АПИБК-2020

МАТЕРИАЛЫ

МЕЖДУНАРОДНОЙ НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ

АКТУАЛЬНЫЕ ПРОБЛЕМЫ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ В КАЗАХСТАНЕ

Алматы

15 января, 2020 года

Мақалада келтірілген АЖ жобалау концепциясы ауқымды ақпарат алмасу платформасының негізі болып табылады. Оқыту процессін автоматтандыру қазіргі күні өте маңызды сұрақтардың біріне айналды, алайда сол оқытылатын ақпарат көзі болатын ғылыми зерттеу процестері мен өзара байланысының автоматтандырылуы анық проблема түрінде көрсетілмеген. Тек ұқсас проблеманы мысалға келтіред. ЖОО білім алушылардың мамандар ретінде деңгейлері төмендеуде. Көріп отырғанымыздай ЖОО, ғылыми зерттеу жүргізетін мекемелер мен кәсіпорындар арасындағы өзара мағлұмат алмасу платформасы жоқ. Сол платформаны құрастыру барысында ең алғаш вуз ішіндегі АЖ ны ғылыми жұмыстардың белсенділігін бақылауға мүмкіндік беретіндей етіп құрастырып, бірнеше университет арасында алмасу процессін ұйымдастыру қажет.

Әдебиеттер

1. Гвоздева, В. А. Основы построения автоматизированных информационных систем/ В.А.Гвоздева, И.Ю. Лаврентьева. - М.: Форум, Инфра-М, 2016. - 320 с.

2. Ипатова, Э. Р. Методологии и технологии системного проектирования информационных систем / Э.Р. Ипатова, Ю.В. Ипатов. - М.: Флинта, 2013**.** - 256 с.

ӘЛЕУМЕТТІК ЖЕЛІДЕГІ ЭКСТРЕМИСТІК МӘТІНДЕРДІ ЖІКТЕУ ДӘЛДІГІН ГРАММАТИКАЛЫҚ ҚАТЕЛЕРДІ АНЫҚТАУ ЖӘНЕ ТҮЗЕТУ АРҚЫЛЫ АРТТЫРУ

Мусиралиева Ш.Ж., Болатбек М.А.
e-mail: mussiraliyevash@gmail.com, bolatbek.milana@gmail.com
әл-Фараби атындағы Қазақ ұлттық университеті,
Қазақстан

Аннотация. *Бұл жұмыста авторлар экстремистік бағыттағы мәтіндерді анықтау мақсатында әлеуметтік желі мәтіндеріне семантикалық талдау жүргізу барысында туындайтын қиындықтардың бірі грамматикалық қателерді автоматты түрде анықтау және түзету мәселесіне тоқталады. Қателерді анықтау және түзету жүйелерінің ағылшын тіліне арналған бірнеше бағдарламалары мен деректер қоры бар. Соңғы уақытта орыс, неміс тілдері үшін де қателерді анықтау жүйелері құрылуда. Берілген жұмыста ағылшын, орыс, неміс, чех тілдеріне арналған грамматикалық қателерді түзету жүйелеріне шолу жасалады. Сондай-ақ, авторлар қазақ тілі үшін аталған сипаттағы жүйелердің болмауына байланысты берілген мәселенің өзекті екендігін көрсетеді.*

Қазіргі таңда халықаралық ақпараттық-коммуникациялық Интернет желісі экстремистік материалдарды таратуда белсенді қолданылады. Бұл жаһандық саяси процестің негізгі қатысушыларының бірі ретінде Қазақстан Республикасы үшін өте маңызды болып табылады. Экстремистік ұйымдар криминалдық қызметте Интернеттің шексіз мүмкіндіктерін белсенді пайдаланады, оның ішінде: қылмыс жасауға дайындық және оны жасау кезінде интернет-ресурстарды пайдалану; қоғамдық қауіпті қызметті басқару және үйлестіру мақсатында жасырын ақпарат алмасу; арнайы құрылған

сайттарда және басқа да интернет-ресурстарда белсенді насихаттау бойынша жоспарланған ақпараттық операцияларды жүзеге асыру; жаңа қатысушыларды аталған ресурстармен заңсыз әрекеттерге тарту және т.б. Нәтижесінде, соңғы жылдары экстремизм проблемасының өршуі байқалуда, оны Қазақстанның ұлттық қауіпсіздігіне қауіп төндіретін мәселе ретінде қарастыруға болады.

Ғаламтор алпауыттары Google, Facebook және Twitter лаңкестік мазмұнды табу және жою үшін жасанды интеллект технологиясын қолдануда. IBM-де әлеуметтік желілердегі барлық деректерді талдайтын Watson бағдарламасы бар. Ресейде Платонның ақпараттық серіктес авторы әлеуметтік желілерді бақылау және қауіптерді болжау жүйесін құрастырды. Германия үкіметі террористік шабуылдардан кейін Интернетте террористермен күресу үшін ZITiS деп аталатын жаңа киберқауіпсіздік бөлімшесін құру туралы жариялады. Қазақстанда мұндай жүйе жоқ. Сол себепті экстремизм көріністерін анықтауға, алдын алуға және жолын кесуге бағытталған тиімді шаралар кешенін іске асыруды қажет ететін бағдарламаларды құру өзекті болып табылады.

