UO‘K 004.85:81.322

MASHINALI O‘QITISH ALGORITMLARI ASOSIDA O‘ZBEK TILI MATNLARIDAGI IMLO XATOLARINI ANIQLASH VA TUZATISH

[+] *Ochilov M.M.*[1], *Narzullayev O.O.*[1], *Xolmatov O.A.*[1]

[1] Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Toshkent, O‘zbekiston

[+] ochilov.mannon@mail.ru

Annotatsiya. Ushbu tadqiqotda o‘zbek tili matnlaridagi imlo xatolarni aniqlash va tuzatish muammosi ko‘rib chiqiladi. O‘zbek tilining murakkab morfologik tuzilishi va agglutinativ xususiyatlari tufayli an’anaviy imlo xatolarni tekshirish usullari yetarli natija bermaydi. Shuning uchun ushbu ishda Levenshteyn masofasi algoritmi yordamida so‘z o‘xshashligini hisoblash va neyron tarmoqlarga asoslangan til modellari orqali kontekstual tuzatish mexanizmlari ishlab chiqildi. Til modeli sifatida KenLM (statistik til modeli), LSTM (Long Short-Term Memory) va BiLSTM (Bidirectional LSTM) yondashuvlari qo‘llanildi. Modelni o‘qitish uchun 80 million so‘zdan iborat matn korpusi yig‘ildi va tahlil qilindi. Test natijalari shuni ko‘rsatdiki, BiLSTM modeli imlov xatolarni tuzatishda eng yuqori samaradorlikni (90.09%) ta’minladi, LSTM modeli esa 84.62% natijani qayd etdi. KenLM modelidan foydalangan holda esa samaradorlik 62.31% ni tashkil etdi.

Kalit so‘zlar: O‘zbek tili, imlo xatolarini tuzatish, tabiiy tilni qayta ishlash (NLP), Levenshteyn masofasi, til modeli, KenLM, LSTM, BiLSTM, neyron tarmoqlar, kontekstual tahlil, mashinali o‘qitish, agglutinativ tillar, imlo tekshiruvi, chuqur o‘rganish, statistik modellar.

1 KIRISH

So‘nggi o‘n yilliklarda dunyo bo‘ylab tadqiqotchilar tomonidan tabiiy tilni qayta ishlash (NLP) sohasi turli masalalarida sezilarli yutuqlarga erishildi, shu jumladan O‘zbek tilida olib borilayotgan tadqiqotlarda ham turli masalalarda sezilarli darajada yutuqlarga erishildi. E’tibor qaratishimiz kerak bo‘lgan masalalardan biri bu O‘zbek tilida yozilgan matn imlo xatolarni to‘g‘irlashdir. So‘zlarda imlo xatolar mavjudligi matnli ma’lumotlar, undan foydalanishda yetarli darajada o‘zining ta’sirini ko‘rsatishi mumkin. Ko‘plab tillarida, xususan, o‘zbek tilida, matnlarda imlo xatolar mavjudligi yoki gaplardagi konteks tog‘ri shakllanmagani sababli insonlar va avtomatlashtirilgan tilni qayta ishlash tizimlari uchun jiddiy muammo yuzaga keltirishi mumkin. Shu sababli murakkab til tuzilishiga ega bo‘lgan, xususan, o‘zbek tilida yozilgan matnlardagi gaplarni tiklay oladigan samarali modellarni yaratish muhim ahamiyat kasb etadi. Ushbu maqolada o‘zbek tilidagi matnlardagi imlo xatolarni tuzatishning bir qancha usullari ko‘rsatib beriladi hamda tadqiqot natijalariga asoslangan holda eng yuqori natijalarni qayt etgan yondashuv taklif etiladi.

NLP sohasida so‘zlarda imlo xatolarni tuzatish masalasi o‘tgan asrning 60 yillarida boshlangan. Xususan so‘zlarda imlo xatolarni to‘g‘irlash uchun so‘z o‘xshashligi usuli yordamida imlo xato mavjud bo‘lgan so‘z bilan asl so‘zlar lug‘atidagi eng o‘xshash so‘zni topish algoritmlari ishlab chiqilgan. Bunday usullarga misol keltirib Levenshteyn masofasini olsak bo‘ladi. Tadqiqot davomida aynan mana shu so‘z o‘xshashligi yordamida imlo xatoligi bor so‘zga yaqin bo‘lgan to‘g‘ri so‘zlar topiladi. Buning uchun o‘zbek tilidagi so‘z turkumlari bilan lug‘ati hosil qilindi. Ammo bu usullarni o‘zigina o‘zbek tilidagi gaplardagi imlo xatolarni to‘g‘ri tuzatish uchun yetarli emas. Chunki tavsiya qilinayotgan so‘zlarning bazilari o‘xshashlik jihatidan imlo xatolikga ega so‘zga bir xil masofada yaqin keladi. Bunda tavsiya etilayotgan so‘zlardan gap konteksiga mos tushadigan eng kerakli so‘z tanlab olish masalasi ham yuzaga keladi. Shu boisdan ham tadqiqotda o‘zbek tili uchun til modeli turli yondashuvlar, jumladan KenLM, LSTM hamda BiLSTM bilan qurib chiqildi hamda tavsiya etilgan so‘zlardan til modeli bo‘yicha eng mos keladigan so‘zni ajratib olishimiz mumkin bo‘ldi.

O‘zbek tili matnlaridagi imlo xatolarini aniqlash va tuzatish bo‘yicha tadqiqotlar so‘nggi yillarda sezilarli rivojlanish kasb etmoqda. Bunda mashinali o‘qitish algoritmlaridan foydalanish matnlarni avtomatik qayta ishlashda yangi imkoniyatlar yaratmoqda. Ushbu bo‘limda aynan mashinali o‘qitish sohadagi ilg‘or ishlanmalar va ularga bog‘liq tadqiqotlar ko‘rib chiqiladi.

Imlo xatolarini aniqlash va tuzatish masalasida turli tillar bo‘yicha amalga oshirilgan ishlanmalar asosida ikki asosiy yondashuvni ajratib ko‘rsatish mumkin: qoidaga asoslangan va statistik yondashuvlar.

