ҒТАХР 16.31.61
ӘОЖ 004.912
JEL C88, O33, D83

https://doi.org/10.46914/1562-2959-2024-1-3-128-138

БАЙТЕНОВА Л.М.,*[1]
э.ғ.д., профессор.
*e-mail: l.baitenova@turan-edu.kz
ORCID ID: 0000-0002-1591-2235
РАХИМОВА Д.Р.,[1,2]
PhD, қауымдастырылған профессор.
e-mail: di.diva@mail.ru
ORCID ID: 0000-0003-1427-198X
ТУРАРБЕК Ә.Т.,[1,2]
PhD, доцент.
e-mail: turarbek_asem@mail.ru
ORCID ID: 0000-0002-4793-0446
АДАЛИ Е.,[3]
PhD, профессор.
e-mail: adali@itu.edu.tr
ORCID ID: 0000-0002-1561-8255
[1]«Тұран» университеті,
Алматы қ., Қазақстан
[2]әл-Фараби атындағы Қазақ ұлттық университеті,
Алматы қ., Қазақстан
[3]Стамбұл техникалық университеті,
Стамбұл қ., Түркия

МЕМЛЕКЕТТІК ТІЛДЕГІ ЖАРТЫЛАЙ ҚҰРЫЛЫМДЫҚ БАСЫЛЫМДАРДАҒЫ ҚАТЕЛЕРДІ АНЫҚТАУДЫҢ ЭКОНОМИКАЛЫҚ АСПЕКТІЛЕРІ

Аңдатпа

Интернеттегі және әлеуметтік желілердегі ақпараттың тез өсуіне байланысты қазіргі уақытта компьютерлік лингвистика саласындағы зерттеулер өте өзекті болып отыр. Адамдар мен машиналар табиғи тілде жасайтын ақпарат көлемін өңдеу, талдау және тексеру қажет. Ол үшін ақпаратты іздеу жүйелері, диалогтық жүйелер, машиналық аударма құралдары қолданылады. Мәтінді автоматты өңдеу жүйелерінің өзі әртүрлі бағыттарды қамтитын өте кең сала. Мәтіндер мен сөздердегі қателерді табу, қате сөздерді анықтау және түзету табиғи тілді өңдеудің (NLP) маңызды міндеттерінің бірі. Мақалада табиғи тілдердегі қате сөздерді анықтаудың жартылай құрылымдық деректері, әдістері мен технологиялары қарастырылған. Қазақ тіліндегі қате сөздерді анықтау тәсілі құрылып, бұл тәсілдің ерекшеліктері мен мүмкіндіктері талданды. Зерттеудің мақсаты – қазақ тіліндегі мәтіндерде, әсіресе ресурстардың шектеулілігі мен құрылымдалмаған деректер жағдайында кездесетін қателерді анықтау және түзетудің тиімді әдісін әзірлеу. Зерттеу машиналық оқыту әдістерін пайдалануды, сондай-ақ осындай шешімдерді әзірлеу және енгізу шығындарын экономикалық талдауды қамтиды. Ұсынылған тәсіл мәтінді тексеруді автоматтандыруға көмектеседі, бұл деректерді қолмен өңдеу құнын айтарлықтай төмендетуге және әртүрлі салаларда, соның ішінде бизнес пен мемлекеттік басқаруда ақпарат сапасын жақсартуға мүмкіндік береді.

Тірек сөздер: экономикалық аспектілер, мемлекеттік тіл, қазақ тілі, жартылай құрылымдық деректер, мәтін, қате, қате сөздер, әлеуметтік желі.

Кіріспе

Интернеттегі және әлеуметтік желілердегі орасан зор ақпарат ағыны табиғи тілді өңдеу саласының, компьютерлік лингвистиканың қарқынды дамуына әкелді. Қазіргі уақытта әртүрлі зерттеу механизмдері пайдаланушылар арасында ақпарат алмасу, ақпаратты машиналық аудару, электрондық поштаны тексеру, сұрақ-жауап жүйесін дамыту сияқты өз жобаларын жасауда [1].

Жалпы, мәтіндер мен ондағы сөздердегі қателерді тауып, түзету міндеті – табиғи тілдегі сөз өңдеудің негізгі міндеттерінің бірі. Жарты ғасырдан астам уақыт бойы бұл тақырып өзектілігін жойған жоқ, жаңа әдістер пайда болды, оның қолданылу аясы кеңейіп келеді.

Интернетте және Instagram, VKontakte, Facebook және басқа да әлеуметтік желілерде қосымшалар хабарламалардағы ақпаратты қабылдау және талдау тұрғысынан өте тартымды, өйткені бұл жүйелердегі ақпарат нақты, олар дәл осы уақытта пайда болады [2]. Дегенмен, интернеттегі мәтін көбінесе тілдің жалпы қабылданған үлгісінен ерекшеленеді. Сөздерді әдейі бұрмалауынан әртүрлі қателер шығады [3]. Осындай қатесі бар сөздерді ақпаратты анықтау-түзету арқылы өңдеуге және талдауға болады, өйткені бұл қателер мәтінді оқып, өңдеуді қиындатады. Табиғи тілді өңдеу сөздің қалыпты үлгілерін қажет етеді, себебі мәтінді қате жазу немесе цифрлау ақпараттық құндылықты төмендетеді. Орфографиялық қате, мысалы, медициналық құжаттардың дерекқорында, диагностика процесінің тиімділігін төмендетеді, ал Интернеттегі қателер пайдаланушылардың қате жазылған пікірлері мен жарияланымдары зерттеу немесе ұйымдастыру процестеріне әсер етуі мүмкін [4].

