MATNDAGI IMLOVIY XATONI AVTOMATIK ANIQLASH USULLARI

Sobirova Nazira G‘anijon qizi
ToshDO‘TAU tayanch doktoranti
nazirasobirova@gmail.com

Annotatsiya. Bugungi raqamli muhitda matnlar bilan ishlashda imloviy xatolarni aniqlash va tuzatish muhim bosqichlardan biridir. Ayniqsa, tilshunoslik, tabiiy tilni qayta ishlash (NLP), mashinaviy o‘rganish va sun’iy intellekt sohalarida bu jarayonlarning avtomatlashtirilgan bo‘lishi katta ahamiyat kasb etadi. Imloviy xatolarni aniqlash algoritmlari foydalanuvchilarga sifatli matnlar yaratishda yordam beradi, matn indeksatsiyasini yaxshilaydi hamda qidiruv tizimlari uchun optimal natijalarni ta’minlaydi. Matnni yozishda inson omili sababli yuzaga keladigan imloviy xatolar informatsion tizimlar, tilni qayta ishlash tizimlari va foydalanuvchi interfeyslarida muhim muammo hisoblanadi. Ayniqsa, ijtimoiy tarmoqlar, bloglar, chatlar, matnli izohlar (kommentariyalar) ko‘paygan sari, avtomatik tahlilga bo‘lgan ehtiyoj ortib bormoqda.

Imlo xatosi matnni tushunishni qiyinlashtiradi va undan ham yomoni, matnni qayta ishlash jarayoniga salbiy ta’sir ko‘rsatadi. **Tabiiy tilni qayta ishlash (NLP)** texnologiyalarida so‘zlarning normalizatsiya qilingan shakli talab qilinadi, chunki noto‘g‘ri yozilgan yoki noto‘g‘ri raqamlashtirilgan matnning informatsion qiymati pasayadi.Shu sababli, imloviy xatolarni **avtomatik aniqlash va tuzatish** texnologiyalari tilshunoslikda muhim yo‘nalishga aylangan. Texnologiyaning jadal rivojlanishi natijasida tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasida katta yutuqlarga erishildi. Bu rivojlanish natijasida matnlarni avtomatik tekshirish va imloviy xatolarni aniqlash imkoniyatlari kengayib bormoqda. Ushbu maqolada imloviy xatolarni aniqlash usullari va ularga oid tadqiqotlar tahlil qilinadi.

Kalit so‘zlar: imlo xatosi, NLP, **ASC, n-gramm, tahrir masofasi, xato turlari, qoidalarga asoslangan usul, lug**‘at tekshiruvi.

Annotation. In today's digital environment, identifying and correcting spelling errors is one of the key stages in working with texts. This is particularly important in fields such as linguistics, natural language processing (NLP), machine learning, and artificial intelligence, where the automation of such processes plays a vital role. Spelling error detection algorithms assist users in producing high-quality texts, improve text indexing, and provide optimal results for search engines. Spelling errors, often caused by human factors during writing, represent a significant problem for information systems, language processing systems, and user interfaces. The growing presence of social media, blogs, chats, and textual comments has increased the demand for automatic analysis.Spelling errors complicate text comprehension and, more critically, negatively affect the text processing pipeline. In NLP

technologies, normalized word forms are required, since incorrectly written or digitized texts lose informational value.

Therefore, automatic detection and correction of spelling errors has become a crucial direction in linguistics. Due to the rapid development of technology, significant achievements have been made in the field of Natural Language Processing (NLP). As a result, the capabilities of automatically checking texts and detecting spelling mistakes are expanding. This article analyzes various methods and research related to spelling error detection.

Keywords: spelling error, NLP, ASC, n-gram, edit distance, error types, rule-based method, dictionary checking.

