ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 3 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2024 Vol. 3 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 101-114

N-GRAMLAR ASOSIDA IMLONI TUZATISH TIZIMINI ISHLAB CHIQISH

Botir Elov[1]
Maftuna Ahmedova[2]

Annotatsiya. Imloni avtomatik tuzatish tabiiy tilni qayta ishlash (NLP) sohasidagi juda muhim vazifalardan biri hisoblanadi. Bu qidiruv tizimlari, kayfiyatni tahlil qilish, matnni qisqartirish kabi turli vazifalarda qo‘llaniladi. Imloni tuzatish jarayonida, kutilganidek, xatolarni aniqlash va to‘g‘rilash amalga oshiriladi. NLP vazifalarida biz ko‘pincha tipografik xatolarni o‘z ichiga olgan ma’lumotlar bilan ishlaymiz. Shunday paytlarda imlo tuzatish tizimi yordamga kelib, modelning samaradorligini oshiradi. Masalan, biz “apple” so‘zini qidirishni istab, “aple” deb yozsak, qidiruv tizimi hech qanday natija bermasdan, “apple”ni taklif qilishini xohlaymiz. N-gram modeli imlo xatoliklarini aniqlash va tuzatish jarayonini samarali bajarishda keng qo‘llaniladi. Ushbu maqolada N-gram modeliga asoslangan imlo tuzatish tizimining ishlash prinsiplarini, afzalliklarini, qiyinchiliklarini va amaliy qo‘llanilishini ko‘rib chiqamiz.

Kalit so‘zlar: *n-gram, imloni tuzatish tizimi, Spellcheckerlar, NLP.*

Kirish

Imlo tuzatish - bu har qanday tabiiy tilda yaratilgan hujjatda xato yozilgan so‘zni eng ko‘p mo‘ljallangan so‘z bilan almashtirish jarayoni [Faili va boshq., 2016]. Bu tabiiy tilni qayta ishlash (NLP) sohasida davom etayotgan tadqiqotlarning diqqat markazida bo‘lmoqda [Faili va boshq., 2016]. **N-gramga asoslangan imloni tuzatish** texnikasi tabiiy tilni qayta ishlashda keng qo‘llaniladi. Ushbu yondashuv, matn ichidagi harflar yoki so‘zlarning ketma-

[1]*Elov Botir Boltayevich – texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.*
E-pochta: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2] *Ahmedova Maftuna – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi mutaxasisligi magistranti.*
E-pochta: MaftunaAhmedova1997@gmail.com

ketligini (N-gram) tahlil qilish orqali noto‘g‘ri yozilgan so‘zlarni aniqlash va ularni to‘g‘ri shakllariga tuzatish imkonini beradi. NLP guruhining vazifasi tabiiy tillarni tahlil qiladigan, tushunadigan va yaratadigan dasturiy ta’minotni yaratish va qurishdir [Jain, 2014]. NLPda matnni umumlashtirish, savollarga javob berish, mashina tarjimasi, tahlil qilish, axborotni qidirish va optik aniqlash kabi turli xil ilovalar qo‘llaniladi. Imloni tekshirishning asosiy maqsadi yozma matndagi muammolarni aniqlashdir. Lug‘atni qidirish va N-gram tahlili xatolarni aniqlashning ikkita usulidir. Xatolarni aniqlash usuli ko‘pincha kiritilgan satr lug‘at mavjud leksema ekanligini yoki yo‘qligini aniqlashni talab qiladi.

Bunday xatolarni aniqlash uchun samarali yondashuvlar ishlab chiqilgan. Lug‘atni qidirish va n-gram yondashuvlari ko‘pchilik imlo tekshiruvchilar tomonidan qo‘llaniladi. Yozma matndagi so‘z xato birlik sifatida aniqlanganda, so‘zni tuzatish yoki to‘g‘ri variantlarni taqdim etish uchun imlo tuzatish tartib-qoidalari qo‘llaniladi. Matn xatolarini tuzatish uchun qoidalarga asoslangan texnikalar, masofaviy tahrirlash texnikasi, N-gram texnikasi va chuqur o‘rganish usullari kabi ko‘plab usullar mavjud.

