O‘ZBEK TILI UCHUN STEMMERNI ISHLAB CHIQISH

Aloyev Narzillo Rahmatullayevich
ToshDO‘TAU tayanch doktoranti,
vip.alayev@gmail.com
Xuydayberganov Nizomaddin Uktambay o‘g‘li
ToshDO‘TAU
Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi stajor-o‘qituvchisi
nizomaddin@navoiy-uni.uz

Annotatsiya. Ushbu maqolada o‘zbek tili leksik birliklarini axborot qidirish tizimlaridan izlashga mo‘ljallangan stemmerni ishlab chiqish usullari taqdim etiladi. U chiziqli hisoblash murakkabligiga ega bo‘lib, uning sinov muvaffaqiyatining koeffitsienti 96,1%ni tashkil qiladi. Ushbu maqolada o‘zak shakllanishidagi ehtimollik ko‘rinishining batafsil tavsifi beriladi.

Kalit so‘zlar: *stemmer, NLP, agglyutinativ til, stem, morfologik analizatorlar, IR, ma’lumot olish tizimlari, o‘zbek stemmeri modeli.*

Abstract. This article presents the methods of developing a stemmer for searching Uzbek lexical units from information search systems. It has linear computational complexity, and its trial success rate is 96.1%. This article provides a detailed description of the probability view in core formation.

Key words: *stemmer, NLP, agglutinative language, stem, morphological analyzers, IR, information retrieval systems, Uzbek stemmer model.*

Аннотация. В данной статье представлены методы разработки стеммеров для поиска узбекских лексических единиц из информационно-поисковых систем. Он имеет линейную вычислительную сложность, а показатель успешных тестов составляет 96,1%. В данной статье дается подробное описание появления вероятности при формировании керна.

Ключевые слова: *стеммер, НЛП, агглютинативный язык, основа, морфологические анализаторы, ИР, информационно-поисковые системы, узбекская стеммерная модель.*

Kirish

Ingliz tili kabi tillarda stemlash jarayoni nisbatan oddiyroq kechadi. Buning sababi esa so‘z shakllarining morfologik o‘zgarishlari cheklangan bo‘ladi. Boshqa tomondan esa, o‘zbek tili kabi agglyutinativ tillarda stemlash hamon jiddiy muammo bo‘lib qolmoqda, chunki ular nazariy jihatdan cheksiz mavjud so‘z shakllarini hosil qilish imkoniyatiga ega. Morfologik analizatorlar agglyutinativ tillarda ma’lumot qidirish tizimlari uchun stemmer sifatida yagona aniq vosita hisoblanadi [1,2]. Ammo bu holda, agglyutinativ tillarda IR tizimlari uchun stemlash, qarama-qarshi talablar juftligi bilan aniqlanishi mumkin bo‘lgan paradoksni keltirib chiqaradi: IR

tizimlaridan agglyutinativ tillar uchun xotiraga saqlash murakkabligini bartaraf etish va ish faoliyatini yaxshilash uchun stemlashni talab qiladi, lekin, morfologik analizatorlardan stemmer sifatida foydalanishda yuzaga keladigan hisoblash murakkabligining yuqori darajasi tufayli IR tizimlarining umumiy ishlash koeffitsienti pasayadi. Biz taklif qilayotgan stemlash usuli, ushbu qiyinchilikni samarali yengish uchun ishlatilishi mumkin.

Ushbu maqolada o‘zbek IR tizimlari uchun taqdim etilgan statistik muhitga asoslangan ehtimollik stemlash modelini taqdim etiladi. Shuningdek, taklif qilingan model tahlillar va tajriba natijalari bilan tasdiqlanadi. Shunday qilib, yangi stemmer **O(n)** hisoblash murakkabligiga ega bo‘lgan holda yuqorida ko‘rsatilgan to‘siqlarni yengib o‘tadi [3].

Stemlash va ma’lumot olish

IR (ma’lumot olish) tizimlari katta hajmdagi elektron hujjatlardan to‘plangan ma’lumotlarni qayta ishlash uchun ishlatiladi. Hujjat haqidagi ma’lumotlar asosan so‘zlarning semantikasidan iborat bo‘ladi. Demak, IR tizimlari haqiqatdan semantika vakillari hisoblangan so‘zlarni boshqaradi/saqlaydi. Hujjatdagi so‘z matn ketma-ketligidagi grammatik qo‘llanishiga qarab turli morfologik shakllarga ega bo‘ladi, lekin o‘zagi bilan ifodalangan semantika o‘zgarishsiz qoladi. Shuning uchun, IR tizimlari odatda xotiraga saqlash muammolarini bartaraf etish va ish faoliyatini oshirish uchun so‘z shakllari o‘rniga stemlardan foydalanadi.

