STATISTIK POS-TEGLASH ALGORTIMLARI (HMM, CRF) VA MATEMATIK MODELLARI

Botir Elov

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

K E Y W O R D S

POS-teglash, yashirin Markov modeli, HMM, shartli tasodifiy maydon, CRF, statistik NLP, Viterbi algoritmi, kam resursli til

A B S T R A C T

Ushbu maqolada o‘zbek tili uchun statistik part-of-speech (POS) teglashning ikki asosiy paradigmasi, yashirin Markov modeli (HMM) va shartli tasodifiy maydon (CRF) matematik jihatdan tahlil qilinadi hamda ularning amaliy samaradorligi tajriba orqali baholanadi. Avvalo har bir modelning ehtimollik-funksional tuzilishi, o‘tish va chiqarish ehtimolliklari (HMM) hamda xususiyat funksiyalari va vazn koeffitsientlari (CRF) formal ravishda keltiriladi. Tadqiqot doirasida 17038, 56616 va 77821 gapdan iborat iborat bo‘lgan, CONLL-U formatidagi qo‘lda POS-teglangan datasetlardan foydalanilgan holda, modellar Laplas smoothing + Viterbi (HMM) va L-BFGS optimizatori + Viterbi (CRF) bilan o‘qitildi. Test to‘plamida HMM 82 % aniqlik, CRF esa 88 % aniqlikka erishdi; CRF ko‘proq kontekst va lingvistik xususiyatlarni aniqlash hisobiga HMMdan 6 punkt yuqori natija ko‘rsatdi. 3 ta datasetga BiLSTM-CRF va BERT-CRF neyron tarmoqli modellarni qo‘llash natijasida 93.4% F1-ko‘rsatkichga erishildi. Natijalar statistik modellar o‘zbek tili kabi aglutinativ va kam resursli tillarda ham barqaror ishlashini, lekin xususiyatlarga sezgirligini ko‘rsatadi. Maqola yakunida model tanlovi, teg ajratishdagi asosiy xatolik turlari va kelgusida chuqur o‘rganishga asoslangan yondashuvlarga o‘tishning afzalliklari muhokama qilinadi.

Kirish

O‘zbek tili so‘z turkumlarini avtomatik belgilash **(Part-of-Speech tagging, POS-teglash)** – matndagi har bir so‘zga uning grammatik turkumini (ot, fe’l, sifat va hokazo) to‘g‘ri tayinlash vazifasidir[1]. Bu tabiiy tilni qayta ishlashning asosiy masalalaridan biri bo‘lib, boshqa yuqori darajadagi lingvistik tahlillar (*sintaktik tahlil, semantik tahlil, ma’lumot izlash, mashina tarjimasi* va hokazo) uchun zarur oraliq bosqich hisoblanadi. Korpus lingvistikasida so‘z turkumlari bilan to‘g‘ri teglangan matnlar til strukturasi bo‘yicha tadqiqotlar olib borishga imkon beradi. Ingliz, rus kabi resurslarga boy tillar uchun POS-teglash bo‘yicha ko‘plab usullar ishlab chiqilgan va ular deyarli 97% aniqlikka erishgan. Biroq, o‘zbek tili kabi kam resursli aglutinativ tillar uchun ushbu masala hali ham dolzarb bo‘lib qolmoqda. O‘zbek tilida morfologik hodisalar juda boy – so‘zlar ko‘plab affikslar qo‘shish orqali yasaladi va ular turli kontekstlarda turli so‘z turkumiga ega bo‘lishi mumkin[2]. Bu esa POS-teglash vazifasini murakkablashtiradi, chunki so‘zning asosi va uning yashirin grammatik holati (turkumi) orasidagi bog‘liqlik ko‘p ma’noli va kontekstga bog‘liq bo‘ladi.

Hozirgacha o‘zbek tilida POS-teglash uchun qoidaga asoslangan yondashuvlar taklif qilingan. Masalan, Sharipov va hammualliflar (2023) UzbekTagger nomli qoidaviy teglashga asoslangan usulni taqdim etdilar[3]. Ushbu usulda so‘zning lug‘atdagi asosiy shakliga qarab yoki maxsus grammatik qoidalarga tayangan holda teglash bajaradi. Biroq, bunday qoidaviy tizimlar cheklangan kontekstni hisobga oladi xolos. Natijada ular neyron modellar darajasida kontektstual ma’lumotni e’tiborga ololmaydi va aniqlik jihatidan ortda qoladi. Shu sababli, so‘nggi yillarda statistik va *ma’lumotga asoslangan metodlar* muhim ahamiyat kasb etmoqda. Statistik yondashuvlar so‘zlarni teglashda korpusdan olingan ehtimollik modellariga tayanadi[4]. Xususan, yashirin Markov modellari (Hidden Markov Model, HMM) va shartli tasodifiy maydonlar (Conditional Random Field, CRF) POS-teglash uchun keng qo‘llaniladigan modellar sirasiga kiradi[5]. Ushbu modellar ma’lumotlarda so‘zlar ketma-ketligi va ularning teglar ketma-ketligi o‘rtasidagi statistik bog‘liqlikni o‘rganib,

yangi gaplardagi teglar ketma-ketligini ehtimoli eng yuqori bo‘lgan tarzda aniqlaydi[6].

Hozirgi kunda chuqur o‘rganishga asoslangan yondashuvlar statistik modellarga yangicha imkoniyatlar qo‘shdi. An’anaviy HMM va CRF modellariga neyron tarmoqlar integratsiya qilinib, yanada samaradorligi yuqori POS-teglagichlar yaratilmoqda. Masalan, BiLSTM-CRF arxitekturasi[7] (ikki yo‘nalmalı LSTM + CRF) ketma-ketliklarni teglashda juda samarali ekani bir qator tadqiqotlarda ko‘rsatildi[8]. Shuningdek, oldindan katta korpuslarda o‘qitilgan transformator modellar (BERT kabi) yordamida BERT-CRF kabi arxitekturasi orqali kam resursli tillarda ham yuqori natijalarga erishish mumkin bo‘lmoqda. O‘zbek tilida neyron modellarga asoslangan POS-teglash borasidagi ilk ishlardan biri Murat va Ali (2024) tadqiqotidir[9]. Ushbu tadqiqotda so‘zga qo‘shimchalar orqali chuqurroq taqdimot hosil qiluvchi hamda **multi-head attention mexanizmi**ni qo‘shuvchi model taklif qilinib, an’anaviy BiLSTM, CNN va CRF modellariga nisbatan aniqlik 4.13% ga oshirilgan (umumiy aniqlik 79.74%). Yaqindan, Bobojonova va boshq. (2025) esa BERT modeliga asoslangan POS-teglash uchun maxsus korpus va modellar (BBPOS)ni yaratib, o‘rtacha 91% aniqlikka erishganini ta’kidlashgan[10].

Ushbu maqolada o‘zbek tili uchun statistik POS-teglash modellarining matematik asoslari va ishlanish prinsiplari yoritiladi. Avvalo, HMM va CRF modellarining formal ta’rifi, ularning chuqur o‘rganishga asoslangan variantlari (klassik CRF vs BiLSTM-CRF vs BERT-CRF) batafsil tushuntiriladi. Har bir modelning matematik modeli formulalar yordamida ifodalanadi, algoritmlari psevdokod ko‘rinishida keltiriladi. So`ngra, natijalar 6528, 56616 va 77821 ta o‘zbekcha gapdan iborat (ConLL-U formatida POS-teglangan korpus)da ushbu modellarni o‘qitib sinovdan o‘tkazamiz va ularning ishlash ko‘rsatkichlarini qiyosiy tahlil qilamiz. Natijalar tahlili jadval va diagrammalar yordamida ko‘rsatiladi, eng yaxshi yondashuvlar aniqlanadi. Yakunda, modellarni o‘zbek tili xususiyatlari nuqtai nazaridan tahlil qilib, ularning afzallik va kamchiliklari, erishilgan yutuqlar va cheklovlar haqida ilmiy xulosa qilinadi.

Metodlar

Yashirin Markov modeli (HMM) asosida POS-teglash

HMM – bu ketma-ketlikli ma’lumotlar uchun generativ statistik model bo‘lib, unda kuzatilgan so‘zlar ketma-ketligi qanday yashirin holatlar (teglar) zanjiridan hosil bo‘lganligi ehtimollik nuqtai nazaridan modellashadi[11]. POS-teglash masalasida HMM quyidagi stoxastik jarayon tasavvur qilinadi: har bir pozitsiyada dastlab ma’lum bir **teg** (masalan, *ot, fe’l, sifat* va hokazo) **Markov zanjiri** qoidalariga binoan oldingi tegdan kelib chiqib tanlanadi (holatlar zanjiri), so‘ngra tanlangan tegga muvofiq ushbu pozitsiyadagi so‘z *“chiqariladi”*. Shuning uchun HMM modeli ikki turdagi ehtimollik parametrlariga ega: **o‘tish ehtimolliklari** va **chiqish ehtimolliklari**[6]. O‘tish ehtimolligi $\boldsymbol{P(t_i|t_{i-1})}$ bir tegdan keyingi tegning paydo bo‘lish ehtimolini ifodalaydi. Masalan, ot so‘z turkumining ortidan fe’l kelish ehtimoli, fe’ldan keyin ko‘makchi kelish ehtimoli aniqlanadi. Ushbu ehtimollarni o‘rgatish uchun Markov taxminiga asosan faqat ketma-ket kelgan teglar juftligi hisobga olinadi (ya’ni $\boldsymbol{t_i}$ faqat $\boldsymbol{t_{i-1}}$ ga bog‘liq, undan oldingi holatlarga bog‘liq emas). Korpusda bu ehtimollar ikki tegning birgalikda kelish chastotalaridan quyidagicha hisoblanadi:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

bu yerda,

$\boldsymbol{C(t_{i-1}, t_i)}$ – $\boldsymbol{t_{i-1}}$ tegidan keyin $\boldsymbol{t_i}$ tegi ketma-ket kelgan hollar soni,

$\boldsymbol{C(t_{i-1})}$ – $\boldsymbol{t_{i-1}}$ tegining umumiy chiqish chastotasi.

