AVTOMATIK IMLO TUZATISH BO‘YICHA TADQIQOTLAR TAHLILI: TAVSIF VA MUAMMOLARI

Ahmedova Maftuna Kaxramon qizi,
magistrant
maftunaAhmedova1997@gmail.com
ToshDO‘TAU

Annotatsiya. Matndagi imloni avtomatik to‘g‘rilashga qaratilgan tadqiqotlar uchta bosqichma-bosqich murakkablashib boruvchi muammoga e’tibor qaratgan: so‘z bo‘lmagan xatoliklarni aniqlash; so‘z xatolarini tuzatish; kontekstga bog‘liq so‘zlarni to‘g‘rilash. Birinchi muammoda berilgan so‘zlar ro‘yxatida uchramaydigan qatorlarni aniqlash uchun samarali andoza moslashtirish (pattern-matching) va n-gramm tahlil usullari ishlab chiqilgan. Ikkinchi muammoda umumiy va maxsus ilovalarga mo‘ljallangan turli xil imlo tuzatish usullari ishlab chiqilgan bo‘lib, ularning ayrimlari imlo xatolari qoliplarini batafsil o‘rganishga asoslangan. Uchinchisida esa tabiiy tilni qayta ishlash vositalari yoki statistik til modellaridan foydalangan holda bir nechta tajribalar o‘tkazilgan.

Ushbu maqolada imlo xatolari qoliplari bo‘yicha tadqiqot natijalari tahlil qilindi, matndagi avtomatik xatolarni tuzatishning barcha yo‘nalishi bilan bog‘liq tadqiqot masalalari muhokama qilindi.

Abstract. Research on automatic spelling correction in text has focused on three progressively more complex problems: detecting non-word errors; correcting word errors; and correcting words in context. In the first problem, efficient pattern-matching and n-gram analysis methods were developed to detect strings that do not appear in a given list of words. In the second problem, various spelling correction methods were developed for general and specific applications, some of which are based on a detailed study of spelling error patterns. In the third, several experiments were conducted using natural language processing tools or statistical language models.

This article analyzes the results of research on spelling error patterns and discusses research issues related to all areas of automatic text correction.

Аннотация. Исследования в области автоматического исправления орфографии в тексте были сосредоточены на трех постепенно усложняющихся проблемах: обнаружение несловесных ошибок; исправление орфографических ошибок; Исправление слов в зависимости от контекста. В первой задаче были разработаны эффективные методы сопоставления с образцом и анализа n-грамм для выявления строк, которые не встречаются в заданном списке слов. Во второй задаче были разработаны различные методы исправления орфографии для общего и частного применения, некоторые из

которых основаны на детальном изучении закономерностей орфографических ошибок. Третий этап включал в себя несколько экспериментов с использованием инструментов обработки естественного языка или статистических языковых моделей.

Kalit so‘zlar: *imlo qoidalari, noleksik birliklar, avtomatik imlo tuzatish, tinish belgilari.*

Imlo tekshiruvchi va tuzatuvchi dasturlar – bu berilgan matndagi grammatik va kontekstual imlo xatolarini aniqlaydigan va ularni maxsus algoritm yoki qoidalar to‘plami asosida to‘g‘rilaydigan dasturiy ilovadir. Ba’zi imlo tekshiruvchi dasturlar noto‘g‘ri yozilgan so‘zlar uchun to‘g‘ri muqobil variantlar ro‘yxatini yoki so‘zlar ketma-ketligi bo‘yicha takliflarni taqdim etadi. Avtomatik imlo tekshirish aksariyat tillarda so‘z protsessorlari (matn muharrirlari) uchun keng tarqalgan xususiyat hisoblanadi. Shuningdek, deyarli barcha veb-brauzerlarda ichki imlo tekshirgichlari mavjud. Avtomatik imlo tuzatish sohasi so‘nggi yillarda ilmiy tadqiqot e’tiborini tortib kelmoqda. Tadqiqot ishlarida ushbu mavzuga oid qisqacha kirish qismi mavjud bo‘lsa-da, hozirgacha nazariy asoslarni umumlashtiruvchi va shu kungacha ishlab chiqilgan yondashuvlarga umumiy ko‘rinish beruvchi tadqiqot yetishmaydi.