Аннотация. В современном цифровом пространстве выявление и исправление орфографических ошибок является одним из важнейших этапов при работе с текстами. Особенно это актуально в таких областях, как лингвистика, обработка естественного языка (NLP), машинное обучение и искусственный интеллект, где автоматизация подобных процессов играет большую роль. Алгоритмы обнаружения орфографических ошибок помогают пользователям создавать качественные тексты, улучшают индексацию текста и обеспечивают оптимальные результаты для поисковых систем. Ошибки, возникающие при написании текста по причине человеческого фактора, становятся серьезной проблемой для информационных систем, систем обработки языка и пользовательских интерфейсов. С ростом популярности социальных сетей, блогов, чатов и текстовых комментариев, потребность в автоматическом анализе становится все более актуальной.Орфографические ошибки затрудняют понимание текста и, что еще хуже, негативно влияют на процессы его обработки. В технологиях обработки естественного языка требуется нормализованная форма слов, поскольку неправильно написанный или ошибочно оцифрованный текст теряет свою информационную ценность.

В связи с этим технологии автоматического выявления и исправления орфографических ошибок становятся важным направлением в лингвистике. Быстрое развитие технологий привело к значительным достижениям в области обработки естественного языка (NLP). В результате расширяются возможности автоматической проверки текста и обнаружения орфографических ошибок. В данной статье рассматриваются методы обнаружения орфографических ошибок и соответствующие исследования.

Ключевые слова: орфографическая ошибка, NLP, ASC, n-граммы, расстояние редактирования, типы ошибок, метод на основе правил, проверка по словарю.

KIRISH

Til – inson muloqotining asosiy vositasi bo‘lib, matnlar orqali axborot almashish keng tarqalgan. Biroq, inson tomonidan yozilgan matnlarda imloviy va grammatik xatolar uchrashi tabiiy holatdir. Ushbu xatolarni aniqlash va tuzatish jarayonini avtomatlashtirish tabiiy tilni qayta ishlash (NLP) sohasida muhim masalalardan biridir. Ilg‘or algoritmlar va sun’iy intellekt texnologiyalarining rivojlanishi bilan imloviy xatolarni avtomatik aniqlash yanada samarali bo‘lib bormoqda.

Matnni turli usullarda yozish mumkin. Ba’zida yozilgan matn muallif yoki o‘quvchi kutganidan farqli ko‘rinishda bo‘ladi. Ayniqsa, ona tili boshqa bo‘lgan insonlar uchun aniq va tushunarli matn yaratish oson emas. Gap tarkibida noto‘g‘ri yozilgan so‘z imlo xatosi hisoblanadi. Ba’zan muallifda imlo xatolarini to‘g‘rilashga vaqt yoki imkoniyat yetarli bo‘lmasligi mumkin. **Avtomatik imlo tuzatish tizimlari (ASC)** ushbu muammoni hal qilishda yordam beradi. Ushbu tizimlar xatoli so‘zlarni aniqlaydi va to‘g‘ri yozilgan so‘z variantlarini tavsiya etadi. Nomzod so‘zlar xatoning turiga va atrofdagi kontekstga qarab saralanadi. Eng mos keladigan tuzatish variantini tizim avtomatik yoki foydalanuvchi ishtirokida tanlaydi.

Matnlardagi imloviy xatoni avtomatik aniqlash usul, yondashuv va algoritmlari

Matnda imlo xatolarini aniqlash va tuzatish bo‘yicha algoritmik usullar kompyuter fanida uzoq va mustahkam tarixga ega[1].Matn ichidagi so‘zlarni avtomatik tarzda to‘g‘rilash uchun algoritmlar va texnikalarni ishlab chiqish muammosi doimiy ilmiy izlanishlar mavzusiga aylangan. Kompyuter yordamida avtomatik imlo tuzatish va matnni tanib olish bo‘yicha ilk tadqiqotlar 1960-yillarda boshlangan bo‘lib, bugungi kungacha davom etib kelmoqda. Ushbu sohadagi tadqiqotlarning doimiy davom etishiga bir nechta muhim sabablar bor: sifat va samaradorlikni oshirish hamda ushbu texnologiyalarning qo‘llanilish doirasini kengaytirish. Ushbu muammo ustida ishlash boshlanganidan beri ko‘plab yondashuvlar qo‘llanilgan:

- tahrirlash masofasi (edit distance) [2],
- qoidalarga asoslangan usullar (rule-based techniques) [3],
- n-grammlar (n-grams) [4],
- ehtimollik usullari (probabilistic techniques) [5],
- neyron tarmoqlar (neural nets) [6],

- o‘xshashlik kalitlari usuli (similarity key techniques) [7,8],
- shovqinli kanal modeli (noisy channel model) [9,10].