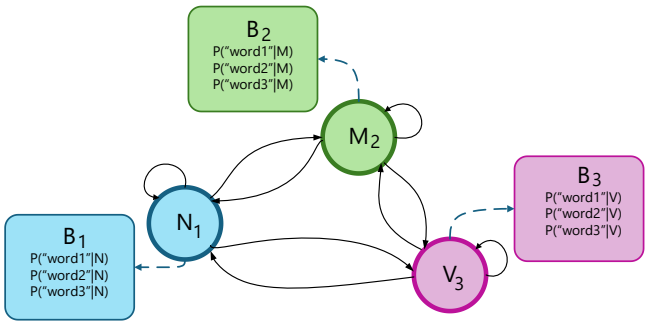
Chiqish ehtimolligi esa $\boldsymbol{P(w_i|t_i)}$ tarzida ifodalanib, berilgan $\boldsymbol{t_i}$ teg ostida aynan $\boldsymbol{w_i}$ so‘zi kuzatilish ehtimolini bildiradi. U ham korpusdagi chastotalar vositasida baholanadi:

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

bunda

$\boldsymbol{C(t_i, w_i)}$ – $\boldsymbol{w_i}$ so‘zi $\boldsymbol{t_i}$ turkumida kelgan hollari soni,

$C(t_i) - t_i$ turkumining korpusdagi jami soni. Ushbu ehtimollar matritsa ko'rinishida **A (o'tish)** va **B (chiqish)** parametrlari sifatida saqlanadi.



1-rasm. Yashirin Markov modeli misoli (uchta yashirin holat: N – ot, M – modal so'z, V – fe'l)

HMM modeli berilgan *"word1 word2 word3"* so'zlar ketma-ketligi orqasida yashirin **N–M–V** teglar zanjiri yotgan deb taxmin qilgan. 1-rasmda **B1, B2, B3** bloklarida har bir holatdan ma'lum bir so'z chiqishi ehtimoli ko'rsatilgan. Masalan B1 blokida $P("word1" | N)$ kabi qiymatlar keltirilgan. Holatlar orasidagi yoylar esa **A** matritsasidagi teglarning o'tish ehtimollarini ifodalaydi. HMM modelining maqsadi, kuzatilgan **W = w1, w2, .., wn** so'zlar ketma-ketligi uchun eng ehtimoli katta teglar ketma-ketligini aniqlashdan iborat. Ya'ni $\hat{T} = t_1, t_2, \dots, t_n$ shundayki $P(W, T)$ maksimum bo'lsin. Generativ model sifatida $P(W, T)$ qiymati quyidagicha hisoblanadi:

$$P(W|T) = \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

bu yerda, $t_0 = <Boshlanish>$ maxsus start holat deb faraz qilinadi.

Tegishli eng ehtimoli katta **T** ketma-ketlikni aniqlash uchun barcha mumkin bo'lgan teglar kombinatsiyalarini tekshirib chiqish ($O(m^n)$ variant, m – teglar soni) amaliy jihatdan imkonsiz. Buning uchun dinamik dasturlashga asoslangan **Viterbi algoritmi** qo'llaniladi. Viterbi yondashuvi rekursiv ravishda har bir pozitsiyada eng optimal tegga kelish yo'lini hisoblab boradi va yakunida eng yaxshi umumiy yo'lni topadi. Quyida HMM asosida POS-teglash uchun Viterbi algoritmining psevdokodi keltirilgan.

1-algoritm: HMM asosida Viterbi dekodingi (POS-teglash)

```
# Kirish: w[1..n] – teggilanishi lozim bo'lgan so'zlar ketma-ketligi
# Chiqish: t[1..n] – eng ehtimolli teglar ketma-ketligi
    # 1. Boshlang'ich ehtimollar (Boshlanish holatdan birinchi teggaa o'tish):
    for each tag s:
        delta[1][s] = P(s | <Boshlanish>) * P(w1 | s)
        backpointer[1][s] = <Boshlanish>
    # 2. Rekursiv hisob-kitoblar:
    for i from 2 to n:
        for each tag s:
        # avvalgi pozitsiyadan holat s ga eng yuqori ehtimolli o'tish yo'lini topish
        max_prob = 0
        arg_max_state = None
        for each tag s_prev:
            prob = delta[i-1][s_prev] * P(s | s_prev) * P(w_i | s)
            if prob > max_prob:
                max_prob = prob
                arg_max_state = s_prev
            delta[i][s] = max_prob              # i-pozitsiyagacha eng yaxshi yo'l ehtimoli
            backpointer[i][s] = arg_max_state   # i-pozitsiya s holati uchun eng yaxshi oldingi
            holat
    # 3. Yakuniy holatni aniqlash:
    max_prob = 0
    last_state = None
    for each tag s:
        if delta[n][s] > max_prob:
            max_prob = delta[n][s]
```

```
24          last_state = s
25  # 4. Eng yaxshi taglar ketma-ketligini orqaga qaytib tiklash:
26  t[n] = last_state
27  for i from n down to 2:
28      t[i-1] = backpointer[i][ t[i] ]
29  return t[1..n]
```

Yuqoridagi algoritmda $\boldsymbol{\delta[i][s]}$ qiymati $\boldsymbol{w_1 \dots w_i}$ so'zlar uchun i-pozitsiyada $\boldsymbol{s}$ teg keladigan eng yuqori ehtimolli qisman yo'lni bildiradi. Backpointer massivlari esa ushbu yo'lni tiklash uchun orqaga ko'rsatkichlarni saqlaydi. Algoritmning murakkabligi $\boldsymbol{O(n \times m^2)}$ (bu yerda m – teglar soni, n – so'zlar soni).

Yashirin Markov Modeli (HMM) asosida POS-teglashning matematik modeli

O'zbek tili matnlarini HMM modeli asosida POS-teglash bir nechta bosqichda amalga oshiriladi.

1-bosqich. Formallashtirish bosqichi. HMM modelini ishlab chiqishda quyidagi belgilanishlarni kiritamiz:

$\boldsymbol{W = w_1^n}$	*uzunligi n ga teng bo'lgan so'zlar ketma-ketligi*
$\boldsymbol{T = t_1^n}$	*yashirin teglar (POS) ketma-ketligi*
$\boldsymbol{V}$	*teglar to'plami (UPOS → 15 ta)*
$\boldsymbol{\pi(t)}$	*boshlang'ich ehtimol* $P(t_1 = t)$
$\boldsymbol{A_{uv} P(t_i = v \mid t_{i-1} = u)}$	*o'tish ehtimoli*
$\boldsymbol{B_v P(w_i = w \mid t_i = v))}$	*chiqish (emissiya) ehtimoli*

Quyidagi HMM farazini shakllantiramiz:

$$P(W,T) = \pi(t_1)B_{t_1}(w_1)\prod_{i=2}^{n} A_{t_{i-1}t_i}B_{t_i}(w_i)$$

HMM dekoding qadami quyidagicha amalga oshiriamiz:

$$T^* = arg\max_T P(W,T) = arg\max_T \left[\pi(t_1)B_{t_1}(w_1)\prod_{i=2}^{n} A_{t_{i-1}t_i}B_{t_i}(w_i)\right]$$

2-bosqich. Parametrlarni baholash (ML o'qitish)

Ushbu bosqichda teglangan korpusdan chastotalar quyidagicha aniqlanadi:

$$\hat{A}_{uv} = \frac{C(u,v)+\alpha}{\sum_{v'} C(u,v') + \alpha|V|}$$

$$\hat{B}_u(w) = \frac{C(v,w)+\beta}{\sum_{w'} C(v,w') + \beta\left||\Sigma|\right|},$$

bu yerda $\boldsymbol{\alpha, \beta}$**>0** – Laplas silliqlash. CONLL-U formatidagi 77821 gapli datasetda **C(·)** lar bevosita sanaladi.

3-bosqich. Viterbi algoritmi (dekoding)

$$\delta_i(v) = \max_{t_1^{i-1}} P\left(t_1^{i-1}, t_i = v, tw_1^i\right) = \max_u [\delta_{i-1}(u)A_{uv}]\, B_v(w_i)\Big)$$

$$\psi_i(v) = arg\, max_u[\delta_{i-1}(u)A_{uv}]$$

Yakunda $t^* = \arg\max_v \delta_n(v)$, so'ng orqaga yurib $\boldsymbol{T}^*$ tiklanadi.

4-bosqich. Viterbi algoritmi uchun o'zbek tilida berilgan sodda gapni ko'rib chiqamiz.

