ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 1 (6)

www.compling.tsuull.uz

MUNDARIJA

ISSN 2181-922X 40–64

O‘ZBEK, TURK VA UYG‘UR TILLARIDA POS TEGLASH VA STEMMING

Botir Elov[1],
Shahlo Hamroyeva[2],
Oqila Abdullayeva[3],
Zilola Xusainova[4],
Nizomaddin Xudayberganov[5]

Annotatsiya

O‘zbek, turk va uyg‘ur tillari agglutinativ tilllar hisoblanib, morfologik jihatdan va so‘z shakllarining miqdori jihatidan murakkab hisoblanadi. Mazkur tillarda o‘zak va qo‘shimchalarni birlashtirish orqali yangi so‘z va so‘z shakllari hosil qilinadi. O‘zakka qo‘shimchalar qo‘shilganda fonetik uyg‘unlik va disgarmoniya yuzaga kelishi oqibatida ham fonetik, ham morfologik o‘zgarishlar yuzaga keladi. Bu vaziyat matnda so‘z shakllarni POS teglash va stemming jarayonida turli hal qilinishi kerak bo‘lgan muammolarni hosil qiladi. Ko‘pgina NLP vazifalarni hal qilishda so‘z shakllarini

[1] *Elov Botir Boltayevich* – texnika fanlari falsafa doktori (PhD), dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648

[2] *Hamroyeva Shahlo Mirdjonovna* – filologiya fanlari doktori (DSc), dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: shaxlo.xamrayeva@navoiy-uni.uz

ORCID: 0000-0002-5429-4708

[3] *Abdullayeva Oqila Xolmo‘minovna* – filologiya fanlari bo‘yicha falsafa doktori.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-pochta: abdullayeva.oqila@navoiy-uni.uz

ORCID: 0000-0002-2524-4832

[4] *Xusainova Zilola Yuldashevna* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti doktoranti.
E-pochta: xusainovazilola@navoiy-uni.uz

ORCID: 0000-0003-4357-7515

[5] *Xudayberganov Nizomaddin Uktamboy o‘g‘li* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi o‘qituvchisi.
E-pochta: nizomaddin@navoiy-uni.uz

ORCID: 0000-0002-6213-3015

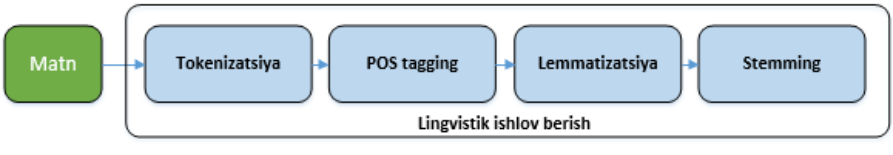
ularning o‘zakkacha qisqartirish (stemlash)ga to‘g‘ri keladi. So‘zdan barcha flektiv affikslarni olib tashlash va so‘zning qolgan qismini lemmatizatsiya qilish tabiiy tilni qayta ishlash (NLP)ning muhim vazifalaridan biri hisoblanib, ushbu jarayon stemming deb yuritiladi. Stemming jarayoni axborot qidirish (IR, Information Retrieval) tizimlarida muhim ahamiyat kasb etadi.

Kalit so‘zlar: *nutq qismlari, teglash, nutq qismlarini teglash, stemming, axborot qidirish, IR, stemming algoritmlari.*

KIRISH

Axborot qidirish va qayta ishlash tizimlarida foydalanuvchi so‘roviga to‘g‘ri natijani qaytarish tezligini oshirish eng muhim masala hisoblanadi. Buni amalga oshirishning eng oson va qulay usuli stemming jarayonidir. NLPda so‘zning turli morfologik variantlarini ularning umumiy shakl (o‘zak, stem)ini aniqlaydigan metod **stemming algoritmi** deyiladi [Paice, 1994]. Axborot qidiruv tizimlarida so‘z shaklning o‘zakkacha bo‘lgan qismini aniqlash uchun uning barcha qo‘shimcha(suffiks va prefix)larini olib tashlash zarur [Anjali, Jivani, 2011].

POS teglash hodisasi – bu berilgan gapdagi har bir so‘zshaklga uning turkum (*ot, fe’l, sifat, son, ravish yoki olmosh*)ga mansubligini belgilash (teglash) vazifasidir. POS teglash tabiiy tilni qayta ishlash (Natural Language Processing, NLP)ning asosiy vazifalaridan biri bo‘lib, pipeline konveyerining muhim bosqichi hisoblanadi (1-rasm).



1-rasm. Matnga boshlang‘ich ishlov berish bosqichlari

Mashina tarjimasi, matnni umumlashtirish, savol-javob va hissiyotlarni tahlil qilish kabi NLP ilovalari uchun POS teglash muhim qadam hisoblanadi. Masalan, “olma” so‘zini boshqa tilga tarjima qilish uning POS tegidan foydalaniladi. “Olma” (apple) ot so‘z turkumiga mansub bo‘lsa predmet bo‘ladi, “ol-ma” (don’t take) fe’l so‘z turkumiga mansub bo‘lsa, u harakatni bildiradi.

Agglutinativ tillarning ba‘zilarida POS teglash jarayonini amalga oshirish uchun so‘zlarning stemlaridan foydalaniladi [Taner Dincer, Karaog’lan, 2003]. O‘zbek, turk va uyg‘ur tillaridagi so‘zlar va

uning stemi turli POS tegga mansub bo‘lishi mumkin.

1-jadval. O‘zbek, turk va uyg‘ur tillaridagi so‘zshaklning lemmasi, stemi va POS tegi

№	so‘z	lemma	POS	Stem	POS	Root	POS
			O'zbek	**tili**			
1	muzladi	muzlamoq	VB	muz	N	muz	N
2	issiqroq	issiq	JJ	isi	VB	isi	VB
3	soddalashtiri-ladi	sodda-lashtirmoq	VB	sodda	JJ	sodda	JJ
4	ixtiyoriy	ixtiyoriy	JJ	ixtiyor	N	ixtiyor	N
5	qo‘llaniladigan	qo‘llamoq	VB	qo‘l	N	qo‘l	N
6	yo‘lakda	yo‘lak	N	yo‘l	N	yo‘l	N
7	qishlog‘im	qishloq	N	qishlog‘	?	qishloq	N
			Turk	**tili**			
7	yetkili	yetkili	ADJ	yetkili	ADJ	yetki	N
8	kurullarimizla	kurul	N	kurul	N	kurul	N
9	teşkilatlarim-izla	teşkilat	N	teşkilat	N	teşkil	N
10	seçimlere	seçim	N	seçim	N	seç	VB
11	futbolcularin	futbolcu	N	futbolcu	N	futbol	N
12	kullandi	kullanmak	F	kulla	F	kulla	F
13	bilgi	bilgi	N	bilgi	N	bil	F
			Uyg'ur	**tili**			
14	tarazichi	tarazichi	N	tarazichi	N	tarazi	N
15	yashaptu	yashamaq	VB	yasha	VB	yash	N
16	yegizligi	yegizlik	N	yegizlig	N	yegiz	VB
17	og’urluqqa	og’urluq	N	og‘urluq	N	og‘ur	N
18	chyshkənligini	chyshkən-liq	N	chyshkən-lig	N	chysh	?

Agglutinativ tillar uchun pos teglash va stemmingni amalga oshirish uchun ba‘zi terminlar izohini keltiramiz:

O‘zak (root) – so‘zning asl ma‘nosini bildirib, boshqa ma‘noli qismlarga bo‘linmaydigan, mustaqil holda leksik ma‘no bildiradigan eng kichik qism. So‘zga turlicha affikslar qo‘shilib kelganda ham, o‘zakning ma‘nosi yo‘qolmaydi, undan yasalgan so‘zlarning ma‘nosi ana shu ma‘no bilan bog‘langan bo‘ladi. Shuningdek, o‘zak boshqa ma‘noviy qism, ya‘ni morfemaga bo‘linmaydigan qism.

