagan matnlarning unigram taqsimotlarini solishtirish*

Ushbu ikki taqsimot o‘rtasidagi sezilarli farqlar quyidagilar:

–*uzb_text2* matnda **"ila"** unigramsi kamroq ishlatilgan, **"men"** va **"bo`lsa"** kabi unigramlar esa ancha keng tarqalgan.

–o‘quv to‘plamida eng keng tarqalgan 100 ta unigramlar orasida bir nechta unigramlar mavjud, ammo *uzb_text2* da bu qiymat 0 ga teng.

Bu barcha farqlar bilan *uzb_text2* matnining o‘rtacha logarifmik ehtimolligi *uzb_text1* matniga qaraganda pastroq bo‘lishi ajablanarli emas, chunki unigram modelini o‘qitish uchun ishlatiladigan matn avvalgisiga qaraganda ikkinchisiga ko‘proq o‘xshaydi.

Unigram modelining interpolyatsiyasi

O‘quv va *uzb_text2* matnlari o‘rtasidagi unigram taqsimotidagi sezilarli farqni hisobga olsak, oddiy unigram modelini qandaydir tarzda

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

yaxshilay olamizmi? Ma'lum bo'lishicha, quyida tavsiflangan model interpolyatsiyasi usulidan foydalanib buni amalha oshirish mumkin.

Laplas silliqlash usuli vositasisa matnidagi har bir unigramga uning ehtimolini hisoblashdan oldin k psevdo-soni qo'shilgan bo'lib, Ushbu formulani quyidagicha talqin qilish mumkin:

$$P_{o`quv}(unigram) = \frac{n_{o`quv}(unigram) + k}{N_{o`quv} + kV}$$

$$= \frac{n_{o`quv}(unigram)}{N_{o`quv} + kV} + \frac{k}{N_{o`quv} + kV}$$

$$= \frac{N_{o`quv}}{N_{o`quv} + kV} \frac{n_{o`quv}(unigram)}{N_{o`quv}} + \frac{kV}{N_{o`quv} + kV} \frac{k}{kV}$$

$$= \frac{N_{o`quv}}{N_{o`quv} + kV} \frac{n_{o`quv}(unigram)}{N_{o`quv}} + \frac{kV}{N_{o`quv} + kV} \frac{1}{V}$$

silliqlanmagan unigram ehtimollik

tekis ehtimollik

Yuqoridagi qayta tartibga solingan formuladan, unigramning silliqlabgan ehtimolligi **1/V** tekis ehtimollik bilan silliqlanmagan unigram ehtimolligining og'irlikdagi yig'indisidan iborat. O'quv matnidagi barcha unigramlar, shu jumladan noma'lum [UNK] unigram uchun bir xil ehtimollik tayinlangan.

Xususan, 100769 tokendan iborat o'quv matnidagi, unigrama lug'ati 7896 va (k=1) add-one silliqlashda, Laplas silliqlash formulasi quyidagi ko'rinishga ega:

$$P_{o`quv}(unigram) = \frac{100769}{100769 + 7896} \frac{n_{o`quv}(unigram)}{N_{o`quv}} + \frac{7896}{100769 + 7896} \frac{1}{V}$$

Ya'ni, add-one silliqlashdagi unigram ehtimolligi 96.4% ni tashkil qiladi. Natijada, Laplasni silliqlash modelini interpolyatsiya qilish usuli sifatida talqin qilinishi mumkin: yakuniy ehtimollik qiymatini aniqlash uchun turli modellardagi taxminlarni bir nechta mos og'irliklar bilan birlashtirish lozim.