Lovins tomonidan o‘zak ta’rifi “bir xil o‘zakli barcha so‘zlarni umumiy shaklga qisqartirish tartibi, odatda har bir so‘zdan uning hosilaviy va flektiv qo‘shimchalarini olib tashlash” orqali berilgan [4]. Ingliz tili kabi analitik tillarda IRni aniqlash agglyutinativ tillarga qaraganda ancha oson kechadi. Misol tariqasida, Porterning ingliz tili uchun algoritmi cheklangan miqdordagi kaskadli qayta yozish qoidalariga asoslanadi va deyarli barcha so‘z shakllarini atigi **1200** ga yaqin morfologik o‘zgarishlarni tugatish orqali qabul qilishi mumkin. Lekin o‘zbek tilda agglyutinativ til sifatida, hatto taxminan **23000** ta stemlar bo‘lsa ham; nazariy jihatdan cheksiz turli xil so‘z shakllari yaratilishi mumkin [5].

Morfologik murakkablikning ushbu darajasida lug‘atni o‘zgartirish modellari va o‘rnatilgan qoidalarni ifodalaydigan morfologik analizatorlari, morfologik tahlil qilish uchun qabul qilingan standart hisoblanadi. Ammo bu morfologik analizatorlar ikki darajali til modeliga asoslanadi. IR tizimlari uchun ikkita to‘siq mavjud: Hisoblashning murakkabligi morfologik analizatoridan stemmer sifatida foydalanishdan kelib chiqadi va aksincha, stemmersiz, ya’ni, saqlash murakkabligi bir xil semantik, ammo morfologiyasi boshqacha bo‘lgan so‘zning har bir turli shaklini indekslashdan kelib chiqadi [3,4].

Shu sababli, xotiraga saqlash hajmini boshqara olish va qoniqarli ish darajasiga erishish uchun agglyutinativ tillar uchun IR tizimlaridan stemlash talab qilinadi. Biroq, o‘z navbatida, IR da samarali foydalanish uchun chiziqli hisoblash murakkabligi, algoritmlarni stemlash uchun muhimdir. Agglyutinativ tillarda stemming zarurligini isbotlash uchun Porter stemmeri ko‘plab inglizcha so‘zlarni o‘z

ichiga olgan **newswire** ma’lumotlar bazasiga qo‘llanilib ma’lum bir siqish darajasiga erishildi va shuningdek, 400000 dan ortiq o‘zbekcha so‘zni o‘z ichiga olgan (kun.uz satidan olingan) siyosiy yangiliklar ma’lumotlar bazasiga qo‘llanilgan stemmer 1-jadvalda o‘z aksini topgan.

1-jadval. O‘zbek tilidagi matnlarni stemmlash ko‘rsatkichlari

korpus	**so‘z tokenlari**	**unikal terminlar**	**siqish (%)**
O‘zbek	422,112	32,370	87,3
inglizcha	650,192	18,384	38,1

Yuqoridagi 1-jadvaldan ko‘rinib turganidek, ingliz tilidagi ma’lumotlar bazasida ko‘proq so‘z tokenlari mavjud bo‘lishiga qaramay; o‘zbek tilidagi alohida atamalar ingliz tilidan ko‘ra ko‘proq. Bu o‘zbek tilidagi so‘zlarning xilma-xilligini ko‘rsatadi hamda, bu siqish nisbati natijalari bilan ham tasdiqlangan. 38,1% siqish darajasi 87,3% dan ancha kichik, bu esa ingliz tili uchun morfologik o‘zgarishlarning past darajasini aniq ko‘rsatadi. Biroq, o‘zbek tili ma’lumotlar bazasi uchun kuzatilgan **87,3%** siqish nisbati esa qanchalik stemlash zarurligini ta’kidlab turadi.