Қазақ тілі аз ресурсты тілдер тобына жататындықтан, аударма жүйелері, сөздіктер, корпус (көп тілді және екі тілді), сөздердегі қателерді тауып, түзететін жүйелер мен бағдарламалар аз екені белгілі. Осыған сәйкес, бүгінгі таңда қазақ тілі сияқты ресурстары шектеулі тілдердің қолданылуын жақсартатын орфографиялық қателерді анықтайтын бағдарламалар мен жүйелерді әзірлеу маңызды.

Жартылай құрылымды деректер – бұл құрылымдық реляциялық деректер қоры модельдеріндегі кестелер мен қатынастардың қатаң құрылымына сәйкес келмейтін деректер. Интернеттегі ақпарат әрқашан белгілі бір білім саласына қатысты бола бермейді. Осыған байланысты көптеген ұйымдар мен ғалымдар білім беру саласына қатысы жоқ мәтін құрылымын құрудың нақты алгоритмдерін жасауда [5].

Жартылай құрылымды деректер зерттеудің маңызды объектісіне айналады, өйткені Интернетті дамыту үшін толық мәтінді құжаттар мен мәліметтер қоры арасындағы байланыстырушы рөл атқаратын деректер пішімі (JSON, XML және т.б.) болу қажет. Жартылай құрылымдық деректері бар жүйелердің мысалдарына Интернетте, веб-сайттарда және әлеуметтік желілерде пайдаланушылар жазған пікірлер, жарияланымдар және мәтіндер жатады [6]. Мұндай жүйелерден алынған деректер зерттеулер мен қолданбалар үшін үлкен қызығушылық тудырады, өйткені нақты уақыт режимінде кез келген мәселе бойынша адамдардың пікірі мен көңіл-күйін жариялауға және ақпараттың артуына ықпал етеді. Ол сондай-ақ адамдардың бизнеске, саясатқа және бүгінгі таңдағы әлеуметтік жүйеге деген көзқарасын өзгертуге көмектеседі. Мәліметтердің әрбір түрінің өзіне тән ерекшеліктері бар, оларды жинау, дайындау, алдын ала өңдеу және объектілерді сипаттау кезінде ескеру қажет.

Зерттеу барысында Интернет пен әлеуметтік желілердегі ақпарат пайдаланылды және бұл деректер жоғарыда айтылғандай жартылай құрылымды, олар зерттеу барысында тәжірибеде қолданылған.

Мәтіндегі орфографиялық қателерді анықтау және түзету мәселелері мен жұмысы шамамен 1960 ж. басталып, бүгінгі күнге дейін жалғасуда. Сапа мен өнімділікті жақсарту, сондай-ақ ықтимал қолданбалар ауқымын кеңейту үшін осы саладағы зерттеулерді жалғастыруға жақсы себептер бар. Мысалы, жүйелік бағдарламалар (процессорлар және т.б.) күрделі бола бастағанымен, олар пайдаланушыға кіріс көзі деректеріндегі көптеген айқын орфографиялық қателерін түзетуге көмектеспейді [7]. Қателерді табу және түзету мәселесін шешудің 50 жылында зерттеушілер көптеген әртүрлі әдістерді қолданып көрді. Таңба кодтары мен n-грамды қабылдау кестелерінен және Дамерау-Левенштейн қашықтықты тікелей қолданудан бастап, сөз туралы фонетикалық ақпаратқа және машиналық аударма әдістеріне әртүрлі машиналық оқыту әдістерін белсенді пайдалануға дейін жүргізді. Ал мәтіндегі қателерді анықтау және түзету жүйесін құру бірқатар іргелі және шешімін таппаған мәселелерге тап болады: сөздіктерді жинақы сақтау, морфологиялық және синтаксистік талдаудың тиімді әдістері, ғылыми-техникалық мәтіндерді өңдейтін ғылыми редакторлар жүйесі, яғни әдеби және ғылыми жұмыстарды жүргізетін тұлға [8]. Ағылшын тіліндегі мәтінді түзету жүйелеріне «Grammarly», «Grammarchecker», «ReversoSpeller» және т.б. кіреді. Орысша мәтінді өңдеу жүйелеріне

мыналар жатады: «Орфограмма», «Адвего», «ORFO», «ЛИНАР» және т.б. Агглютинативті тілдер үшін (мысалы, түрік, қырғыз және т.б.) қателерді тексеру жүйелері бар, мысалы, MS Word бағдарламасында орнатылған орфографияны тексеру құралы. Бірақ бұл жүйелер қазақ тіліне жарамайды. Өкінішке орай, қазақ тіліне қатысты жүйелердің аналогтары (жалпыға қолжетімді) жоқ.

Зерттеу және талдау барысында әртүрлі стильдегі мәтіндер, интернет пен әлеуметтік желілердегі мәтіндер, хабарламалар қарастырылды. Сонымен қатар, ағылшын және орыс тілдеріндегі мәтінді тексеру және түзету жүйелеріне талдау жасалды. Жүйелердің артықшылықтары мен кемшіліктері 1-кестеде көрсетілген.

Танымал мәтінді түзету жүйелеріне талдау жүргізу барысында мәтінді түзету жүйелерінің кемшілігі анықталды, ол жүйелерді қазақ тіліне қолдану мүмкін еместігі, себебі агглютинативті тіл болғандықтан морфологиялық және лексикалық түрі күрделі болады [1].