Gap: *"Yaxshi ovqat yesang."*

i	w_i	kuzatilgan teglar	B misol-eht.	izoh
1	*yaxshi*	ADJ (0.6), ADV (0.4)	B_{ADJ}(yaxshi)= 0.6	lug'atda ikki ma'noli

2	*ovqat*	NOUN (1.0)	$B_{NOUN} = 1$	
3	*yesang*	VERB (1.0)	$B_{VERB} = 1$	-sang → fe'l

Boshlang'ich $\boldsymbol{\pi}$**(ADJ)=0.3,** $\boldsymbol{\pi}$**(NOUN)=0.4,** $\boldsymbol{\pi}$**(VERB)=0.1**

Trigram o'rniga bigram o'tishlar natijasi:

$A_{ADJ,NOUN} = 0.55$ ("sifat→ot" keng tarqalgan)

$A_{NOUN,VERB} = 0.30$

$A_{ADV,VERB} = 0.60$

Viterbi hisoblarini amalga oshiramiz:

$\delta_1(ADJ) = \pi(ADJ)B_{ADJ}(yaxshi) = 0.18,$

$\delta_1(ADV) = 0.12, \quad \delta_1(NOUN) = 0.40$

2-pozitsiyagi $w_2 = (ovqat)$ so`zi uchun hisoblashlar:

$$\delta_1(NOUN) = max[\delta_1(ADJ)A_{ADJ,NOUN}], \quad \delta_1(ADV) = A_{ADV,NOUN}, \quad \delta_1(NOUN) = A_{NOUN,NOUN} \times 1.0 = 0.18 \cdot 0.55 = 0.099$$

Shunday qilib, $\boldsymbol{\psi_2(NOUN) = ADJ}$

3-pozitsiya $w_2 = (yesang), VERB$:

$$\delta_1(VERB) = max[0.99 \cdot A_{NOUN,VERB,\dots}] \times 1.0 = 0.099 \cdot 0.30 = 0.0297$$

Orqaga harakatlanib, $\boldsymbol{T^* = (ADJ, NOUN, VERB)}$ ketma-ketligini aniqlaymiz.

Demak,

"yaxshi" → sifat,

"ovqat" → ot,

"yesang" → fe'l;

kabi to'g'ri teglandi.

O'zbek tilida teglar soni nisbatan kichik (masalan, Universal Tagset bo'yicha 17 ta), shu bois Viterbi usuli juda samarali ishlaydi. Biroq, HMM modelining katta cheklovi shundaki, u **faqat qo'shni teglar va so'zlarni hisobga oladi**. Ya'ni, biror so'zning tegi faqat oldingi so'zning tegiga bog'liq, so'zning o'ziga esa faqat uning tegi orqali bilvosita bog'liq (emissiya ehtimoli orqali). Shuningdek, HMM har bir so'z uchun faqat bitta chiqish ehtimoli bilan ishlaydi, natijada korpusda uchramagan yangi so'zlar **(out-of-vocabulary)** uchun model aniq bashorat berishda qiynaladi. Murakkab aglutinativ tillarda esa so'z shakllari juda ko'p bo'lishi mumkin. HMM bunday yangi so'zlarni odatda ehtimoli nol deb baholab, xatolik kiritadi. Bu kamchiliklarni bartaraf etish uchun quyida ko'rib chiqiladigan zamonaviy yondashuvlar qo'llaniladi.

Shartli tasodifiy maydon (CRF) asosida POS-teglash

HMM modeli kuzatilgan ketma-ketliklar uchun qo'shma ehtimollikni modellashtirsa, shartli tasodifiy maydonlar (CRF) bevosita teglar ketma-ketligining kuzatilgan so'zlar ketma-ketligiga shartli taqsimotini modellashtiradi[7]. Chiziqli zanjirli CRF modelida teglar o'rtasidagi bog'lanishlar HMM dagi kabi qo'shni holatlar bilan cheklansa-da, u **yo'nalishsiz graf (undirected graph)** ko'rinishida tasvirlanadi. Bunda har bir pozitsiyadagi teg uchun butun kirish gap bo'yicha xususiyatlar majmuasi hisobga olinadi va ularga mos $\boldsymbol{\lambda}$ vaznlar o'rganiladi. CRF modelining shartli ehtimollik funksiyasi quyidagicha ifodalanadi:

$$P(W|T) = \frac{1}{Z(W)} exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j\,(t_{i-1}, t_i, W, i)\right)$$

Bu yerda

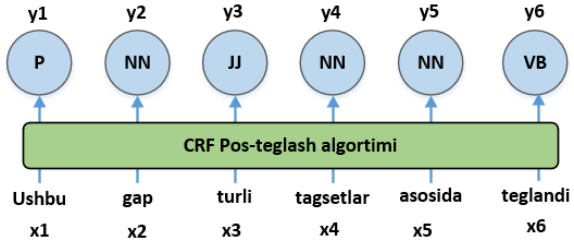
$\boldsymbol{f_j(t_{i-1}, t_i, W, i)}$ – i-pozitsiyada, $\boldsymbol{t_{i-1}, t_i}$ qo'shni teglar va butun kirish $\boldsymbol{W}$ ketma-ketligi (masalan, $\boldsymbol{w_i}$ yoki atrofidagi so'zlar) kontekstida aniqlanuvchi j-xususiyat funksiyasi.

$\boldsymbol{\lambda_j}$ – ushbu xususiyatning modeldagi og'irlik koeffitsienti (o'qitish jarayonida o'rganiladigan parametr).

$\boldsymbol{Z(W)}$ – normallashtiruvchi omil bo'lib, u $\boldsymbol{W}$ uchun barcha mumkin bo'lgan $\boldsymbol{T}$ ketma-ketliklar

bo'yicha $\boldsymbol{exp}(\cdot)$ yig'indisini hisobga olib, $\boldsymbol{P(T|W)}$ ni $[\mathbf{0}, \mathbf{1}]$ oralig'iga keltiradi.

CRF modelini o'qitishda berilgan etiketli korpus bo'yicha maksimal shartli ehtimollik (Maximium Conditional Likelihood) mezoni optimallashtiriladi, ya'ni $\boldsymbol{\lambda}$ parametrlari $\boldsymbol{L(\lambda) = \prod_{(W,T)\in D} P_{\lambda}(T \mid W)}$ ni maksimal qiladigan tarzda gradient usullari bilan tanlanadi. Bu diskriminativ o'qitish usuli bo'lib, HMM ga nisbatan bevosita $\boldsymbol{P(T|W)}$ni maximallashtiradi. Natijada CRF modelida mustaqillik taxminlari kamroq, ya'ni xususiyatlar orqali turli bog'liq ma'lumotlar (kontekstdagi boshqa so'zlar, so'zning morfologik tuzilishi, bosh harf bilan boshlanishi, so'nggi harflari va h.k.) hisobga olinishi mumkin. Aynan shu bois CRF modeli kontekstni hisobga olgan holda har bir pozitsiya uchun teglar ketma-ketligini baholaydi va HMMga qaraganda aniqroq natijalar beradi.x_1



2-rasm. Chiziqli zanjirli shartli tasodifiy maydon (CRF) modeli strukturasining soddalashtirilgan ko'rinishi

Ushbu rasmning pastki qatordagi $\boldsymbol{x_1, x_2, \dots, x_n}$ tugunlar kirish so'zlar ketma-ketligini ifodalaydi, yuqori qatordagi $\boldsymbol{y_1, y_2, \dots, y_n}$ tugunlar esa ularning teglar ketma-ketligini ifodalaydi. Chiziqli CRFda chiqish (teg) tugunlari o'zaro qo'shni bog'langan (yo'nalishsiz bog'lanishlar) bo'lib, bu ularga bir-birining ta'sirini hisobga olgan holda birgalikda optimal holatlarni tanlash imkonini beradi. Kirish tugunlari orasida to'g'ridan-to'g'ri bog'lanish yo'q – barcha ta'sir faqat teglar orqali modellashtiriladi. Shu tariqa, CRF modeli qo'shni teglar orasidagi bog'liqlikni hamda har bir pozitsiyadagi so'zning turli xususiyatlarini (masalan, uning ma'nosi, morfologik affikslari, qo'shni so'zlar) birlashtirib, yakuniy teglashga erishadi.

CRF modelida ham eng ehtimolli taglar ketma-ketligini topish uchun dinamik dasturlash (masalan, Viterbi yoki Forward-Backward algoritmi) qo'llaniladi. Bunda har bir pozitsiyada har bir tag uchun score (ball) hisoblanadi. U o'tgan pozitsiyadan shu tegga o'tish va pozitsiyada teg uchun xarakterli xususiyatlarning $\boldsymbol{\lambda}$ vaznlar bilan baholangan yig'indisidan hosil bo'ladi. Kerakli formulalar HMM dagiga o'xshash rekursiya tarzida amalga oshadi, faqat unda $\boldsymbol{P(t_i|t_{i-1})}$ va $\boldsymbol{P(w_i|t_i)}$ ko'rinishidagi tayyor ehtimollar o'rniga $\boldsymbol{exp(\sum_j \lambda_j f_j(t_{i-1}, t_i, W, i))}$ tarzidagi baholar ishlatiladi. Quyida CRF modeli uchun eng yaxshi teglar ketma-ketligini topish algoritmining soddalashtirilgan psevdokodi keltirilgan (Viterbi usuli bilan)

2-algoritm: CRF modelida eng yaxshi taglar ketma-ketligini topish

```
# Kirish: X = [w1, w2, ..., wn] – so'zlar ketma-ketligi
# Chiqarish: Y = [y1, y2, ..., yn] – eng ehtimolli taglar ketma-ketligi
    # 1. Birinchi so'z uchun boshlang'ich ballar:
    for each possible tag t:
        score[1][t] = \sum_j \lambda_j * f_j(<Boshlanish>, t, X, 1)
        # (boshlanish holatdan t ga o'tish + w1 ga tegishli xususiyatlar xatoligi)
        backpointer[1][t] = <Boshlanish>
    # 2. Ketma-ket dinamik dasturlash:
    for i from 2 to n:
        for each tag t:
            max_score = -∞
            best_prev_tag = None
            for each tag t_prev:
                s = score[i-1][t_prev] + \sum_j \lambda_j * f_j(t_prev, t, X, i)
                if s > max_score:
                    max_score = s
                    best_prev_tag = t_prev
```