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

Interpolyatsiya og'irligi qiymatini o`zgartirish

Bugungi kunda unigrama modellarini 96.4%:3.6% nisbatda birlashtirish kerakligini tasdiqlovchi hech qanday qoida mavjud emas. Darhaqiqat, unigrama va uniforma modellarining turli kombinatsiyalari quyidagi jadvalda ko'rsatilganidek, turli xil **k** pseudo-hisoblarga mos keladi:

Silliqlanmagan unigram modelining interpoliyasion vazni

$$\rightarrow a_{unigram} = \frac{N_{o`quv}}{N_{o`quv} + kV} \leftrightarrow k = \frac{N_{o`quv} - N_{o`quv}\, a_{unigram}}{V\, a_{unigram}}$$

$a_{unigram}$	k
0	∞
0.2	106
0.4	40
0.6	18
0.8	7
1	0

Yuqoridagi jadvaldan shuni ko'ramiz:

–Agar k=0 bo'lsa, unigramning boshlang'ich modeli saqlanib qoladi. Bu esa interpolatsiyada og'irligi 1 ga teng bo'lgan silliqlanmagan unigram modeliga teng.

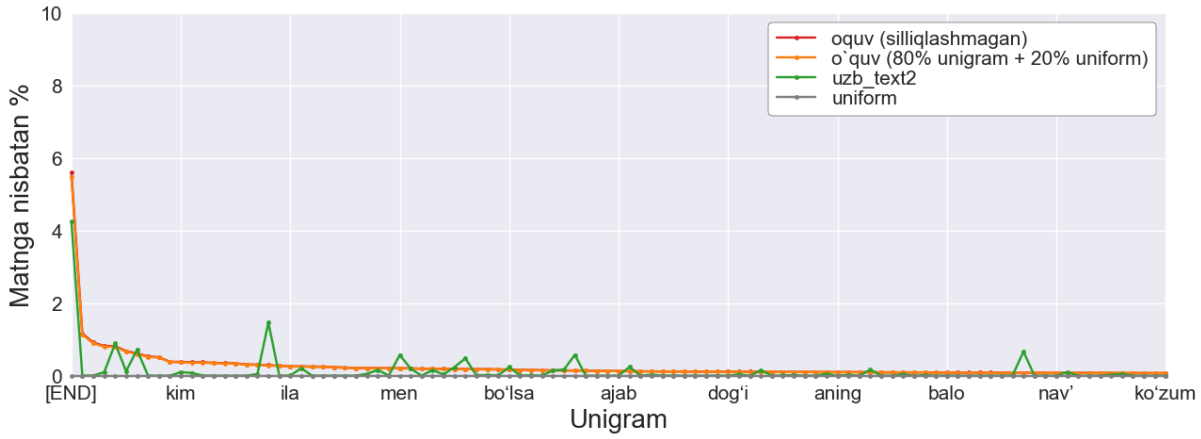
–K qiymat ortishi bilan unigram taqsimotining silliqlanishi ortadi: umumiy unigramlardan unikal unigramlarga ko'proq ehtimollar olinadi va barcha ehtimollarni silliqlanadi. Natijada, birlashtirilgan model tobora kamroq unigram taqsimotiga o'xshab qoladi va barcha unigramlarga bir xil ehtimollik berilgan yagona modelga o'xshaydi.

–Nihoyat, unigram modeli to'liq silliqlashganda, uning interpolyatsiyadagi og'irligi nolga teng bo`ladi. Bu har bir unigramga cheksiz psevdo-sonni qo'shishga tengligi sababli, ularning ehtimoli imkon qadar teng/bir xil bo'ladi.

Laplas silliqlash va **model interpolyatsiyasi** bir tanganing ikki tomoni ekanligini tushunganimizdan so'ng, unigram modeli samaradorligini oshirish/yaxshilash uchun ushbu usullarni qo'llashni ko'rib chiqamiz.

Unigram modelining qayta o`qitishni kamaytirish

Unigram modelini silliqlash, ya'ni uniform bilan ko'proq interpolyatsiya qilish natijasida, model o'quv ma'lumotlariga kamroq mos keladi. Bu jarayonni quyida keltirilagan (to'q sariq chiziq) 80-20 tekis interpolyatsiyli unigram model uchun ko'rish mumkin. Bunda silliqlanmagan unigram modelidan (qizil chiziq) tekis modelga (kulrang chiziq) qarab uzoqlasha boshlaydi.