Кесте 1 – Мәтінді тексеру және түзету жүйелерінің салыстырмалы сипаттамалары (ағылшын және орыс тілдеріне арналған жүйелер)

Мәтінді тексеру және түзету жүйелері	Кемшіліктер	Артықшылықтар	Бағамы
Advego	тыныс белгілерін тексермейді.	емлені тексеру; қателерді, жетіспейтін немесе артық әріптерді, бос орындарды анықтау; кеңейтілген SEO параметрлері (тоқтату сөздері, таңбалар саны, сөздер); үлкен көлемдер (бір уақытта 100 000 дейін); 20 тілдегі мақалаларды тексеру.	тегін
LanguageTool	тыныс белгілерін тексермейді.	грамматикалық қателерді тексеру; стильдік қателерді табу; тыныс белгілерін тексеру; мәтіндік редакторлармен біріктіру; танымал браузерлерде орнату; 30 тілге дейін қолдайды; дұрыс нұсқаларды ұсынады.	тегін/ақылы
Istio	тыныс белгілерін тексермейді.	емлені тексеру; веб-парақты тексеру; шектеусіз мәтінді тексеру; SEO мәтінін талдау; ауыстыру опциялары.	Тегін
Orfogramka –https://orfogrammka.ru/	тек ақылы нұсқасы.	орфографияны, пунктуацияны, стильді, типографияны, семантиканы тексеру; тавтологиялардың болуы; какофонияны тексеру; қателерді түзету; SEO мәтінін талдау; ережелері бар ұсыныстар; үнемі жаңартылатын сөздік.	Ақылы
Text.ru	тіркелмеген пайдаланушылар үшін растауды күтудің ұзақ уақыты; тегін нұсқада мақаланың шегі 15 000 таңбаға дейін; сөздердің толық емес сөздігі.	орфография мен грамматиканы тексеру; мәтіннің бірегейлігін талдау; регистрді, жақшаларды, бос орындарды, қайталауларды дұрыс пайдаланбауын анықтау; SEO параметрлері (спам, таңбалар санын санау); Қатені ауыстыру нұсқалары ұсынылады.	тегін/ақылы

1-кестенің жалғасы

Орфограф	түзету нұсқаларының болмауы; минималды параметрлер; тыныс белгілерін тексермейді; шектеулі сөздік.	орфографияны тексеру; веб-беттегі мәтіндегі қателерді тексеру; маркер дизайнын теңшеу.	Тегін
ORFO – https://online.orfo.ru/	жеке сөздерді қолдануда жақсы нәтиже көрсетеді, үлкен мәтіндерде кейбір қателерді таба алмайды, тыныс белгілерін тексермейді.	орфографияны тексеру; 50 қолжетімді тіл; бірегейлікті және басқа SEO көрсеткіштерін талдау; жеке веб-парақты және тұтастай сайтты тексеру мүмкіндігі;	тегін/ақылы
Ескертпе: Авторлармен [9] дереккөз негізінде құрастырылған.			

Мәтіндегі қателерді анықтап, оларды қазақ тілінде түзету жүйесін жасау үшін тілдің ерекшелігіне тоқталған жөн. Қазақ тілі морфологиялық және синтаксистік заңдылықтар қатысатын агглютинативті тіл тобы және сөйлем құрылысына қарай семантикасы бар тіл.

Осы жұмыстың логикалық нәтижесі ретінде қазақ тілінің электрондық сөздігі мен өндірістік қолданыс деңгейіне жеткен қазақ тіліндегі мәтіндердің дұрыстығын тексеру жүйесі пайда болды. Дегенмен, қазақ тіліндегі жартылай құрылымды мәтіндердің дәлдігін тексеруге арналған жүйелер жалпыға қолжетімді емес, тіпті коммерциялық бағдарламалық қамтамасыз етуді Интернеттен табу қиын [10].

Жалпы орфографиялық қателер екі классқа бөлінеді: типографиялық және танымдық. Танымдық қателер (сөздіктегі жоқ қателер) сөздердің фонетикалық немесе орфографиялық ұқсастығы; адам сөзді қалай жазуды білмейді. Типографиялық қателер (нақты сөз қателері) пернетақтадағы екі әріптік пернелердің жақындығына байланысты орфографиялық қателері орын алатын пернетақта мен қол/саусақ қозғалысына қатысты. Қателер тек осы түрлерімен шектелмейді. Бүгінгі таңда жартылай құрылымдалған мәліметтерде қателердің бірнеше түрі анықталды [10, 11].

Зерттеу барысында жартылай құрылымдалған мәліметтерден алынған қазақ тіліндегі сөздерден қате сөздер анықталған кезде келесі жағдайлар анықталды, яғни сөздердегі қателіктердің келесі түрлері көрсетіледі:

- типографиялық қателер (кітап (kitap) – кіап (kiap);
- орфографиялық қателер (мұхит (muhit) – мухит (mýhit);
- сөздердің әдейі бұрмалануы (алғааа (alǵaaa), тағда (taǵda);
- грамматикалық қателер;
- тыныс белгілерінің қателері;
- сөздерді орыс алфавитімен жазу;
- сөздерді латын әліпбиімен жазу;
- қысқартулар және т.б.