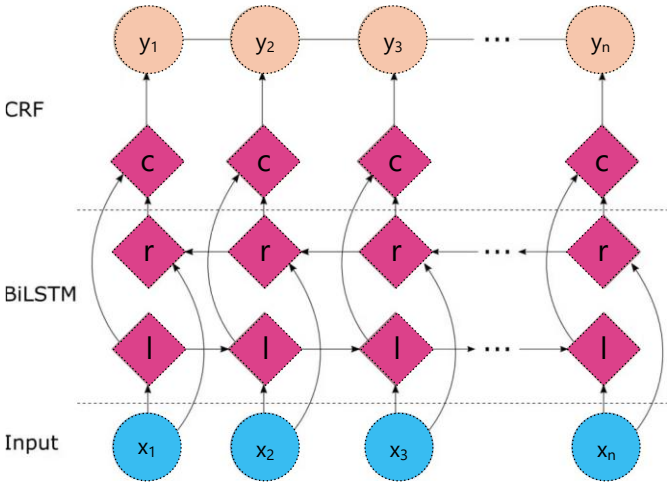
```
        score[i][t] = max_score
        backpointer[i][t] = best_prev_tag
# 3. Yakuniy eng yaxshi y_n tegini topish:
best_score = -∞
best_last_tag = None
for each tag t:
    if score[n][t] > best_score:
        best_score = score[n][t]
        best_last_tag = t
# 4. Orqaga qaytish orqali butun ketma-ketlikni tiklash:
y[n] = best_last_tag
for i from n down to 2:
    y[i-1] = backpointer[i][ y[i] ]
return Y
```

Yuqoridagi algoritmda $\boldsymbol{f_j(t_{i-1}, t_i, X, i)}$ lar orasida, masalan, *"t_i tagi = Ot ∧ w_i so'zi oxiri -ni affiksi bilan tugaydi"* kabi qoidaviy xususiyat bo'lishi mumkin. Klassik CRF modelida bunday xususiyatlar dizayni til mutaxassisi tomonidan qo'lda teglanadi va korpusdan olinadi. O'zbek tilining xususiyati shundaki, so'z oxiridagi qo'shimchalar uning grammatik turkumini aniqlashda muhim rol o'ynaydi. Masalan, fe'lga *"-gan"* qo'shimchasi qo'shilsa, sifat (o'tgan zamon sifatdosh) yasalishi mumkin, yoki *"-lik"* qo'shimchasi sifatga qo'shilib ot yasalishi mumkin. Shunday ekan, CRF modelida bunday affikslar bo'yicha xususiyatlarni kiritish uning aniqligini oshiradi. Shu bilan birga, CRF har bir so'z uchun qo'shni so'zlardan tortib, matn darajasigacha kontekstni modellashtirishga imkon beradi, bu esa kontekstga bog'liq ko'p ma'nolilik masalasini hal etishda muhimdir. HMM modeli kontektstni faqat oldingi tag orqali bilvosita hisobga olgan bo'lsa, CRF modeli qo'shni so'zlarning taglari, hatto o'zlari yoki boshqa lingvistik xususiyatlarini bevosita kiritishga imkon beradi. CRF modelining kamchiligi – o'qitishning nisbatan murakkabligidir. Yuqorida ko'rsatilgan eng yaxshi ketma-ketlikni topish (dekoding) jarayoni $\boldsymbol{O(n \times m^2)}$ bo'lsa, modelni o'qitish (parametrlarni aniqlash) uchun har bir iteratsiyada forward-backward algoritmi yordamida barcha mumkin ketma-ketliklar bo'yicha ehtimollik taqsimotini hisoblash talab etiladi. Bu esa katta korpuslar uchun sekin ishlashi mumkin. Shunga qaramay, CRF nazariy jihatdan kuchli model bo'lib, etarli hajmdagi belgilangan ma'lumotlar mavjud bo'lganda HMMdan ancha yuqori natija beradi.

BiLSTM-CRF neyron tarmog'i modeli

So'nggi yillarda chuqur o'rganish usullari tabiiy tilni qayta ishlashda katta yutuqlarga erishdi. Xususan, **ikki tomonlama uzun-qisqa xotirali tarmoq (BiLSTM)**lar ketma-ketlikdagi uzoq bog'lanishlarni o'rganish qobiliyati tufayli matnlarni lingvistik jihatdan chuqurroq modellashtira oladi. BiLSTM-CRF arxitekturasi esa an'anaviy CRF modelini neyron tarmoq bilan birlashtirib, har bir so'z uchun uning chap va o'ng kontekstidan hosil bo'lgan boy taqdimot (embedding)ga asoslangan holda eng yaxshi teglar zanjirini tanlaydi[12].

BiLSTM-CRF modelining tuzilishini quyidagicha tavsifash mumkin: avval kirish so'zlar sonli vektorlar ko'rinishida ifodalanadi[13]. Ushbu vektorlar ikkita LSTM tarmog'iga uzatiladi va ulardan biri gapni boshidan oxirigacha (chapdan o'ngga) o'qib chiqadi, ikkinchisi oxiridan boshigacha (o'ngdan chapga) o'qib chiqadi[14]. Natijada har bir so'z pozitsiyasi uchun ikki yo'nalishli kontekstni o'zida mujassam etgan yashirin holat vektorlari olinadi: $\boldsymbol{\overrightarrow{h_i}(chap\ kontekst)}$, $\boldsymbol{\overleftarrow{h_i}}$ **(o'ng kontekst)**. Ular birlashtirilib $\boldsymbol{c_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]}$ tarzida to'liq kontekstual vakillik vektorini hosil qiladi. Keyinchalik har bir $\boldsymbol{c_i}$ uchun teglar ehtimolliklarini chiquvchi bir qatlam (masalan, oddiy softmax yoki logistik regressiya) o'rniga CRF qatlami qo'yiladi. CRF qatlamida ketma-ketlikdagi teglar o'rtasidagi bog'liqlik va o'zaro moslik cheklovlari hisobga olinib, butun gap bo'yicha eng mos keluvchi teglar zanjiri tanlab olinadi.



3-rasm. BiLSTM-CRF arxitekturasi

Pastki qatlamda kirish so‘zlar $\boldsymbol{x_1, x_2, \dots, x_n}$ birlashtirilgan so‘z embeddinglari ko‘rinishida neyron tarmoqqa uzatiladi. O‘rta qatlamda ikki yo‘nalmali LSTM mavjud: chap tomonga qarab **(l nodes)** va o‘ng tomonga qarab **(r nodes)** yuradigan tarmoqlar har bir pozitsiya uchun ikki xil kontekst ma’lumotini to‘playdi. Keyin ular birlashtirilib, yuqori qatlamda har bir pozitsiya uchun birlashtirilgan kontekst vektorlar $\boldsymbol{c_1, c_2, \dots, c_n}$ olinadi. Nihoyat, $\boldsymbol{c_i}$ lar CRF qatlamiga uzatilib, u yerda $\boldsymbol{y_1, y_2, \dots, y_n}$ teglar ketma-ketligi uchun optimal yechim topiladi. BiLSTM-CRF modelining afzalligi shundaki, u **butun gap kontekstini** (chap va o‘ng qo‘shnilar, uzoqdagi bog‘liqliklar) hamda kuzatilgan ketma-ket teglarning global mosligini birvaqtning o‘zida hisobga oladi. Ayniqsa, uzun gaplarda yoki murakkab so‘z juftliklarida BiLSTM-CRF modeli bir taraflama CRFga qaraganda aniqroq yechim beradi. Chunki oddiy CRF faqat oldingi tegdan kelib chiqqan holda joriy tegni tanlasa, BiLSTM-CRF butun gap bo‘ylab oldingi va keyingi so‘zlardan ham signal oladi.

BiLSTM-CRF arxitekturasini o‘zbek tiliga qo‘llaganda, biz qo‘shimcha yana bir muhim jihatni yodda tutish kerak: **morfologik murakkablik**. So‘zlarning tarkibiy qismlari (*o‘zak va affikslar*) ma’noga kuchli ta’sir ko‘rsatadi va ularni teglashda yordam beradi. BiLSTM-CRF modelida odatda har bir so‘zni ifodalashda nafaqat tayyor so‘z embeddinglari, balki harf yoki bo‘g‘in darajasidagi embeddinglar ham kiritilishi mumkin (masalan, har bir so‘zning bosh harfi katta/kichikligi, so‘nggi harflari, bo‘g‘in tuzilishi va h.k.). Tadqiqotlarda turk tillari uchun bunday mexanik xususiyatlarni (feature engineering) kiritish an’anaviy modellar samaradorligini oshirishi ko‘rsatilgan. Chuqur o‘rganish yondashuvida esa model bunday xususiyatlarni o‘zi affikslar orqali o‘rganishi ham mumkin. Murat va Ali (2024) o‘zbek, qozoq, qirg‘iz tillarida aynan affikslarga oid chuqur taqdimotlarni kiritish orqali POS-teglash aniqligini sezilarli oshirishga erishganlarini ta’kidlaydilar[9]. Shu bois, BiLSTM-CRF arxitekturasini o‘zbek tilida qo‘llaganda, agar ma’lumotlar hajmi yetarli bo‘lsa, u an’anaviy HMM/CRF modellardan ustun natija ko‘rsatishini kutish mumkin.