Stemlash jarayoni bog‘liq masalalar

IR maqsadlari uchun foydalaniladigan stemlash algoritmlarini to‘rtta guruhga bo‘lish mumkin:

- *table lookup;*
- *successor variety;*
- *n-gram;*
- *affix removal.*

Table Lookup algoritmlarida, indekslangan terminlar sifatida ishlatilishi mumkin bo‘lgan barcha atamalar va ularning asoslari jadvalda saqlanadi. Shunday qilib, so‘rovlar va hujjatlardan atamalarni **juda tez stemlash mumkin.** Ammo amalda ingliz tili uchun bunday unikal terminlar jadvali mavjud emas va boshqa tabiiy tillar uchun esa ba’zi indeks atamalar domen yoki kontekstga bog‘liq bo‘ladi. Shuningdek, shu tartibdagi jadvaldagi saqlashga ketadigan resurslar muammoligicha qolaveradi.

Successor Variety algoritmlari asosida olingan natijalar katta qismidan kuzatilgan fonemalarning taqsimlanishi bilan birga so‘z yoki morfema chegaralarini aniqlashga asoslangandir.

N-gramm asosidagi stemmer so‘rov va hujjat o‘rtasidagi bo‘shliq shartlari o‘xshashlik o‘lchovidan foydalanadi. Uzunligi m bo‘lgan atama m-n+1 n-grammga ega. Har bir shart juftligi uchun har ikkala shartning yagona n-grammasidan va o‘xshashlik matritsasidan hisoblangan koeffitsientlar tuziladi. O‘xshashlik matritsasi tuzilgandan so‘ng, atamalar yagona bo‘g‘inli klasterlash usuli yordamida klasterlanadi.

Affikslarni ajratishga asoslangan stemmerlarda stemlarni aniqlash uchun atamalardan qo‘shimchalar va/yoki prefikslarni olib tashlaydigan algoritm mavjud. Yuqorida eslatib o‘tilganidek, Porterning stemmeri affikslarni olib tashlash algoritmlariga yaxshi misol bo‘la oladi. To‘rtta stemmerlar guruhining barchasi ingliz tili kabi analitik tillar yoki fransuz va nemis kabi ba’zi Hind-Yevropa tillari uchun ishlab chiqilgan hisoblanadi. Ular turkiy, turkman, ozarbayjon yoki azeri, qashqay va o‘g‘uz tillarini ham o‘z ichiga olgan janubi-g‘arbiy yoki o‘g‘uz guruhidagi turkiy agglyutinativ tillar kabi inflektiv paradigmaga ega bo‘lgan tillarda yuqori darajadagi morfologiyaga dosh bera olmaydi.

Shuningdek, o‘zbek tili uchun ishlab chiqilgan bir qator stemlash algoritmlari mavjud bo‘lib, ular morfemalarni belgilash va aniqlash kabi ba’zi usullarni kerakli aniqlik darajasiga tushirish orqali morfologik analizatorni IR tizimiga moslashtirishga qaratilgan harakat bo‘lib hisoblanadi [6]. Ushbu turdagi algoritmlar oldindan tuzilgan elektron leksikondan ehtimoliy o‘zakni izlaydilar, so‘ngra esa, o‘zbek tilining morfotaktikasiga ko‘ra qo‘shimchalarning birikmalarini qidirib, so‘z shaklining qolgan qismini aniq qo‘shimchalar qatoriga moslashtirish orqali uning to‘g‘riligini tekshiradi. Chunki ularning tayanch tushunchasi morfologik tahlillar bo‘lib hisoblanadi. Tekshirishning keyingi bosqichining qat’iylik darajasi ham aniqlikni, ham, ayniqsa, hisoblash murakkabligini belgilaydi; yuqori darajadagi tekshirish qat’iyligi bilan, aniqlik yuqori bo‘ladi, lekin hisoblash murakkabligi NP-murakkabligiga yaqin bo‘ladi va aksincha. Shuning uchun, bu stemmerlar uchun hisoblash murakkabligi darajasi chiziqli nuqtadan NP-Hardga yaqin nuqtagacha, ular foydalanadigan morfologik bilim darajasi qanchalik oshsa shunga qadar yana oshadi [7,8].