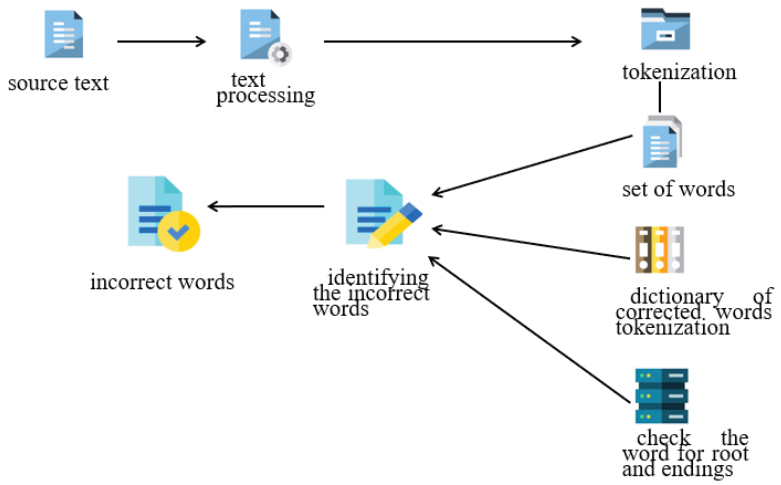
Материалдар мен әдістер

Орфографиялық қатені анықтау сөздегі қатені анықтаумен байланысты. Сөйлемдегі немесе сөздегі қатені анықтаудың бірінші әдісі – орфографияны сөздікпен тексеру. Орфографияны тексеру әдістеріне сөздерді салыстыратын және оларды тілдік сөздікке орналастыратын сөзді іздеу әдістері жатады. Сөздікте сөздің анықталмауы сөзде орфографиялық қате бар екенін көрсетеді. Орфографияны анықтау үшін қолданылатын әдістерге n-грамм алгоритмдері, морфологиялық талдау және машиналық оқыту алгоритмдері жатады. Орфографияны тексеру үшін гибридті әдістер жиі қолданылады [12].

Сөздік – бұл қазақ тіліндегі барлық сөздерді қамтитын файл. Сөздер алфавиттік ретпен орналасады, әр сөз жаңа жолда орналасады. Сөздерді сөздікпен тексеру – мәтіндегі қателерді анықтаудың ең танымал әдісі. Тексеру сөздіктегі сөзді жүйелі түрде іздеу арқылы жүзеге асырылады. Егер сөздің барлық әріптері сөздіктегі сөзге сәйкес келсе, онда бұл дұрыс сөз. Егер мұндай сөз болмаса, онда ол дұрыс емес немесе қатесі бар сөз [13]. Ережелер жиынтығын қолданатын сөздік барлық сөздердің тіл ережелерін қолдана отырып дұрыс жазылуын тексереді.

Екінші әдіс – сөздіктің көмегінсіз орфография тексеру, оған сөйлемнің басындағы бас әріпті тексеру, сөздердің қайталануын тексеру және n-граммен тексеру кіреді. Бас әріппен жазу, яғни нүктеден кейінгі әрбір әріп автоматты түрде бас әріпке айналуы керек. Бір сөздің барлық әріптерінің басқа сөздің әріптерімен сәйкестігі тексеріледі, бірақ егер олар толық сәйкес келсе, бұл қате болып саналады. n-граммдық талдау мәтіндік жиынға қате жазылған сөздерді табу әдісі ретінде тұжырымдалған. Мәтіндегі әрбір сөзді сөздікпен салыстырудың орнына тек n-грамм тексеріледі. Егер жоқ немесе сирек кездесетін n-грамм табылса, сөз қате жазылған деп белгіленеді, әйтпесе ол дұрыс. Бұл әдіс тілге тәуелді емес, өйткені ол қолданылатын тілді білуді қажет етпейді [14, 15].

Сонымен қатар, тілдегі қате сөздерді анықтайтын модельдер бойынша зерттеулер жүргізілді. Нәтижесінде қазақ тіліндегі тура емес сөздерді анықтау тәсілі жасалды. 1-суретте қате сөздердің жақындауы көрсетілген.



Сурет 1 – Қате сөздерді анықтау тәсілінің схемасы

Ескертпе: Авторлармен [13] дереккөз негізінде құрастырылған.

Көрсетілген тәсілде, алдымен мәтін жартылай құрылымдалған деректерден жиналады, содан кейін мәтін алдын ала өңделеді, мәтінді сөйлемдерге, сөйлемдерді сөздерге бөледі, сол сөздерден сәйкесінше сөздер жиынтығы жасалады.

Экономикалық талдау.

Қазіргі ақпараттық технологиялардың дамуында қателерді автоматты түрде анықтау жүйелері, қазақ тілі сияқты ресурсы шектеулі тілдер контексінде шешуші рөл атқарады. Мұндай жүйелерді сәтті әзірлеу және енгізу үшін тек техникалық аспектілерді ескеріп қана қоймай, сонымен бірге мұқият экономикалық талдау жүргізу қажет. Себебі, тілдік технологияларды жасауға бағытталған жобалар ғылыми-зерттеу, тәжірибелік-конструкторлық, сынақтан өткізу және енгізу үшін қомақты шығындарды талап етеді. Бұл мақалада машиналық оқыту әдістерін және жартылай құрылымдық мәтінді өңдеуді қолдана отырып, қазақ тілінде қателерді анықтау жүйелерін енгізудің экономикалық құрамдас бөлігі талданады.

1. Ғылыми-зерттеу және тәжірибелік-конструкторлық шығындар.

Қазақ тіліндегі қателерді автоматты түрде тексеру және түзету жүйесін зерттеу және әзірлеу бірнеше негізгі кезеңдерді қамтиды:

- Қазақ тіліндегі мәтіндер корпусының жинағы мен аңдатпасы;

◆ Морфологияны, синтаксисті және семантиканы талдау алгоритмдерін жасау және тестілеу;

◆ Жіктеу және қателерді анықтау үшін машиналық оқыту үлгілерін орнату.