Ushbu barcha yondashuvlarning asosiy g‘oyasi noto‘g‘ri yozilgan so‘z va lug‘atdagi to‘g‘ri so‘zlar o‘rtasidagi o‘xshashlikni hisoblashga asoslangan [11].

TADQIQOTLAR METODOLOGIYASI

NLPning uch asosiy yondashuvi:

- ✓ **Qoidalarga asoslangan yondashuv**– Lingvistik qonun-qoidalarga asoslangan. Bunday tizimlar lug‘aviy bazaga (dictionary-based) va grammatik qoidalarga asoslangan bo‘lib, ular foydalanuvchi matnini belgilangan qoidalar orqali tekshiradi. Misol uchun, Ingliz tili uchun ishlab chiqilgan GNU Aspell, Hunspell, va LanguageTool kabi tizimlar an’anaviy qoidaviy yondashuvlarga asoslanadi. Bu tizimlar foydalanuvchi matnidagi har bir so‘zni oldindan belgilangan lug‘at bilan solishtirib, nomuvofiq elementlarni aniqlaydi.
- ✓ **Mashinani o‘rganish yondashuvi** – Statistik tahlillarga asoslangan. Bu usullar N-gram modellari, yashirin Markov modellari (HMM), Bayes klassifikatorlari asosida imloviy xatolarni aniqlaydi. Jumladan, Kernighan tomonidan taklif etilgan ehtimollik modeliga asoslangan imloviy xatolarni tuzatish algoritmi bugungi kungacha mashhurligini saqlab kelmoqda. Shuningdek, Islam and Inkpen tomonidan ishlab chiqilgan “Spelling Correction Using Google Web 1T 5-Gram” usuli juda katta hajmdagi korpusga tayanadi va Google qidiruv tizimining statistik ma’lumotlarini tahlil qiladi.
- ✓ **Neyron tarmoqlar yondashuvi** – Sun’iy, qayta bog‘lovchi va konvolyutsion neyron tarmoqlar algoritmlariga asoslangan. So‘nggi yillarda chuqur o‘rganish (deep learning) texnologiyalarining rivojlanishi natijasida neyron tarmoqlarga asoslangan imloviy xato aniqlash tizimlari paydo bo‘ldi. Transformer arxitekturasiga asoslangan BERT va RoBERTa kabi til modellari matn kontekstini chuqur tushunib, xatoni aniqlash va hatto to‘g‘rilash imkonini beradi. Xitoy tili, Koreys tili va boshqa tillar uchun ishlab chiqilgan modelar bu yondashuvlarning yuqori samaradorligini isbotladi.

Kukich ham matndagi noma’lum so‘z xatolarini aniqlash texnikalarini ikki turga ajratdi:

1. Lug‘at qidiruvi texnikasi;
2. N-gramma tahlili texnikasi.

Lug‘at/Wordnet – bu ma‘lum bir tilning to‘g‘ri so‘zlar ro‘yxatini o‘z ichiga olgan leksik manba hisoblanadi. Xato so‘zlarni lug‘atga solishtirib, har bir so‘zni tekshirib, osonlik bilan aniqlash mumkin. Lekin uning o‘ziga yarasha kamchiliklari ham bor. Ya’ni, bunda lug‘atni yangilash va matndagi barcha so‘zlarni qamrab olish darajasi yetarli, keng va to‘liq emas.

Imlo xatolarini tuzatish jarayonida n-gram – bu so‘z yoki qator ichidagi n ta harfdan tashkil topgan ketma-ketlik hisoblanadi. N-gramma modeli ikki qator (so‘z) orasidagi o‘xshashlikni aniqlash uchun ishlatiladi. Bu o‘xshashlik ikkala qator ichidagi umumiy n-grammalar sonini hisoblash orqali aniqlanadi. Qanchalik ko‘p umumiy n-grammalar mavjud bo‘lsa, ikkita qator shunchalik o‘xshash bo‘ladi.