O‘zbek tilidagi stemmerni ishlab chiqish maqsadida turkiy tillar oilasiga mansub tillarda olib borilgan ilmiy tadqiqotlar o‘rganib chiqildi. Jumladan, turk tilidagi stemlash muammolariga birinchilardan bo‘lib Köksal tomonidan yechim topilishi harakati boshlangan [9]. Bu nolingvistik nuqtayi nazaridan olib qaraganda boshqalardan butunlay farq qiladi. U so‘z shaklining o‘zagi sifatida boshidanoq so‘z shaklining o‘zgarmas uzunlikdagi qismini qabul qiladigan stemmerni ishlab chiqdi. Dastlab, Köksalning ta’kidlashicha, 5 ta stem uzunligi sinovlarida eng yaxshi natija beradi. Ammo amalda barcha hujjatlar to‘plami uchun umumiy belgilangan uzunlik mavjud emas. Muayyan mavzu bo‘yicha ma‘lum bir hujjat to‘plami bilan yaxshi ishlaydigan qiymat boshqa mavzu ostidagi hujjatlar to‘plami bilan ishlaydigan IR tizimi ish unumdorligini kamayishiga olib kelishi mumkin.

Taklif etilayotgan modelimizda biz ilgari tahlil qilingan matn ma’lumotlar bazasidan to‘plangan umumiy morfologik bilimlar statistikasidan modelni takomillashtirish uchun pragmatik sifatida foydalanamiz. Oldindan hisoblangan statistik ma’lumotlardan foydalangan holda, biz chiziqli hisoblash murakkabligining saqlanishini ham ta‘minlashga erishdik.

O‘zbek stemmeri modeli

Har qanday agglyutinativ tildagi so‘z shaklini $\boldsymbol{S_n = h_1 h_2 \dots h_n}$ bo‘lgan satr orqali bilan belgilash mumkin. Bunda har bir $\boldsymbol{h_i\ (i = 1, 2, \dots, n)}$ tegishli A alifbosining elementi bo‘lib, **n** esa harflar sonini ifodalaydi. Ushbu tadqiqotda **26** harfdan iborat o‘zbek alifbosi va bo‘sh belgi uchun pastki chiziq "_" dan quyidagi tarzda foydalanamiz:

$$A = \{a, b, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, x, y, z, o', g', sh, ch, ng, ', '\_'\}$$

Quyidagi yozuvni har qanday $\boldsymbol{S_n}$ satrning $\boldsymbol{1 \le i \le j \le n}$ qanoatlantiruvchi satr ostisini belgilashda moslashtiramiz,

$$\boldsymbol{S_n[i : j] = h_i h_{i+1} \dots h_j}$$

$$\boldsymbol{S_n[: j] = h_i h_{i+1} \dots h_j} \textbf{ va } \boldsymbol{S_n[i :] = h_{i+1} \dots h_n}$$

Yuqoridagi belgilashga asoslangan holda,

$\boldsymbol{S_n[i : i+1] = h_i h_{i+1}}$ maxsus satr osti $\boldsymbol{(h_i, h_2)_i}$ harflarining tartiblangan juftligi bilan belgilanadi.

Bu yerda $\boldsymbol{i\ (i = 1, 2, \dots, n-1)}$ sub-indeks bo‘lib, ushbu qatordagi tartiblangan juftlikning boshlang‘ich pozitsiyasini va $\boldsymbol{h_1 = h_i, h_2 = h_{i+1} \in A}$ ni bildiradi.

$\boldsymbol{i = n}$ uchun juftlik oldindan tayyor qilib qo‘yilgan $\boldsymbol{(h_n, '\_')_{i=n}}$ orqali aniqlanadi. Shunday qilib, tadqiqotda har qanday $\boldsymbol{S_n = h_1 h_2 \dots h_n}$ satr **n** ta tartiblangan juftlikka ega bo‘ladi.

Har qanday o‘zbek tilidagi so‘z shaklda $\boldsymbol{1 \le j \le n_{max}}$ pozitsiyasida paydo bo‘lishi mumkin bo‘lgan $\boldsymbol{(h_i, h_2)_j}$ harflarining tartiblangan juftligi (**nmax** – o‘zbek so‘z shakli ega bo‘lishi mumkin bo‘lgan maksimal harflar soni) va $\boldsymbol{n \ge j}$ uchun $\boldsymbol{S_n = h_1 h_2 \dots h_n}$ bilan belgilangan ma’lum bir so‘z shakl, $\boldsymbol{(h_i, h_2) \in S_n}$ yozuvi $\boldsymbol{i = j}$ uchun $\boldsymbol{h_1, h_2 = (h_1, h_2)_j}$ sharti bilan $\boldsymbol{S_n}$ da $\boldsymbol{i\ (1 \le i \le n)}$ pozitsiyasida tartiblangan $\boldsymbol{(h_i, h_2)_i}$ juftlik mavjudligini anglatadi.