Lemma (leksema) – faqat o‘zakdan yoki o‘zak+so‘z yasovchi qo‘shimcha shaklidan iborat bo‘ladi. Leksema (yun. lexis – so‘z, ifoda) – til qurilishining leksik ma‘no anglatuvchi lug‘aviy birligi.

Leksema bildiradigan ma‘no so‘zning material qismi: ma‘lum tovush kompleksini ma‘lum obyektiv voqelikka bog‘lash bilan kishi ongida yuzaga keladigan mazmun-mundarija.

Stem – so‘zshaklning qo‘shimchalarini kesib tashlashdan hosil bo‘luvchi qism bo‘lib, ba‘zi hollarda ma‘no anglatmasligi mumkin. Shuningdek, stem so‘zning morfologik o‘zagi bilan aynan mos bo‘lmasligi yoki mos tushishi mumkin.

Turk va uyg‘ur tillarida stemming jarayoniga quyidagicha ta‘rif berilgan:

Stemming (turk va uyg‘ur) – bu so‘zga qo‘shilgan flektiv qo‘shimchalarni olib tashlash orqali uning o‘zagigacha qisqartirish vazifasidir. Biroq o‘zbek tili uchun stemming jarayoni quyidagicha ta‘riflanadi:

Stemming (o‘zbek) – bu so‘zga qo‘shilgan *derivatsion* va *flektiv* qo‘shimchalarni olib tashlash orqali uning o‘zagigacha qisqartirish vazifasidir.

O‘zbek, turk va uyg‘ur tillarida stem va lemmaga qarash turlicha. O‘zbek tilida lemma tub yoki yasama so‘z shaklida bo‘ladi: *kitob, kitobxon, bilim, bilimdon.* Demak, o‘zbek tilida lemma lug‘atda mavjud leksemaga teng keladi. O‘zbek tilida o‘zakdosh (asosdosh) so‘zlar alohida-alohida lemma sanaladi.

O‘zbek tilida stemmingni amalga oshirish uchun so‘zshakldagi o‘zakkacha bo‘lgan barcha qo‘shimchalar kesib tashlanadi. ***Maktab****+dosh+lar+imiz* so‘zshaklida so‘z yasovchi va shakl yasovchi qo‘shimcha mavjud. O‘zbek tilida stemming jarayonida shu qo‘shimchalarning barchasi kesib tashlanadi:

So‘zshakl: ***maktab****+{dosh}+(lar)+(imiz)*

Lemma: ***maktabdosh***

Stem: ***maktab***

O‘zak (root): ***maktab***

Turk tilida stemlash jarayonida so‘zshakldagi faqat sintaktik va lug‘aviy shakl yasovchi qo‘shimchalar kesiladi, ammo so‘z yasovchilar qoldiriladi. Masalan:

So‘zshakl: ***seçim****+(ler)+(e)*

Stem: seçim

Ko‘rinadiki, turk tilida stem tarkibida so‘z yasovchi qo‘shimcha qoladi, o‘zak bilan stemning farqi so‘z yasovchi qo‘shimchaning mavjudligidadir.

So‘zshakl: ***seçim****+(ler)+(e)*

Lemma: ***seçim***

Stem: ***seçim***

O‘zak (root): ***seç***

Uyg‘ur tilida stemlash jarayonida so‘zshakldagi sintaktik va lug‘aviy shakl yasovchi qo‘shimchalar kesiladi, ammo so‘z yasovchilar qoldiriladi.

oqutquchi
So‘zshakl: ***oqut*** + *(qu) + (chi)*
Lemma: ***oqut***
Stem: ***oqut***
O‘zak (root): ***o***

STEMMING JARAYONIDAGI MUAMMOLAR

Stemming jarayonida 3 turdagi muammoni qiyosiy tahlil qilamiz. Bular:

1) o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi;
2) so‘zning tovush o‘zgarishiga uchrashi;
3) neologizm va NERlarni stemmlash.

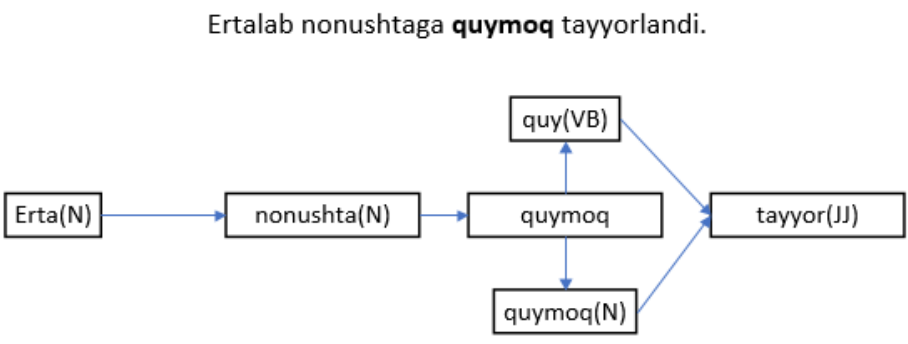
O‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi

An’anaviy stemming usul (algoritm)lari – bu qo‘shimchalar va ba‘zi morfologik qoidalarga asoslangan bo‘lib, stemming jarayoni natijasida stemdagi noaniqlik yuzaga kelishi mumkin. Ko‘p ma‘noli o‘zakni aniqlash ancha murakkab jarayon bo‘lib, stemming jarayonida gap darajasidagi semantik ma‘lumotlar e’tiborga olinmaydi. Ba‘zida so‘zning POS tegi uning o‘zagining POS tegi bilan bir xil bo‘lmasligi mumkin.

2-jadval. O‘zbek, turk va uyg‘ur tillaridagi so‘zshaklning stemi va qo‘shimchalari

Til	So‘zshakl	1-ma‘nosi	2-ma‘nosi
Turk	gelecek	keladi (will come)	kelajak (future)
Uyg‘ur	alma	ol+ma (don’t take)	olma (apple)
O‘zbek	quymoq	quy+moq (pour)	quymoq (panke)

O‘zbek tilidagi gaplarda stemdagi noaniqlikni quyidagi 3-rasmda ko‘rish mumkin:



3-rasm. O‘zbek tilidagi gaplarda stemdagi noaniqlik

Turk tilida:
1-ma‘noda: Kış yine **gelecek.**
2-ma‘noda: **Gelecek** hakkında ne düşünüyorsunuz?
Uyg‘ur tilida:
1-ma‘noda: Qalamni qolunga **alma.**
2-ma‘noda: U bazardin **alma** setiwaldi.
O‘zbek tilida: (quymoq)
1-ma‘noda: Zarifa mehmonlarga choy **quymoqchi** bo‘ldi.
2-ma‘noda: Ertalab nonushtaga **quymoq** tayyorlandi.

Misol uchun, koyun (turkcha) so‘zini agar gapda fe’l sifatida kelsa koy-(mak) ko‘rinishida stemmlash mumkin. Agar gapda ot sifatida kelsa koyun (qo‘y) ko‘rinishida stemmlash lozim. POS teglash jarayonida turli muammolar yuzaga kelishi mumkin. Ulardan biri POS teglashdagi noaniqlikdir. So‘zlar gapdagi sintaktik roliga qarab turli so‘z turkumlarga mansub bo‘lishi mumkin. So‘zning aniq/to‘g‘ri POS tegi uning o‘zagini ham topishga yordam beradi.

Misol uchun,

1. *Aydınlık gelecek günler bizi bekliyor.* (Kelajakda bizni yorqin kunlar kutmoqda).

2. *Ahmet birazdan gelecek.* (Ahmad tez orada keladi);

Birinchi gapdagi *gelecek – sifatdosh,* o‘zak esa **gelecek (kelajak)** bo‘ladi. Ikkinchi gapda *gelecek – fe’l,* o‘zak esa **gel-(mek) (kelmoq)**. Yuqoridagi fikr mulohazalardan, POS teglash jarayoni stemmingda muhim rol ekanligini qayd etish mumkin.