BERT-CRF modeli

Hozirgi zamon NLP tadqiqotlarida Transformator arxitekturasi va ularga asoslangan oldindan o‘qitilgan modellarning (BERT, RoBERTa, XLM-RoBERTa va h.k.) muvaffaqiyati kuzatilmoqda. **BERT (Bidirectional Encoder Representations from Transformers) modeli** katta hajmdagi matnlarda kuzatuvsiz o‘rganish (self-supervised) orqali tilning ichki xususiyatlarini o‘zida aks ettirgan kontekstual embeddinglarni hosil qila oladi. BERTning ikki yo‘nalmali transformer bazasi matnni to‘liq konteksti bilan tahlil qilib, har bir so‘z (aniqrog‘i, so‘z qismlari, subword tokenlar) uchun qudratli yagona yashirin vakillik (embedding) hosil qiladi[10]. Shu sababli, BERTdan keyingi maxsus vazifalar uchun fine-tune qilish, ayniqsa kam resursli tillar uchun eng samarali yechimlardan biri ekani ko‘rsatib kelinmoqda[15]. POS-teglash vazifasida BERT modelini qo‘llash odatda quyidagicha amalga oshiriladi: oldindan o‘qitilgan BERT modeliga o‘zbek tilidagi belgilangan matnlar **qo‘shimcha o‘qitiladi (fine-tuning)** va BERT chiqishida har bir pozitsiya uchun vektorlar ustiga teglarni bashoratlovchi qatlam qo‘yiladi. Eng oddiy holda, bu qatlam mustaqil softmax klassifikatorlari bo‘lishi mumkin (har bir pozitsiyada tegni alohida tanlash), biroq amalda teglar ketma-ketligining izchilligini saqlash uchun yuqoriga yana CRF qatlamini qo‘shish yanada yaxshi natija beradi.

BERT-CRF arxitekturasi BiLSTM-CRFga o‘xshash, faqat LSTM o‘rniga transformer encoder qatlamlari kontekstni modellashtiradi[16]. BERT modeli o‘zbek tilida mavjud bo‘lmagan so‘zlarni ham qismlarga ajratib (WordPiece tokenization), ularning ma’nosini

atrofidagi kontekst yordamida aniqlashi mumkin. Bu ayniqsa o‘zbek tilining qo‘shimchalar orqali yangi so‘zlar hosil qilish xususiyatida juda qo‘l keladi. Misol uchun, “bormaganlardanmisiz” so‘zi BERT modelida bir necha qismlarga ajratiladi va har bir qism o‘z kontekstida tahlil qilinadi; modelning e’tibor mexanizmi tufayli uzoqdagi bog‘lanishlar ham hisobga olinadi. Natijada, BERT shunday murakkab so‘z tarkibidagi grammatik ma’lumotlarni ham (masalan, *fe’l + inkor -magan + ko‘plik -lar + dan + so‘roq mi + hurmatli siz*) to‘g‘ri talqin qilib, so‘zga to‘g‘ri tegni qo‘ya oladi. An’anaviy qoidaviy teggerlar bunday hollarda qiynalishini hisobga olsak, BERT modeli kontekst sezgirligi va morfologik xabardorlik borasida juda katta ustunlikka ega.

O‘zbek tilida bir necha BERT modellari yaqinda ishlab chiqildi: UzBERT (Mansurov va Mansurov, 2021) – 142 mln so‘zli toza matnlarda, TahrirchiBERT (Mamasaidov va Shopulatov, 2023) – 5 mlrd. tokenli nisbatan *“shovqinli”* matnlarda (bloglar, OCR kitoblar) o‘qitilgan. Shuningdek, ko‘p tilli mBERT (Devlin va boshq., 2019) ham mavjud[17]. Bu modellar hozircha faqat **MLMA (Masked Language Modeling)** yordamida baholangan, ammo yaqinda Bobojonova va boshq. (2025) ushbu modellarni POS-teglash vazifasida sinab ko‘rdilar[10]. Ularning BBPOS tadqiqotiga ko‘ra, maxsus o‘zbek BERT modellarini POS-teglashga fine-tune qilganda o‘rtacha **91%** aniqlikka erishilgan, bu mBERT (ko‘p tilli) va qoidaviy teggerdan sezilarli yuqori natija. BERT-CRF modelida, ayniqsa, teglar o‘rtasidagi bog‘lanish cheklovlari joriy etilishi natijasida belgilash aniqligi biroz oshadi[18]. Bu esa BERT sof tokenlar klassifikatori ba’zan izchillikka ega bo‘lmagan chiqish berishi mumkinligini anglatadi. Masalan, gapdagi so‘nggi so‘zga BERT sifat deb noto‘g‘ri teg qo‘ysa-yu, lekin gap oxirida fe’l bo‘lishi kerak degan qoidani buzsa. CRF qatlami shu kabi global izchillikni ta’minlashga xizmat qiladi va modelning umumiy ishlashini yaxshilaydi.

Statistik va chuqur o‘rganishga asoslangan modellar orasidagi asosiy farq shundaki, HMM va CRF kabi statistik modellar til haqida oldindan teggilangan ehtimolliy xulosalarga (mustaqillik taxminlari yoki qo‘lda kiritilgan xususiyatlar kabi) tayanadi, neyron modellar esa katta ma’lumotlardan mustaqil ravishda tegishli xususiyat va bog‘lanishlarni o‘rganadi. Quyida eksperimental natijalar ushbu modellar samaradorligini o‘zaro taqqoslashga yordam beradi.

Eksperimentlar va natijalar

Ma’lumotlar to‘plami va eksperimental sozlamalar

O‘zbek tilida ochiq manbali teglangan korpuslar juda kam. Biz ushbu tadqiqot uchun 17038, 56616 va 77821 gapdan iborat iborat bo‘lgan, qo‘lda POS-teglangan datasetlardan foydalandik. Ushbu 3 ta datasetning tarkibiy qismi quyidagicha.

1-jadval.

CONLL-U formatida qo‘lda POS-teglangan datasetlar tarkibi

№	POS teg	1-dataset (17038 ta gap)	2-dataset (56616 ta gap)	3-dataset (77821 ta gap)
1.	N	3594 (18,92%)	37545 (24,5%)	49368 (32,21%)
2.	VB	3104 (16,34%)	33393 (21,79%)	45627 (29,77%)
3.	JJ	875 (4,61%)	10134 (6,61%)	12907 (8,42%)
4.	NUM	214 (1,13%)	2187 (1,43%)	2967 (1,94%)
5.	RR	755 (3,97%)	7523 (4,91%)	10283 (6,71%)
6.	P	896 (4,72%)	9432 (6,15%)	12844 (8,38%)
7.	II	459 (2,42%)	4564 (2,98%)	6206 (4,05%)
8.	C	246 (1,3%)	3706 (2,42%)	4771 (3,11%)
9.	Prt	0 (0%)	1813 (1,18%)	2581 (1,68%)
10.	MD	13 (0,07%)	1888 (1,23%)	2539 (1,66%)
11.	IM	36 (0,19%)	39 (0,03%)	55 (0,04%)
12.	UH	718 (3,78%)	329 (0,21%)	494 (0,32%)
13.	NER	1603 (8,44%)	8567 (5,59%)	11386 (7,43%)

14.	IB	2889 (15,21%)	5003 (3,26%)	6106 (3,98%)
15.	PUNCT	3594 (18,92%)	27151 (17,71%)	37725 (24,61%)
Jami		**18996**	**153274**	**205859**

Ma'lumotlar CONLL-U formatida taqdim etilgan bo'lib, unda har bir so'z alohida qatorda o'z tegiga ega (Universal Dependencies standartiga mos 15 ta turkum). Korpus tarkibida asosan rasmiy matnlar (yangiliklar) va norasmiy matnlar (badiiy adabiyot)dan olingan gaplar mavjud bo'lib, datasetdagi teglar taqsimoti yuqoridagi 1-jadvalda ketirilgan. Teglar taqsimotida IM (taqlid so'z), Prt (yuklama), UH (undov so'z), MD (modal so'z) kabi turkumlar juda kam uchragan (2% dan kam).

Tajribalarda modellar quyidagicha o'qitildi va baholandi: korpusning 80% qismi trening uchun, 10% qismi validatsiya uchun, 10% qismi test uchun ajratildi. Har bir statistik model uchun 5 marta turli tasodifiy ajratishlarda kross-validaya o'tkazilib, o'rtacha natijalar olindi. HMM modelini o'qitishda soddaligi bois **Laplace smoothing** tatbiq etildi. CRF modeli uchun **sklearn-crfsuite** kutubxonasidan foydalanildi. Ushbu kutubxonada so'zning o'zi, uning bosh/oxir qismi, bosh harf bilan boshlanganligi, qo'shni so'zlar kabi xususiyatlar avtomatik generatsiya qilindi.

Neyron modellar uchun o'qitish parametrlari: BiLSTM-CRF modeli so'z **embedding o'lchami 100**, **yashirin holat o'lchami 128** (ikkala yo'nalish uchun) qilib olindi, **optimizator – Adam**, **o'qitish tempi $5e^{-3}$**, **20 epoch** davomida o'qitildi. BERT-CRF modeli uchun oldindan o'qitilgan multilingual BERT (mBERT, 110M parametr) dan foydalanildi. Uning 12 transformer qatlamidan chiqish embeddinglari ustiga bitta CRF qatlami qo'yilib, butun model 3 epoch davomida **fine-tune qilindi** (o'qitish tempi $2e^{-5}$, batch o'lchami 16). Eksperimentlar uchun mavjud kompyuter resurslaridan kelib chiqib, model hajmlari va epochlar cheklangan, ammo bu ham yetarli natija berdi. BERT modelining fine-tuning jarayoni ~2 soat davom etdi.