N-gramm modeli matndagi birliklarning (so‘zlar yoki harflar) ketma-ketligini tahlil qilishga asoslangan bo‘lib, unigram (1-gramm), bigram (2-gramm), trigram (3-gramm) kabi turli variantlari mavjud.

Unigram: N = 1, ya’ni har bir so‘z mustaqil ravishda baholanadi.

Bigram: N = 2, ya’ni ikkita ketma-ket so‘zlar o‘rtasidagi bog‘liqlik hisoblanadi.

Trigram: N = 3, ya’ni uchta ketma-ket so‘zlar o'rtasidagi bog‘lanish hisoblanadi.

4-gram, **5-gram** va hokazo: N > 3, bu ko‘proq so‘zlar ketma-ketligini tahlil qilishni anglatadi.Imlo tuzatish algoritmlaridan eng ko‘p o‘rganilganlari – minimal tahrir masofasi (minimum edit distance) hisoblashga asoslangan texnikalardir.

Minimal tahrir masofasi – bu noto‘g‘ri yozilgan so‘zni lug‘atdagi so‘zga aylantirish uchun talab qilinadigan minimal tahrir amallarining (qo‘shish, o‘chirish, almashtirish) soni. Ushbu tushuncha Wagner tomonidan ta’riflangan. Birinchi minimal tahrir masofasi asosida imlo tuzatish algoritmi Damerau [12] tomonidan ishlab chiqilgan.

Xuddi shu davrda, *Levenshtein* o‘chirishlar, qo‘shishlar va joy almashtirishlarni (transpozitsiyalarni) tuzatish uchun o‘xshash algoritmni ishlab chiqdi. *Wagner* imlo tuzatish muammosiga dinamik dasturlash texnikalarini tatbiq etish tushunchasini birinchi bo‘lib kiritdi va bu hisoblash samaradorligini oshirish uchun mo‘ljallangan edi. *Wagner* va *Fischer* Levenshtein algoritmini umumlashtirib, uni bir nechta xatolarni o‘z ichiga olgan noto‘g‘ri yozilgan so‘zlar uchun ham tatbiq etdilar.

Lowrance va *Wagner* tomonidan ishlab chiqilgan yana bir algoritmik variant qo‘shimcha o‘zgarishlarni (masalan, bir-biriga yaqin bo‘lmagan harflarning almashinishi) hisobga oldi.

Bundan tashqari, *Wong* va *Chandra*, Okuda va boshqalar, va Kashyap va Oommen tomonidan boshqa variantlar ishlab chiqildi. Bu texnikalar hisoblaydigan metrika Damerau-Levenshtein metrikasi deb ham yuritiladi, chunki u ikki asosiy tadqiqotchining ishlanmalariga asoslangan.

Hawley ushbu metrikalarning ayrimlarini Unix buyruq satri interfeysi uchun imlo tuzatish vositasida sinovdan o‘tkazdi. *Aho* o‘zining sharh maqolasida Unix fayl taqqoslash vositasi (“diff”) dinamik dasturlashga asoslangan minimal tahrir masofasi algoritmidan foydalanishini ta’kidladi. Shuningdek, maqolada boshqa dinamik dasturlash algoritmlarining vaqt murakkabligi bo‘yicha tavsifi keltirilgan. Mor va Fraenkel o‘chirishlar eng keng tarqalgan xatoliklardan biri ekanligiga asoslanib samarali qidirish usulini ishlab chiqdilar:

1. Har bir lug‘at so‘zi lug‘atda (so‘z uzunligi + 1) marta saqlanadi.
2. Har safar har bir harf o‘chirilgan holda saqlanadi.
3. Noto‘g‘ri yozilgan so‘z hash funksiya orqali qidirilib, mos keladigan variantlar topiladi.

Teskari minimal tahrir masofasi usuli (“Reverse” Minimum Edit Distance Technique). Ushbu yondashuv Gorin tomonidan DEC-10 imlo tekshirgichida qo‘llangan. Durham va boshqalar buyruq satri tuzatish vositasida ishlatgan. Kernighan va boshqalar va Church va Gale bu usulni ehtimollik asosida ishlovchi imlo tekshirgichlarida qo‘llashgan.