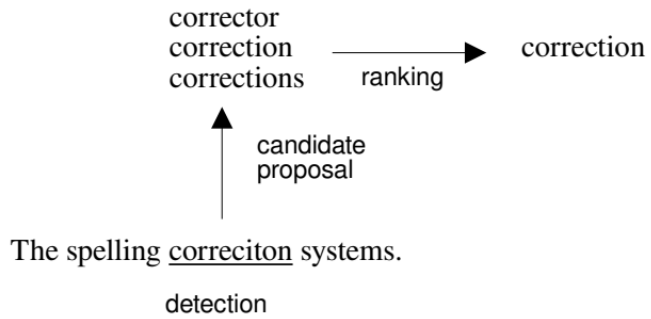
O‘zbek tili kam resursga ega bo‘lgan tillardan biri hisoblanib, unga oid tadqiqotlar, ma’lumotlar va vositalar hali rivojlanish bosqichida. Ma’lumotlar bazasida katta hajmdagi matnlarning yetishmasligi yangi muammolarni keltirib chiqaradi. Avtomatik imlo tuzatish tizimi tabiiy tilni qayta ishlash tizimining (NLP) bir qismi hisoblanadi. Matnlarni interaktiv tarzda tuzatish juda qiyin bo‘lganligi sababli, ma’lumotlar bazasidagi matnlar avtomatik ravishda to‘g‘rilanishi kerak. Imlo tuzatish tizimi oldingi va keyingi matn kontekstiga asoslanib, eng mos tuzatish variantini tanlaydi. Imlo tekshiruvchi dasturlar bo‘yicha tadqiqotlar 1950-yillarning oxirlariga borib taqaladi va hozirda ular ingliz, nemis va xitoy tillari kabi ko‘plab tillar uchun yaxshi rivojlangan. 1992-yilda Kukich[1]tomonidan taqdim etilgan izlanish keng qamrovli tadqiqotlarni boshlab berdi. Imlo xatolarini avtomatik tuzatish bo‘yicha ko‘p tadqiqotlar mavjud. Mitton tomonidan yozilgan kitob imlo xatolari turlarini tahlil qilib, avtomatik imlo tuzatish tizimini yaratish yondashuvlarini tasvirlab bergan [2]. Yannakoudakis va Favtroplar aksariyat imlo xatolari fonologik va ketma-ketlikka asoslangan maxsus qoidalarga bo‘ysunishini yozgan. Bu mualliflar maqolasida imlo xatolarini uchta asosiy kategoriya bo‘yicha tasniflab, 1377 ta imlo xatosi shakllarining tahlil natijalarini taqdim etgan[3]:

1. Undosh harflarga oid xatolar
2. Unli harflarga oid xatolar
3. Ketma-ketlik xatolari

Ba’zi yozuv tizimlari: arab, vetnam yoki slovak kabi yozuvlarda turli xil harf variantlaridan foydalanadi, bu esa so‘zning ma’nosini o‘zgartirishi mumkin. Gimenes va Roman braziliya-portugal tilida diakritik belgilarning tushirib

qoldirilishi imlo xatolarining keng tarqalgan turi ekanligini tasdiqlagan[4]. Zamonaviy adabiy arab tilida matnlar odatda diakritik belgilarsiz yoziladi[5]. Bu tipografik xatolik bo‘lib, muallif qo‘shimcha belgilardan foydalanmaydi va o‘quvchi asl ma’noni o‘zi tushunishini kutadi. Yo‘qolgan diakritik belgilar odatda qisqa unli tovushlarni yoki harflarning o‘zgarishini bildiradi. Ular harflarning tepasiga yoki tagiga joylashtiriladi. Arab matnlariga unli tovushlarni va boshqa diakritik belgilarni qo‘shish jarayoni diakritizatsiya yoki vokalizatsiya deb ataladi[6]. Azmi va Almajed arab tilidagi diakritizatsiya muammosiga e’tibor qaratib, uning baholash mezonini taklif qilgan[7]. Asahiah va boshqalar[8] esa arab tilidagi diakritizatsiya texnikalari bo‘yicha sharh nashr etgan.

Avtomatik imlo tuzatish tizimi imlo xatolarini aniqlaydi va tuzatish uchun nomzod variantlar to‘plamini taklif qiladi (1-rasmga qarang).



1-rasm. Avtomatik imlo tuzatish jarayoni[9]

So‘z yangi yoki shunchaki kam uchraydigan, kam tanilgan nom (NER) bo‘lishi yoki boshqa tilga tegishli leksema bo‘lishi mumkin. Biroq to‘g‘ri yozilgan so‘z gapda semantik jihatdan noto‘g‘ri bo‘lishi ham ehtimoldan xoli emas. Avtomatik imlo tuzatish tizimini ishlab chiqish uchun imlo xatosining paydo bo‘lish jarayonini tushunish zarur. Kukich imlo xatolarini to‘g‘ri so‘zlar lug‘atiga asoslanib quyidagi guruhlarga ajratgan:

Haqiqiy so‘z xatolari – so‘z noto‘g‘ri yozilgan bo‘lsa-da, uning to‘g‘ri shakli so‘zlar lug‘atida mavjud bo‘ladi.