Norvig (2007) tomonidan taklif etilgan qoidaga asoslangan yondashuv oddiy va samarali usullardan biri bo‘lib, so‘zlar orasidagi ehtimoliy moslikni aniqlash uchun Levenshtein distance kabi algoritmlardan foydalanadi[1]. Ushbu yondashuvlarning afzalligi oddiyligi va izohli ma’lumotlarga bo‘lgan ehtiyojning kamligidir. Ammo uning asosiy kamchiliklaridan biri bu tilning kontekstual xususiyatlarini inobatga olmasligidir. Ushbu yondashuv faqat individual so‘zlarni tahlil qiladi va ularning kontekst bilan bog‘liqligini hisobga olmaydi. O‘zbek tili kabi boy morfologiyaga ega agglutinativ tillar uchun bunday usul yetarli bo‘lmasligi mumkin, chunki imlo xatolari ko‘pincha qo‘shimcha yoki so‘z shakllari bilan bog‘liq bo‘ladi.

So‘nggi yillarda neyron tarmoqlarga asoslangan yondashuvlar, xususan, Recurrent Neural Networks (RNN) va Transformer arxitekturasi asosida ishlovchi modellar ommalashmoqda. Devlin va boshqalar [2] tomonidan taklif etilgan BERT modeli matnlarni kontekstual tahlil qilish orqali noaniqliklarni aniqlashda samarali natijalar ko‘rsatdi. Biroq u asosan katta hajmdagi ma’lumotlar va resurslarga ega tillar uchun moslashtirilgan. O‘zbek tilining nisbatan kichik korpusi ushbu modelning to‘liq imkoniyatlaridan foydalanishni qiyinlashtiradi. Shuningdek, BERT kabi yondashuvlarni qo‘llash uchun katta hisoblash resurslari talab qilinadi, bu esa uni har doim samarali tanlovga aylantirmaydi. O‘zbek tilida ushbu modelni moslashtirish uchun katta hajmdagi ma’lumotlarning yetishmasligi muammosi ham mavjud.

O‘zbek tili matnlarini tahlil qilish bo‘yicha nisbatan cheklangan bo‘lsa-da, boshqa turkiy tillar uchun yaratilgan modellar va korpuslar mazkur sohada foydalanish uchun yaxshi asos bo‘lib xizmat qilmoqda. Masalan, Eryiğit turkiy tillar uchun morfologik analiz va sintez vositalarini ishlab chiqqan [3]. Ushbu yondashuvlarning ba’zilari o‘zbek tili uchun moslashtirilishi mumkin. Lekin uni to‘g‘ridan-to‘g‘ri o‘zbek tiliga tatbiq etish masalasi muammo sanaladi. Turkiy tillarning umumiy xususiyatlariga qaramay, o‘zbek tilining morfologik tuzilishi va leksik xususiyatlari o‘ziga xosdir. Shu bois, ushbu vositalar va modellarni to‘liq moslashtirish uchun qo‘shimcha modifikatsiya qilish zarur.

Shuningdek, o‘zbek tili uchun maxsus korpuslarning yetishmasligi ushbu yo‘nalishdagi asosiy muammolardan biri hisoblanadi. Abduazimov va boshqalar [4] o‘zbek tilidagi matnlar uchun korpus yaratish va uni mashinali o‘qitish uchun tayyorlash bo‘yicha tadqiqotlar olib bordilar. Bu korpus mashinali o‘qitish modellarini o‘qitishda foydalanish uchun muhim ahamiyatga ega. Ammo uning hajmi va qamrovi cheklangan. Korpusdagi ma’lumotlarning turlicha sohalarni qamrab olmasligi yoki ma’lumotlarning noaniqligi mashinali o‘qitish modellari samaradorligini pasaytirishi mumkin. Bundan tashqari, ushbu korpusda kontekstual xatolarni tahlil qilish uchun maxsus belgilangan qatorlar yetarli darajada ko‘rsatilmagan bo‘lishi mumkin.

Ko‘pgina tadqiqotlarda o‘zbek tilidagi imlo xatolarini aniqlash va tuzatishning murakkabligi o‘zbek tili agglutinativ xususiyatlari, boy morfologik strukturasiga bog‘liq ekani qayd etiladi. Shu sababli, maxsus yondashuvlar va zamonaviy algoritmlar asosida matnlarni qayta ishlash strategiyalarini ishlab chiqish dolzarb masala hisoblanadi.

2 ASOSIY QISIM

Ushbu qismda o‘zbek tili uchun imlo xatolarni tuzatish uchun dastlab o‘quv to‘plami (Dataset) ya’ni o‘zbek tilidagi gaplar to‘plamini yig‘ish hamda imlo xatolarni esa tuzatish uchun o‘quv to‘plam asosida o‘zbek tilining til modelini (Language model) statiktik usul (KenLM) hamda chuqur o‘qitish neyron tarmoqlari xususan RNN ning LSTM hamda BiLSTM arxitekturalari yordamida 3 ta til modeli quriladi va ularnining samaradorligi solishtirib chiqiladi. Bundan tashqari neyron tarmoqlari asosida til modellarini qurish jarayonida o‘zbek tilidagi so‘zlar uchun maxsus tokenizer ishlab chiqiladi. Shuningdek imlo xato bor so‘zga eng yaqin so‘zni topishda Levenshtein masofasi algoritmidan foydalaniladi. Levenshtein distance algoritmda xato so‘zga to‘g‘ri so‘zlarni tavsiya qilish maqsadida o‘zbek tilidagi so‘zlardan iborat lug‘atni tavsiflanadi. Umumiy holda o‘zbek tilidagi matnlarda imlo xatolarni tuzatuvchi algoritmni quyidagi tartibda loyihalanadi (1-rasm).

Keltirlgan algoritmga ko‘ra:

- x_i - xatolik mavjud bo‘lgan so‘z;
- $[x_1, x_2, x_3 ... x_n]$ - xatolik mavjud bo‘lgan so‘z uchragan gap;
- $[h_1, h_2, h_3 ... h_n]$ - so‘z o‘xshashligi bilan topilgan xatolik mavjud bo‘lgan so‘zga eng yaqin masofadagi tavsiya etilgan so‘zlar;
- h_i – tavsiya etilgan so‘zlar orasidan til modeli yordamida olingan gap ketma-ketligiga mos keluvchi eng yuqori ehtimoldagi so‘z;
- $x_1, x_2, x_3 ... h_i ... x_n$ - to‘g‘irlangan gap.