Бұл кезеңдердің әрқайсысы маңызды ресурстарды, адам мен қаржылық ресурстарды талап етеді. Негізгі шығындарға мыналар кіреді:

A. Мамандардың жалақысы: Мұндай жүйелерді әзірлеу жоғары білікті мамандарды, соның ішінде лингвистерді, бағдарламашыларды және машиналық оқыту мамандарын тартуды талап етеді. Орташа алғанда, машинаны оқытудағы жалақы біліктілік деңгейі мен аймаққа байланысты айына $5,000-дан $15,000-ға дейін болуы мүмкін.

B. Инфрақұрылымдық шығындар: Машиналық оқыту үлгілерін әзірлеу және оқыту серверлерді, бұлттық қызметтерді және басқа АТ инфрақұрылымын қоса алғанда, жоғары өнімді есептеу ресурстарын талап етеді. Оқыту үлгілеріне арналған серверлерді немесе бұлттық шешімдерді жалға алу құны айына 1000 АҚШ доллардан 5000 АҚШ долларға дейін болуы мүмкін.

C. Лицензиялау және бағдарламалық қамтамасыз ету: арнайы мәтінді өңдеу құралдарын және бағдарламалық шешімдерді, соның ішінде машиналық оқытуды және лингвистикалық талдау пакеттерін пайдалану лицензиялау үшін айтарлықтай шығындарды қажет етуі мүмкін. Мысалы, белгілі бір кітапханалар үшін лицензиялық алымдар бір пайдаланушы үшін 1000 АҚШ долларға дейін болуы мүмкін.

D. Деректерді жинау және анықтау: Жүйелерді сәтті дамыту үшін қазақ тіліндегі мәтіндердің үлкен корпусы болуы қажет. Дегенмен, мұндай деректерді жинау және анықтау күрделі және қымбат процесс болуы мүмкін, әсіресе қолжетімді ресурстар шектеулі болған кезде. Мұндай корпусты жасау құны деректер көлеміне және аннотация мәліметтерінің деңгейіне байланысты 20 000 АҚШ долларынан 50 000 АҚШ долларына дейін болуы мүмкін.

2. Іске асыру шығындары.

Даму кезеңі аяқталғаннан кейін жүйені бизнес, мемлекеттік органдар және оқу орындары сияқты әртүрлі салаларда енгізу процесі жүреді. Бұл процесс сонымен қатар айтарлықтай шығындармен бірге жүреді:

A. Техникалық интеграция: жүйелерді бар АТ инфрақұрылымдарына енгізу уақыт пен күш-жігерді қажет етеді. Мысалы, қателерді анықтау жүйесін мазмұнды басқару жүйелеріне немесе дерекқорларға біріктіру тәжірибелі әзірлеушілер мен инженерлердің қатысуын талап етеді. Мұндай интеграцияның орташа құны 10 000 доллардан 30 000 долларға дейін болуы мүмкін.

B. Қызметкерлерді оқыту: Жүйелерді тиімді пайдалану үшін олармен жұмыс істейтін қызметкерлерді оқыту қажет. Бұған жүйеге қолдау көрсететін техникалық қызметкерлер де, оны пайдаланудан пайда көретін соңғы пайдаланушылар да кіреді. Оқыту құны жүйенің күрделілігіне және пайдаланушылар санына байланысты 5 000 доллардан 15 000 долларға дейін болуы мүмкін.

C. Техникалық қолдау және жаңартулар: Жүйеге енгізілгеннен кейін оның жұмысын жақсарту және жаңа деректерді орналастыру үшін жүйелі техникалық қызмет көрсету және жаңартулар қажет. Жылдық техникалық қызмет көрсету шығындары жүйенің күрделілігіне және қолдау көрсетілетін деректер көлеміне байланысты 2 000 доллардан 10 000 долларға дейін болуы мүмкін.

3. Экономикалық пайда.

Қателерді анықтау жүйелерін әзірлеуге және енгізуге бастапқы шығындар айтарлықтай болуы мүмкін болса да, оларды пайдалану болашақта айтарлықтай экономикалық пайда әкеледі:

A. Мәтінді қолмен тексеру құнын төмендету: Автоматтандырылған жүйелер, әсіресе үлкен көлемдегі мәтіндерді қолмен тексеруге және түзетуге жұмсалатын уақыт пен күш-жігерді айтарлықтай қысқартуы мүмкін. Мысалы, күн сайын мыңдаған мәтіндерді (құжаттар, есептер, т.б.) өңдейтін компанияларда процестерді автоматтандыру деректерді өңдеуге кететін шығындардың 50% дейін үнемдеуге мүмкіндік береді.

B. Жақсартылған деректер сапасы: жоғары сапалы жазу бизнес нәтижелеріне тікелей әсер етеді. Мысалы, маркетингтік немесе заңды құжаттарда мәтіндердің дәлдігі компанияның кірісіне немесе заңды жауапкершілігіне тікелей әсер етуі мүмкін. Мәтінді тексеру процесін

автоматтандыру қателер ықтималдығын азайтады, бұл өз кезегінде бизнестің тиімділігін арттырады.

С. Уақытты оңтайландыру: Автоматтандырылған жүйелерді енгізу мәтінді өңдеуге байланысты тапсырмаларды орындау үшін қажетті уақытты айтарлықтай қысқартуы мүмкін. Бұл әсіресе жоғары бәсекелестік жағдайында жұмыс істейтін компаниялар үшін және нарық сұранысына тез жауап беру қажеттілігі үшін маңызды.

Математикалық шығын моделі.