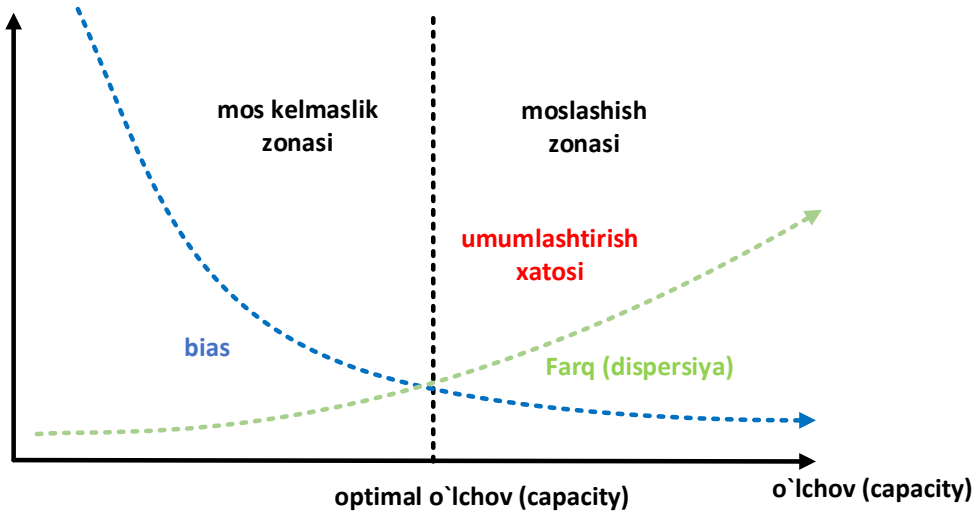


4-rasm. *Unigram modellarining qiyosiy taqqosi*

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

Biroq, bunday interpolyatsiyaning afzalligi shundaki, model o‘quv ma'lumotlariga kamroq mos keladi va yangi ma'lumotlarga yaxshiroq umumlasha oladi. Yuqoridagi grafikdan yangi *uzb_text2* model (yashil chiziq)ning unigram taqsimotini asl modelga qaraganda yaqinroq ekanligini qays etish mumkin.

Dispersiya farqini kamaytirish



5-rasm. *Haddan tashqari silliqlash*

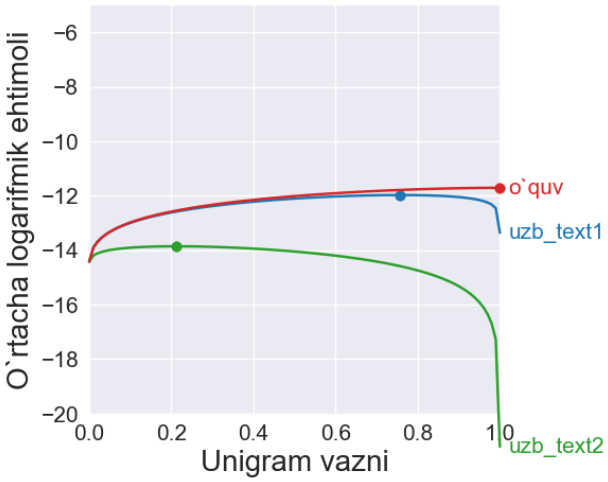
Haddan tashqari silliqlashning bunday qisqarishini boshqa ob'ektivda ko‘rish mumkin:

–Chap tomondagi model bizning ma'lumotlarimizga yomon yaqinlashadi (yuqori taraflama), ammo uning natijalari o‘qitilgan ma'lumotlardan qat'iy nazar barqaror bo‘lib qoladi.

–O‘ng tomonda, haddan tashqari silliqlasgan mos model bizning ma'lumotlarimizga juda mos keladi, lekin uning natijalari boshqa ma'lumotlar to‘plamida o‘qitilgan bo‘lsa (yuqori tafovut) boshqacha bo‘lishi mumkin. Natijada, o‘rganilgan ma'lumotlarga haddan tashqari mos keladigan model boshqa ma'lumotlar to‘plamida baholanganda juda yomon natija berishi mumkin.