2.1 Levenshtein masofasi algoritmi

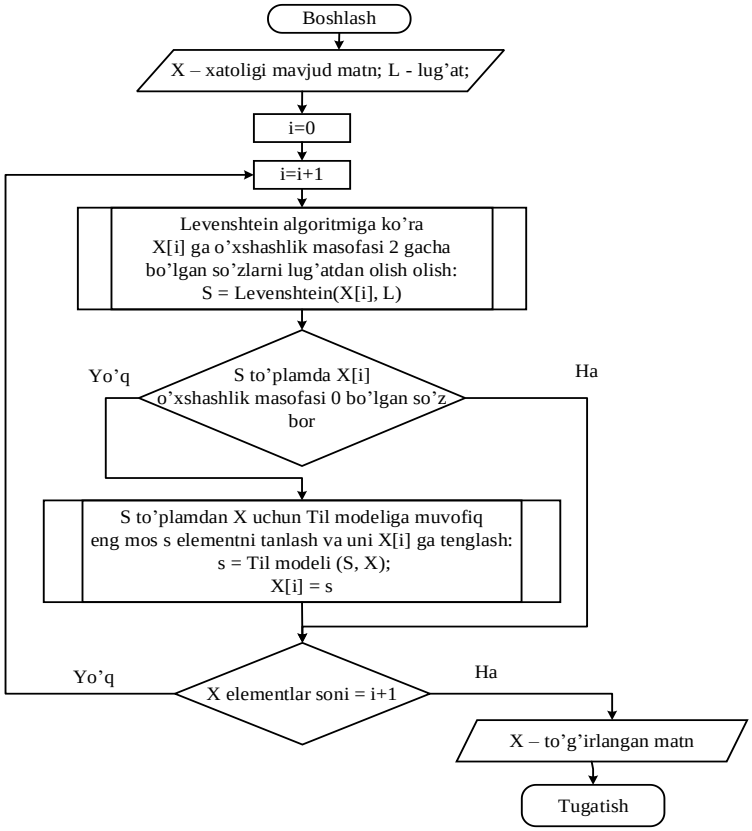
Levenshtein masofasi algoritmi – ikki matn so‘z o‘rtasidagi farqni o‘lchash uchun ishlatiladigan matematik usuldir. Ushbu masofa bitta so‘zni boshqasiga aylantirish uchun zarur bo‘lgan minimal tahrirlar sonini hisoblaydi. Bu tahrirlar qo‘shish, o‘chirish va almashtirish operatsiyalaridan iborat. Qo‘shish bu bir belgi qo‘shish, o‘chirish bir belgi o‘chirish, almashtirish bu bir belgi boshqasi bilan almashtirish hisoblanadi. Algoritm birinchi marta 1965-yilda rus matematigi Vladimir Levenshtein tomonidan taklif qilingan. Hisoblash formulasi quyidagicha:

Agar $D(i,j)$ A va B so‘zlarining mos ravishda dastlabki i va j belgilaridan iborat qismlari orasidagi minimal tahrir masofasi bo‘lsa, quyidagi rekursiv tenglama ishlatiladi:

$$D\,|\,i\,\|\,j\,|=\begin{cases} i,\ agar\ j=0 \\ j,\ agar\ \ i=0 \\ \min(D|i-1\|\,j|+1, D|i\|\,j-1|+1, D|i-1|+c(i.j)) \end{cases}. \quad (1)$$

$c(i,j)$ – agar $A[i]=B[j]$ bo‘lsa 0 aks holda 1 qiymatga ega yani harflar almashtirish qiymati. Algoritmga ko‘ra ba’zi namunalari quyidagilar:

- “kitob” va “kitoblar” so‘zlari Levenshtein masofasi 3 ga (3 ta harf qo‘shish);
- “gul” va “pul” so‘zlarining o‘xshashligi 1 ga (bitta harf almashtirish);
- “Saida” va “Said” so‘zlarining o‘xshashligi esa 1 ga (bitta so‘z o‘chirishga) tengdir.



1-rasm. O‘zbek tilidagi matnlarda imloviy xatoliklarini to‘g‘irlash algoritmi

2.2 O‘quv to‘plami va lug‘at

Har qanday til modelini yaratishda o‘quv to‘plamining o‘rni katta. Tadqiqotlar shuni ko’rsatadiki, kattaroq ma’lumotlar to’plamlari sun’iy intellekt modellarining aniqligini oshiradi. Uzluksiz o’zbek tilidagi nutqni aniqlash tizimi uchun til modellarini ishlab chiqish bo’yicha sa’y-harakatlar doirasida taxminan 15 million jumladan, jami 80 million so’zdan iborat matn korpusini tuzildi. Bu jumlalar orasida taxminan 68 000 tasi takrorlanmaydi [5]. Boshqa bir ishlarda esa o’zbek tilida nutq tahlili tizimlarini yaratishda neyron tarmoq modelini o’rgatish uchun nutq korpusini yaratish uchun 2 millionga yaqin gaplardan foydalanilgan [6,7,8]. Bundan tashqari, davlatning rasmiy veb-saytlaridan (masalan, lex.uz huquqiy hujjatlar platformasi [9]) matnlar to’plangan. Til modelini qurish uchun korpusda quyidagi qayta ishlar amalga oshirildi:

- Gaplardagi harflarning barchasini quyi registrga o‘tkazildi;
- Gaplarni punktuatsiya belgilaridan tozalandi;
- Takrorlangan hamda til modeli uchun xos bo‘lmagan gaplarni olib tashlandi.

Bizning to‘plangan matn korpusi namunaviy o‘qitishda bo‘linish nisbati 1-jadvalda keltirilgan.