Ushbu texnika quyidagicha ishlaydi:

Birinchi bosqich – noto‘g‘ri yozilgan so‘zning har bir mumkin bo‘lgan yagona xato kombinatsiyasi yaratib chiqiladi.

Ikkinchi bosqich – ushbu variantlar lug‘at bilan taqqoslanadi.

Misol uchun, agar noto‘g‘ri yozilgan so‘z uzunligi n bo‘lsa va alfavit hajmi 26 bo‘lsa,tekshirilishi kerak bo‘lgan satrlar soni quyidagicha bo‘ladi:

- (26 × (n + 1)) qo‘shishlar (insertions)
- n o‘chirishlar (deletions)
- (25 × n) almashtirishlar (substitutions)
- (n - 1) joy almashtirishlar (transpositions)

Minimal tahrir masofasi algoritmlari Damerau va Levenshtein tomonidan ishlab chiqilgan asosiy imlo tuzatish texnikalaridan biridir. Wagner va Fischer uni umumlashtirgan va dinamik dasturlash texnikalarini qo‘llagan. Ehtimollik

modellari yoki fonetik jihatdan o‘zgartirilgan versiyalar ham ishlab chiqilgan.Mor va Fraenkel qidirish vaqtini optimallashtirish uchun lug‘at so‘zlarini oldindan qayta ishlashga asoslangan yondashuvni taklif qilishdi. “Teskari minimal tahrir masofasi” texnikasi noto‘g‘ri yozilgan so‘zlarning barcha variantlarini lug‘at bilan taqqoslash orqali ishlaydi. Trie daraxtlari ham minimal tahrir masofasi algoritmlariga muqobil sifatida o‘rganilgan.

O‘xshashlik kaliti texnikalarining asosiy g‘oyasi har bir satrni shunday kalitga bog‘lashki, o‘xshash yozilgan satrlar bir xil yoki o‘xshash kalitlarga ega bo‘lishi kerak. Shunday qilib, noto‘g‘ri yozilgan satr uchun kalit hisoblanganda, bu kalit lug‘atdagi barcha o‘xshash yozilgan so‘zlar (nomzodlar) uchun ko‘rsatkich bo‘lib xizmat qiladi. O‘xshashlik kaliti texnikalari tezlik bo‘yicha ustunlikka ega,chunki noto‘g‘ri yozilgan satrni lug‘atdagi har bir so‘z bilan to‘g‘ridan-to‘g‘ri taqqoslashning hojati yo‘q.

Eng dastlabki va ko‘p keltiriladigan o‘xshashlik kaliti texnikalaridan biri bu SOUNDEX tizimidir,u Odell va Russell tomonidan fonetik imlo tuzatish ilovalari uchun patentlangan. SOUNDEX tizimi aviakompaniya bron qilish tizimi va boshqa ilovalarda qo‘llanilgan. Bu tizim so‘z yoki noto‘g‘ri yozilgan satrni maxsus kalitga bog‘laydi. Kalit so‘zning birinchi harfi va raqamlar ketma-ketligidan iborat bo‘ladi. SOUNDEX tizimi juda qo‘pol (aniqlik past) yondashuvga ega.Bundan tashqari, SOUNDEX tizimi nomzodlarni saralash funksiyasiga ega emas, ya’ni barcha nomzod so‘zlar oddiygina foydalanuvchiga taqdim etiladi. Qoidalarga asoslangan usullar mashinani o‘rganish va ma’lumotlarni tahlil qilishning mashhur texnikalaridan biri hisoblanadi.

Imlo tekshirgichning asosida tilning qoidalari yotadi. Bu qoidalar grammatik, morfologik va sintaktik tuzilmalarga asoslanadi. Matn tahlil qilinadi va oldindan belgilangan qoidalar asosida imlo xatolari aniqlanadi. Bu jarayon so‘zlarni ma’lumotlar bazasidagi to‘g‘ri shakllar bilan solishtirish orqali amalga oshiriladi. Imlo tekshirgich matndagi noto‘g‘ri yozilgan so‘zlarni qoidalar asosida aniqlaydi. Misol uchun:

- “yozmoq” o‘rniga “yozmok” deb yozilgan so‘z noto‘g‘ri deb topiladi.
- Qo‘shib yozilishi kerak bo‘lgan so‘zlar ajratib yozilganda xato sifatida qayd etiladi (masalan, “bir kun” o‘rniga “birkun”).
- Belgilar orasida bo‘sh joylar noto‘g‘ri deb belgilanishi mumkin (masalan, “kitob , qalam” -> “kitob, qalam”).