Hozirgi vaqtda imloni tuzatish ko‘plab dasturlarda keng qo‘llaniladi. Shaxsiy kompyuterlarda ham, tarmoq kompyuterlarida ham ishlov berish hajmi oshgani sababli, imloni avtomatik tuzatish tabiiy holatga aylandi. So‘nggi o‘n yil ichida imloni tuzatish tobora ko‘proq avtomatlashtirildi va kompyuterlarning oddiy foydalanuvchisi bunday imlo tuzatuvchilari ilovalarga, ya’ni ofis to‘plamlariga o‘rnatilganligini biladi. Oddiy foydalanuvchi bilmaydigan narsa - tuzatish tizimining qanday ishlashi va u qanday modelga amal qiladi.

Matnlardagi so‘zlarni avtomatik ravishda to‘g‘rilash uchun algoritmlar va usullarni ishlab chiqish masalasi doimiy tadqiqot muammolaridan biri bo‘lib kelmoqda. Avtomatik imlo tuzatish va avtomatik matnni aniqlash uchun kompyuter texnikasi bo‘yicha ishlar 1960-yillarda boshlangan va hozirgi kungacha davom etib kelmoqda. Ushbu sohadagi tadqiqotlarning davomiyligining asosiy sabablari sifatida sifat va samaradorlikni oshirish hamda qo‘llash imkoniyatlarining spektrini kengaytirishni ko‘rsatish mumkin [Kukich, 1992].

Masalan, tizim dasturlari (til protsessorlari, operatsion tizimlar va boshqalar) tobora kuchliroq va murakkabroq bo‘lib borayotganiga qaramay, ular (kamdan-kam istisnolarni hisobga olmaganda) foydalanuvchiga kirish qismidagi ko‘plab aniq imlo xatolarini tuzatishda yordam bermaydi.

So‘z xatolari ikki turga bo‘linadi: **haqiqiy so‘z xatosi** va **noto‘g‘ri so‘z xatosi**.

• **Haqiqiy so‘z xatosi** deganda ma’noga ega bo‘lgan va lug‘atda mavjud bo‘lgan noto‘g‘ri yozilgan so‘zlar tushuniladi.

• **Noto‘g‘ri so‘z xatosi** esa ma’nosiz bo‘lgan va lug‘atda mavjud bo‘lmagan so‘zlarni anglatadi.

Maqolada o‘rganiladigan va taklif qilingan algoritm bilan noto‘g‘ri so‘z xatolarini tuzatishga e’tibor qaratamiz. Damerau (1964) tadqiqotlariga ko‘ra, noto‘g‘ri so‘z xatolarining 80 foizi bitta harfni qo‘shish, o‘chirish, almashtirish yoki joyini almashtirish natijasida yuzaga keladi [Damerau, 1964]. Shuning uchun bunday oddiy operatsiyalarni hisobga oluvchi tuzatish algoritmlarini yaratish mantiqiydir. Ammo sof n-gram statistikasi (bu operatsiyalarni bilvosita hisobga oladi) asosida ishlovchi yondashuvlar ham yaxshi natijalar berganini ko‘rsatdi [Kukich, 1992; Hodge, Austin, 2003].

Imlo tekshiruvchi ba’zi yondashuvlar

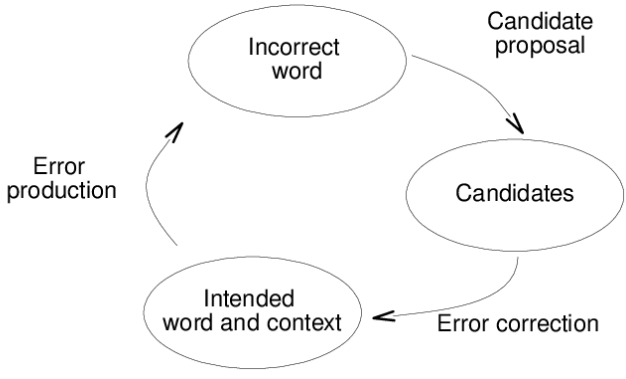
Matndagi imlo xatolarini aniqlash va tuzatish bo‘yicha algoritmik texnikalar informatika sohasida uzoq va mustahkam tarixga ega [Kukich, 1992]. Ushbu muammoni hal qilish bilan shug‘ullanish boshlanganidan buyon ko‘plab yondashuvlar qo‘llanilgan. Masalan, **tahrir masofasi** [Wagner, Fisher, 1974], **qoidaga asoslangan usullar** [Yannakoudakis, 1983], **n-gramlar** [Zhan va boshq., 1988], **ehtimollik usullari** [Church, Gale, 1991], **neyron tarmoqlar** [Hodge, Austin, 2003], **o‘xshashlik kalitlari texnikalari** [Pollock, Zamora, 1983] va **shovqinli kanal modeli** [Brill, Moore, 2000; Toutanova, Moore, 2002] kabi texnikalar taklif qilingan. Ushbu usullar lug‘atdagi so‘zlar va noto‘g‘ri yozilgan so‘zlar orasidagi o‘xshashlikni hisoblash g‘oyasiga asoslanadi. Quyida biz eng mashhur yondashuvlardan biri (Aspell) va portugal tiliga moslashgan yaqinda taklif qilingan yondashuv (TST) haqida qisqacha ma’lumot beramiz.