1-jadval. O‘quv to‘plami xususiyatlari

To‘plam	Foizi (%)	Gaplar	So‘zlar	Takrorlanmas so‘zlar
Train to‘plami	80	17 323 316	77 799 044	63 684
Test to‘plami	20	4 330 829	19 449 761	15 921
Jami	100	21 654 145	97 248 805	79 605

O‘xshash so‘zlarni qidirib topish uchun o‘zbek tilidagi so‘zlarining lug‘atini tuzildi. Bunda turli xil lug‘at kitoblaridan (masalan, Inglizcha-o‘zbekcha va o‘zbekcha-inglizcha lug‘at [10]) foydalanildi. Lug‘at so‘zlarni o‘z ichiga olish bilan bir qatorda so‘zlar turkumlarini ham o‘z ichiga oladi. So‘zlar yig‘ib bo‘linganidan so‘ng so‘zning turkumiga qarab ushbu so‘z olishi mumkin bo‘lgan qo‘shimchalar (masalan, ot so‘z turkumi uchun egalik, kelishik qo‘shimchalari so‘zlarning birlik va ko‘plik shakli) bilan yangi so‘zlarni yasaldi va lug‘at kengaytirildi. 2-jadvalda lug‘atning so‘z turkumlari bo‘yicha asosiy va qo‘shimchalar bilan kengaytirilgan holatdagi taqsimoti keltirilgan.

2-jadval. Lug‘atlar taqsimoti

Lug‘at	Jami	Ot	Sifat	Son	Fe’l	Olmosh	Ravish
Asosiy	15150	9886	3002	26	1755	49	432
Kengaytirilgan	292103	227378	12008	547	43955	1127	9088

2.3 Til modeli

Til modeli — bu kompyuter dasturi yoki algoritm bo‘lib, u matn yoki tilning statistik va kontekstual xususiyatlarini o‘rganib, tabiiy tilni tushunish, qayta ishlash va hosil qilish imkonini beradi. Til modeli matndagi so‘zlar, iboralar yoki jumlalarning bir-biri bilan bog‘liqligini aniqlash uchun ishlatiladi. Til modellarini qurish va rivojlantirishda turli yondashuvlar mavjud. Ushbu usullar o‘zaro tilning murakkabligiga, ma’lumotlar hajmiga, texnik imkoniyatlarga va tadqiqotning asosiy maqsadlariga bog‘liq.

2.3.1. Qoidaga asoslangan yondashuvlar (Rule-Based Approaches). Bu usul lingvistik qoidalar, grammatika qoidalari va qo‘lda tuzilgan lug‘atlar asosida ishlaydi. Bu usulning afzalliklari:

- Oddiy va ma’lumotga uncha bog‘liq emas;
- Tilda mavjud bo‘lgan qoidalar o‘zgaruvchan emasligi sababli moslashuvchanlik talab qilinmaydi.

Ammo usulning kamchiligi yirik va murakkab tillar uchun qoidalar to‘plamini qo‘lda ishlab chiqish qiyinligi hamda kontekstni tushunish imkoniyati cheklanganligidandir. Shuning uchun ham bu usul til modellarini qurishning zamonaviy usuli hisoblanmaydi.

2.3.2. Statistik til modellarini qurish (Statistical Language Models - SLMs). Statistik til modellarida ma’lumotlar korpusidagi so‘zlar ketma-ketligi va ularning ehtimolliklari asosida til modeli quriladi. Masalan, N-gram modellar ushbu usulga misol bo‘la oladi. Ushbu usulning afzalliklari:

- Oddiy implementatsiya.
- To‘plangan ma’lumotlar asosida tilni modellashtirish.

Biroq bu usullarda ma’lumotlar hajmi cheklangan bo‘lsa, samaradorlik pasayadi hamda kontekstning uzoq masofali bog‘liqligini inobatga olmaslik yuzaga keladi.

2.3.3. Neyron tarmoqlar asosidagi modellar (Neural Network-Based Models). Hozirda til modellarini yaratishda eng ommabop usul hisoblanadi. Neyron tarmoqlar kontekstual xususiyatlarni avtomatik aniqlash va chuqur o‘rganishga imkon beradi. Xususan RNN (Recurrent Neural Networks) asosida ishlovchi modellar, masalan, LSTM va GRU uzoq muddatli bog‘lanishlarni o‘rganishga qodir. Ushbu usullarning afzalliklari:

- Kontekstni yaxshi tushunadi.
- Katta hajmdagi ma’lumotlar bilan ishlashda samaradordir.

2.4 KenLM asosida til modeli qurish

KenLM — Statistik til modellarini qurish uchun ishlatiladigan ochiq kodli kutubxona bo‘lib, u asosan N-gram til modelini samarali o‘rganish va qo‘llash uchun mo‘ljallangan. Bu kutubxona N-gramlar asosidagi ehtimollik modellarini o‘qitish va ulardan foydalanish uchun maxsus optimallashtirilgan

yondashuvlarni taqdim etadi. KenLM N-gram modellarni qurish uchun quyidagi asosiy qadamlarni amalga oshiradi:

- *Ma’lumotlarni o‘qitish.* Berilgan matn korpusida so‘zlar va ularning ketma-ketliklari (N-gramlar)ni hisoblaydi. Har bir N-gram uchun ehtimollikni hisoblab, uni xotirada yoki diskda samarali saqlash formatida yozadi.

- *Ehtimollikni hisoblash*. Modellar berilgan so‘zlar ketma-ketligining (masalan, so‘zlar orasidagi bog‘liqlik) ehtimolini hisoblaydi. Ushbu ehtimollik keyingi so‘zlarni bashorat qilish, avtomatik to‘ldirish yoki xatolarni aniqlash kabi vazifalarda qo‘llaniladi;

- *Optimallashtirilgan xotira ishlatilishi*. KenLM tezlik va samaradorlik uchun maxsus algoritmlar bilan ishlab chiqilgan. U N-gram modellarni katta hajmdagi ma’lumotlar bilan ishlashda ham samarali qiladi.

Endi KenLM kutubxonasining qanday ishlashiga misol. Aytaylik *"Kitob do‘konga borib kitob sotib oldi."* kabi korpus bo‘lsin. N=3 sifatida olaylik va 3-gramlar 3-jadvalda ko‘rsatilganidek bo‘ladi.