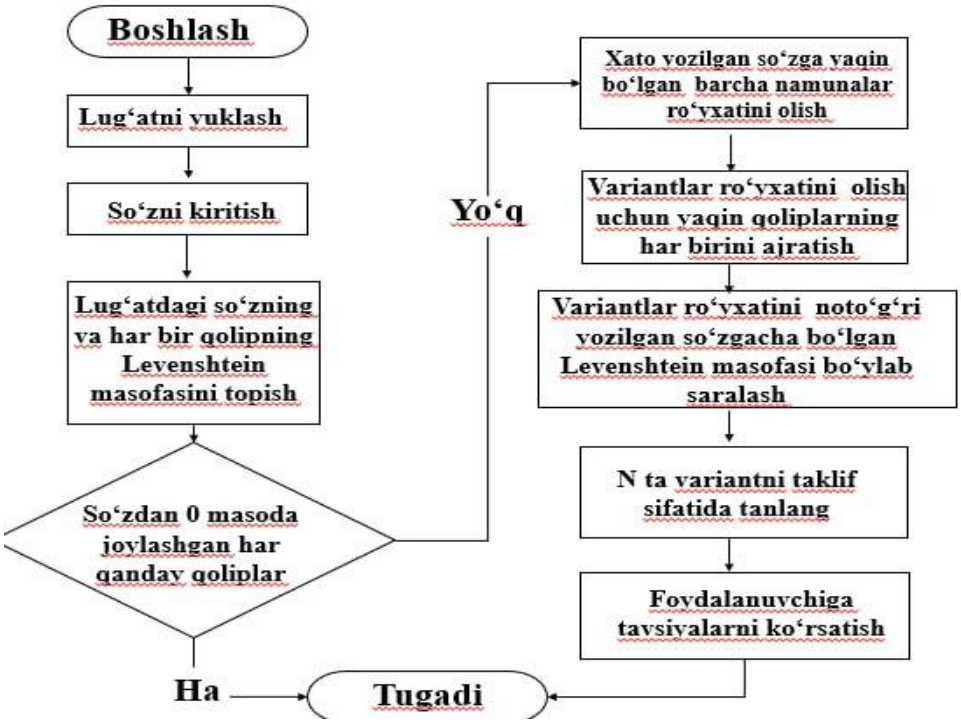
So‘z bo‘lmagan xatolar – noto‘g‘ri yozilgan so‘z shakli to‘g‘ri so‘zlar lug‘atida mavjud emas.

Ko‘pgina imlo tuzatish tizimlari so‘z bo‘lmagan xatolarni aniqlash uchun so‘zni to‘g‘ri so‘zlar lug‘atidan qidiradi. Bu bosqich so‘zlarni qidirish va tahlil qilish uchun mo‘ljallangan hash jadvali (hash table) yoki qidiruv daraxti kabi tezkor qidiruv usullarini talab qiladi. Ko‘pgina so‘z bo‘lmagan xatolarni tuzatish tizimlari ochiq kodli oldindan belgilangan imlo tuzatish tizimlaridan foydalanadi, masalan, **Aspell** yoki **Hunspell** xatolarni aniqlash, tuzatish nomzodlarini yaratish va dastlabki nomzodlarni saralash uchun qo‘llaniladi.

Kukich[10]hamda Pirinen va Linden[11] ushbu jarayonni uch bosqichga ajratgan:

1. Xatoni aniqlash
2. Tuzatish variantlarini yaratish
3. Tuzatish variantlarini tartiblash

Tadqiqotlarda avtomatik imlo tekshiruvchi dasturlarning ishlash tizimining Blok-sxema algoritm bosqichlarining umumiy ko‘rinishlari taklif qilingan. 2-rasmda taklif qilingan imlo tekshirish algoritmining blok-sxemasini ko‘rish mumkin.