GNU Aspell

GNU Aspell, odatda faqat Aspell deb ataladi, GNU dasturiy tizimi uchun standart imlo tekshirish dasturi hisoblanadi. Unda taxminan 70 til uchun lug‘atlar mavjud. GNU Aspell bepul va ochiq manba kodli bo‘lib, uni http://aspell.sourceforge.net/ manzilidan yuklab olish mumkin. Ispell bilan solishtirganda, u kichik tahrir masofasiga ega so‘zlarni taklif qilsa, Aspell buning ustiga tovushga o‘xshash ekvivalentlarni ham (ingliz so‘zlari uchun metaphone algoritmi [Deorowicz, Ciura, 2005] yordamida hisoblab) berilgan tahrir masofasigacha solishtiradi.

Ternary Search Trees (TST)

Ternary Search Trees (TST) yondashuvi bu matnli kalitlar bilan ishlaydigan lug‘at ma’lumotlar tuzilmasidir. U ushbu kalitlarni tezda topish, o‘chirish va qo‘shish imkoniyatini beradi hamda qisman mosliklarni osongina qidirishga imkon beradi. Bundan tashqari, yaqin moslik funksiyalarini amalga oshirish mumkin. Bu funksiyalar noto‘g‘ri yozilgan so‘zlar uchun muqobil variantlarni taklif qilish imkonini beradi.



1-rasm. Xatolarni ishlab chiqish va tuzatishning interfaol jarayonlari

N-gramga asoslangan imloni tuzatish qanday ishlaydi?

1. N-gramlarni yaratish

1. N-gramlar noto‘g‘ri yozilgan so‘zdan va lug‘atdan olingan barcha so‘zlardan yaratiladi.

2. Har bir lug‘atdagi so‘z uchun N-gramlar to‘plami hosil qilinadi.

2. O‘xshashlikni hisoblash

• Noto‘g‘ri yozilgan so‘z va lug‘atdagi har bir so‘zning N-gramlari o‘zaro taqqoslanadi.

• O‘xshashlik darajasi odatda **Jaccard koeffitsienti** yoki boshqa statistik usullar yordamida hisoblanadi:

$$\delta(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|}$$

• bu yerda s_1 va s_2 — ikki so‘zning N-gramlar to‘plami.

3. Tavsiya etiladigan so‘zlar ro‘yxatini tuzish

• O‘xshashlik darajasiga qarab, lug‘atdagi so‘zlar tartiblanadi.

• Eng yuqori o‘xshashlikka ega so‘zlar noto‘g‘ri yozilgan so‘zning tavsiya etiladigan tuzatishlari sifatida ro‘yxatga kiritiladi.

4. Statistik modellarni qo‘llash

• Eng ko‘p uchraydigan N-gramlar asosida so‘zlar ustuvorlik bilan tartiblanadi.

• Ushbu jarayon uchun **N-gram chastotalari** korpusdan yoki boshqa tahliliy ma’lumotlar to‘plamidan olinadi.

N-gram yondashuvining afzalliklari

1. Tilga bog‘liq emas: Ushbu yondashuv tilga xos grammatik qoidalarga bog‘liq bo‘lmay, faqat so‘z yoki harflar ketma-ketligini tahlil qiladi.

2. Tezkorlik: N-gram modellarini qurish va o‘xshashlikni hisoblash kompyuterda nisbatan tez amalga oshiriladi.

3. Tuzatish imkoniyati: N-gramlar avtomatik ravishda eng ko‘p uchraydigan ildizlar va o‘xshashliklarni topishda yordam beradi.

Kamchiliklari

1. Korpusga bog‘liq bo‘lishi: N-gram chastotalarining to‘g‘ri ishlashi uchun katta va sifatli korpus zarur.

2. Kontekstdan foydalanmaslik: Izolyatsiyada ishlaydigan N-gramlar matn konteksti bo‘yicha ma’noni to‘g‘ri anglab ololmaydi.