Shuningdek, har qanday so‘z shaklini barcha $\boldsymbol{1 \le m \le n}$ uchun $\boldsymbol{S_n^m = (g_m, e_m)}$ bo‘yicha satr ostining tartiblangan juftligi sifatida ifodalash uchun yana ikkita $\boldsymbol{g_m = s_n, [: m]}$ **va** $\boldsymbol{e_m = s_n, [m :]}$ belgilashlarni kiritamiz.

Namuna maydoni va tartiblangan juftlik ehtimoli

Faraz qilaylik, **L** to‘plam $\boldsymbol{i = 1, 2, \dots, n_{max}}$ pozitsiyalari uchun har qanday o‘zbek tilidagi so‘z shaklida paydo bo‘lishi mumkin bo‘lgan barcha mumkin bo‘lgan tartiblangan $\boldsymbol{(h_1, h_2)_i}$ harf juftliklarining yig‘indisi bo‘lsin. Demak, **L** namuna maydonini quyidagicha aniqlash mumkin:

$$\boldsymbol{L = \{(h_1, h_2)_i | = h_1, h_2 \in A\ va\ 1 \le i \le n_{max}\}}$$

$\boldsymbol{G_k, E_k, T_k \in L\ va\ 1 \le k \le n_{max}}$ bo‘lgan $\boldsymbol{G_k, E_k\ va\ T_k}$ to‘plamlari aniqlangan eventlarni quyidagicha ifodalash mumkin:

$$\boldsymbol{G_k = \{(h_1, h_2)_i |\ i = k\ va\ (h_1, h_2)_i \in g_m\ va\ 1 \le m \le n_{max}\}}$$

$$\boldsymbol{E_k = \{(h_1, h_2)_i |\ i = k\ va\ (h_1, h_2)_i \in e_m\ va\ 1 \le m \le n_{max}\}}$$

$$\boldsymbol{T_k = \{(h_1, h_2)_i |\ i = k, h_1 = S_n[k : k],\ h_2 = S_n[k+1, k+1], 1 \le i \le n_{max}\}}$$

Shunday qilib, $S_n = h_1 h_2 \dots h_n$ bilan belgilangan har qanday berilgan so‘z shaklining $i = 1, 2, \dots, n$ pozitsiyalarida tartiblangan har bir $(h_1 h_2)_i$ juftligi uchun yuqoridagi uchta to‘plamda bo‘lish ehtimolini quyidagicha aniqlash mumkin,

$$Pr(S_n[i: i+1] \in G_i) = Pr(\ (h_1, h_2)_i\ \in G_i) = P_G((h_1, h_2)_i) \quad (1)$$

$$Pr(S_n[i: i+1] \in E_i) = Pr(\ (h_1, h_2)_i\ \in E_i) = P_E((h_1, h_2)_i) \quad (2)$$

$$Pr(S_n[i: i+1] \in T_i) = Pr(\ (h_1, h_2)_i\ \in T_i) = P_T((h_1, h_2)_i) \quad (3)$$

Bu yerda **(1)** tenglama $(h_1, h_2)_i$ tartiblangan juftlik berilgan so‘z shaklining o‘zak qismida bo‘lish ehtimolini bildirsa, (2) tenglama ham $(h_1, h_2)_i$ tartibli juftlik berilgan so‘z shaklining qo‘shimcha qismida bo‘lish ehtimolini bildiradi va nihoyat, **(3)** tenglama $(h_1, h_2)_i$ tartiblangan juftlik berilgan so‘z shaklining o‘zak qismi va qo‘shimcha qismi o‘rtasida bo‘lish ehtimolini bildiradi.

Ehtimoliy stemlash tasdig‘i

Ushbu tadqiqotdagi ehtimollik modelimiz aniqligi quyida keltirilgan tajriba orqali tekshiriladi.

1-taklif. $[P_E((h_1, h_2)_m) > P_G((h_1, h_2)_m]\ va\ P_T\ ((h_1, h_2)_{m-1}) \geq \alpha$ tenglamani qanoatlantiradigan $1 \leq m \leq n$ butun qiymat mavjud bo‘lsa,

Bu hodla $S_n^m = (g_m, e_m)$ so‘z shakli va $0 \leq \alpha \leq 1$ o‘zgarmas uchun, g_{m-1} so‘z shaklining **ehtimoliy o‘zagi** deb aytiladi.