Uyg‘ur tilida ham xuddi shunga o‘xshash holatni kuzatishimiz mumkin. Masalan *alma* so‘zi olma mevasi ma‘nosida olma shaklida stemmlash, *ol-ma* fe’l sifatida esa olmoq shaklida stemmlanadi. Stemmlashda so‘z shakllardagi POS teglashdagi farqni kelgüsi so‘zida ham kuzatish mumkin.

1. *Kelgüsi ishimni planladim*. (Kelajak ishlarimni reja qildim.)
2. *Bala ete kelgüsi*. (Bola ertaga keladi.)

Birinchi gapda *kelgüsi* – sifat, o‘zak esa kelgüsi (kelajak) bo‘ladi. Ikkinchi gapda *kelgüsi* – kelasi zamon shaklidagi fe’l, o‘zak esa kel-(mek) (kelmoq) shaklidadir.

O‘zbek tilida o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi va POS teglash, stemmini aniqlashdagi murakkabliklarni ko‘plab misollarda kuzatish mumkin. Misol uchun, *tortma, olma, yozma, o‘sma* va hakozo so‘zshakllarda. Bu so‘zlar tortma – tort-(moq), olma-ol-(moq), yozma-yoz-(moq), o‘sma-o‘s-(moq) stemmlari shaklida bo‘lib, POS tegi ot va fe’l deb belgilanadi. Masalan:

1. *Sen bozordan kitob olma.*
2. *Akbar kecha olma yedi.*

Bu yerda birinchi gapda *olma* – inkor ma‘nosidagi fe’l, o‘zak ol-(moq) shaklida bo‘lsa, ikkinchi gapda *olma* – ot, o‘zak ham olma bo‘ladi.

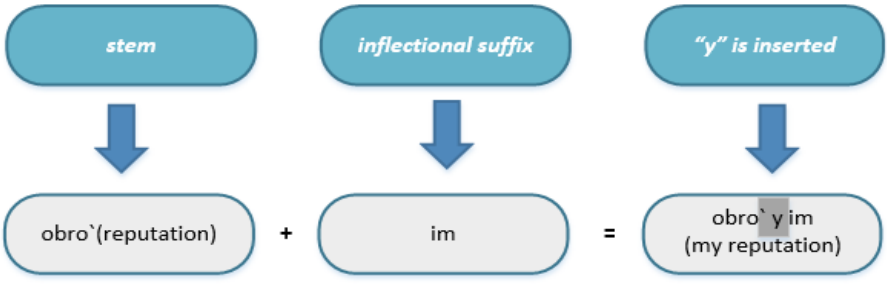
Yuqoridagi fikr mulohazalardan bilish mumkinki, uchala turkiy tilda ham o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi holati uchraydi va bu vaziyatda POS teglash jarayoni stemmingda muhim rol ekanligini qayd etish mumkin.

So‘zning tovush o‘zgarishiga uchrashi

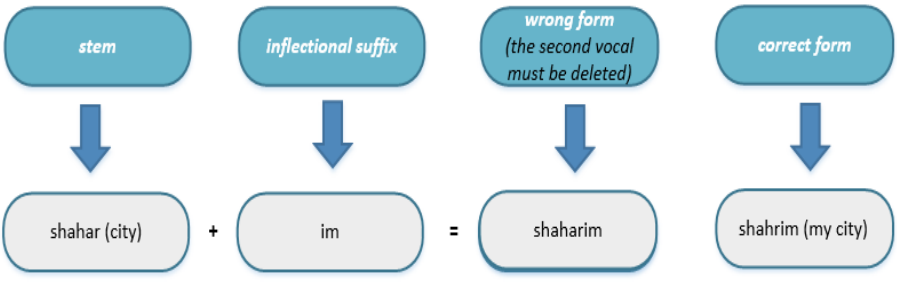
Shakl yasovchi qo‘shimchalarni o‘zakning oxirgi harflariga qo‘shish natijasida ba‘zi hollarda so‘zda fonetik o‘zgarishlar yuzaga kelishi mumkin [Tsygankin, Ivanova, 2019, Mirtojiyev, 2013]. Agglutinativ tillarda so‘zda *tovush ortishi, tushishi* va *almashinishi* kabi uch xil fonetik o‘zgarishlar amalga oshirilishi mumkin (5-jadval).

3-jadval. Agglyutinativ tillarida stemming jarayonidagi kamchiliklar

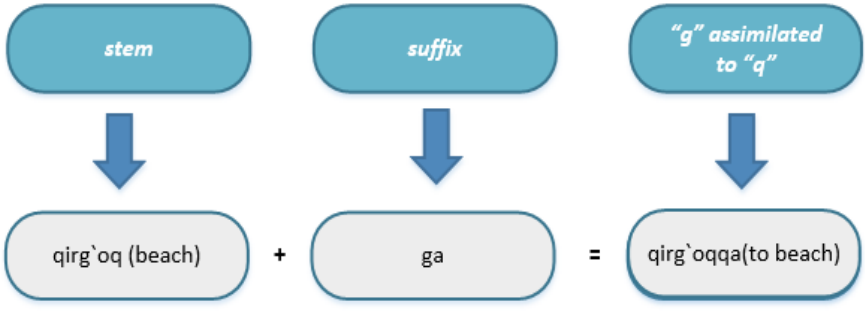
O'zbek		**Turk**		**Uyg'ur**	
to‘g‘ri	**xato**	**to‘g‘ri**	**xato**	**to‘g‘ri**	**xato**
lavozim+ida	boshlig‘+i	yara+landi-g'ini	ögre-nlere	yoğ+an	binay+im
ish+lagan	san+aydi	belirt+ti	jandar+ma-lig'ini	eshek+-medeği	oghl+um
hafta+larida	tarog‘+ini	ara+sinda	rastla+diği	chaplish+ivalidu	yot+im
bo‘lim+i	me+ning	koşul+lar-da	iznin+e	bash+lapti	yurag+im
hokim+ining	obro'y+imiz	gösteri+ci-nin	geti+riliyor	dep+ti	shahr+im
ish+lagan	achch+iq	belir+li		ini+sinin	



4-rasm. So‘zda tovush ortishi



5-rasm. So‘zda tovush tushishi



6-rasm. So‘zda tovush almashinishi

Stemni aniqlashdagi tovush o‘zgarishiga uchrashi muammosini hal qilish uchun, birinchi bosqichda o‘zak va qo‘shimchalarning chegaralari aniqlanadi, ikkinchi bosqichda esa lemmatizatsiya amalga oshiriladi. Lemmatizatsiya natijasida xato hosil qilingan stemlar lug‘atda mavjud root (o‘zak)ga o‘zgartiriladi.

Neologizm va NERlarni stemmlash

NERlarni stemlash muammolari

-lik qo‘shimchasi, asosan ot, sifat va ravish turkumiga oid so‘zlardan ot yasaydi. Ot turkumiga oid leksemalardan yasalgan so‘zlar yasovchi asos anglatgan narsa-predmetning xususiyati bilan bog‘liq holda turli ma’nolarni bildiradi:

1) shaxs (inson) bildiradigan so‘zlardan yasalganda: qarin-

doshlik, *(otalik, onalik, tog‘alik, o‘g‘illik, farzandlik, erlik, xotinlik)*; umrning ma’lum davrini bildiradigan so‘zdan yasalgan otlar (*bolalik, yigitlik, qizlik, o‘smirlik, kelinlik, kuyovlik*); kasb, amal-unvon egasini anglatuvchi so‘zlardan yasalgan otlar (*mudirlik, o‘qituvchilik, qassoblik, chorvadorlik, tabiblik, suvchilik, sartaroshlik, savdogarlik, rassomlik, shofyorlik, aktyorlik*);

2) asosdan anglashilgan narsa ishg‘ol etgan obyektni bildiruvchi ot (*botqoqlik, qumlik, muzlik*);

3) yer sathining yasovchi asosdan anglashilgan qismini bildiruvchi ot (*jarlik, do‘nglik, qiyalik, pastlik, ichkarilik, yalanglik*).