3. Resurs talabchanligi: Katta lug‘atlar yoki korpuslar bilan ishlashda hisoblash jarayoni ko‘p resurs talab qilishi mumkin.

Qo‘llanilishi

• **Imloni tekshirish va tuzatish**: Google, Microsoft Word kabi dasturlarda ishlatiladi.

• **So‘zlar tavsiyasi**: Avtokorrektor funksiyalarida.

• **Matnni avto-tahlil qilish**: Ma’lumotlarni tuzatishda, masalan, OCR tizimlarida.

Tilni qayta ishlashda n-gramlardan foydalanish g‘oyasi birinchi marta Shannon [Shannon, 1951] tomonidan muhokama qilingan. Ushbu dastlabki ishlardan so‘ng, n-gramlardan foydalanish g‘oyasi so‘zlarni taxmin qilish, imlo tuzatish, nutqni tanib olish, tarjima qilingan so‘zlarni tuzatish va qatorlarni qidirish kabi ko‘plab muammolarga tatbiq qilindi. N-gram usulining asosiy afzalligi uning tilga bog‘liq emasligidir.

Imlo tuzatish vazifasida n-gram deganda, so‘z yoki satrda ketma-ket joylashgan n ta harfning ketma-ketligi tushuniladi. N-gram modeli ikki satr o‘rtasidagi o‘xshashlikni hisoblash uchun ishlatilishi mumkin, bunda ularning umumiy n-gramlari soni hisobga olinadi. Ikki satr o‘rtasida qancha ko‘p umumiy n-gram mavjud bo‘lsa, ular bir-biriga shunchalik o‘xshash bo‘ladi. Shu asosda o‘xshashlik

koeffitsienti aniqlanishi mumkin.

O‘xshashlik koeffitsienti δ quyidagi tenglama yordamida aniqlanadi:

$$\delta(\alpha, \beta) = \frac{\alpha \cap \beta}{\alpha \cup \beta}$$

Bu yerda:

• α va β — taqqoslanayotgan ikki so‘zning n-gram to‘plamlari.

• |α∩β| — α va β dagi umumiy n-gramlar soni.

• |α∪β| — α va β ning birlashmasidagi noyob n-gramlar soni.

Imlo tekshirish texnikalari sezilarli darajada rivojlandi, masalan, xatolarni aniqlash va tuzatish usullari. Xatolarni aniqlashning ikki umumiy yondashuvi - lug‘at orqali qidirish va n-gram tahlilidir. Adabiyotlarda tasvirlangan ko‘pchilik imlo tekshiruvi usullari lug‘atlarni to‘g‘ri yozilgan so‘zlar ro‘yxati sifatida ishlatadi, bu algoritmlarga maqsadli so‘zlarni topishda yordam beradi. Turli xil yechimlar taklif qilingan, masalan, Gokhan Dalkilich va Yalchin Chebi matn massasidagi noto‘g‘ri yozilgan so‘zlarni aniqlash uchun n-gram tahliliga asoslangan usulni taklif qilishgan.

N-gramni ishlatishning birinchi bosqichi korpus yordamida tilga xos n-gramni aniqlashdir. Ammo biror korpus barcha mumkin bo‘lgan so‘z n-gramlarini qamrab olish uchun yetarli darajada katta bo‘lolmaydi. Orqaga qaytish silliqlash usuli korpusdagi noma’lum n-gramning chastotasini taxmin qilish usullaridan biridir. Agar mavjud bo‘lmagan n-gram topilsa, so‘z noto‘g‘ri yozilgan deb aniqlanadi.

Lug‘at ma’lum bir tilga xos to‘g‘ri so‘zlarning ro‘yxatini o‘z ichiga olgan leksik manbadir. Lug‘atga asoslangan usullar [de Amorim, 2009] o‘zining ichki arxitekturasi tufayli hali ham ishlash cheklovlariga ega, bu ishlash chekloviga umumiy muqobil esa lug‘atlarni Cheklangan Holatlar Avtomati (Finite State Automata - FSA) sifatida tashkil qilishdir.