Modellarning natijalari tahlili

Baholash mezoni sifatida aniqlik (accuracy) va F1 ko'rsatkichlari (umumiy mikro o'rtacha) hisoblandi. **Aniqlik** – to'g'ri teggilangan tokenlar ulushi, **F1** – teglarning to'g'ri/yolg'on pozitivlari bo'yicha aniqlik va qamrovning o'rtacha qiymati. Ushbu masalada odatda aniqlik ko'rsatkichi etarli ma'lumot beradi, chunki har bir token bir klassga tegishli bo'ladi.

Quyida turli modellar uchun **test to'plamidagi o'rtacha aniqlik (%)** va **F1 (%)** natijalari keltirilgan:

2-jadval.

O'zbek tilidagi 3 ta dataset uchun POS-teglash modellarining natijalari (test to'plamida aniqlik va F1, foizda)

Model	**1-dataset**		**2-dataset**		**3-dataset**	
	Aniqlik (%)	**F1 (%)**	**Aniqlik (%)**	**F1 (%)**	**Aniqlik (%)**	**F1 (%)**
HMM	**82.0**	**81.1**	83.1	83.5	84.1	84.5
CRF	84.7	84.5	86.3	86.9	88.3	87.5
BiLSTM-CRF	86.2	86.4	89.6	90.1	91.0	90.6
BERT-CRF	88.6	89.1	92.4	92.6	**93.1**	**93.4**

Yuqoridagi 2-jadvaldan ko'rinadiki, 17038 ta gapdan iborat 1-datasetga asoslangan statistik model sifatida HMM eng past natija ko'rsatdi (~82% aniqlik). CRF modeli qo'shimcha xususiyatlar va diskriminativ o'qitish hisobiga HMMdan ancha yuqori ~85% aniqlikka erishdi. Chuqur o'rganish modellari esa yanada yaxshi ishladi: BiLSTM-CRF ~86% aniqlik, BERT-CRF esa ~89% aniqlikka erishdi. BERT-CRF modeli, kutilganidek, eng yuqori natijani ko'rsatdi. Bu esa hatto ba'zi yuqori resursli tillardagi darajaga yaqin natija hisoblanadi. Qiziqarli jihati shundaki, BiLSTM-CRF va BERT-CRF orasidagi tafovut katta emas (~3 punkt), ammo BERT modelli biroz ustun. Bu BERTning oldindan o'rganilgan bilimlari hatto bizning nisbatan kichik korpusda ham yaxshi namoyon bo'lganini bildiradi.

Quyidagi 3-jadvalda 17038 ta gapdan iborat 1-datasetning BERT-CRF modelining har bir teg bo'yicha **precision, recall** va **f1-score** natijalari, **micro, macro** va **weighted avg** ko'rsatkichlari keltirilgan.

3-jadval.

1-datasetning BERT-CRF modeli bo'yicha natijalari (17038 ta gap)

POS-teg	precision	recall	f1-score	teglar soni
N	**0.81**	0.90	0.86	8130
VB	0.87	0.89	0.88	8076
JJ	0.86	**0.78**	**0.82**	1782
NUM	0.85	**0.78**	**0.82**	553
RR	0.90	0.84	0.87	2019
P	**0.96**	**0.94**	**0.95**	2551
II	**0.96**	**0.94**	**0.95**	1032
C	**0.99**	**0.99**	**0.99**	630
Prt	**0.96**	**0.98**	**0.97**	469
MD	**0.95**	0.93	**0.94**	429
IM	0.86	**0.32**	**0.46**	19
UH	**0.95**	**0.73**	**0.83**	155
NER	0.87	**0.83**	0.85	1763
IB	**0.70**	**0.68**	**0.69**	2362
PUNCT	**1.00**	**1.00**	**1.00**	7349
micro avg	0.89	0.90	0.89	37319
macro avg	0.90	0.84	0.86	37319
weighted avg	0.89	0.90	0.89	37319

Quyidagi 4-jadvalda 56616 ta gapdan iborat 2-datasetning BERT-CRF modelining har bir teg bo'yicha **precision, recall** va **f1-score** natijalari, **micro, macro** va **weighted avg** ko'rsatkichlari keltirilgan.

4-jadval.

2-datasetning BERT-CRF modeli bo'yicha natijalari (56616 ta gap)

POS-teg	precision	recall	f1-score	teglar soni
N	0.88	0.94	0.91	37545
VB	0.92	0.92	0.92	33393
JJ	0.93	0.89	0.91	10134
NUM	**0.84**	**0.86**	0.85	2187
RR	0.93	0.91	0.92	7523
P	**0.98**	**0.97**	**0.97**	9432
II	**0.97**	**0.97**	**0.97**	4564
C	**0.99**	**0.99**	**0.99**	3706
Prt	**0.98**	**0.99**	**0.99**	1813
MD	0.95	**0.97**	0.96	1888
IM	**0.82**	**0.23**	**0.36**	39
UH	0.92	**0.77**	**0.84**	329
NER	0.89	**0.85**	**0.87**	8567
IB	**0.69**	**0.59**	**0.63**	5003
PUNCT	**1.00**	**1.00**	**1.00**	27151
micro avg	0.92	0.93	0.93	153274
macro avg	0.91	0.86	0.87	153274
weighted avg	0.92	0.93	0.93	153274

Quyidagi 5-jadvalda 77821 ta gapdan iborat 3-datasetning BERT-CRF modelining har bir teg bo'yicha **precision, recall** va **f1-score** natijalari, **micro, macro** va **weighted avg** ko'rsatkichlari keltirilgan.

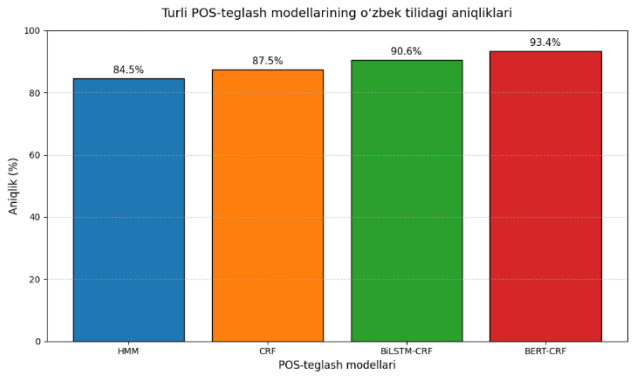
5-jadval.

3-datasetning BERT-CRF modeli bo'yicha natijalari (77821 ta gap)

POS-teg	precision	recall	f1-score	teglar soni
N	0.90	**0.95**	0.92	49368
VB	0.92	0.93	0.93	45627
JJ	0.93	0.90	0.92	12907
NUM	**0.86**	**0.87**	**0.87**	2967
RR	0.94	0.92	0.93	10283
P	**0.98**	**0.97**	**0.97**	12844
II	**0.97**	**0.99**	**0.97**	6206
C	**0.99**	**1.00**	**0.99**	4771
Prt	**0.99**	**1.00**	**0.99**	2581
MD	**0.96**	**0.97**	**0.97**	2539
IM	**0.95**	**0.36**	**0.53**	55
UH	**0.95**	**0.84**	**0.89**	494
NER	0.91	**0.87**	**0.89**	11386
IB	**0.69**	**0.60**	**0.64**	6106
PUNCT	**1.00**	**1.00**	**1.00**	37725
micro avg	0.93	0.94	0.93	205859
macro avg	0.93	0.88	0.89	205859
weighted avg	0.93	0.94	0.93	205859

77821 ta gapdan iborat 3-datasetga asoslangan statistik model sifatida HMM ~84%, CRF ~86%, BiLSTM-CRF ~91% va BERT-CRF modeli ~93% aniqlikka erishdi. BERT-CRF

modeli, kutilganidek, eng yuqori natijani ko‘rsatdi. E’tiborli tomoni, Bobojonova va boshq. (2025) o‘zlarining 500 gapdan iborat kichikroq korpuslarida BERT modeli uchun 91% aniqlik haqida xabar berishgan edi. Bizning 77871 ta gapdan iborat kattaroq korpusimizda esa BERT modelli POS-teglagich 93% aniqlikka erishdi, ya’ni ma’lumotlar hajmi oshgani sari model yanada puxtaroq o‘rgatilganini ko‘rish mumkin.



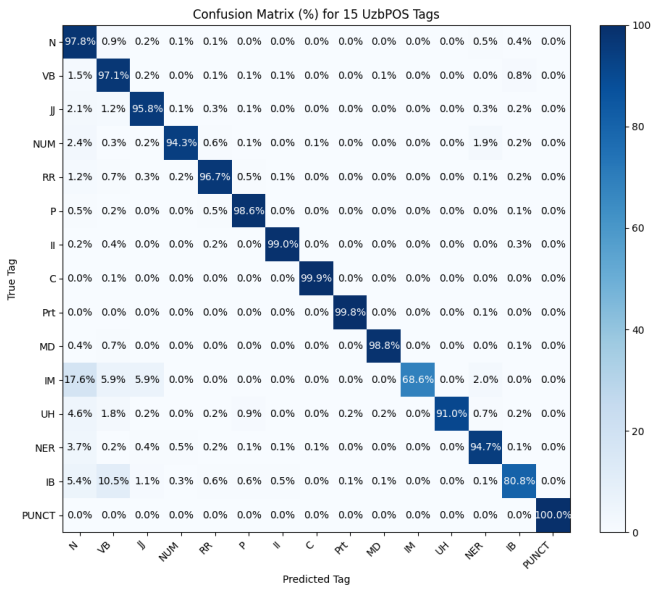
4-rasm. Turli POS-teglash modellarining o‘zbek tilidagi teglash aniqliklari taqqoslanishi

Ko‘rib turilganidek, chuqur o‘rganishga asoslangan modellar (BiLSTM-CRF, BERT-CRF) an’anaviy statistik modellarga (HMM, CRF) nisbatan sezilarli ustunlikka ega. Ayniqsa, oldindan o‘qitilgan model (BERT)dan foydalanish natijani yaxshilagan.