Әзірлеу және енгізу шығындарын бағалау үшін келесі математикалық модельді қолдануға болады:

$$C_{total} = C_{research} + C_{implementation} + C_{suppor}t - B_{efficiency} - B_{quality} - B_{time}$$

мұнда:

C_{total} – жүйені әзірлеуге және енгізуге кеткен жалпы шығындар;

$C_{research}$ – ғылыми-зерттеу және тәжірибелік-конструкторлық шығындар (соның ішінде мамандардың жалақысы, инфрақұрылымдық шығындар және т.б.);

$C_{implementation}$ – жүйені енгізу шығындары (интеграция, персоналды оқыту және т.б.);

$C_{support}$ – жүйені іске асырғаннан кейін оны қолдауға және қызмет көрсетуге арналған шығындар;

$B_{efficiency}$ – мәтінді тексеру процесінің тиімділігін арттыру арқылы шығындарды үнемдеу;

$B_{quality}$ – деректер сапасын жақсарту есебінен үнемдеу;

B_{time} – процесті автоматтандыру арқылы уақытты үнемдеу.

Нәтижелер мен талдау

Тәжірибе үшін [10]–да сипатталған және қол жетімді бағдарлама қолданылды, сонымен қатар қазақ тілінің тілдік ресурстары, атап айтқанда тоқтау сөздерінің сөздігі, жалғаулардың толық жиынтығы қолданылды [14]. Бағдарлама бірнеше рет сынақтан өтті. 3-кестеде тәжірибе нәтижелері көрсетілген.

Кесте 2 – Алгоритмнің тәжірибелік нәтижелері

Тексерілген сөздер саны	Дәлдігі, %
190	89
687	92
1299	90
3438	93
998	94
1431	96
Ескертпе: Авторлармен құрастырылған.	

Бағдарламамен жұмыс істегеннен кейін тәжірибе нәтижелері талданды. Оқу алгоритмін қолдана отырып, сөздердің түбірлері мен жалғаулары алынды. Кіріс деректегі сөздерді белгілеу процесінен кейін нәтижелерді талдау кезінде келесі қателер орын алды:

- жалғаулар (аффикс) дұрыс анықталмады;
- алгоритмі дұрыс орындалмады, себебі кейбір жалғаулар аффикс деректер қорында табылмады.

Қазақ тіліндегі қателерді анықтау жүйелерін қолданудың алғашқы практикалық нәтижелері мәтінді тексеруді автоматтандыру орта есеппен құжаттарды өңдеу уақытын 40–60%-ға қысқартатынын көрсетеді. Коммерциялық компанияларда бұл мәтінді тексеруге қатысатын персонал шығындарының 30–50%-ға төмендеуіне әкелді. Мемлекеттік секторда мұндай жүйелерді пайдалану құжаттар мен есептердің дәлдігін арттырды, түзетулер санын және ақпаратты қайта өңдеуді азайтты.

Болашақта әзірлеу және енгізу шығындары мен экономикалық пайда арасындағы оңтайлы теңгерімге қол жеткізу үшін келесі әдістерді қолдануға болады:

1. Әзірлеуге модульдік тәсіл: негізгі функционалдық модульдерден бастап жүйені кезеңкезеңімен әзірлеу бастапқы шығындарды азайтуға және болашақта жүйені масштабтауға икемділікті қамтамасыз етуге мүмкіндік береді;

2. Бұлтты технологияларды пайдалану: бұлттық шешімдер инфрақұрылымдық шығындарды азайтуы мүмкін, өйткені бұлттық қызметтерді жалға алу әдетте жеке жабдықты сатып алу мен жөндеуден гөрі арзанырақ;

3. Инвестициялардың кірістілігін (ROI) талдауы: егжей-тегжейлі ROI талдауын жүргізу жүйеге таза пайда әкеле бастау үшін қанша уақыт қажет болатынын бағалауға көмектеседі. Бұл ресурстарды ұтымды бөлуге және іске асыру туралы негізделген шешімдер қабылдауға мүмкіндік береді.

Қорытынды

Жүргізілген зерттеулер қазақ тіліндегі қателерді анықтаудың автоматтандырылған жүйесін енгізу айтарлықтай экономикалық тиімділікке қол жеткізуге болатынын растайды. Жартылай құрылымдық мәтіндерді өңдеуде машиналық оқыту әдістерін қолдану тілді тексеру сапасын жақсартып қана қоймайды, сонымен қатар ақпаратты түзету және тексеру шығындарын азайтады. Практикалық нәтижелер бұл технологияларды компаниялар мен мемлекеттік органдарда қолдану мәтіндік деректерді өңдеуге қажетті уақыт пен қаржылық ресурстарды қысқартуға болатынын көрсетеді. Қазақ тіліндегі қателерді автоматты түрде анықтау жүйелерін енгізудің экономикалық талдауы айтарлықтай бастапқы инвестиция арқылы жүйе бизнес пен мемлекеттік органдар үшін тиімді құрал бола алатынын көрсетеді. Мәтінді қарау процестерін оңтайландыру, қателерді түзету шығындарын азайту және ақпарат сапасын жақсарту осындай шешімдерді әзірлеу және енгізу шығындарын негіздейтін ұзақ мерзімді экономикалық пайда әкеледі. Бұл саладағы болашақ әзірлемелер мазмұнды басқару жүйелерімен тереңірек интеграцияны және қазақ тілінің әртүрлі контексттері мен диалектілік ерекшеліктерін есепке алу үшін жетілдірілген үлгілерді қамтуы мүмкін.

Қаржыландыру туралы ақпарат. Бұл зерттеу Қазақстан Республикасы ғылым және жоғары білім министрлігінің AP23487753 «Қазақ тіліндегі мәтіндерді автоматтандырылған түзетуге арналған инновациялық технологиялар: машиналық оқыту және морфологиялық талдау» жобасының грантымен орындалды және қаржыландырылды.