FSA morfologik jihatdan boy tillar, masalan, vengr, fin va turk tillari uchun ayniqsa qiziqarlidir. Lug‘atlarni FSA shaklida tashkil qilish bo‘yicha imlo tekshirish tadqiqotiga [Hulden, 2009] o‘xshashlik mezonlari, masalan, minimal tahrir masofasi asosida, finite-state avtomatdagi qatorlarni taxminiy moslashtirish algoritmini taqdim etdi. Algoritm ikki so‘z orasidagi masofani aniqlash uchun turli xil metrikalardan foydalanishi mumkin va so‘z bilan katta so‘zlar ro‘yxati orasidagi eng yaqin moslikni topish juda talabchan vazifa ekanligini ta’kidlaydi.

V. Ramaswamy va H. A. Girijamma finite avtomatni fuzzy avtomatga o‘tkazish usulini taqdim etdi, chunki fuzzy avtomat individual belgilarning yoki belgilar ketma-ketligining o‘xshashlik darajalari aniqlanganda, qatorlarni taqqoslash uchun finite avtomatga qaraganda samaraliroqdir. Finite avtomat berilgan qator qabul qilinganmi yoki yo‘qligini aniqlashda foydali bo‘lsa, fuzzy avtomat qatorning qanchalik darajada qabul qilinganligini aniqlaydi. Ushbu usul noto‘g‘ri yozishlar uchun hisoblanadigan masofalar sonini kamaytiruvchi va shuning uchun ishlov berish tezligini oshiruvchi FSA-ga asoslangan lug‘atlarga muqobil bo‘lib xizmat qiladi. Tegishli ishlarga ko‘ra, imlo xatolari ikki turga bo‘linadi: kognitiv xatolar va tipografik xatolar. **Tipografik xatolar**: Damerau tomonidan o‘tkazilgan tadqiqotda tipografik xatolarning 80%i quyidagi to‘rtta kategoriya ichiga kiradi:

- yolg‘on harf qo‘shish; masalan, “ccomputer” o‘rniga “computer” yozish. - yolg‘on harf o‘chirish; masalan, “cmputer” o‘rniga “computer” yozish. - yolg‘on harf almashtirish; masalan, “compoter” o‘rniga “computer” yozish. - ikki yonma-yon harfni joyidan o‘zgartirish; masalan, “cumpoter” o‘rniga “computer” yozish.

Kognitiv xatolar: Bu xatolar so‘zning to‘g‘ri yozilishi ma’lum bo‘lmaganda yuzaga keladi. Ushbu turdagi xatolarda, noto‘g‘ri yozilgan so‘zni talaffuzi to‘g‘ri yozilgan so‘zga o‘xshash yoki unga yaqin bo‘ladi. Masalan, “piece” o‘rniga “peace” yozish.

Shannon, tilni qayta ishlashda n-gramdan foydalanish g‘oyasini muhokama qilgan. Ushbu ishlardan so‘ng, n-gramdan foydalanish g‘oyasi nutqni tanib olish, tarjima qilingan so‘z, to‘g‘ri so‘z tuzish, oldindan aytish va imlo xatolarini tuzatishda qo‘llanildi. Ushbu texnika, to‘liq statistik bo‘lib, hujjatning tilini bilishni talab qilmaydi. N-gramning yana bir afzalligi – eng ko‘p uchraydigan ildizlarni avtomatik tarzda aniqlashdir. N-gram ikki yo‘lda qo‘llanilishi mumkin: birinchidan, so‘z lug‘atisiz, ikkinchidan, so‘zlik bilan birga.

N-gramni so‘zlikdan foydalanmasdan qo‘llashda, xato so‘zdagi xatolikni aniqlash uchun ishlatiladi. Agar xato so‘zni shunday o‘zgartirishning maxsus usuli bo‘lsa, u faqat to‘g‘ri n-gramni o‘z ichiga olsa, bu tuzatish sifatida qabul qilinadi. Bu usulning ishlash samaradorligi past, ammo u sodda usul bo‘lib, so‘zlikni talab qilmaydi. Boshqa yo‘lda, so‘zlik bilan birga, n-gram so‘zlar orasidagi masofani aniqlash uchun ishlatiladi, ammo so‘zlar har doim so‘zlikka nisbatan tekshiriladi. Bu ko‘p usullarni talab qiladi, masalan, noto‘g‘ri yozilgan so‘z va so‘zlikdagi so‘zlar o‘rtasidagi qancha

umumiy n-gram borligini tahlil qilish, so‘z uzunligiga qarab og‘irlik berish. Leksik resurslar tabiiy tillardagi so‘zlar haqidagi lingvistik ma’lumotlarni taqdim etadi. Ushbu ma’lumotlar oddiy ro‘yxatlardan tortib, ko‘p turdagi lingvistik ma’lumotlar va bog‘lanishlar bilan murakkab tuzilmalarda taqdim etilishi mumkin. Leksik resurslar til va kompyuter lingvistikasi uchun ishlatiladi. Bu, kompyuterlar va insonlar tomonidan zarur bo‘lgan ma’lumotlarni tayyorlash, qayta ishlash va boshqarishda rol o‘ynaydi.