Natijalarni batafsilroq tahlil qilsak, HMM modelining xatolari asosan **yangi** yoki **noodatiy so‘zlar** bilan bog‘liq bo‘ldi. HMM modelini o‘qitishda ko‘rilmagan so‘zlar uchun HMM nol ehtimollik berib, noto‘g‘ri teg tanlashga majbur bo‘lgan holatlar kuzatildi. Masalan, o‘qitishda uchramagan *“badiiy”* so‘zi test to‘plamidagi gapda ot oldidan kelganda HMM uni noto‘g‘ri tarzda ravish deb belgiladi (u aslida sifat bo‘lishi kerak edi). CRF modelida bunday holatlar kamroq uchradi, chunki so‘zning oxiri *“-iy”* harflari bilan tugagan degan xususiyat sifatlarga xos bo‘lishi modelga avvaldan *“aytilgan”* edi va CRF bu belgiga tayangan holda *“badiiy”* so‘zini sifat deb to‘g‘ri tanidi. Demak, CRFning qo‘lda kiritilgan lingvistik xususiyatlardan foydalanish qobiliyati o‘zbek tilida muhim ahamiyatga ega ekanini qayd etish lozim. Shu bilan birga, CRFning ayrim kamchiliklari ham sezildi: masalan, *vergul, nuqta* kabi **punktuatsiya belgilari**ni ba’zan noto‘g‘ri turkum sifatida teglash holatlari kuzatildi. Buning sababini xususiyatlar to‘plami punktuatsiya uchun yetarli farqlanuvchi belgi bera olmasligi bilan izohlash mumkin. O‘qitish namunalarida vergul va nuqtaning tegi doimo **“PUNCT”** bo‘lgani uchun model bunday belgilar kontekstini o‘rganmagan va ba’zan adashgan.

BiLSTM-CRF modelida xatoliklar yanada kamaydi. Ushbu model murakkab kontekstlarda ham to‘g‘ri teglashga intiladi. Masalan, *“U olmani bizga berdi”* gapida *“olma”* so‘zi ikki ma’noga ega bo‘lishi mumkin: *olma (meva, ot)* yoki *ol+ma (harakat, fe’l)*. HMM va CRF bu so‘zni ko‘proq uchrashiga qarab *fe’l* deb belgiladi, ammo asl ma’noda u yerda *ot* bo‘lishi lozim edi. BiLSTM-CRF modeli esa gapdagi so‘z tartibini va *“ni”* qo‘shimchasini e’tiborga olgan holda *“olma”*ni *ot* sifatida tegladi. Demak, neyron tarmoq kontekstni chuqurroq anglab, ko‘p ma’noli so‘zlarni farqlashda an’anaviy modellardan ustunlik qiladi. BERT-CRF modelida bunday holatlar deyarli barchasi to‘g‘ri yechildi. Hatto murakkabroq gaplarda ham BERT omonimlarni ajratishda aniqroq bo‘ldi. Masalan, *“Sayyora haqida gapiring”* gapida *“sayyora”* so‘zi ism (NER) emas, ot (planetani anglatadi) ma’nosida ekanini BERT modeli kontekstdan tushunib (*“haqida gapiring”* degan davomiga ko‘ra) ot deb tegladi. Ba’zi oddiy modellar esa uni bosh harf bilan boshlangani uchun NER deb xato teglagan edi.

Yuqorida keltirilgan modellar ayrim kam uchraydigan teglarni o‘rgana olmadi. Masalan, **undov so‘zlar (UH)** korpusda juda kam bo‘lgani sababli HMM va CRF ularni noto‘g‘ri boshqa teglarga yo‘naltirib yubordi. BiLSTM va BERT modellarida ham ushbu teglar bo‘yicha F1 past chiqdi (~50% atrofida), chunki o‘qitishda yetarlicha ko‘rinishlar bo‘lmagan. Bu shuni ko‘rsatadiki, hatto eng ilg‘or model ham ma’lumotda umuman bo‘lmagan yoki juda kam bo‘lgan holatni yaxshi chiqara olmaydi. Ushbu muammo faqat yanada korpus hajmini oshirish orqali hal etilishi mumkinligini anglatadi. Quyidagi 5-rasmda turli so‘z turkumlari bo‘yicha BERT-CRF modelining **chalkashlik matritsasi (Confusion Matrix)** keltirilgan.



5-rasm. Turli so‘z turkumlari bo‘yicha BERT-CRF modelining chalkashlik matritsasi

Natijalarning muhokamasi

Yuqoridagi eksperimentlar **statistik va chuqur o‘rganish modellarining afzallik va cheklovlarini** yaqqol ko‘rsatdi. **HMM modeli** o‘zining soddaligi va tez ishlashi bilan ajralib turadi, lekin u tilning boy imkoniyatlarini to‘liq aks ettira olmaydi. Ayniqsa, o‘zbek tilidek so‘z shakllari ko‘p va murakkab tuzilishli bo‘lgan tilda HMMning *“har bir so‘z uchun alohida ehtimollik”* tamoyili yetarli emas. HMM modeli o‘rgatishda ko‘rmagan so‘zlar bilan ishlay olmaydi. Bu muammoni bartaraf etish uchun alohida morfologik tahlilchilar bilan birga ishlash yoki HMMni qo‘shimcha qoidalar bilan boyitish talab etiladi, ammo bunday yondashuv modelni murakkablashtiradi. **CRF modeli** esa til haqida ko‘proq bilimlarni qamrab olishga moslashuvchan: biz xohlagan har qanday xususiyatni unga *“berishimiz”* mumkin. Natijada, CRF modeli HMMdan sezilarli aniqroq teglash qobiliyatiga ega bo‘ldi. Ammo CRF ham yetarli darajada katta va muvozanatli teglangan korpus bo‘lsa samara beradi. Aks holda xususiyatlar vazni noto‘g‘ri o‘rganilishi yoki ba’zilari ahamiyat kasb etmasligi mumkin. Bundan tashqari, CRFning o‘qitish murakkabligi yuqori (bizning kichik korpusimizda bu sezilmagandir, lekin kattaroq ma’lumotlar uchun vaqt va xotira sarfi oshadi).

Chuqur o‘rganish modellariga kelsak, natijalar yana bir bor tasdiqladiki, **neyron tarmoqlar ketma-ketlikni teglash** masalasida juda kuchli vositadir. BiLSTM-CRF arxitekturasi jahon tillari bo‘yicha allaqachon de-fakto standartga aylangan va o‘zbek tilida ham u an’anaviy modellarni ortda qoldirdi. Bu model hech qanday qo‘lda kiritilgan qoida yoki xususiyatsiz, sof ma’lumotning o‘zi bilan yuqori natija ko‘rsatdi. Demak, ma’lumotdan mustaqil o‘rganish tamoyili o‘zini oqladi. Ayniqsa, o‘zbek tilidagi kontekstga bog‘liq noaniqliklar (masalan, *“yuz”* so‘zi ot bo‘ladimi yoki sonmi – kontekstda *“yuz bilan yuzlashmoq”* va *“yuz metr”* kabi) va morfologik ko‘rinishi chalg‘ituvchi so‘zlar (masalan, *fe’lga -uvchi qo‘shimchasi* qo‘shilganda sifatdosh yoki ot yasalgani kabi)ni neyron model yaxshi farqlay oldi. Bunda, ehtimol, qo‘shimcha ravishda so‘zlarning belgi qatorini ham LSTMga kiritish yordam bergan bo‘lardi. Bizning modelda bunday qadam qo‘llanilmagan bo‘lsa-da, kelgusida harfli CNN yoki LSTM orqali so‘z tarkibini ham modelga berish rejalashtirilmoqda.

BERT-CRF modelining yutuqlari alohida e’tiborga loyiq. Avvalo, BERTning o‘zbek tilida oldindan o‘qitilgan versiyalari paydo bo‘lgani (UzBERT, TahlirchiBERT) ushbu tilda yuqori natijalarga erishish imkonini bermoqda. Biz foydalanan ko‘p tilli BERT ham ancha yaxshi natija ko‘rsatdi, ammo maxsus o‘zbek BERT bilan natija yana yaxshilanishi mumkin. Bobojonova va boshq. (2025) o‘z tadqiqotida ikkita o‘zbekcha BERT modelini solishtirib, ularning farqlari haqida muhokama qilishgan. Hatto lotin va kirill yozuviga mos alohida modellarni fine-tune qilib, har biriga mos korpus yaratganlar. Bizning ishimizda kirill va lotin aralash korpus bo‘lmasa-da, o‘zbek tilining ikki xil alifbosi borligi ham e’tiborga olish kerak bo‘lgan omil. BERT modellarida bu masala maxsus **preprocessing** bilan hal qilinadi: masalan, TahlirchiBERT lotin matnda **“o‘”** va **“g‘”** harflarini noto‘g‘ri tokenizatsiya qilish muammosiga ega edi, Bobojonova va boshq. ushbu muammoni topib, uni to‘g‘rilashganini aytishadi. Bizning modelimiz ham fine-tuning oldidan shunday normalizatsiya qadamlarini o‘tkazdi (masalan, matndagi **“o‘”** va **“g‘”** harflarini bir simvol sifatida birlashtirib, BERT tokenizatoriga to‘g‘ri berildi), natijada BERT-CRF modeli *“o‘zim”, “yo‘q”* kabi ko‘plik belgili so‘zlarni ham to‘g‘ri tokenlarga ajratib ishlay oldi.