Sifat va ravishlarga qo‘shilib, belgi oti yasaydi: *qizillik, semizlik, xursandlik, aniqlik*. Bu kabi holatlarda ularning tarkibidagi so‘z yasovchi qo‘shimchalar kesiladi, qolgan qism stem sanaladi.

Ammo joy nomini bildiruvchi atoqli otlarga ***–lik*** qo‘shimchasi qo‘shilganda, ular turdosh otga aylanadi va kichik harf bilan yoziladi: *samarqandlik, buxorolik, amerikalik, o‘zbekistonlik, turkiyalik, arabistonlik.* Bunday holatda ***–lik*** qo‘shimchasi kesiladi, qolgan qism stem deb tushuniladi, bosh harfga aylantirilib, NER sifatida tan olinadi.

samarqandlik = Samarqand~~lik~~

amerikalik = Amerika~~lik~~

kanadalik = Kanada~~lik~~

NERlarning stemini topish muammosi yuzaga kelganda shakl yasovchilar kesiladi, so‘z yasovchi shaklidagi qo‘shimcha yoki so‘zning qismi qoldiriladi, shu qism NER sanaladi: *O‘zbekistondan* so‘zshaklining stemi *O‘zbekiston*.

Shunday qo‘shimchalar borki, ular so‘z yasovchi va shakl yasovchi vazifasida keladi (5-jadval).

5-jadval. So‘z yasovchi va shakl yasovchi omonim qo‘shimchalar

Shakl yasovchi va so‘z yasovchi qo‘shimchalar		
-ay	-k	-chak
-gi	-ka	-chiq
-da	-kin	-choq
-i	-la	-qa
-in	-lab	-qin
-im	-m	-sa
-ir	-ma	-siz
-iq	-moq	-xon
-y	-cha	

Bunday qo‘shimchalar bosh harf bilan yozilgan so‘zlarning tarkibida kelganda, shakl yasovchilar va so‘z yasovchilar tarkibida bo‘lsa, so‘z shakl tarkibida qoldiriladi va shu shaklida stem deb olinadi. Masalan, *Jon Kennedi* so‘zning tarkibida –***i*** harfi bor. Dastur o‘zakni bilmaganligi sababli, ya’ni mazkur so‘z o‘zbek tili lug‘atida mavjud emasligi sababli o‘zakni ajratolmay qoladi, natijada –**i** qo‘shimchasini kesib, ***Kenned*** so‘zini o‘zak deb olishi mumkin. Bunday holatdan qochish maqsadida shakl yasovchi va so‘z yasovchilar orasida omonimiya hosil qiluvchi qo‘shimcha bilan shakldosh bo‘lgan har qanday birlik so‘zshakl tarkibida qoldiriladi.

Neologizmlarni stemlash muammolari

Neologizm yun. “neos” — yangi, “logos” — so‘z — jamiyat taraqqiyoti, hayotning talab-ehtiyoji bilan paydo bo‘lgan yangi narsa va tushunchalarni ifodalovchi so‘zlar. Neologizmlarning yangiligi dastlab paydo bo‘lgan vaqtlardagina sezilib turadi: vaqt o‘tgach, ular “yangilik” xususiyatini yo‘qotib, odatda, faol so‘zlar qatoriga o‘tadi.

Neologizmning shakliy neologizm, semantik neologizm, funksional neologizm, ijtimoy neologizm, texnologik neologizm, stilistik neologizm kabi turlari mavjud [Qo‘ziboyeva, 2022].

Neologizmlarning paydo bo‘lish yo‘llari xilma-xil bo‘lib, ular tilning mavjud lug‘aviy tarkibi va grammatik qonun-qoidalari asosida yangi so‘z yasash yo‘li, shuningdek, mavjud so‘zning lug‘aviy ma’nolaridan birini yangi ma’noda qo‘llash yo‘li bilan va boshqa tildan so‘z qabul qilish orqali hosil qilinadi.

Neologizmlar tarkibida –izm (neologism), -ik (dalton**ik**), -la (gul**la**sh) kabi qo‘shimchalar uchraydi.

Neologizmlar lug‘atda mavjud bo‘lmaganligi sababli ularni stemlashda muammolar yuzaga chiqadi. Ularning tarkibidagi qo‘shimchalar, so‘zning bir qismining qo‘shimchaga o‘xshab qolishi muammolari shular shumlasidan. Bunday holatda shakl yasovchi qo‘shimchalar bazasida mavjud qo‘shimchalar kesiladi. Qolgan qism stemga teng keladi. O‘zbek tilidagi neologizmlar va NERlarga mos stem yuqoridagi 2-rasmda keltirilgan turk va uyg‘ur tillaridagi stem ta’rifiga mos keladi.

Neologizmlar keng jamoatchilik tomonidan ma’lum vaqt oralig‘ida faol qo‘llanib og‘zaki va yozma nutqqa ko‘chganda tilning leksik boyligi sifatida rasman e’tirof etilishi, ya’ni lug‘atga kiritilishi mumkin. Bu jarayondan so‘ng ularni stemlash lug‘atdagi so‘zlarni stemlash qoidasi asosida amalga oshiriladi.

POS TEGGING VA STEMMING MUAMMOLARINING O‘RGANILISHI

Avvalo POS teglash va stemming vazifalarini ikkita alohida vazifa sifatida mustaqil ravishda ko‘rib chiqamiz. So‘ngra, POS teglash vazifasini morfologik segmentatsiya kabi boshqa vazifalar bilan birlashtirgan POS tegining qo‘shma modellarini ko‘rib chiqamiz.

POS teglashga oid tadqiqotlar

Korpus matnlarini POS teglash NLPdagi klasterlash muammosi sifatida keng tarqalgan. Brown so‘zlarning sintaktik sinflarini o‘rganish uchun *murakkab iyerarxik klasterlash* algoritmi asosida klasslarga asoslangan n**-gram modelini** taqdim etgan [Brown, Della Pietre de Souza, 1992]. Tadqiqotda kontekstli ma‘lumotlar n-gramm shaklida kiritilgan bo‘lib, boshlang‘ich holatda har bir so‘z bitta sinfga mansub bo‘ladi. So‘ngra, o‘rtacha minimal yo‘qotishni beradigan har bir klaster juftligi barcha klasterlar bitta klaster ostida birlashtirilgunga qadar umumlashtiriladi. Keyingi qadamda, sintaktik kategoriyalar orasidagi iyerarxiyani ifodalovchi binar daraxt shakllantiriladi.

Schutze har bir so‘zning ikkita chap va ikkita o‘ng qo‘shni so‘zidan olingan so‘z vektorlari tomonidan tuzilgan kontekst matritsasining o‘lchamini kamaytirish uchun **Singular Value Decomposition (SVD)**dan foydalangan [Schutze, 1993]. Keyingi qadamda kontekstli ma‘lumotlardan foydalangan holda so‘zlarni klasterlash uchun Buckshot klasteri qo‘llanilgan [Cutting, Kupiec, Pedersen, Sibun, 1992].

Biemann to‘rtta so‘zli kontekst oynalari va eng ko‘p uchraydigan so‘zlarni ishlatib, “Chinese Whispers” grafli klasterlash algoritmini qo‘llagan [Biemann. 2006].

Ba‘zi boshqa yondashuvlarda POS teglash amali gapdagi so‘zlarni ketma-ketlikdagi belgilash/teglash muammosi sifatida qaraladi. Bunday turdagi yondashuvlar asosidagi algoritmlarda ko‘p hollarda Yashirin Markov Modeli (Hidden Markov Models, HMMs) dan foydalaniladi.