Qoidalarga asoslangan tizimlarning dastlabki misollaridan biri **General Problem Solver (GPS)** bo‘lib, u 1950-yillarda *Herbert A. Simon* va *Allen Newell* tomonidan ishlab chiqilgan. GPS inson kabi masalalarni kichikroq bo‘laklarga ajratib hal qilish uchun mo‘ljallangan edi.

Yana bir muhim misol – **MYCIN**, 1970-yillarda bakterial infeksiyalarni tashxislash va antibiotiklar tavsiya qilish uchun ishlab chiqilgan ekspert tizimi. MYCIN o‘z sohasida inson mutaxassislariga teng darajada ishlash qobiliyatini namoyish etib, SI ning amaliy qo‘llanilishi imkoniyatlarini ko‘rsatdi.

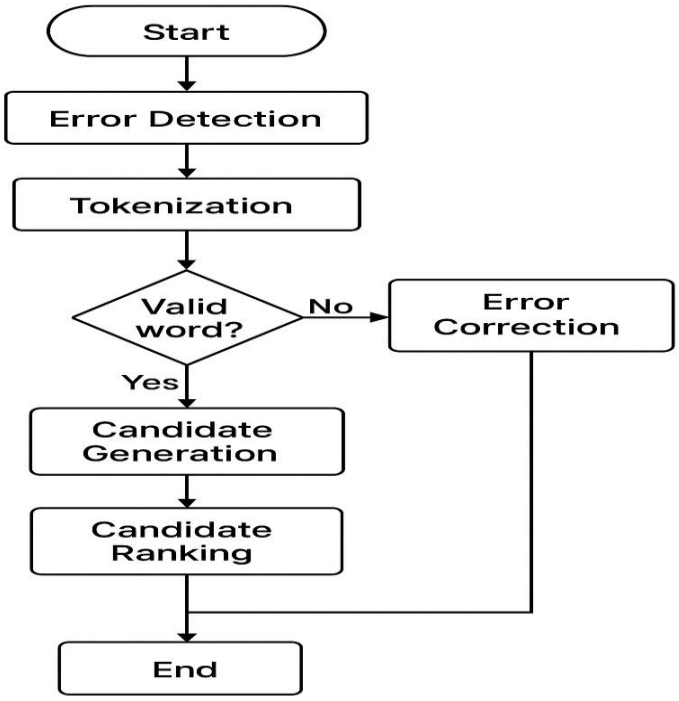
Yannakoudakis va Fawthrop [13] tomonidan ishlab chiqilgan umumiy imlo tuzatish dasturi qoidalarga asoslangan texnikalardan foydalanishga misol bo‘la oladi. Ushbu dastur lug‘atdan foydalanadi, so‘zlarni birinchi harflari va uzunliklariga qarab bo‘linadi. Qoidalar noto‘g‘ri yozilgan so‘zning ehtimoliy uzunligini belgilashga asoslangan. Takliflarni hosil qilish uchun lug‘atning maxsus qismlari qidiriladi va qoidalarga mos kelgan takliflar beriladi. Agar bir nechta takliflar topilsa, ular qoidalarning paydo bo‘lish ehtimolliklariga qarab saralanadi.

Maxsus qoidalarga asoslangan imlo tuzatish tizimi Means tomonidan ishlab chiqilgan. Ushbu tizimda qoidalar morfologik ma'lumotlarni, masalan, "ing" qo‘shimchasini qo‘shishdan oldin oxirgi undoshni ikki marta yozish qoidalarini o‘z ichiga oladi. Ushbu tizim kengaytirilgan qoidalarga asoslangan abbreviatura transformatsiyasini ham amalga oshiradi. Agar bu qadam muvaffaqiyatsiz tugasa, noto‘g‘ri so‘zning barcha mumkin bo‘lgan oddiy xatolarini (qo‘shish, o‘chirish, almashtirish va transpozitsiyalar) ko‘rib chiqadi.