ӘДЕБИЕТТЕР

1 Рахимова Д.Р. Компьютерная обработка казахского языка: сборник научных трудов (материалов) // Қазақ университеті. – Алматы, 2020. – 146 с.

2 Han B., Baldwin T. Lexical normalisation of short text messages: Makn sens a# twitter // 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011. Vol. 1. P. 368–378.

3 Farra N. et al. Generalized Character-Level Spelling Error Correction. Association for Computational Linguistics. 2014. No. 2. P. 161–167.

4 Hladek D. et al. Survey of Automatic Spelling Correction // Electronics. 2020. Vol. 9. No. 10. P. 1–29.

5 Peter B. Semistructured data // Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. 1997. P. 117–121.

6 Brill E., Moore R.C. An improved error model for noisy channel spelling correction // Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. 2000. P. 1–10.

7 Farag A., Ernesto W., Andreas N. Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. 2009. No. 40. P. 117–121.

8 Kaufmann M., Kalita J. Syntactic normalization of twitter messages // International conference on natural language processing. Kharagpur, India. 2010. P. 266

9 Лучшие программы для исправления ошибок в тексте. URL: https://www.rush-analytics.ru/blog/programmy-dlya-ispravleniya-v-tekste-razbor-primerov-i-osnovnye-osobennosti (accessed: 22.08.2024)

10 Shaalan K., Aref R., Fahmy A. An approach for analyzing and correcting spelling errors for non-native Arabic learners // Computer Science. The 7th International Conference on Informatics and Systems. 2010. P. 53–59.

11 Такташкин Д.В., Мокроусова Е.А. Методы и алгоритмы проверки орфографии тестовых документов // Электронный научно-практический журнал «Современные научные исследования и инновации». 2017. № 5. URL: https://web.snauka.ru/issues/2017/05/72892 (дата обращения: 12.08.2023)

12 Rakesh K., Minu B. and Kumar S. A study of spell checking techniques for indian languages // JK Research Journal in Mathematics and Computer Sciences. 2018. Vol. 1. No. 1. P. 105–111.

13 Tukeyev U., Turganbaeva A. Lexicon-free stemming for the Kazakh language. Materials of the International Scientific Conference «Computer science and Applied Mathematics» dedicated to the 25th anniversary of the Independence of the Republic of Kazakhstan and the 25th anniversary of the Institute of Information and Computational Technologies. Part II. Almaty. 2016. September 21–24. P. 84–88.

14 Tukeyev U., Turganbaeva A., Karibayeva A., Amirova D., Abduali B. Language_Resources_for_Kazakh_language. URL: https://github.com/NLPKazNU / Language_Resources_for_Kazakh_language.2020. (accessed: 12.08.2024)

15 Recent advances in Apertium, a free/open-source rul-based machine translation platform for low-esource languages. 2021. URL: https://turkic.apertium.org/index.kaz.html?choice=kaz&qA=%D0%9C%D0%B5%D0%BD%D1%96%D2%A3%20%D0%BE%D2%9B%D1%83%D1%88%D1%8B%D0%BC%20#analyzation (accessed: 29.07.2024)

REFERENCES

1 Rahimova D.R. (2020) Komp'juternaja obrabotka kazahskogo jazyka: sbornik nauchnyh trudov (materialov) // Қазақ universiteti. Almaty, 146 p. (In Russian).

2 Han B., Baldwin T. (2011) Lexical normalisation of short text messages: Makn sens a# twitter // 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. V. 1. P. 368–378. (In English).

3 Farra N. et al. (2014) Generalized Character-Level Spelling Error Correction. Association for Computational Linguistics. No. 2. P. 161–167. (In English).

4 Hladek D. et al. (2020) Survey of Automatic Spelling Correction // Electronics. V. 9. No. 10. P. 1–29. (In English).

5 Peter B. (1997) Semistructured data // Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. P. 117–121. (In English).

6 Brill E., Moore R.C. (2000) An improved error model for noisy channel spelling correction // Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. P. 1–10. (In English).

7 Farag A., Ernesto W., Andreas N. (2009) Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. No. 40. P. 117–121. (In English).

8 Kaufmann M., Kalita J. (2010) Syntactic normalization of twitter messages // International conference on natural language processing. Kharagpur, India. P. 266. (In English).

9 Luchshie programmy dlja ispravlenija oshibok v tekste. URL: https://www.rush-analytics.ru/blog/programmy-dlya-ispravleniya-v-tekste-razbor-primerov-i-osnovnye-osobennosti (accessed: 22.08.2024). (In Russian).

10 Shaalan K., Aref R., Fahmy A. (2010) An approach for analyzing and correcting spelling errors for non-native Arabic learners // Computer Science. The 7th International Conference on Informatics and Systems. P. 53–59. (In English).

11 Taktashkin D.V., Mokrousova E.A. (2017) Metody i algoritmy proverki orfografii testovyh dokumentov // Jelektronnyj nauchno-prakticheskij zhurnal «Sovremennye nauchnye issledovanija i innovacii». No. 5. URL: https://web.snauka.ru/issues/2017/05/72892 (data obrashhenija: 12.08.2023). (In Russian).

12 Rakesh K., Minu B. and Kumar S. (2018) A study of spell checking techniques for indian languages // JK Research Journal in Mathematics and Computer Sciences. V. 1. No. 1. P. 105–111. (In English).