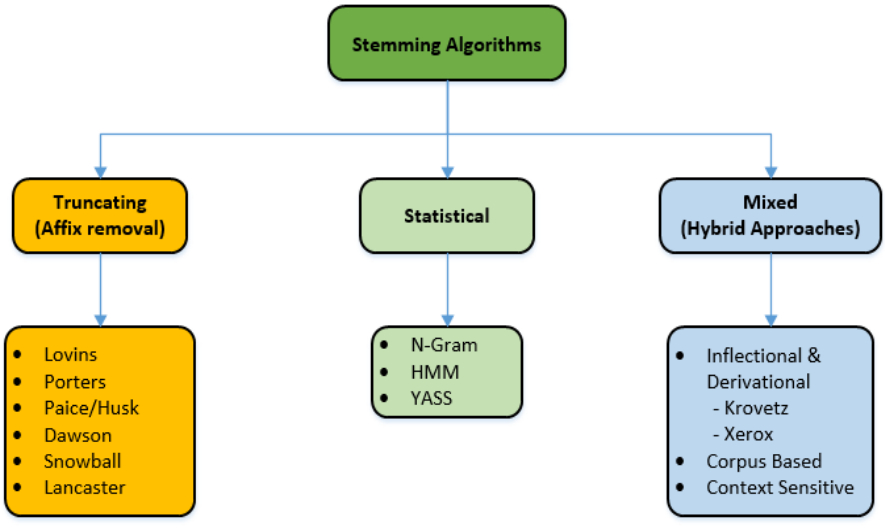
Merialdo uch sinfdan iborat Markov modelini taqdim etgan bo‘lib, tadqiqotda korpusdagi o‘quv ma‘lumotlarining turli o‘lchamlari uchun turli parametrlarni baholash usullari taqqoslangan [Merialdo, 1994]. Korpus ma‘lumotlari uchun nisbiy chastotali trening ishlatilgan va tegsiz ma‘lumotlar uchun **Maksimal ehtimollik (Maximum Likelihood)** usuli qo‘llanilgan.

Banko va Moore har bir so‘zni faqat joriy so‘z “teg”idan emas, balki oldingi va keyingi so‘z “teg”lari (valentlik)ni o‘z ichiga olgan uchta qo‘shni teg asosida kontekstli HMM teggerini taqdim etadilar [Banko, Moore, 2004]. Ushbu model asosiy HMM bilan solishtirganda ko‘proq kontekstli ma‘lumotlarni o‘z ichiga olgan va samarali natija qaytargan.

Jonson **HMMga asoslangan POS teglash**da ishlatiladigan turli parametrlarni solishtirgan. Shu maqsadda **Expectation Maximization (EM), Variatsion Bayes** va **Gibbs** namunalaridan foydalangan [Haghighi, Klein, 2006]. Tadqiqot Gibbs namunasi va variatsion Bayes baholovchisi bilan solishtirganda **EM** algoritmining past samaradorligini ko‘rsatib bergan.

Stemmingga oid tadqiqotlar

Zamonaviy stemming algoritmlari odatda uchta sinfga bo‘linadi: *qoidaga asoslangan, statistik* va *gibrid* algoritmlar (7-rasm). Qoidalarga asoslangan stemmerlar avtomatik bo‘lmagan qoidalardan foydalangan holda stemmlarni aniqlashga qaratilgan. Qoidalarga asoslangan ommabop stemmerlar sifatida Lovins [Lovins, 1968], Porter [Porter, 2006; Porter, 2001] va Krovets [Krovetz, 2000]larni keltirish mumkin. Qoidalarga asoslangan stemmlash algoritmlari odatda nazorat qilinadi.



7-rasm. Stemming algoritmlarining tasnifi

Statistik stemmlash algoritmlari stemmlarni o‘rganish uchun statistik usullardan foydalanadi. Xu va Croft [Xu, Croft, 1998], Porter stemmerining [Porter, 2006] kamchiliklarini bartaraf etish uchun tasodifiy yuzaga keladigan statistik so‘zdan foydalanadigan usulni

taqdim etishgan. Tasodifiy statistik ma‘lumotlariga asoslanib, ular Porter stemmeri tomonidan yaratilgan sinflar sonini kamaytirish uchun *grafni qismlarga ajratish algoritmini* amalda qo‘llashgan [Porter, 2006].

Gibrid stemmlash algoritmlari qoidalarga asoslangan va statistik usullarni yagona tizimga birlashtiradi. Ba‘zi gibrid stemmlash algoritmlari Shrivastava [Manish, Nitin, Bibhuti], Goweder [Goweder, Alhami] va Adam [Adam, 2010] tomonlaridan ishlab chiqilgan.

Goldsmith [Goldsmith, 2001; Goldsmith, 2006], minimal tavsif uzunligi (Minimum Description Length, MDL) tamoyiliga asoslangan, nazoratsiz stemming modelini taklif qilgan. Model, asosan morfologik segmentatsiya uchun mo‘ljallangan bo‘lib, undan stemmer sifatida ham foydalanish mumkin. Har bir so‘zdagi segmentatsiya nuqtalari, korpusning umumiy hajmini qisqartirish uchun mo‘ljallangan.

Graflarga asoslangan stemming algoritmi Bacchin tomonidan taklif qilingan [Bacchin, Ferro, Melucci, 2002]. Algoritm satr ostilar to‘plamini aniqlash uchun birinchi bosqichda har bir so‘zni barcha mumkin bo‘lgan bo‘linish nuqtalariga ajratadi. Ikkinchi bosqichda satr ostilar to‘plamidan foydalangan holda yo‘naltirilgan graf hosil qilinadi. Nihoyat grafdagi satr ostidagilar chastotasiga qarab prefiks va suffiks ballari hisoblanib, stem aniqlanadi.

Melucci va Orio **HMM asosidagi stemmer**ni taqdim etadilar [Melucci, Orio, 2003]. HMM holatlari prefiks va suffikslarga mos kelib, holatlar orasidagi o‘tishlar grammatik qoidalarga mos keladi. Parametrlarni baholash uchun **Expectation Maximization (EM)** algoritmidan foydalanilgan. Parametrlar baholangandan so‘ng, maksimal ehtimollikka ega yo‘nalish bo‘yicha segmentatsiya amalga oshirilgan va stem aniqlangan.

McNamee va Mayfield **n-grammlarga asoslangan muqobil stemmlash algoritmi**ni taqdim etishgan [McNamee, Mayfield, 2004]. Har bir so‘z uchun korpus asosida barcha bigramma va trigrammalar generatsiya qilingan bo‘lib, o‘xshash so‘zlar n-grammning asosiy qismini tashkil etgan.

Bacchin graflar asosidagi stemmer modelini kengaytirib [Baschin, Ferro, Melucci, 2005], dastlabki qadamda har bir so‘zni barcha pozitsiyalarga bo‘lish orqali mumkin bo‘lgan satr ostilar to‘plami aniqlangan. So‘ngra, satr ostilarni ifodalaydigan yo‘nalishli graf shakllantirilgan. Agar z so‘zi, z = xy ni qanoatlantirsa, x va y tugunlar orasiga yo‘nalishli qirra hosil qilingan. Affiks ballarini

baholash **HITS** algoritmi [Kleinberg, Kumar, Raghavan, 1999] asosida hisoblangan. Prefiks va sufiks ballari asosida so‘zlarga tegishli bo‘lgan prefiks va qo‘shimchalar juftlarining ehtimolini maksimal darajaga oshirish orqali eng katta ehtimoliy bo‘linish nuqtasi aniqlangan.

Majumder tomonidan **YASS** (Yet Another Suffix Striper) [Majumder, Mitra, Parui, 2007] deb nomlangan stemmlash algoritmi taqdim etilgan bo‘lib, u satrlar orasidagi masofa o‘lchovidan foydalanadigan klasterlash algoritmiga asoslangan. Satrlar orasidagi masofa o‘lchovi so‘zlar orasidagi morfologik o‘xshashlikni baholash uchun ishlatilgan.

So‘zlar orasidagi o‘xshashlikni aniqlashga asoslangan stemmer Peng tomonidan ishlab chiqilgan [Peng, Ahmed, Li, Lu, 2007]. Ushbu stemmer qidiruv tizimi natijalarini yaxshilash uchun IR vazifalari uchun qo‘llanilgan va yuqori samaradorlikni bergan.