3-jadval. KenLM kutubxonada til modelini qurish uchun 3-gramlarni yaratish

kitob	do‘konga	borib
do‘konga	borib	kitob
borib	kitob	sotib
kitob	sotib	oldi

Har bir 3-gram uchun ehtimollik hisoblanadi. Misol:

$$P = \frac{Soni(so‘z_1, so‘z_2, so‘z_3)}{Soni\left(so‘z_1, so‘z_2\right)} = \frac{Soni(do‘konga, borib, kitob)}{Soni\left(do‘konga, borib\right)}. \quad (2)$$

Ushbu ehtimollik $So‘z_1$ va , $So‘z_2$ dan keyin $So‘z_3$ ni kelish ehtimolligidir ya’ni “do‘konga borib” dan keyin “kitob” ni kelish ehtimoli. KenLM N-gramlarni samarali indekslash va tezkor qidirish uchun maxsus saqlash formatidan foydalanadi.

Til modeli so‘zlar ketma-ketligining ehtimolini hisoblaydi va keyingi so‘zni bashorat qiladi yoki matnning grammatika bo‘yicha to‘g‘riligini baholaydi.

Misol: Korpusda quyidagi o‘qitilgan jumlalar bo‘lsin:

- "Men bugun bozorga bordim."
- "Men bugun kitob o‘qidim."

Berilgan so‘zlar: "Men bugun bozorga..." endi undan keyin keluvchi sifatida “bordim” va “o‘qidim” ni baholaydi:

$$P\left(bor\,\dim \mid Men, bugun, bozorga\right) > P\left(o‘qi\,\dim \mid Men, bugun, bozorga\right).$$

Model yuqori ehtimollikda "bordim" so‘zini tanlaydi, chunki u o‘rgatilgan korpusda ko‘proq uchraydi.

KenLM katta hajmdagi ma’lumotlar bilan ishlash uchun xotirada samarali saqlash mexanizmlarini qo‘llaydi. Bu optimallashtirilgan xotira ishlatilishini ta’minlaydi.

Yuqorida ko‘rsatilgan tartibda KenLM til modelini o‘quv to‘plamimiz yordamida qurildi. Bunda uch grammlardan foydalanildi.

2.5 LSTM va BiLSTM yordamida til modeli qurish

LSTM (Long Short-Term Memory) — bu Recurrent Neural Network (RNN) tarmog‘ining maxsus turi bo‘lib, uzoq muddatli bog‘liqlikni (long-term dependencies) saqlab qolish uchun ishlab chiqilgan. LSTM RNN larning an’anaviy muammolaridan biri bo‘lgan *"vanishing gradient"* (yo‘qoluvchi gradient) muammosini hal qiladi. U vaqt qatorlari, tabiiy tilni qayta ishlash (NLP), ovoz tanish va boshqa ketma-ketlikka bog‘liq vazifalar uchun ishlatiladi. LSTM oddiy RNNlardan farqli ravishda uchta asosiy darvoza (*gate*) strukturasiga ega. Ushbu darvozalar tarmoq ichidagi xotirani nazorat qilish uchun ishlatiladi:

2.5.1. *Forget Gate (unitish darvozasi, f_t).* Ushbu darvoza muhim bo‘lmagan ma’lumotlarni unutishga mas’ul. U oldingi xotira hujayrasidan $\left(C_{t-1}\right)$ keraksiz qismlarni olib tashlaydi,

$$f_t = \sigma\left(w_f \times \left|h_{t-1}, x_t\right| + b_f\right), \quad (3)$$

bu yerda σ sigmoid faollashtirish funksiyasi bo‘lib, 0 dan 1 gacha qiymatlar oladi. 0 - hamma narsa unutilishi kerak, 1 - hamma narsa saqlanishi kerak.

2.5.2. *Input Gate (kirish darvozasi i_t).* Bu darvoza yangi ma’lumotlarni qanday qabul qilishni hal qiladi,

$$i_t = \sigma\left(w_i \times |\, h_{i-1}, x_i \,| + b_i\right). \quad (4)$$

Shuningdek, kandidat yangi xotira yaratish uchun tanh funksiyasi ishlatiladi:

$$\tilde{C}_t = \tanh\left(W_c \times |\, h_{i-1}, x_i \,| + b_c\right). \quad (5)$$

Oxirgi yangilangan xotira quyidagicha hisoblanadi:

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t. \quad (6)$$

2.5.3. ***Output Gate (Chiqish darvozasi,*** o_t**).** Bu darvoza qaysi qismlar chiqish sifatida chiqarilishi kerakligini hal qiladi.

$$o_t = \sigma\left(w_0 \times |\, h_{t-1}, x_t \,| + b_0\right), \quad (7)$$

$$h_t = o_t \times \tanh(c_t), \quad (8)$$

bu yerda, σ – (sigmoid funksiyasi) – Natijani 0 va 1 oralig‘iga o‘tkazadi, W_f, W_i, W_c, W_o – og‘irlik (weight) matritsalari. Ular model o‘qitish jarayonida tahlil qilinadi va o‘zgaradi, h_{t-3} – oldingi vaqt bosqichidagi yashirin (hidden) holat, x_t – joriy vaqt bosqichidagi kirish ma’lumoti, b_f, b_i, b_c, b_o -bias (og‘ish) parametri. Modelni yanada moslashuvchan qilishga yordam beradi, tanh – tangens giperbolik (tanh) funksiyasi, U - 1 dan 1 gacha bo‘lgan qiymatlarni qaytaradi, h_t – yangi yashirin holat. Bu keyingi vaqt bosqichiga uzatiladi va chiqish sifatida ham ishlatilishi mumkin.

BiLSTM (Bidirectional Long Short-Term Memory) – bu LSTM ning ikki tomonlama versiyasidir. Oddiy LSTM vaqt qatorlarini faqat oldinga (pastdan yuqoriga) o‘qiydi, BiLSTM esa ma’lumotlarni ikkala yo‘nalishda (oldinga va orqaga) qayta ishlaydi. Bu kelajak va o‘tgan vaqt ma’lumotlarini birgalikda tahlil qilishga imkon beradi.