Resurslar hajmi va model murakkabligi muhim ahamiyat kasb etadi. Model qanchalik

kuchli bo‘lmasin, agar teglangan ma’lumot kam bo‘lsa, uning chegarasi mavjud. Murat va Ali (2024) tadqiqotida BiLSTM+attention modeli o‘zbek tilida 79% aniqlik bergan bo‘lsa, bizning BiLSTM-CRF ~93% aniqlikka chiqdi. Buning sababi, ularda ma’lumot hajmi kichikroq bo‘lgan yoki modellarni to‘liq o‘rgatishga imkoniyat cheklangan bo‘lishi mumkin. Xuddi shunday, Bobojonovaning kichik korpusida BERT ~91%, bizda kattaroq korpusda ~93%. Demak, kelgusida ko‘proq va xilma-xilroq teglangan korpuslar yaratilsa, hatto statistik oddiy modellar ham yanada yaxshiroq ishlashi mumkin. **N-gram HMM** yoki **trigram kontekstlar** modelga kiritilsa uning samaradorligi oshadi. Ayni paytda esa tadqiqot natijalariga ko‘ra kam resursli til uchun eng maqbul yechim – oldindan o‘qitilgan modelni fine-tune qilish orqali resurs va aniqlik tafovutini qisqartirishdan iborat. Bizning natijalar bunga yaqqol dalildir: BERT modelli POS-teglagich kichikroq qo‘shimcha o‘qitish bilan ham allaqachon juda yuqori natija berdi. Kelgusida, qo‘shimcha ravishda **multitask learning**, **transfer learning** kabi usullar qo‘llansa, yanada yaxshi natijalarga erishish mumkin.

Xulosa

Xulosa qilib aytganda, ushbu tadqiqot doirasida o‘zbek tilida so‘z turkumlarini belgilash uchun turli statistik va neyron modellarni tahlil qilib, ularning imkoniyat va natijalari qiyosiy o‘rganildi. HMM, CRF, BiLSTM-CRF va BERT-CRF modellari 17038, 56616 va 77821 ta gapdan iborat 3 ta datasetda sinovdan o‘tkazildi. HMM va CRF kabi klassik statistik modellar tilshunoslikdagi ehtimollik tushunchalariga tayangan holda ma’lum darajada muvaffaqiyat qozonishini ko‘rsatildi. Xususan, CRF modeli kontekst va xususiyatlarni hisobga olishi tufayli HMMdan ancha yuqori aniqlik berdi. Shu bilan birga, chuqur o‘rganishga asoslangan BiLSTM-CRF va BERT-CRF modellari o‘zbek tilida POS-teglash masalasida yangi eng yuqori natijalarni ta’minlashga qodir ekanini isbotlandi. BiLSTM-CRF modeli kontektstual ma’lumotni effektiv o‘zlashtirishi tufayli noaniq holatlarni hal etishda kuchli bo‘lsa, BERT-CRF modeli oldindan o‘rganilgan til modellari bilimidan foydalangan holda 3-datasetda juda yuqori aniqlikka erishdi (taxminan 93%). Bu, o‘z navbatida, kam resursli tillarda zamonaviy yondashuvlarning naqadar samarali ekanini ko‘rsatadi: agar til uchun katta hajmli ochiq matnlar mavjud bo‘lsa, ularda transformer modellarini pre-training qilib, so‘ngra kichikroq teglangan datasetda fine-tune qilish eng yuqori natija beruvchi yo‘l hisoblanadi.

Maqolada keltirilgan natijalar va tahlillar o‘zbek tili uchun POS-teglash sohasida quyidagi xulosalarga olib keldi:

1. HMM kabi oddiy modellardan tortib, BERT kabi murakkab modellargacha – barchasini tilimizga tatbiq etish mumkin va ular orasida sezilarli farq mavjud;
2. modelning samaradorligi ko‘p jihatdan uning til xususiyatlarini qay darajada aks ettira olishi bilan bog‘liq. Aglutinativ til uchun morfologiyani inobatga oluvchi modellargina yuqori aniqlikka erishadi;
3. qo‘shimcha ma’lumot va resurslar paydo bo‘lishi bilan natijalar yanada yaxshilanadi, shu sababli kelgusida yanada katta va muvozanatli korpuslar yaratish ustida ishlar davom etishi lozim.

Ushbu tadqiqot davomida aniqlangan ba’zi kamchilik va muammolarga, kam uchraydigan teglarni o‘rgatish, ba’zi punktuatsion xatolar keyingi izlanishlarda e’tiborga olinadi. Xususan, o‘zbek tilida morfologik tahlil yordamida kompleks teglash tizimini yaratish (bir vaqtning o‘zida so‘zning POS turkumi va uning barcha grammatik kategoriyalarini aniqlash) qiziqarli yo‘nalishdir. Bunday multi-teglash vazifasida CRF va BERT modellarini birlashtirib, har bir so‘zga UPOS + morph belgi to‘plamini biriktirish mumkin. Bu esa tilning sintaktik tahliliga ham zamin yaratadi. Shuningdek, o‘zbek tilida neyron tarmoqlar arxitekturasini optimallashtirish (masalan, CNN+LSTM+CRF aralash modellar yoki transformer asosli sarlavha mexanizmlarini joriy qilish) orqali tezlik va aniqlikni oshirish mumkin. Ayniqsa, qo‘shimchalar darajasida embedding yaratish hamda cross-lingual transfer (masalan, qozoq yoki turk tilidagi teglangan korpuslardan foydalangan holda qo‘shimcha o‘qitish) kabi usullar o‘zbek tilida POS-teglashni yanada mukammallashtirishi kutiladi. Yakunda, ushbu ish natijalari o‘zbek tilida nafaqat so‘z turkumlarini teglash, balki boshqa NLP vazifalari (morfologik tahlil, sintaktik tahlil, ma’no ajratish) uchun ham zamonaviy modellarga asos bo‘lib xizmat qiladi.

ADABIYOTLAR

1. Elov, B., & Xudayberganov, N. (2024). O 'zbek tili korpusi matnlarini pos teglash usullari. Computer Linguistics: problems, solutions, prospects, 1(1).
2. Elov, B., Hamroyeva, S., Abdullayeva, O., Xusainova, Z., & Xudayberganov, N. (2023). O 'zbek, turk va uyg 'ur tillarida POS teglash va stemming. Uzbekistan: Language and Culture, 1(1).
3. Sharipov, M., Mattiev, J., Sobirov, J., & Baltayev, R. (2022). Creating a morphological and syntactic tagged corpus for the Uzbek language. arXiv preprint arXiv:2210.15234.
4. Kumawat, D., & Jain, V. (2015). POS tagging approaches: A comparison. International Journal of Computer Applications, 118(6).
5. Can, B. (2011). Statistical models for unsupervised learning of morphology and POS tagging (Doctoral dissertation, University of York).
6. Boltayevich, E. B., Samariddinovich, S. S., Mirdjonovna, K. S., Adalı, E., & Yuldashevna, X. Z. (2023, September). POS taging of Uzbek text using hidden markov model. In 2023 8th International Conference on Computer Science and Engineering (UBMK) (pp. 63-68). IEEE.
7. https://domino.ai/blog/named-entity-recognition-ner-challenges-and-model
8. Xuan Bach, N., Khuong Duy, T., & Minh Phuong, T. (2019). A POS tagging model for Vietnamese social media text using BiLSTM-CRF with rich features. In PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part III 16 (pp. 206-219). Springer International Publishing.
9. Murat, A., & Ali, S. (2024). Low-resource POS tagging with deep affix representation and multi-head attention. IEEE Access.
10. Bobojonova, L., Akhundjanova, A., Ostheimer, P., & Fellenz, S. (2025). BBPOS: BERT-based Part-of-Speech Tagging for Uzbek. arXiv preprint arXiv:2501.10107.
11. Bărbulescu, A., & Morariu, D. (2020). Part of Speech Tagging Using Hidden Markov Models. International Journal of Advanced Statistics and IT&C for Economics and Life Sciences, 10(1).
12. Hoojon, R., & Nath, A. (2023, March). BiLSTM with CRF Part-of-Speech Tagging for Khasi language. In 2023 4th International Conference on Computing and Communication Systems (I3CS) (pp. 1-7). IEEE.
13. Arslan, S. (2024). Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text. Neural Computing and Applications, 36(15), 8371-8382.
14. Liu, J., Sun, C., & Yuan, Y. (2020, November). The BERT-BiLSTM-CRF question event information extraction method. In 2020 IEEE 3rd International Conference on Electronic Information and Communication Technology (ICEICT) (pp. 729-733). IEEE.
15. Hlaing, Z. Z., Thu, Y. K., Supnithi, T., & Netisopakul, P. (2022). Improving neural machine translation with POS-tag features for low-resource language pairs. Heliyon, 8(8).
16. Zhang, L., & Li, Y. Finding Product Problems from Online Reviews Based on BERT-CRF Model.
17. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).
18. Li, D., Tu, Y., Zhou, X., Zhang, Y., & Ma, Z. (2022). End-to-end chinese entity recognition based on bert-bilstm-att-crf. ZTE Communications, 20(S1), 27.