Paik tomonidan graflarga asoslangan **GRAS** (GRAph-based Stemmer) stemmeri ishlab chiqilgan bo‘lib, u so‘zlarni guruhlash uchun leksik ma‘lumotlardan foydalanadigan statistik stemmer hisoblanadi. Ushbu algortimda so‘zlar grafning tugunlari sifatida ifodalangan. Algoritm grafni parchalash orqali so‘zlar o‘rtasidagi bog‘lanishni aniqlaydi.

Brychcin va Konopik tomonidan taklif etilgan yuqori aniqlikdagi stemmer (High Precision Stemmer, HPS)da imlo va semantik ma‘lumotlardan so‘zlarni o‘zak va qo‘shimchalarga bo‘lish xususiyati sifatida foydalanilgan. Usul ikki bosqichdan iborat:

- orfografik va semantik jihatdan o‘xshash so‘zlar maksimal o‘zaro ma‘lumot (Maximum Mutual Information, MMI) yordamida klasterlash;

- birinchi bosqichdan olingan klasterlar yordamida maksimal entropiya tasniflagichini amalga oshirish.

Turk tilidagi stemmingni amalga oshirish usuli Köksal tomonidan kiritilgan [Brown, Della Pietra, 1992]. Ushbu usul dastlabki 5-6 harfni o‘zak deb hisoblashga asoslangan. Kut va boshqalar o‘z tadqiqotlarida L-M (Longest Match) nomli usulni ishlab chiqqanlar [Schutze, 1993]. So‘z o‘zaklari va ularning mumkin bo‘lgan shakllarini o‘z ichiga olgan lug‘atdan foydalangan holda, usul chapdan o‘ngga harflarni lug‘atda joylashgan so‘zlar bilan o‘zak so‘zni taqqoslaydi. Eng uzun mos kelgan so‘z o‘zak hisoblanadi.

Solak va Can [Paice, 1994] stemlarni aniqlashda ildizlar

lug‘atidan foydalanganlar. Har bir ildiz chapdan o‘ngga stem hosil qilish usullariga mos keladigan 64 xususiyatga ega deb qayd etilgan. Harf birliklari ildiz leksikasiga chapdan o‘ngga tartibda moslashtirilgan va agar mos stem aniqlansa, tizim qo‘shimcha qoidalar asosida mumkin bo‘lgan stemlarni aniqlaydi. **AF algoritmi** deb ataladigan ushbu tadqiqot, asosan, Oflazer [Banko, Moore, 2004] tomonidan ishlab chiqilgan morfologik tahlil usulining variant hisoblanadi.

FindStem - bu Sever va Bitirim [Goldwater, Griffiths, 2007] tomonidan ishlab chiqilgan stemming usuli bo‘lib, asosan uchta amalni o‘z ichiga oladi: *ildizni aniqlash, o‘zakni morfologik tahlil qilish va aniqlash.* Usul so‘zlarning morfologik va POS xususiyatlarini, sintaktik qoidalarini o‘z ichiga olgan lug‘atdan foydalanadi. Sever va Bitirimning ta‘kidlashicha, FindStem algoritmi AF va L-M algoritmlariga qaraganda yaxshiroq va samaraliroq ishlaydi.

Turk tilidagi so‘zlarning o‘zagini aniqlashga oid boshqa tahliliy usullarga Akin [Cutting, Kupiec, Pedersen, Sibun, 1992] tomonidan ishlab chiqilgan **“zemberek”** va Childen [Haghighe, Klein, 2006] tomonidan ishlab chiqilgan **“snowball”** algoritmlarini keltirish mumkin. Shuningdek, Dincher [Merialdo, 1994] o‘zak va qo‘shimchalar orasidagi chegarani n-gramm statistikasidan foydalangan holda hal qilish usulini taklif qilgan. Ushbu tadqiqot amaliyotga qo‘llanilishi natijasida samaradorlik 95,8% ni tashkil qilgan.

Batuer Aisha va Maoshistal Sun statistik yondashuvga asoslangan tokenizatsiya ishlab chiqdilar [Aisha, Sun, 2009]. Ushbu tokenizatsiya **UC uyg‘ur tili denoted korpusi** ustiga qurilgan. UC korpusi **594172 ta** uyg‘ur tili so‘zlari va UC ikkita korpusdan tashkil topgan, ya‘ni UCS denoted qo‘lda **stemlangan** va **lemmalangan** korpusi. Keyinchalik uyg‘ur tili tokenizatsiyasi uchun ikki bosqichli jarayondan foydalandilar.

Aishan Vumaier va boshqa tadqiqotchilar 2009 - yilda yangi uyg‘ur tili ot so‘z turkumi stemming usulni ishlab chiqdilar[Maimaiti, Wumair, 2017]. Uyg‘ur ot so‘z turkumi stemming usulini 2 bosqichda amalga oshirdilar:

- Uyg‘ur tili **FSM** qo‘shimchalari yordamida;
- Uyg‘ur tili **FSM** qo‘shimchalari bilan yuzaga kelgan noaniqliklarni bartaraf qilish uchun **CRF** metodi yordamida.

Birinchi bosqichda, uyg‘ur tili ot so‘z turkumi stemming jarayonini FSM ot qo‘shimchalari yordamida ishlab chiqdilar. Stemming jarayoni **55625** kiritilgan so‘z ustida amalga oshirilib,

natijada **6239** noto‘g‘ri over stemming keltiruvchi so‘zlar aniqlandi. Ikkinchi bosqichda, stemming jarayonida yuzaga kelgan noaniqliklarni CRF metodi yordamida aniqlab, **55125** ta so‘zli korpus qurildi. Korpus **17317** ta noaniq qo‘shimchali so‘zlar, **6239** ta correctsiz qo‘shimchalar va **11078** ta correct qo‘shimchali so‘zlardan tashkil topgan. Algoritm natijasi shuni ko‘rsatadiku, jarayonda FSM qo‘shimchalaridan foydalanilsa eslatma (recall rate) **88.78%**ni, FSM va CRFdan foydalanilsa eslatma (recall rate) **94,04%** aniqlikka erishildi. Xulosa o‘rnida CRF usulidan foydalanish qayta tiklash stavkasini **5,26%** yaxshilaydi.

2012-yilda Azragul, Qixiangjwei va Yusupulla Uyg‘ur tili stemmerini ishlab chiqdilar [Azragul, Qi, Yusup, 2012]. Ular lug‘atga asoslangan usuldan foydalanganlar. Algoritm ishlash jarayonida kiritilgan so‘z stem lug‘atidan qidiriladi. Bunda, qo‘shimchalar lug‘ati yordamida so‘z ajratiladi va qo‘shimchalarni o‘chirish bilan ajratilgan so‘z nomzod so‘zi lug‘atdan izlanadi.

Amalga oshirilgan tadqiqotlar shuni ko‘rsatadiki, oldingi tadqiqotlarda to‘liq bo‘lmagan lug‘atdan (ochiq lug‘at) foydalanilgan va stemming natijasida hosil bo‘lgan noqniqliklar keyinchalik boshqa usullar orqali hal qilingan.

XULOSA

Lug‘at orqali POS teglash va stemmingni amalga oshirish tabiiy tilni qayta ishlashning ko‘plab vazifalariga qiyinchilik tug‘diradi. POS teglash va stemlash uchun til korpusidan foydalanish orqali lug‘at bilan bo‘ladigan muammolar hal qilinadi. Til korpusi ustida o‘tkazilgan turli tajribalar shuni ko‘rsatadiki, o‘zak ma‘lumotlarini sintaktik vazifa bilan birlashtirish morfologik jihatdan boy til uchun POS teglash natijasini yaxshilaydi, bu esa NLP vazifasining hal qilish samaradorligini oshirishga xizmat qiladi. Maqolada korpusda turli xil bog‘liqliklarni qabul qiladigan bir nechta turli xil qo‘shma modellar taqdim etildi. Umumiy eksperimental natijalar shuni ko‘rsatadiki, neural word embeddingsdan foydalanadigan Bayes HMM modeli POS teglash vazifasi uchun boshqa modellardan ustundir. Shuningdek, flektiv morfologiyani aniqlash uchun o‘zak va so‘zlar o‘rtasidagi semantik o‘xshashlikdan foydalanishda flektiv qo‘shimchalar so‘zning ma‘nosini o‘zgartirmaydi. Shu maqsadda word2vecdan olingan neural word embeddings metodidan foydalanish lozim. Natijalar shuni ko‘rsatadiki, semantik ma‘lumotlardan foydalanish stemlash va POS teglashni sezilarli darajada yaxshilaydi.