BiLSTM ikkita alohida LSTM tarmog‘ini bir vaqtda ishlatadi:

Forward LSTM – Ma’lumotlarni odatiy tartibda (t=1 dan t=n gacha) qayta ishlaydi.

Backward LSTM – Ma’lumotlarni teskari tartibda (t=n dan t=1 gacha) qayta ishlaydi.

Bu ikki yo‘nalishli chiqishlar birlashib (concatenate), yakuniy natijani hosil qiladi.

BiLSTM ikkita LSTM dan tashkil topgani uchun har bir vaqt bosqichi ikkita yashirin holat (h_t) ishlab chiqaradi:

$$\overrightarrow{h_t} = LSTM_{\,forward}\left(x_t, \overrightarrow{h_{t-1}}\right), \quad (9)$$

$$\overleftarrow{h_t} = LSTM_{backward}\left(x_t, \overleftarrow{h_{t+1}}\right), \quad (10)$$

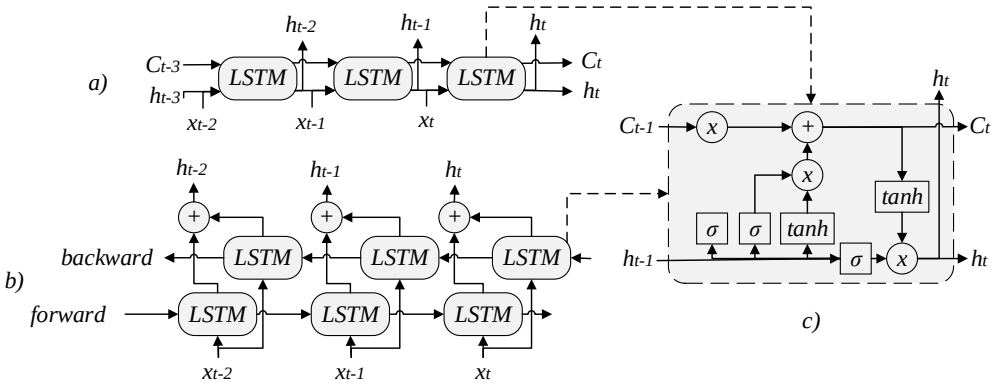
$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right], \quad (11)$$

bu yerda, $\overrightarrow{h_t}$ - Oldinga yo‘nalishdagi LSTM ning yashirin holati, $\overleftarrow{h_t}$ - Orqaga yo‘nalishdagi LSTM ning yashirin holati, h_t - Ikki yo‘nalishli yashirin holatlar qo‘shilib (concatenate) umumiy chiqishni hosil qiladi. 4-jadvalda LSTM va BiLSTM ning asosiy farqlari keltirilgan, 2-rasmda LSTM va BiLSTM ning arxitekturalari keltirilgan.

4-jadval. LSTM va BiLSTM asosiy farqlari

Xususiyat	LSTM	BiLSTM
Ma’lumotni o‘qish tartibi	Faqat oldinga	Ikkala yo‘nalishda
Kontekst	Faqat o‘tgan vaqtni hisobga oladi	O‘tgan va kelajakni hisobga oladi
Hisoblash resurslari	Tezroq ishlaydi	Sekinroq ishlaydi, lekin aniqroq
NLP vazifalarida samaradorlik	Yaxshi	Juda yaxshi

Har ikkala neyron tarmog‘ida ham til modellarini qurib chiqildi. Bunda xuddi KenLM kutubxonasida foydalangan usulimdan foydalanildi 3 gram asosida til modelini qurildi. Ya’ni gapdagi 3 ta so‘z hamda undan keyin va oldin keluvchi so‘zlarni topiladi. Misol uchun shunday gap bor “LSTM bu RNN tarmog‘ining maxsus turi hisoblanadi.” ushbu gapni quyidagicha bo‘laklarga ajratib har ikkala LSTM va BiLSTM asosidagi neyron tarmoq modelini o‘qitamiz. Neyron tarmoq modelini o‘qitish va testlash uchun lug‘atdagi va korpusdagi barcha 343672 ta so‘zlardan tokenizer ishlab chiqildi.



2-rasm. LSTM va BiLSTM tuzilishi: a - LSTM zanjiri, b - BiLSTM zanjiri, c- LSTM xotira hujayrasi

5-jadval. LSTM va BiLSTM da til modelini qurish uchun 3-gramlarni yaratish

Kiruvchi ma’lumotlarni	Chiquvchi ma’lumotlar	
	Oldingiz so‘z	**Keyingi so‘z**
LSTM bu RNN	-	tarmog‘ining
bu RNN tarmog‘ining	LSTM	maxsus
RNN tarmog‘ining maxsus	bu	turi
tarmog‘ining maxsus turi	RNN	hisoblanadi
maxsus turi hisoblanadi	turi	-

LSTM va BiLSTM neyron tarmoq modellari uchun parametrlarni quyidagicha belgilab oldik:
- Embedding o‘lchami – 100;
- LSTM yashirin qatlamlar o‘lchami – 128;
- Chiqish o‘lchami – 343672 (barcha takrorlanmas so‘zlar soni).

3 NATIJALAR TAHLILI

Mazkur ishda ishlab chiqilhan algoritm yuqoridagi 1 – rasmda ko‘rsatilgandek ikki qismga ajratiladi:

1. Levenshteyn masofasiga ko‘ra har bir so‘z uchun eng yaqin so‘zni lug‘atdan topish. So‘z uchun so‘z o‘xshashligi ko‘pi bilan 2 teng bo‘lgan tavsiyaviy so‘zlar ro‘yxati keltiriladi. Aga tavsiyaviy so‘zlar orasidan so‘z o‘xshashligi 0 ga teng so‘z bo‘lsa (yani kiruvchi so‘zning o‘zi) bo‘lsa bu so‘zda imlo xato mavjud emas, aks holda so‘zda imlo xato mavjud deb topildi;

2. Tavsiya qilingan so‘zlarning gapdagi so‘zlar ketma-ketligiga eng mos keluvchi tavsiya qilingan so‘zni til modeli bo‘yicha topiladi. Bunda biz ketma-ketlikning imlo xatolik bor so‘zning oldingi va keyin so‘zlari bo‘yicha til modeli bilan tavsiya etilgan so‘zlarni baholanadi va eng katta natija ko‘rsatgan tavsiyaviy so‘zni to‘g‘irlangan so‘z sifatida qabul qilinadi.