O‘xshashlikni hisoblash

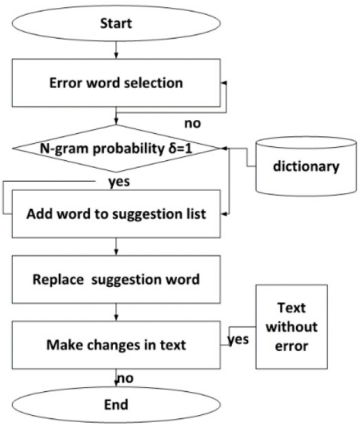
O‘xshashlik kalitlari, minimal tahrir masofasi, neyron tarmoqlar va n-gram asosida bir necha yondashuvlar mavjud. Ehtimollik xato tuzatish vazifasini bajarish uchun taklif etilgan. N-gram, satr harflarini solishtirish uchun ishlatiladi. Bu tilga bog‘liq emas, bu texnika faqat so‘zlarning harflarini solishtiradi, foydalangan tilga qarab emas, u ikki satr o‘rtasidagi o‘xshashlikni ular o‘rtasida o‘xshash n-gramlarning sonini sanash orqali hisoblaydi. O‘xshashlik koeffitsienti asosida, ikki satr o‘rtasida necha o‘xshash n-gram mavjud bo‘lsa, ular shuncha o‘xshashdir. Umuman olganda, o‘xshashlik koeffitsienti δ quyidagi tenglama orqali hisoblanadi.

Bu yerda s_1 va s_2 ikki so‘zning n-gram to‘plamlari bo‘lib, ular solishtiriladi. $| s_1 \cap s_2 |$ s_1 va s_2 o‘rtasidagi o‘xshash n-gramlar sonini ko‘rsatadi, va $| s_1 \cup s_2 |$ s_1 va s_2 ning birlashmasidagi noyob n-gramlar sonini ko‘rsatadi.

“Camputer” so‘zining xatoliklari va “computer” to‘g‘ri so‘zi o‘rtasidagi o‘xshashlik koeffitsienti, n=2 (bigram) bilan n-gram yordamida hisoblash misoli sifatida 1-jadvalda ko‘rsatilgan, shuningdek, 1-rasmda n-gramning amalga oshirilishi tasvirlangan.

1-jadval. Ikki so‘z o‘rtasida bigram o‘xshashlik koeffitsientini hisoblashga misol

bi-grams	**Computer**	**Computer**
Co	1	-
Om	1	-
Mp	1	1
Pu	1	1
Ut	1	1
Te	1	1
Er	1	1
Ca	-	1
Am	-	1
Similarity coefficient	**5/9 = 0.55**	



2-rasm. Taklif etilayotgan usulda n-gram usuli ketma-ketligi

Xatoni topgandan so‘ng, odatda imlo tekshirgich (spellchecker) tuzatish uchun qisqa tavsiyalar ro‘yxatini taqdim etadi, bu ro‘yxatda eng yaxshi taxmin yuqorida bo‘ladi. Kundalik foydalanishda imlo tekshirgich butun matn parchalari bo‘ylab xatolarni tekshirish uchun chaqiriladi, bunda xatolar kontekstda paydo bo‘ladi. Ammo, imlo tekshirgich kontekstdan foydalanmasdan ham sezilarli natijalarga erishishi mumkin va, aslida, ko‘plab imlo tekshirgichlar kontekst mavjud bo‘lsa ham faqat alohida so‘zlarni tuzatish bilan cheklanadi [Kukich, 1992].

Kontekstsiz ham imlo tekshirgich tavsiyalar ro‘yxatini shakllantirishda turli xil ma’lumot turlaridan foydalanishi mumkin. Eng oddiy holatda, imlo tekshirgich noto‘g‘ri yozilgan so‘zni to‘g‘ri yozilgan so‘zlar lug‘ati bilan taqqoslaydi, bu yerda so‘zlarni oddiy harf qatorlari sifatida ko‘radi. Lekin, shuningdek, tipik imlo xatolari yoki so‘zlarning boshqa jihatlari, masalan, talaffuz yoki so‘zlarning ishlatilish chastotasi haqidagi ma’lumotlardan ham foydalanishi mumkin. Buning uchun tekshirgichni bosqichma-bosqich rivojlantirib, turli xil ma’lumot turlarini hisobga oladigan modullarni qo‘shish va har birining hissasini baholash mumkin.