Ushbu o‘xshashlikni bizning muammomizga qo‘llagan holda, yagona model mos kelmaydigan model ekanligi ayon bo‘ladi: u har bir unigramma uchun bir xil ehtimollikni belgilaydi, shuning uchun o‘quv ma'lumotlarini butunlay e'tiborsiz qoldiradi. Boshqa tomondan, silliqlashmagan unigram modeli haddan tashqari mos modeldir: u o‘quv matnidagi unigramlar uchun aniqrioq ehtimollik qiymatlarini beradi, lekin boshqa matndagi unigramlar uchun belgini o‘tkazib yuboradi.

Bir ekstremaldan ikkinchisiga o‘tishni tasavvur qilish uchun bizga berilgan uchta matnning o‘rtacha logarifmik ehtimolligini unigram modellari o‘rtasidagi turli interpolyatsiyalarni ko`rishimiz mumkin.



6-rasm. *Berilgan matnlarning o‘rtacha logarifmik ehtimolliginining unigram modellari*

Quyoridahi grafikdan quyidagi xulosalarni qayd etish mumkin:

1. Sof uniform model (grafikning chap tomoni) uchta matn uchun juda past o‘rtacha logarifmik ehtimolga ega, ya'ni **yuqori tarafkashlik (high bias)**. Biroq, har 3 ta matn ham modeldan bir xil o‘rtacha logarifmik ehtimolga

© Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

ega. Ya`ni, ehtimollik baholarining dispersiyasi nolga teng, chunki yagona model barcha unigrammalarga bir xil ehtimollikni oldindan belgilab beradi.

2. Interpolatsiyaga tobora ko‘proq unigram modeli qo‘shilganligi sababli, har bir matnning o‘rtacha logarifmik ehtimoli ortadi. Biroq, uchta matn o‘rtasidagi o‘rtacha logarifmik ehtimolligi farqlana boshlaydi, bu tafovutning ortishidan dalolat beradi.

3. Nihoyat, interpolyatsiyalangan model sof unigram modeliga yaqinlashganda, o‘quv matnining o‘rtacha logarifmik ehtimolligi tabiiy ravishda maksimal darajaga yetadi. Bundan farqli o‘laroq, *uzb_text1* va *uzb_text2* baholash matnlarining o‘rtacha logarifmik ehtimolligi sezilarli darajada pastga siljiy boshlaydi, bu esa modelning ishlashidagi farqni yanada oshiradi.

Uzb_text1 matni uchun, unigrammaning 91%, uniformaning 9% bilan interpolyatsiya qilinganda, uning o‘rtacha log ehtimolligi maksimal darajaga yetadi. *uzb_text2* matni uchun unigram-uniform modelining ideal nisbati 81:19 ni tashkil qiladi. Bu xuddi shunga o‘xshash nisbatga ega (80:20) silliqlangan unigram modeli silliqlanmagan modelga qaraganda *uzb_text2* matniga yaxshiroq mos kelishi haqidagi oldingi kuzatishimizga mos keladi.

Darhaqiqat, baholash matni o`quv matnidan qanchalik farq qilsa, bizning unigram modelimizni uniform bilan interpolyatsiya qilishimiz kerak. Bu mantiqan to‘g‘ri, chunki biz unigram modelining haddan tashqari mosligini sezilarli darajada kamaytirishimiz kerak, shunda u o‘rganilganidan juda farq qiladigan matnni yaxshiroq umumlashtirishi mumkin.