Imloviy xatolarni aniqlash algoritmlarini ishlab chiqishdan oldin, avvalo ularning turlarini to‘g‘ri aniqlash kerak:

To‘g‘ri yozilmagan (non-word) xatolar – Leksik jihatdan mavjud bo‘lmagan so‘zlar. Masalan, “yangi” o‘rniga “ynagi”.

1. ***To‘g‘ri yozilgan, ammo kontekstdan xato (real-word errors***) – Leksik jihatdan mavjud, ammo kontekstga mos kelmaydigan so‘zlar. Masalan, “siz bizning raxbarimizsiz” so‘zida “siz” ikkita ma’noga ega bo‘lishi mumkin.
2. ***Fonetik xatolar*** – So‘z tovushiga ko‘ra noto‘g‘ri yozilgan. Masalan, “quyosh” o‘rniga “quyos”.
3. ***Klavish xatolari (keyboard errors)*** – Klaviatura tugmalari yaqinligidan yuzaga kelgan xatolar. Masalan, “nimadir” o‘rniga “namdir”.



Yuqorida ko‘rsatilganidek, ko‘pchilik imlo tekshirgichlari va tuzatuvchilar ikki qatlamli yondashuvga amal qiladilar:

*xatolarni aniqlash ;

* xato tuzatish.

Xato aniqlash uchun har bir so‘z jumlada tokenizator yordamida tokenlashtiriladi va uning to‘g‘riligiga tekshiriladi. Nomzod so‘z ma’noga ega bo‘lsa, u to‘g‘ri so‘z deb hisoblanadi, aks holda u xato deb topiladi.Xatoni tuzatish ikki bosqichdan iborat:

a) **nomzod tuzatishlarning generatsiyasi**;

b) **nomzod tuzatishlarni reytinglash**.

Nomzodlar generatsiyasi jarayoni,odatda, to‘g‘ri n-gramlar jadvali yordamida bir yoki bir nechta tuzatishlarni aniqlashni amalga oshiradi.

Reytinglash jarayoni esa, xato yozilgan satr va nomzodlar o‘rtasidagi leksik o‘xshashlik o‘lchovini yoki tuzatishning ehtimolligini hisoblash orqali nomzodlarni reytinglashni amalga oshiradi.

Ushbu ikki bosqich ko‘pincha alohida jarayon sifatida qaraladi va ketma-ket bajariladi. Ba’zi usullar ikkinchi jarayonni o‘tkazib yuborishi mumkin, reytinglash va yakuniy tanlovni foydalanuvchiga qoldiradi.

So‘zlarni avtomatik tuzatish bo‘yicha tadqiqotlar uchta asosiy muammoga qaratilgan:

1. ***Noto‘g‘ri so‘zlarni aniqlash (nonword error detection)*** – ushbu yondashuv lug‘atda mavjud bo‘lmagan so‘zlarni aniqlashga asoslanadi.
2. ***Alohida so‘zlardagi xatolarni tuzatish (isolated-word error correction)*** – so‘z xatolarini kontekstga bog‘liq bo‘lmagan holda tuzatish jarayoni.
3. ***Kontekstga asoslangan so‘zlarni tuzatish (context-dependent word correction)*** – matn kontekstiga asoslangan holda xatolarni aniqlash va tuzatish.

Xatolarni aniqlash (error detection) va tuzatish (error correction) o‘rtasidagi farqni aniq tushunish kerak. So‘zlar ro‘yxati, lug‘at yoki leksikonda mavjud bo‘lmagan qatorlarni aniqlash uchun samarali texnikalar ishlab chiqilgan. Ammo noto‘g‘ri yozilgan so‘zni to‘g‘ri variant bilan almashtirish ancha murakkab muammo. Faqatgina nomzod so‘zlarni topish va tartibga solish jarayoni ham muhim vazifadir.