13 Tukeyev U., Turganbaeva A. (2016) Lexicon-free stemming for the Kazakh language. Materials of the International Scientific Conference «Computer science and Applied Mathematics» dedicated to the 25th anniversary of the Independence of the Republic of Kazakhstan and the 25th anniversary of the Institute of Information and Computational Technologies. Part II. Almaty. September 21–24. P. 84–88. (In English).

14 Tukeyev U., Turganbaeva A., Karibayeva A., Amirova D., Abduali B. (2020) Language_Resources_for_Kazakh_language. URL: https://github.com/NLPKazNU / Language_Resources_for_Kazakh_language. (accessed: 12.08.2024). (In English).

15 Recent advances in Apertium, a free/open-source rul-based machine translation platform for low-esource languages. 2021. URL: https://turkic.apertium.org/index.kaz.html?choice=kaz&qA=%D0%9C%D0%B5%D0%BD%D1%96%D2%A3%20%D0%BE%D2%9B%D1%83%D1%88%D1%8B%D0%BC%20#analyzation (accessed: 29.07.2024). (In English).

БАЙТЕНОВА Л.М.,*[1]
д.э.н., профессор.
*e-mail: l.baitenova@turan-edu.kz
ORCID ID: 0000-0002-1591-2235
РАХИМОВА Д.Р.,[1,2]
PhD, ассоциированный профессор.
e-mail: di.diva@mail.ru
ORCID ID: 0000-0003-1427-198X
ТУРАРБЕК Ә.Т.,[1,2]
PhD, доцент.
e-mail: turarbek_asem@mail.ru
ORCID ID: 0000-0002-4793-0446
АДАЛИ Е.,[3]
PhD, профессор.
e-mail: adali@itu.edu.tr
ORCID ID: 0000-0002-1561-8255
[1]Университет «Туран»,
г. Алматы, Казахстан
[2]Казахский национальный
университет им. аль-Фараби,
г. Алматы, Казахстан
[3]Стамбульский технический университет,
г. Стамбул, Турция

ЭКОНОМИЧЕСКИЕ АСПЕКТЫ ИДЕНТИФИКАЦИИ ОШИБОК В ПОЛУСТРУКТУРИРОВАННЫХ ПУБЛИКАЦИЯХ НА ГОСУДАРСТВЕННОМ ЯЗЫКЕ

Аннотация

В связи с быстрым ростом информации в Интернете и социальных сетях в данное время исследования в области компьютерной лингвистики стали весьма актуальны. Объем информации, которую создают люди и машины на естественном языке, нуждается в обработке, анализе и проверке. Для этого используются информационно-поисковые системы, диалоговые системы, средства машинного перевода. Сам спектр систем автоматической обработки текстов весьма широк, он охватывает различные задачи. Поиск ошибок в текстах и словах, выявление и исправление некорректных слов являются одной из важнейших задач обработки естественного языка (NLP). В статье дается обзор полуструктурированных данных, методов и технологий выявления некорректных слов на естественных языках. Цель исследования – разработка эффективного подхода для обнаружения и исправления ошибок, возникающих в казахскоязычных текстах, особенно в условиях ограниченных ресурсов и неструктурированных данных. Исследование включает использование методов машинного обучения, а также экономический анализ затрат на разработку и внедрение таких решений. Предлагаемый подход способствует автоматизации проверки текстов, что может значительно сократить затраты на ручную обработку данных и повысить качество информации в различных сферах, включая бизнес и государственное управление.

Ключевые слова: экономические аспекты, государственный язык, казахский язык, полуструктурированные данные, текст, ошибка, неправильные слова, социальная сеть.

BAITENOVA L.M.,*[1]
d.e.s., professor.
*e-mail: l.baitenova@turan-edu.kz
ORCID ID: 0000-0002-1591-2235
RAKHIMOVA D.R.,[1,2]
PhD, associate professor.
e-mail: di.diva@mail.ru
ORCID ID: 0000-0003-1427-198X
TURARBEK A.T.,[1,2]
PhD, associate professor.
e-mail:turarbek_asem@mail.ru
ORCID ID: 0000-0002-4793-0446
ADALI E.,[3]
PhD, professor.
e-mail: adali@itu.edu.tr
ORCID ID: 0000-0002-1561-8255
[1]Turan University,
Almaty, Kazakhstan
[2]Al Farabi Kazakh National University,
Almaty, Kazakhstan
[3]Istanbul Technical University
Istanbul, Turkey

ECONOMIC ASPECTS OF ERROR IDENTIFICATION IN SEMI-STRUCTURED PUBLICATIONS IN THE STATE LANGUAGE

Abstract

Due to the rapid growth of information on the Internet and social networks, research in the field of computational linguistics has become very relevant. The volume of information that people and machines create in natural language needs to be processed, analyzed and verified. Information retrieval systems, dialog systems, and machine translation tools are used for this. The range of automatic text processing systems is very wide, it covers various tasks. Finding errors in texts and words, identifying and correcting incorrect words is one of the most important tasks of natural language processing (NLP). The article provides an overview of semi-structured data, methods and technologies for identifying incorrect words in natural languages. The paper gives an overview of semi-structured data, methods and techniques for detecting incorrect words in natural languages. The aim of the research is to develop an effective approach for detecting and correcting errors occurring in Kazakh-language texts, especially in the context of limited resources and unstructured data. The research includes the use of machine learning techniques as well as economic analysis of the costs of developing and implementing such solutions. The proposed approach facilitates the automation of text verification, which can significantly reduce the cost of manual data processing and improve the quality of information in various spheres, including business and public administration.

Key words: economic aspects, state language, kazakh language, semi-structured data, text, error, wrong words, social network.

Мақаланың редакцияға түскен күні: 09.09.2024