Foydalanilgan adabiyotlar

Paice, C. D. (1994). An evaluation method for stemming algorithms. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994. https://doi.org/10.1007/978-1-4471-2099-5_5

Anjali, M., & Jivani, G. (2011). A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl., 2(6).

Gao, J., & Johnson, M. (2008). A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL. https://doi.org/10.3115/1613715.1613761

Goldwater, S., & Griffiths, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.

van Gael, J., Vlachos, A., & Ghahramani, Z. (2009). The infinite HMM for unsupervised PoS tagging. EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009. https://doi.org/10.3115/1699571.1699601

Taner Dinçer, B., & Karaoğlan, B. (2003). Stemming in agglutinative languages: A probabilistic stemmer for Turkish. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2869. https://doi.org/10.1007/978-3-540-39737-3_31

Ҳожиев А. Ўзбек тилида сўз ясалиши. – Ташкент, 2005.

Porter, M. F. (2006). An algorithm for suffix stripping. Program, 40(3). https://doi.org/10.1108/00330330610681286

https://www.turkedebiyati.org/yapim-ekleri/

https://tilachar.ru/ru/grammar/24-grammar

Polus, M. E., & Abbas, T. (2021). Development For Performance Of Porter Stemmer Algorithm. Eastern-European Journal of Enterprise Technologies, 1(2(109)). https://doi.org/10.15587/1729-4061.2021.225362

Memon, S., Mallah, G. A., Memon, K. N., Shaikh, A., Aasoori, S. K., & Dehraj, F. U. H. (2020). Comparative study of truncating and statistical stemming algorithms. International Journal of Advanced Computer Science and Applications, 2. https://doi.org/10.14569/ijacsa.2020.0110272

Khyani, D., S, S. B., M, N. N., & M, D. B. (2021). An Interpretation of

Lemmatization and Stemming in Natural Language Processing. Journal of University of Shanghai for Science and Technology, 22(10).

Tsygankin, D. v., & Ivanova, G. S. (2019). Assimilative potential of vowel harmony in languages of agglutinative type. Bulletin of Ugric Studies, 9(3). https://doi.org/10.30624/2220-4156-2019-9-3-510-518

Миртожиев М.М. Ўзбек тили фонетикаси. – Тошкент, Фан нашриёти, 2013. – 424 бет.

P.~Brown, V.~Della Pietra, de Souza, P., J.~Lai, & R.~Mercer. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18.

Schütze, H. (1993). Part-of-speech induction from scratch. https://doi.org/10.3115/981574.981608

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A practical part-of-speech tagger. https://doi.org/10.3115/974499.974523

Biemann, C. (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop. https://doi.org/10.3115/1557856.1557859

Merialdo, B. (1994). Tagging English text with a probabilistic model. Computational Linguistics, 20(2).

Banko, M., & Moore, R. C. (2004). Part of speech tagging in context. COLING 2004 - Proceedings of the 20th International Conference on Computational Linguistics. https://doi.org/10.3115/1220355.1220435

Haghighi, A., & Klein, D. (2006). Prototype-driven learning for sequence models. HLT-NAACL 2006 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference. https://doi.org/10.3115/1220835.1220876

Lovins, J. B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(June).

Porter M.F. “Snowball: A language for stemming algorithms”. 2001.

Krovetz, R. (2000). Viewing morphology as an inference process. Artificial Intelligence, 118(1–2). https://doi.org/10.1016/S0004-3702(99)00101-0

Xu, J., & Croft, W. B. (1998). Corpus-based stemming using cooccurrence of word variants. ACM Transactions on Information Systems, 16(1). https://doi.org/10.1145/267954.267957

Manish Sh., Nitin A., Bibhuti M. Morphology based natural language processing tools for indian languages. In Proceedings of the 4th Annual Inter Research Institute Student Seminar in Computer

Science, IIT, Kanpur, India, April. Citeseer.

A Goweder, H Alhami, Tarik Rashed, and A Al-Musrati. A hybrid method for stemming Arabic text. Journal of computer Science

Adam, G., Asimakis, K., Bouras, C., & Poulopoulos, V. (2010). An efficient mechanism for stemming and tagging: The case of Greek language. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6278 LNAI(PART 3). https://doi.org/10.1007/978-3-642-15393-8_44

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. Computational Linguistics, 27(2).
https://doi.org/10.1162/089120101750300490

Goldsmith, J. (2006). An algorithm for the unsupervised learning of morphology. Natural Language Engineering, 12(4).
https://doi.org/10.1017/S1351324905004055

Bacchin, M., Ferro, N., & Melucci, M. (2002). The effectiveness of a graph-based algorithm for stemming. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2555. https://doi.org/10.1007/3-540-36227-4_12

Melucci, M., & Orio, N. (2003). A novel method for stemmer generation based on Hidden Markov Models. International Conference on Information and Knowledge Management, Proceedings. https://doi.org/10.1145/956863.956889

McNamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. Information Retrieval, 7 (1–2). https://doi.org/10.1023/b:inrt.0000009441.78971.be

Bacchin, M., Ferro, N., & Melucci, M. (2005). A probabilistic model for stemmer generation. Information Processing and Management, 41(1). https://doi.org/10.1016/j.ipm.2004.04.006

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The web as a graph: Measurements, models,and methods. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1627. https://doi.org/10.1007/3-540-48686-0_1

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., & Datta, K. (2007). YASS: Yet another suffix stripper. ACM Transactions on Information Systems, 25(4). https://doi.org/10.1145/1281485.1281489

Peng, F., Ahmed, N., Li, X., & Lu, Y. (2007). Context sensitive stemming for web search. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07. https://doi.org/10.1145/1277741.1277851

Aisha, B., & Sun, M. (2009). A statistical method for Uyghur Tokeniza-

tion. 2009 International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2009.
https://doi.org/10.1109/NLPKE.2009.5313764

Maimaiti, M., Wumaier, A., Abiderexiti, K., & Yibulayin, T. (2017). Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging. Information (Switzerland), 8(4). https://doi.org/10.3390/info8040157

Azragul, X. Qi and A. Yusup, “Website Phrasal Survey Based Modern Uighur Stem Extraction and Application Study”, Computer Applications and Software, vol.29, no.3, (2012), pp.32-34.

Bölücü, N., & Can, B. (2019). Unsupervised joint PoS tagging and stemming for agglutinative languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 18(3). https://doi.org/10.1145/3292398

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2006). Distributed Representations of Words and Phrases and their Compositionality. Neural Information Processing Systems, 1.

https://uz.wikipedia.org/wiki/Neologizmlar

Qo‘ziboyeva G. Tilimizga kirib kelgan neologizmlar va ularning tahlili // International scientific journal. 2022, № 3. – 78-83-б.