Ushbu maqola doirasida, biron bir tarzda xato aniqlanganini faraz qilamiz va tavsiyalar ro‘yxatini shakllantirishga e’tibor qaratamiz. Biroq, bu bilan xatolarni aniqlash, tuzatishdan farqli ravishda, ahamiyatsiz jarayon ekanligini anglatmoqchi emasman. Aslo unday emas. Bu imlo tekshirgichlarining hozirgi versiyalari qisman muvaffaqiyatga erishgan murakkab masala bo‘lib, haqiqiy so‘z xatolarini aniqlashga katta e’tibor qaratilgan - masalan, “shoh” o‘rniga “shox” kabi noto‘g‘ri so‘zdan foydalanish. Bunday holatlarda kontekstdan foydalanish muhim ahamiyatga ega.

N-gramlar asosida imloni tuzatishda natijalar quyidagilarni o‘z ichiga oladi:

1. Aniqlangan xatolar va ularning to‘g‘rilanishi

N-gramlar asosida ishlaydigan tizimlar noto‘g‘ri yozilgan so‘zlarni aniqlab, ularning ehtimoliy to‘g‘ri variantlarini topadi. Natijada:

• Xato so‘zlar aniqlanadi.

• Ushbu so‘zlar uchun taklif qilingan tuzatishlar ro‘yxati hosil qilinadi.

• Eng mos keladigan tuzatish (masalan, Jaccard koeffitsienti asosida) birinchi o‘ringa qo‘yiladi.

Misol: ingliz tilida computer so‘zi yozilishi kerak bo‘lsa:

• Noto‘g‘ri yozilgan so‘z: *camputer*

• Lug‘atdagi so‘zlar: *computer*, *chapter*, *capture*

• N-gram tahlili natijasida, eng yuqori o‘xshashlikka ega bo‘lgan *computer* birinchi bo‘lib tavsiya qilinadi.

Shunday qilib, avtomatik so‘z tuzatishni o‘rganish, tasviriy maqsadlar uchun uchta asta-sekin kengayadigan mavzuga e’tibor qaratilgan deb qaralishi mumkin: 1) So‘z bo‘lmagan xatolarni aniqlash, 2) Yolg‘iz so‘z xatolarini tuzatish va 3) Kontekstga asoslangan so‘z tuzatish. 1970-yillarning boshlaridan 1980-yillargacha birinchi masala bo‘yicha ish olib borildi. Ushbu davrda asosan tez string taqqoslash va naqshlarni moslashtirish yondashuvlarini tekshirishga qaratilgan ishlar olib borildi, buning orqali berilgan kirish stringi lug‘at yoki so‘zlar ro‘yxatida topilishi aniqlanadi. 1970-yillardan hozirgi kungacha ikkinchi muammo ustida ko‘proq vaqt sarflandi. O‘sha paytda, imlo xatolarini tuzatish uchun turli umumiy va maxsus metodlar ishlab chiqildi, ularning ba’zilari imlo xatosi naqshlari bilan birgalikda ishlatilgan. 1980-yillarning boshlarida avtomatik NLP modellarini yaratish uchinchi muammoga qiziqish uyg‘otdi va statistik til modellari uning rivojlanishini yana jonlantirdi.

Xulosa

N-gramga asoslangan imlo tuzatish metodi, so‘zlarni tuzatishda samarali va universal yondashuvlardan biri sifatida ishlatiladi. Ushbu metod, xato yozilgan so‘zlar va ularning to‘g‘ri shakllarini aniqlashda, so‘zlar orasidagi o‘zaro bog‘lanishlarni hisobga oladi. N-gramlar (masalan, bigramlar yoki trigramlar) so‘zlarning yoki harflarning ketma-ketligini tahlil qilib, matnning tabiiy strukturasini aniqlashga yordam beradi. Biroq, n-gram asosidagi metodning ba’zi cheklovlari ham mavjud. Masalan, katta

o‘lchamdagi n-gramlar ko‘proq hisoblash resurslarini talab qiladi va ba’zan kam uchraydigan so‘zlar yoki xatolar uchun muvaffaqiyatli ishlamasligi mumkin. Shunday qilib, n-gram asosida imlo tuzatishni qo‘llashda uning samaradorligini oshirish uchun qo‘shimcha usullar va resurslardan foydalanish zarur bo‘lishi mumkin.