Levenshteyn masofasini hisoblash uchun pythonda python-Levenshtein [11] kutubxonasidan foydalanib so‘z uchun maksimal o‘xshashlik masofasi 2 ga teng bo‘lgan barcha so‘zlarni olish dasturini tuzildi.(3-rasm).

```
get_suggestions('brgalikda', vocabulary)

[('birgalikda', 1),
 ('birgalikda', 1),
 ('bevalikda', 2),
 ('birgalikga', 2),
 ('birgalikdan', 2),
 ('egalikda', 2),
 ('erkalikda', 2)]
```

3-rasm. python-Levenshtein kutubxonasidan foydalanib so‘z uchun maksimal o‘xshashlik masofasi 2 ga teng bo‘lgan barcha so‘zlarni olish

Yuqorida takidlaganidek 3 ta til modeli mos ravishda KenLM, LSTM, BiLSTM yordamida qurildi.

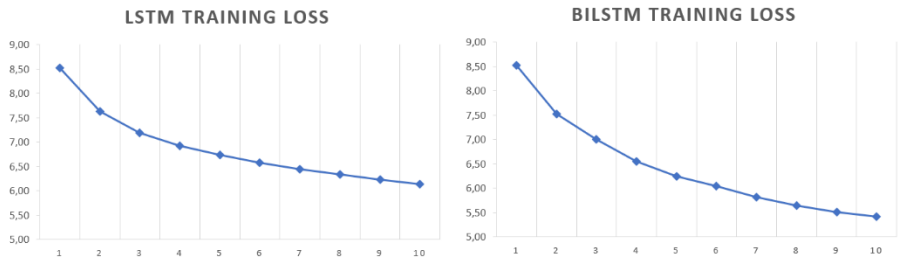
KenLM kutubxonasini yordamida til modelini qurish uchun korpusning training qismini txt kengaytmasidagi faylga saqlab undan binary formatidagi til modelini qurildi.

LSTM va BiLSTM asosidagi neyron tarmoq modellarimizni ham train ma'lumotlar to'plami yordamida 10 epochda o'qitildi. (4-rasm)

Ishlab chiqilgan dasturda har bir til modelida oddiy tarzda test qilib chiqildi. Buning uchun test to'plamidagi gaplardan foydalanib algoritimni testladik.

Dastlab test to'plamdagi gaplar tarkibidagi so'zlarni sun'iy ravishda 3 xil yo'l bilan imloviy xatolarga uchratildi yani:

- *O'chirish*. Misol: "kitob" so'zidan "k" harfini olib tashlandi – "itob";
- *Qo'shish*. Misol: "oyna" so'zida "y" harfini yana qo'shildi – "oyyna";
- *Almashtirish*. Misol: "talaba" so'zidagi "t" harfini "d" harfi bilan almashtirildi – "dalaba".



4-rasm. LSTM va BiLSTM asosida til modellarini qurishdagi training loss o'zgarish grafigi

Algoritmni testlaganganida oddiy usuldan foydalanildi: Agar algoritm gapdagi barcha imlo xatolari to'g'irlasa 1 aks holda 0 deb oldik. (6-jadval) Algoritimimizda BiLSTM asosidagi til modeli qo'llanilganda eng yaxshi natija (90.09%) olindi.

6-jadval. Til modellarini testlash

Tanlangan til modeli tashkil etuvchisi	Gaplar			Samaradorlik foizi
	Jami	To'g'ri tuzatilganlar	Noto'g'ri tuzatilganlar	
KenLM	**4 330 829**	2 698 965	1 631 864	62.31
LSTM		3 664 747	666 082	84.62
BiLSTM		3 902 076	428 753	90.09

4 XULOSA

Ushbu tadqiqotda o'zbek tili matnlaridagi imlo xatolarni aniqlash va tuzatishning samarali usullari o'rganildi. Tadqiqot davomida Levenshteyn masofasi algoritmi yordamida so'z o'xshashligi asosida xatolarni aniqlash va til modeli yordamida kontekstual tuzatish mexanizmlari ishlab chiqildi. O'zbek tilining agglutinativ tuzilishi sababli an'anaviy qoidalarga asoslangan usullar yetarli natija bermasligi sababli, chuqur o'rganish modellari, xususan, LSTM va BiLSTM tarmoqlari tadqiq qilindi.

Eksperimentlar uchun o'zbek tilidagi 80 million so'zdan iborat matn korpusi yaratildi va undan til modellari o'qitishda foydalanildi. Model samaradorligini baholash uchun imlo xatolar sun'iy ravishda kiritilgan test to'plami yaratildi. Test natijalariga ko'ra, KenLM modeli 62.31% aniqlikka ega bo'ldi, LSTM modeli 84.62% natijani qayd etdi, BiLSTM esa eng yuqori natija — 90.09% aniqlikni ta'minladi. Ushbu natijalar shuni ko'rsatadiki, neyron tarmoqlardan foydalangan holda o'zbek tili uchun kontekstual tahlil qilish xatolarni tuzatishda sezilarli yaxshilanishga olib keladi.

Kelajakda tadqiqot doirasida transformer modellaridan, masalan, BERT va GPT kabi ilg'or yondashuvlardan foydalanish, o'zbek tiliga moslashtirilgan annotatsiyalangan korpuslarni kengaytirish hamda til modelining resurs talabini optimallashtirish ustida ishlash rejalashtirilmoqda. Ushbu usullar o'zbek tili uchun yanada samarali imlo xatolarini tuzatish tizimini yaratishda muhim rol o'ynashi kutilmoqda.

ADABIYOTLAR

[1] *Norvig, P.* (2007). How to write a spelling corrector.