XULOSA

Matnlardagi imloviy xatolarni avtomatik aniqlash bugungi kunda til texnologiyalarining eng dolzarb yo‘nalishlaridan biri hisoblanadi. Ushbu maqolada mavjud yondashuvlar, metodlar tahlil qilindi. Qoidaviy (rule-based) tizimlar yuqori aniqlikka ega bo‘lishi mumkin, ammo ularning samaradorligi ko‘p jihatdan qo‘llanilgan lug‘at hajmi va qoidalar bazasining to‘liqligiga bog‘liq. Ayniqsa, morfologik jihatdan boy tillarda, masalan, o‘zbek tilida bu kabi yondashuvlar ko‘p hollarda yetarli natija bermaydi. Statistik modellar, masalan N-gram asosidagi tizimlar, real so‘zlar ichida xatolik aniqlashda nisbatan muvaffaqiyatliroqdir. Biroq ularning kontekstni chuqur anglash imkoniyati cheklangan. Neyron tarmoqlarga asoslangan yondashuvlar esa matnning semantik mazmunini hisobga olish orqali yanada aniqlik bilan xatolarni aniqlay oladi. Masalan, “bilaman” o‘rniga “bolaman” deb yozilgan so‘zlar faqat kontekstual model yordamida to‘g‘ri aniqlanishi mumkin. Bundan tashqari, so‘zning morfologik shakllari va urg‘u (accent) farqlari ham o‘zbek tilida xatolik aniqlashni murakkablashtiradi. Masalan, “boraman” va “boramanmi” so‘zlarining tahlili uchun morfologik analizator zarur bo‘ladi. Bu esa modelni faqat statistik emas, balki tilshunoslikka asoslangan yondashuvlar bilan boyitish zarurligini ko‘rsatadi.

O‘rganilgan ilmiy manbalar va ilgari taklif etilgan metodlardan ko‘rinib turibdiki, matndagi imloviy xatolarni aniqlashning har bir usulda o‘ziga xos afzallik va cheklovlari mavjud bo‘lib, ayniqsa morfologik jihatdan murakkab va o‘ziga xos tillar — xususan o‘zbek tili uchun maxsus moslashtirilgan yondashuvlar zarur.

Foydalanilgan adabiyotlar

1. K. Kukich, “Techniques for automatically correcting words in text,” *ACM Computing Surveys*, 24(4), 377–439, 1992.
2. R. A. Wagner and M. J. Fisher, “The string to string correction problem,” *Journal of Assoc. Comp. Mach*., 21(1):168-173, 1974
3. E. J. Yannakoudakis and D. Fawthrop, “An intelligent spelling error corrector,” *Information Processing and Management,* 19:1, 101-108, 1983..
4. Jin-ming Zhan, Xiaolong Mou, Shuqing Li, Ditang Fang, “A Language Model in a Large-Vocabulary Speech Recognition System,” in *Proc. Of Int. Conf. ICSLP98*, Sydney, Australia, 1998.
5. K. Church and W. A. Gale, “Probability scoring for spelling correction,”*Statistics and Computing*, Vol. 1, No. 1, pp. 93–103, 1991.
6. V. J. Hodge and J. Austin, “A comparison of standard spell checking algorithms and novel binary neural approach,” *IEEE Trans. Know. Dat. Eng.*, Vol. 15:5, pp. 1073-1081, 2003.
7. J. J. Pollock and A. Zamora, “Collection and characterization of spelling errors in scientific and scholarly text,” *Journal Amer. Soc. Inf. Sci*., Vol.34, No. 1, pp. 51–58, 1983.
8. “Automatic spelling correction in scientific and scholarly text,” *Comm. ACM*, Vol. 27, No. 4, pp. 358–368, 1984.
9. E. Brill and R. C. Moore, “An improved error model for noisy channel spelling correction,” in *Proc. 38th Annual Meet. of the Assoc. for Comp. Ling*., Hong Kong, 2000, pp. 286–293.
10. K. Toutanova and R. C. Moore, “Pronunciation modeling for improved spelling correction,” in *Proc. 40th Annual Meeting of the Assoc. for Comp. Ling*., Hong Kong, 2002, pp. 144–151.
11. Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger, “Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness”, 2009
12. F. J. Damerau, “A technique for computer detection and correction of spelling errors,” *Communications of ACM*, 7(3):171-176.7, 1964.
13. E. J. Yannakoudakis and D. Fawthrop, “An intelligent spelling error corrector,” 19(2):101-108; 19(2):87-99, 1983