POS TAGGING AND STEMMING IN UZBEK, TURKISH AND UYGHUR LANGUAGES

Botir Elov[1],
Shahlo Xamroyeva[2],
Oqila Abdullayeva[3],
Zilola Xusainova[4],
Nizomaddin Xudayberganov[5]

Abstract

Uzbek, Turkish and Uyghur languages are considered agglutinative languages and are complex in terms of morphology and number of word forms. In these languages, new words and word forms are formed by combining stems and suffixes. When additions are added to the root, both phonetic and morphological changes occur as a result of phonetic harmony and disharmony. This situation creates various problems that need to be solved in the process of POS tagging and stemming of word forms in the text. When solving many NLP tasks, it is necessary to reduce word forms to their root (stemming). Removing all inflectional affixes from a word and lemmatizing the rest of the word is considered one of the important tasks of

[1] *Elov Botir Boltayevich* – doctor of philosophy of technical sciences (PhD), associate professor. Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.
E-mail: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648

[2] *Hamroyeva Shahlo Mirdjanovna* – doctor of philological sciences, associate professor, etc. Alisher Navoi Tashkent State University of Uzbek Language and Literature.
Email: shaxlo.xamrayeva@navoiy-uni.uz
ORCID: 0000-0002-5429-4708

[3] *Abdullayeva Oqila Xolmo'minovna* – doctor of philosophy in philology, Senior teacher. Alisher Navoi Tashkent State University of Uzbek Language and Literature.
Email: abdullayeva.oqila@navoiy-uni.uz
ORCID: 0000-0002-2524-4832

[4] *Xusainova Zilola Yuldashevna* – PhD student of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.
E-mail: xusainovazilola@navoiy-uni.uz
ORCID: 0000-0003-4357-7515

[5] *Xudayberganov Nizomaddin Uktamboy o'g'li* – Teacher of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.
E-mail: nizomaddin@navoiy-uni.uz
ORCID: 0000-0002-6213-3015

natural language processing (NLP), and this process is called stemming. The stemming process is important in information retrieval (IR) systems.

Key words: *parts of speech, POS tagging, tagging, stemming, information retrieval, IR, stemming algorithms.*

References

Paice, C. D. (1994). An evaluation method for stemming algorithms. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994. https://doi.org/10.1007/978-1-4471-2099-5_5

Anjali, M., & Jivani, G. (2011). A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl., 2(6).

Gao, J., & Johnson, M. (2008). A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL. https://doi.org/10.3115/1613715.1613761

Goldwater, S., & Griffiths, T. L. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.

van Gael, J., Vlachos, A., & Ghahramani, Z. (2009). The infinite HMM for unsupervised PoS tagging. EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009. https://doi.org/10.3115/1699571.1699601

Taner Dinçer, B., & Karaoğlan, B. (2003). Stemming in agglutinative languages: A probabilistic stemmer for Turkish. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2869. https://doi.org/10.1007/978-3-540-39737-3_31

Hojiyev A. O‘zbek tilida so‘z yasalishi. – Toshkent, 2005.

Porter, M. F. (2006). An algorithm for suffix stripping. Program, 40(3). https://doi.org/10.1108/00330330610681286

https://www.turkedebiyati.org/yapim-ekleri/

https://tilachar.ru/ru/grammar/24-grammar

Polus, M. E., & Abbas, T. (2021). Development For Performance Of Porter Stemmer Algorithm. Eastern-European Journal of Enterprise Technologies, 1(2(109)). https://doi.org/10.15587/1729-4061.2021.225362

Memon, S., Mallah, G. A., Memon, K. N., Shaikh, A., Aasoori, S. K., & Dehraj,

F. U. H. (2020). Comparative study of truncating and statistical stemming algorithms. International Journal of Advanced Computer Science and Applications, 2. https://doi.org/10.14569/ijacsa.2020.0110272

Khyani, D., S, S. B., M, N. N., & M, D. B. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. Journal of University of Shanghai for Science and Technology, 22(10).

Tsygankin, D. v., & Ivanova, G. S. (2019). Assimilative potential of vowel harmony in languages of agglutinative type. Bulletin of Ugric Studies, 9(3). https://doi.org/10.30624/2220-4156-2019-9-3-510-518

Миртожиев М.М. Ўзбек тили фонетикаси. – Тошкент, Фан нашриёти, 2013. – 424 бет.

P.~Brown, V.~Della Pietra, de Souza, P., J.~Lai, & R.~Mercer. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18.

Schütze, H. (1993). Part-of-speech induction from scratch. https://doi.org/10.3115/981574.981608

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A practical part-of-speech tagger. https://doi.org/10.3115/974499.974523

Biemann, C. (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop. https://doi.org/10.3115/1557856.1557859

Merialdo, B. (1994). Tagging English text with a probabilistic model. Computational Linguistics, 20(2).

Banko, M., & Moore, R. C. (2004). Part of speech tagging in context. COLING 2004 - Proceedings of the 20th International Conference on Computational Linguistics. https://doi.org/10.3115/1220355.1220435

Haghighi, A., & Klein, D. (2006). Prototype-driven learning for sequence models. HLT-NAACL 2006 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference. https://doi.org/10.3115/1220835.1220876

Lovins, J. B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(June).

Porter M.F. “Snowball: A language for stemming algorithms”. 2001.

Krovetz, R. (2000). Viewing morphology as an inference process. Artificial Intelligence, 118(1–2). https://doi.org/10.1016/S0004-3702(99)00101-0

Xu, J., & Croft, W. B. (1998). Corpus-based stemming using cooccurrence

of word variants. ACM Transactions on Information Systems, 16(1). https://doi.org/10.1145/267954.267957

Manish Sh., Nitin A., Bibhuti M. Morphology based natural language processing tools for indian languages. In Proceedings of the 4th Annual Inter Research Institute Student Seminar in Computer Science, IIT, Kanpur, India, April. Citeseer.

A Goweder, H Alhami, Tarik Rashed, and A Al-Musrati. A hybrid method for stemming Arabic text. Journal of computer Science

Adam, G., Asimakis, K., Bouras, C., & Poulopoulos, V. (2010). An efficient mechanism for stemming and tagging: The case of Greek language. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6278 LNAI(PART 3). https://doi.org/10.1007/978-3-642-15393-8_44

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. Computational Linguistics, 27(2).

https://doi.org/10.1162/089120101750300490

Goldsmith, J. (2006). An algorithm for the unsupervised learning of morphology. Natural Language Engineering, 12(4).

https://doi.org/10.1017/S1351324905004055

Bacchin, M., Ferro, N., & Melucci, M. (2002). The effectiveness of a graph-based algorithm for stemming. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2555. https://doi.org/10.1007/3-540-36227-4_12

Melucci, M., & Orio, N. (2003). A novel method for stemmer generation based on Hidden Markov Models. International Conference on Information and Knowledge Management, Proceedings. https://doi.org/10.1145/956863.956889

McNamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. Information Retrieval, 7(1–2). https://doi.org/10.1023/b:inrt.0000009441.78971.be

Bacchin, M., Ferro, N., & Melucci, M. (2005). A probabilistic model for stemmer generation. Information Processing and Management, 41(1). https://doi.org/10.1016/j.ipm.2004.04.006

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The web as a graph: Measurements, models,and methods. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1627. https://doi.org/10.1007/3-540-48686-0_1

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., & Datta, K. (2007). YASS: Yet another suffix stripper. ACM Transactions on Information Systems, 25(4). https://doi.org/10.1145/1281485.1281489

Peng, F., Ahmed, N., Li, X., & Lu, Y. (2007). Context sensitive stemming for web search. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07. https://doi.org/10.1145/1277741.1277851

Aisha, B., & Sun, M. (2009). A statistical method for Uyghur Tokenization. 2009 International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2009.

https://doi.org/10.1109/NLPKE.2009.5313764

Maimaiti, M., Wumaier, A., Abiderexiti, K., & Yibulayin, T. (2017). Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging. Information (Switzerland), 8(4). https://doi.org/10.3390/info8040157

Azragul, X. Qi and A. Yusup, "Website Phrasal Survey Based Modern Uighur Stem Extraction and Application Study", Computer Applications and Software, vol.29, no.3, (2012), pp.32-34.

Bölücü, N., & Can, B. (2019). Unsupervised joint PoS tagging and stemming for agglutinative languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 18(3). https://doi.org/10.1145/3292398

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2006). Distributed Representations of Words and Phrases and their Compositionality. Neural Information Processing Systems, 1.

https://uz.wikipedia.org/wiki/Neologizmlar

Qo'ziboyeva G. Tilimizga kirib kelgan neologizmlar va ularning tahlili // International scientific journal. 2022, № 3. – 78-83-б.