Foydalanilgan adabiyotlar

Faili, H., Ehsan, N., Montazery, M., & Pilehvar, M. T. (2016). Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language. Digital Scholarship in the Humanities, 31(1), 95-117.

Jain, A., & Jain, M. (2014, September). Detection and correction of non word spelling errors in Hindi language. In 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC) (pp. 1- 5). IEEE

Kukich, K. “Techniques for automatically correcting words in text,” ACM Computing Surveys, 24(4), 377–439, 1992.

Damerau, F. J. “A technique for computer detection and correction of spelling errors,” Communications of ACM, 7(3):171-176.7, 1964.

Hodge V. J. 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.

Wagner R. A. and M. J. Fisher, “The string to string correction problem,” Journal of Assoc. Comp. Mach., 21(1):168-173, 1974.

Yannakoudakis E. J. and D. Fawthrop, “An intelligent spelling error corrector,” Information Processing and Management, 19:1, 101-108, 1983.

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.

Church K. and W. A. Gale, “Probability scoring for spelling correction,” Statistics and Computing, Vol. 1, No. 1, pp. 93–103, 1991.

Hodge V. J. 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.

Pollock J. J. 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.

Brill E. 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.

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.

Deorowicz S. and M. G. Ciura, "Correcting Spelling Errors by Modelling Their Causes," Int. Journal of Applied Mathematics and Computer Science, 15(2):275–285, 2005.

Shannon, C. E. "Prediction and entropy of printed English," Bell Sys. Tec. J. (30):50–64, 1951.

DEVELOPMENT OF A SPELL CORRECTION SYSTEM BASED ON N-GRAMS

Botir Elov[1],
Maftuna Ahmedova[2]

Abstract. Automatic spelling correction is one of the most crucial tasks in the field of natural language processing (NLP). It is applied in various tasks such as search engines, sentiment analysis, and text summarization. As expected, the spelling correction process involves identifying and correcting errors. In NLP tasks, we often work with data containing typographical errors. In such cases, a spelling correction system comes to the rescue and enhances the model's efficiency. For example, if we want to search for the word "apple" but mistakenly type "aple," we want the search engine to suggest "apple" without yielding any irrelevant results. The N-gram model is widely used to effectively perform spelling error detection and correction. In this article, we will explore the operating principles, advantages, challenges, and practical applications of the spelling correction system based on the N-gram model.

Keywords: *n-gram, spelling correction system, spellcheckers, NLP.*

References

Faili, H., Ehsan, N., Montazery, M., & Pilehvar, M. T. (2016). Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language. Digital Scholarship in the Humanities, 31(1), 95-117.

Jain, A., & Jain, M. (2014, September). Detection and correction of non word spelling errors in Hindi language. In 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC) (pp. 1- 5). IEEE

Kukich, K. "Techniques for automatically correcting words in text," ACM Computing Surveys, 24(4), 377–439, 1992.

[1]*Elov Botir Boltayevich – doctor of philosophy in technical sciences, docent. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.*
E-pochta: elov@navoiy-uni.uz
ORCID: 0000-0001-5032-6648
[2] *Ahmedova Maftuna – Master of degree. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.*
E-pochta: MaftunaAhmedova1997@gmail.com

Damerau, F. J. "A technique for computer detection and correction of spelling errors," Communications of ACM, 7(3):171-176.7, 1964.

Hodge V. J. 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.

Wagner R. A. and M. J. Fisher, "The string to string correction problem," Journal of Assoc. Comp. Mach., 21(1):168-173, 1974.

Yannakoudakis E. J. and D. Fawthrop, "An intelligent spelling error corrector," Information Processing and Management, 19:1, 101-108, 1983.

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.

Church K. and W. A. Gale, "Probability scoring for spelling correction," Statistics and Computing, Vol. 1, No. 1, pp. 93–103, 1991.

Hodge V. J. 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.

Pollock J. J. 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.

Brill E. 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.

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.

Deorowicz S. and M. G. Ciura, "Correcting Spelling Errors by Modelling Their Causes," Int. Journal of Applied Mathematics and Computer Science, 15(2):275–285, 2005.

Shannon, C. E. "Prediction and entropy of printed English," Bell Sys. Tec. J. (30):50–64, 1951.

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga **.**.****-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.