1-ta‘rif. $S_n = h_1 h_2 \dots h_n (1 \leq n \leq n_{max})$ bilan belgilangan har qanday so‘z shaklning **ehtimollik asosi** e_{m-1} satr ostini ajratib olish va g_{m-1} satr ostini ma‘lum $0 \leq \alpha \leq 1$ uchun **1-taklifni** qanoatlantiradigan **m** pozitsiyasida ushbu so‘z shaklining o‘zagi sifatida qabul qilishdan iborat.

Tajriba va natijalar

Tajriba va test ma’lumotlar bazalari o‘zbek tili korpusidagi matnli ma’lumotlar bazasidan tanlandi. Bu o‘zbekcha xabar matnlari to‘plami bo‘lib, 1 milliondan ortiq so‘z shakllariga ega. Tanlangan ma’lumotlar bazalarining xossalari va natijalari quyidagi 2-jadvalda keltirilgan.

2-Jadval. Test ma’lumotlar bazasi va tajribalar xususiyati

Ma’lumotlar bazasi	So‘zlar soni	Unikal so‘zlar	Unikal o‘zaklar	Noma’lum o‘zaklar	Bajarilgan (to‘g‘ri)	Juftlik soni	Samaradorlik (%)
Tajribaviy	210,129	32,122	9,568	0	0	5,648	0
TEST	211,489	32,563	9,253	3,212	8,918	-	95.8

Ikkala ma’lumotlar bazasidagi terminlar morfologik tahlil natijalariga ko‘ra o‘zak va qo‘shimcha qismlarni aniqlash uchun oldindan qayta ishlangan. Bunda, ma’lumotlar bazalari boshqa qayta ishlanmaydi, masalan: *nomuhim so‘zlarni olib tashlash, noto‘g‘ri yozilgan so‘zlarni aniqlash va hokazolar*.

Shuni ta’kidlash lozimki, agar morfologik analizator ma’lumotlar bazasi uchun stemmer sifatida ishlatilsa, siqilish nisbatining mutlaq yuqori chegarasi **74%** ni tashkil qiladi (ya’ni, 32,563 ta alohida terminlarning barchasi ularning **9,253** haqiqiy stemiga to‘liq qisqartirilganda) va taklif qilingan o‘zak siqish nisbati **72%** ga yetishi mumkin.

(1), (2) va (3) tenglamada berilgan ehtimollar quyidagi formulalar bilan hisoblanadi:

$$\boldsymbol{P_G((h_1,h_2)_i) = f_{g,i} * w_{g,i} / N ;\ P_E((h_1,h_2)_i) = f_{e,i} * w_{e,i} / N ;}$$
$$\boldsymbol{P_T((h_1,h_2)_i) = f_{t,i} * w_{t,i} / N}$$

Bu yerda

– $\boldsymbol{f_{g,i}}$, $\boldsymbol{f_{e,i}}$ va $\boldsymbol{f_{t,i}}$ - tartiblangan harf juftlarining mos ravishda o‘zak qismida, qo‘shimcha qismida va so‘z shaklining o‘zak va qo‘shimcha qismi o‘rtasidagi tranzitda paydo bo‘lishi eventlari bo‘yicha tajriba bazasidan kuzatilgan chastotalar soni;

– $\boldsymbol{w_{g,i}}$, $\boldsymbol{w_{e,i}}$ va $\boldsymbol{w_{t,i}}$ tuzatish omillari ilgari mos keladigan $\boldsymbol{f_{g,i}}$, $\boldsymbol{f_{e,i}}$ va $\boldsymbol{f_{t,i}}$ chastotalar uchun ijobiy haqiqiy qiymatlar hisoblanadi.

– **N** - tajriba ma’lumotlar bazasida buyurtma qilingan juftliklarning umumiy soni.

Tajriba ma’lumotlar bazasiga ishlov berilgandan so‘ng, **L** namunaviy maydoni uchun umumiy soni **7148** ta unikal tartiblangan harflar juftligi kuzatildi. Ulardan **3142** juft $\boldsymbol{G_k}$ to‘plamiga, **1381** juft $\boldsymbol{E_k}$ to‘plamiga va **367** juft $\boldsymbol{T_k}$ ga tegishli.

$\boldsymbol{G_k}$ va $\boldsymbol{E_k}$ to‘plamlar kesishmasida **1397** ta unikal tartiblangan juftliklar mavjud. Darhaqiqat, bu kuzatuvlar asosida barcha tartiblangan juft harflarning **50%** i faqat o‘zak qismida, **19%** faqat qo‘shimcha qismida va **24%** o‘zak va qo‘shimcha qismlarida bo‘lishlarini ko‘rsatadi.