2-rasm. Taklif qilinadigan imlo tekshirish algoritmining blok-sxemasi

Ba’zi tadqiqotlarda Avtomatik imlo tekshiruvchi dasturlarning noto‘g‘ri birlikni aniqlash va to‘g‘ri variantini taklif etish tizimi farqli ekanligini uchratish mumkin. To‘g‘ri tuzilgan ma’lumotlar bazasi, matnlar resursi yetarli bo‘lmagan tillar uchun xatolari oldindan belgilangan qoidalar va modellar asosida aniqlaydigan va tuzatadigan tizimlar taklif qilinadi. Bu tizim **A priori imlo tuzatish tizimi** deb nomlanadi. Qoidalarga asoslangan bu tizim – **Lug‘at asosida ishlaydi**, ya’ni so‘zlarning to‘g‘ri yoki noto‘g‘ri ekanligini oldindan tuzilgan lug‘atga qarab aniqlaydi, shuningdek, qoidalar asosida xatolarni tuzatadi: tilga xos grammatik va orfografik qoidalar asosida tuzatishlar kiritadi. Aynan mana shu tizim o‘zbek tilida ham imloni tuzatuvchi dasturlarning ishlab chiqilishida ilk bosqich hisoblanadi. Ammo qoidalarga asoslangan modellarda lug‘at bazasi cheklanganligi va so‘zni individual tahlil qilganligi sababli kontekstga ko‘ra tuzatishda xatoliklarga sabab bo‘ladi. Dastlab Google yoki Microsoft Word ning dastlabki imlo tekshiruvchilari ham lug‘at asosida ishlagan, lekin hozirgi zamonaviy tizimlar kontekstga asoslangan neyron tarmoqlar yordamida yanada rivojlangan.

Xulosa sifatida shuni aytish mumkinki, avtomatik imlo tuzatish ham mashina tarjimasiga o‘xshash jarayon hisoblanadi. Xatolarni o‘z ichiga olgan matn eng ehtimoliy to‘g‘ri shaklga “tarjima qilinadi”. Ushbu yondashuv butun natijaviy jumlani hisobga oladi. Bugungi kunda ko‘pchilik norasmiy xat va hujjatlardagi imloviy xatolarni tekshirish uchun ko‘p vaqt sarflaydi. Imloni tekshiruvchi tizimlarning yaratilishi esa barcha shu xizmatlarni yagona kutubxona sifatida taqdim etadi, shuningdek, foydalanuvchilar uni o‘z dasturlariga qo‘shishlari mumkin.

Bu quyidagi afzalliklarni ta’minlaydi:

1. **Vaqtni tejash:** Taklif etilgan algoritm kompyuterda bir soniyada minglab so‘zlarning imlosini tekshirishi va har soniyada o‘nlab so‘zlar uchun tuzatish tavsiyalarini berishi mumkin.
2. **Aniqlik:** Algoritm 70-80% hollarda so‘zni to‘g‘ri yozilishini birinchi tavsiya sifatida taklif qiladi, 90% hollarda esa to‘g‘ri yozilgan so‘zning variant yuqori beshlik ichida bo‘ladi.
3. **Xarajatlarni tejash:** Imlo tekshirish jarayonini avtomatlashtirish natijasida kompaniyalar,nashiryot va davlat tashkilotlari qo‘lyozmalarni tekshiruvchi xodimlar uchun xarajatlarini kamaytirishi mumkin.

Foydalanilgan adabiyotlar:

1. Kukich, K. Techniques for automatically correcting words in text. Acm Comput. Surv. 1992, 24, 377–439.
2. Mitton, R. English Spelling and the Computer; Longman Group: Harlow, Essex, UK, 1996; p. 214
3. Yannakoudakis, E.J.; Fawthrop, D. The rules of spelling errors. Inf. Process. Manag. 1983, 19, 87–99
4. Gimenes, P.A.; Roman, N.T. Spelling error patterns in Brazilian Portuguese. Comput. Linguist. 2015, 41, 175–184
5. Zitouni, I.; Sarikaya, R. Arabic diacritic restoration approach based on maximum entropy models. Comput. Speech Lang. 2009, 23, 257–276.
6. Zitouni, I.; Sarikaya, R. Arabic diacritic restoration approach based on maximum entropy models. Comput. Speech Lang. 2009, 23, 257–276.
7. Azmi, A.M.; Almajed, R.S. A survey of automatic Arabic diacritization techniques. Nat. Lang. Eng. 2015, 21, 477–495.
8. Asahiah, F.O.; Odéjobi, O.A.; Adagunodo, E.R. A survey of diacritic restoration in abjad and alphabet writing systems. Nat. Lang. Eng. 2018, 24, 123–154
9. Daniel Hládek, Ján Staš, Matúš Pleva. Survey of automatic Spelling correction. Electronics 2020, 9, 1670; doi:10.3390/electronics9101670.
10. Kukich, K. Techniques for automatically correcting words in text. Acm Comput. Surv. 1992, 24, 377–439.

11. Pirinen, T.A.; Lindén, K. State-of-the-art in weighted finite-state spell-checking. In Computational Linguistics and Intelligent Text Processing, Proceedings of the CICLing 2014, Kathmandu, Nepal, 6–12 April 2014; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8404, Part 2, pp. 519–532.