[2] *Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.* (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

[3] *Eryiğit, G.* (2014). The impact of morphology in named entity recognition: Detecting mentions of people, locations and organizations in Turkish. Turkish Journal of Electrical Engineering & Computer Sciences, 22(6), 1356–1371.
[4] *Abduazimov, D., Mamatov, A., & Usmonov, U.* (2020). Development of an annotated corpus for Uzbek language processing. In 2020 International Conference on Artificial Intelligence and Data Engineering (AIDE) (pp. 45–50). IEEE.
[5] *Mukhamadiyev, A.; Mukhiddinov, M.; Khujayarov, I.; Ochilov, M.; Cho, J.* Development of Language Models for Continuous Uzbek Speech Recognition System. Sensors 2023, 23, 1145. https://doi.org/10.3390/s23031145.
[6] *Musaev, M., Khujayarov, I., Ochilov, M.* (2023). Speech Recognition Technologies Based on Artificial Intelligence Algorithms. In: Zaynidinov, H., Singh, M., Tiwary, U.S., Singh, D. (eds) Intelligent Human Computer Interaction. IHCI 2022. Lecture Notes in Computer Science, vol 13741. Springer, Cham. https://doi.org/10.1007/978-3-031-27199-1_6.
[7] *Abdullaeva, M.I., Juraev, D.B., Ochilov, M.M., Rakhimov, M.F.* (2023). Uzbek Speech Synthesis Using Deep Learning Algorithms. In: Zaynidinov, H., Singh, M., Tiwary, U.S., Singh, D. (eds) Intelligent Human Computer Interaction. IHCI 2022. Lecture Notes in Computer Science, vol 13741. Springer, Cham. https://doi.org/10.1007/978-3-031-27199-1_5.
[8] *Musaev, M., Mussakhojayeva, S., Khujayorov, I., Khassanov, Y., Ochilov, M., Atakan Varol, H.* (2021). USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. In: Karpov, A., Potapova, R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science(), vol 12997. Springer, Cham. https://doi.org/10.1007/978-3-030-87802-3_40.
[9] https://lex.uz/
[10] *Butayev Sh.* English-uzbek uzbek-english dictionary 80 000 words. “O‘qituvchi” nashriyot-maanba ijodiy uyi. Toshkent – 2013.
[11] https://pypi.org/project/python-Levenshtein/
[12] https://github.com/kpu/kenlm

Поступила в редакцию 08.01.2025

Citation: *Ochilov M.M., Narzullayev O.O., Xolmatov O.A.* (2025). Mashinali o‘qitish algoritmlari asosida o‘zbek tili matnlaridagi imlo xatolarini aniqlash va tuzatish. Raqamli texnologiyalarning nazariy va amaliy masalalari xalqaro jurnali. 8(1). – B. 85-94. https://doi.org/10.62132/ijdt.v8i1.235.

DETECTION AND CORRECTION OF SPELLING ERRORS IN UZBEK TEXTS BASED ON MACHINE LEARNING ALGORITHMS

[+] *Ochilov M.M.*[1], *Narzullaev O.O.*[1], *Kholmatov O.A.*[1]

[1] Tashkent university of information technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

[+] ochilov.mannon@mail.ru

Abstract. This study addresses the problem of detecting and correcting spelling errors in Uzbek texts. Due to the complex morphological structure and agglutinative nature of the Uzbek language, traditional spell-checking methods do not provide sufficient accuracy. Therefore, this research employs the Levenshtein distance algorithm to measure word similarity and utilizes neural network-based language models for contextual correction. KenLM (a statistical language model), LSTM (Long Short-Term Memory), and BiLSTM (Bidirectional LSTM) approaches were used as language models. A text corpus of 80 million words was collected and analyzed for model training. The test results indicate that the BiLSTM model achieved the highest accuracy (90.09%) in correcting spelling errors, while the LSTM model recorded 84.62% accuracy. The KenLM model demonstrated an accuracy of 62.21% as well. These findings highlight that deep learning models capable of contextual analysis can significantly improve the automatic detection and correction of spelling errors in the Uzbek language. Based on the study results, future research plans include the application of transformer models, the expansion of annotated corpora, and the development of models that consider various morphological characteristics of the Uzbek language.

Keywords: Uzbek language, spelling correction, natural language processing (NLP), Levenshtein distance, language model, KenLM, LSTM, BiLSTM, neural networks, contextual analysis, machine learning, agglutinative languages, spell checking, deep learning, statistical models.

ОБНАРУЖЕНИЕ И ИСПРАВЛЕНИЕ ОРФОГРАФИЧЕСКИХ ОШИБОК В ТЕКСТАХ НА УЗБЕКСКОМ ЯЗЫКЕ НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

[+]Очилов М.М.[1], Нарзуллаев О.О.[1], Холматов О.А.[1]

[1] Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий, Ташкент, Узбекистан

[+] ochilov.mannon@mail.ru

Аннотация. В данном исследовании рассматривается проблема обнаружения и исправления орфографических ошибок в текстах на узбекском языке. Из-за сложной морфологической структуры и агглютинативных особенностей узбекского языка традиционные методы проверки орфографии не обеспечивают достаточной точности. Поэтому в данной работе применяется алгоритм вычисления расстояния Левенштейна для определения схожести слов, а также используются языковые модели на основе нейронных сетей для контекстуальной коррекции. В качестве языковых моделей использованы KenLM (статистическая языковая модель), LSTM (Long Short-Term Memory) и BiLSTM (Bidirectional LSTM). Для обучения модели был собран и проанализирован корпус текстов объемом 80 миллионов слов. Результаты тестирования показали, что модель BiLSTM достигла наивысшей точности (90,09%) при исправлении орфографических ошибок, модель LSTM показала точность 84,62%, а KenLM продемонстрировал точность 62,31%. Эти результаты подчеркивают, что модели глубокого обучения, способные к контекстному анализу, могут значительно улучшить автоматическое обнаружение и исправление орфографических ошибок в узбекском языке. В будущем планируется применение трансформерных моделей, расширение аннотированных корпусов и разработка моделей, учитывающих различные морфологические особенности узбекского языка.

Ключевые слова: узбекский язык, исправление орфографических ошибок, обработка естественного языка (NLP), расстояние Левенштейна, языковая модель, KenLM, LSTM, BiLSTM, нейронные сети, контекстный анализ, машинное обучение, агглютинативные языки, проверка орфографии, глубокое обучение, статистические модели.