Ushbu statistik ma’lumotlar biz taklif qilayotgan model nima uchun muvaffaqiyatli bo‘lganini aniq tushuntirib beradi. Test ma’lumotlar bazasida **9253** ta alohida stem mavjud bo‘lib, ulardan **3212** tasi eksperimental ma’lumotlar bazasida yo‘q. Bu nisbat juda katta hisoblanadi va taxminan **40%** noma‘lum stemlarni tashkil qiladi. Stemmer **32563** ta test shartlari uchun **68962** ta mumkin bo‘lgan stemlarni ishlab chiqardi. Shunday qilib, har bir atama uchun taxminan **1,5** ta stem ishlab chiqarilgan.

Aslida, bu **68962** hosil bo‘lishi mumkin bo‘lgan o‘zaklarning aksariyati tan olingan o‘zbek so‘z shakllaridir. Lekin ular kerakli o‘zakdan farqli semantikaga ega bo‘lgan root (o‘zak)ga tushiriladi yoki kerakli o‘zakdan farqli bo‘lgan hosila yoki flektiv shaklga o‘zlashtiriladi.

Hisob-kitob natijasida, kerakli stemming aniq mosligini qabul qildik. Natijada, biz taklif qilgan stemmer sinov shartlarining **94,2%** uchun to‘g‘ri stemlarni ishlab chiqarishga erishdi. Bizning stemmer **32563** tadan **672** tasi turlicha bo‘lgan **1055** ta atamani stemlarga ajrata olmadi. Biroq, o‘tkazib yuborilgan atamalarning aksariyati

oddiy IR tizimida ishlatiladigan har qanday matn protsessorlari tomonidan yo‘q qilinishi mumkin bo‘lgan **neologizmlar**, **qisqartmalar** yoki **olmoshlardir**.

Xulosa

Ushbu maqolada agglyutinativ tillar uchun ehtimollikka asoslangan yangi stemming modeli taqdim etildi. Taklif qilingan model chiziqli hisoblash murakkabligiga ega bo‘lib, **94,2%** aniqlik darajasiga ega. Maqolada keltirilgan stemming modelini fleksiyon paradigmalariga ega bo‘lgan boshqa agglyutinativ tillarga qo‘llash mumkin.

Foydalanilgan adabiyotlar ro‘yxati:

1. Hakkani-Tür, D. Z., Oflazer, K., & Tur, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Language Resources and Evaluation*, *36*(4). https://doi.org/10.3115/990820.990862
2. Jurafsky, D., & Martin, J. (2014). Speech and Language Processing. In *Speech and Language Processing.* (Vol. 3).
3. B.Elov, Sh.Hamroyeva, X.Axmedova. (2022). Methods for creating a morphological analyzer, *14th International Conference on Intellegent Human Computer Interaction, IHCI 2022, 19-23 October 2022, Tashkent.*
4. Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, *11*(June).
5. Porter, M. F. (2006). An algorithm for suffix stripping. *Program*, *40*(3). https://doi.org/10.1108/00330330610681286
6. B.B.Elov, Sh.M.Khamroeva, Z.Y.Xusainova. Yashirin markov modeli vositasida o‘zbek tili matnlarini POS teglash (2022).
7. Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash. O‘zbek amaliy filologiyasi istiqbollari / Respublika ilmiy-amaliy konferensiya to‘plami. Elektron nashr / ebook. – Toshkent: ToshDO‘TAU, 26.10.2022. 159-163 b.
8. Muljono, Afini, U., & Supriyanto, C. (2017). Morphology analysis for Hidden Markov Model based Indonesian part-of-speech tagger. *Proceedings - 2017 1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, *2018-January*. https://doi.org/10.1109/ICICOS.2017.8276368
9. Köksal, A.: Bilgi Eri im Sorunu ve Bir Belge Dizinleme ve Eri im Dizgesi Tasarım veGerçekle tirimi. Doçentlik Tezi. Fen Bilimleri Enstitüsü, Bilgisayar BilimleriMühendisli i Anabilim Dalı, Hacettepe Üniversitesi, Ankara (1979) *(12) (PDF) Stemming in Agglutinative Languages: A Probabilistic Stemmer for Turkish*.