Xulosa

Til modeli – bu tabiiy tildagi so‘zlar ketma-ketligini bashorat qilish modeli bo`lib, asosiy maqsad tilda mavjud bo‘lgan o‘ziga xos tuzilma va shablonlarni aniqlash va matnni yaratishdan iborat. Til modellari turli vazifalar uchun foydalidir, jumladan, nutqni tanib olish, mashina tarjimasi, matnlarni yaratish, optik belgilarni aniqlash, qo‘lyozmalarni tanib olish. Katta til modellari (Large language models, LLM) odatda katta hajmli ma'lumotlar to‘plamlari (ko‘pincha til korpuslari) va neyron tarmoqlar asosdida shakllantiriladi. LLM modellari avvalo statistik modellar, xususan, n-gram til modellari vositasida shakllantirilgan. Bugun kunda ular o‘rnini takrorlanuvchi neyron tarmog‘iga (Recurrent Neural Network, RNN) asoslangan modellardan foydalanilmoqda. Ushbu maqolada o‘zbek tili matnlari uchun n-gramm til modellarining xususiy holi sanalgan (n=1) unigram til modelini amalga oshirish haqiqatan ham tabiiy tilni qayta ishlash sohasida mashinali o‘qitishning klassik hodisalari, masalan, haddan tashqari moslashish va qarama-qarshilik almashinuvi qanday namoyon bo‘lishi haqidagi mulohazalar keltrildi. Maqola keltirilgan unigram modeli va oddiy ehtimollik qoidalaridan foydalanadigan, soddalashtirilgan til modelimiz ma'noli bo‘lishi mumkin bo‘lgan yangi matn yaratishga qodir. Albatta, bunday sodda model odamga o‘xshash maqolalar yoza olmaydi yoki GPT-3 kabi ishlay olmaydi, chunki uning kamchiliklari ko‘p. Ammo bu tabiiy tilni qayta ishlash va tilni modellashtirish haqidagi boshlangi`ch bilimlarga ega bo`lish va soda modellarni ishlab chiqishda foydalanish mumkin.

Foydalanilgan adabiyotlar

1. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.
2. Mayfield, J., & McNamee, P. (2003, July). Single n-gram stemming. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 415-416).
3. 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).
4. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*(3), 853-860.
5. Zitouni, I. (2007). Backoff hierarchical class n-gram language models: effectiveness to model unseen events in speech recognition. *Computer Speech & Language*, *21*(1), 88-104.
6. Aouragh, S. L., Yousfi, A., Laaroussi, S., Gueddah, H., & Nejja, M. (2021). A new

(cc) BY © Botir Elov, Nizomaddin Xudayberganov, Mastura Primova

estimate of the n-gram language model. *Procedia Computer Science*, *189*, 211-215.
7. Fanani, A. M., & Suyanto, S. (2021). Syllabification Model of Indonesian Language Named-Entity Using Syntactic n-Gram. *Procedia Computer Science*, *179*, 721-727.
8. Costa-Jussa, M. R., & Fonollosa, J. A. (2009). An Ngram-based reordering model. *Computer Speech & Language*, *23*(3), 362-375.
9. Jurafsky, D., & Martin, J. H. (2019). Chapter 3: N-Gram Language Models N-Gram Language Models. Speech and Language Processing.
10. 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.
11. Bahl, L., Baker, J., Cohen, P., Jelinek, F., Lewis, B., & Mercer, R. (1978, April). Recognition of continuously read natural corpus. In *ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 3, pp. 422-424). IEEE.
12. Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., ... & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, *6*(1), 158.
13. Lee, R. S. T. (2024). N-Gram Language Model. In *Natural Language Processing*. https://doi.org/10.1007/978-981-99-1999-4_2
14. Colla, D., Delsanto, M., Agosto, M., Vitiello, B., & Radicioni, D. P. (2022). Semantic coherence markers: The contribution of perplexity metrics. *Artificial Intelligence in Medicine*, *134*, 102393.
15. Mimi, R. J., Masud, M. A., Rahman, R., & Dina, N. S. (2022, February). Text Prediction Zero Probability Problem Handling with N-gram Model and Laplace Smoothing. In *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)* (pp. 1-6). IEEE.

© *Botir Elov, Nizomaddin Xudayberganov, Mastura Primova*