Бұл жұмыста авторлар әлеуметтік желілердегі экстремистік бағыттағы материалдарды анықтау мақсатында семантикалық үлгілерді құру барысында туындаған қиындықтардың бірін сипаттайды. Аталған мәселені шешу үшін авторлар бес кезеңнен тұратын үлгі құрастырды. Үлгінің екінші кезеңі бойынша жинақталған мәліметтер арасында қазақ тіліндегі мәтіндер ішінде орфографиялық қателері бар сөздердің көптігі анықталды. Берілген жұмыста құрастырылған корпус ішіндегі орфографиялық қателіктерді түзету әдістеріне шолу жасалады.

Грамматикалық қателіктерді түзету — мәтіндегі сөздердің қате қолданылуын және дұрыс құрастырылмаған грамматикалық құрылымдарды автоматты түрде анықтау және түзету тапсырмасы. Грамматикалық қателерді анықтау және түзету тапсырмасын көп классты классификация есебі ретінде қарастыруға болады [1].

Грамматикалық қателіктерді түзету мәселесі ең көп қарастырылған тіл ағылшын тілі болып табылады. Ал басқа тілдерге келетін болсақ, қателіктерді анықтау және түзетуге қатысты жасалған жұмыстар саны өте шектеулі болып келеді. Атап айтатын болсақ, [2] неміс, [3] орыс, [4] чех тілі үшін грамматикалық қателіктерді түзету жүйесін құрып, аталған тілдер үшін корпус құрастырған. Сондай-ақ, [5] қытай, [6] жапон, [7] араб тілдері үшін оқытуға арналған аннотациялық корпустарды құруға қатысты зерттеу жұмыстарын жүргізген [8].

Экстремистік бағыттағы мәтіндерді классификациялау үшін ең алдымен машиналық оқыту әдістерін жаттықтыруға арналған корпус қажет. Қазақ тілі үшін экстремистік бағыттағы мәтіндердің ашық корпусы жоқ. Аталған мәселені шешу мақсатында зерттеу жұмысының алғашқы кезеңінде авторлар әлеуметтік желідегі қазақ тіліндегі экстремистік мәтіндерді қамтитын корпус құрастырды. Корпуста тікелей экстремистік бағыттағы мәтіндер және "бейтарап" мәтін ретінде жіктелетін экстремистік іс-әрекеттер жайлы жаңалықтар, экстремистік әрекеттерге қарсы сипаттағы мәтіндер және жалпылама лексиканы қамтитын мәтіндер қамтылған. Құрастырылған корпустағы негізгі кілттік сөздер TF-IDF әдісі бойынша анықталды [9]. Аталған кілттік сөздер бойынша "Вконтакте" әлеуметтік желісіндегі 170 топқа талдау жасалды. Олардың ішіндегі 25 топ парсинг жасау бағдарламасында қолданылды. Құрастырылған парсинг жасау бағдарламасы топтағы соңғы алты ай ішіндегі жазбаларды көшіру арқылы корпусты толықтырып отырады. Басқа тілдегі жазбалар Google Translator арқылы аударылды [10].

Корпусқа талдау жүргізу барысында көптеген қолданушылардың қазақ тіліне тән әріптерді кирилл әріптерімен алмастыратыны жиі байқалды, мысалы "соғысқа" сөзінің орнына "согыска", "шайқасқа" сөзінің орнына "шайкаска" деп жазылады.

Осы тұста классификациялау дәлдігін арттыру мақсатында корпустағы орфографиялық қателіктерді түзету мәселесі туындайды. Қазақ тілі үшін енгізілген сөздің қате немесе дұрыс жазылғандығын анықтайтын жүйелер бар [11]. Алайда аталған жүйелер грамматикалық қателіктерді анықтағанымен, оларды автоматты түрде түзетуді қарастырмайды.

Ғалымдар зерттеу жұмыстарында орфографияны түзетудің әр түрлі әдістерін қарастырады. Алғашқы жүргізілген жұмыстардың көбі қателерді түзету үшін сөздікте жылдам іздеу жүргізіп, тіркестерді алмастыру мәселесіне арналған кандидаттарды тиімді іздеуге арналды, бұл агглютинативті және полисемантикалық тілдер үшін маңызды болып табылады.

Notepad ++ немесе MS Word сияқты мәтіндік редакторлардағы түзету жүйелері қолданушыға қате енгізілген сөз үшін таңдауға арналған бірнеше кандидат ұсынады. Алайда орташа ұзындықтағы сөйлемге арналған нұсқалар саны өте үлкен болып табылады, ал бұл орфографияны дұрыс тексерудің тек түзетулерді шығарып қана қоймай, сонымен қатар берілген мәнмәтіндегі ең жақсы нұсқаны да таңдай білу керектігін білдіреді (мысалы, piece / peace немесе компания / кампания). Мұндай мәселелер дұрыс жазуды мәнмәтіндік түзету саласының зерттеу пәні болып табылады [12].

Орфографияны түзету күрделілігі қолданылу саласы мен бастапқы тілге де байланысты болып табылады. Шынында да, егер морфология жүйесі барынша дұрыс бөлінген болса, онда іздеу мен кандидаттарды таңдауды қиындататын сөздік те соғұрлым үлкен болады. Бұл оқытуға арналған корпустың үлкен болуы керектігін білдіреді [13].

[8] жұмыста авторлар чех тіліне арналған жаңа AKCES-GEC корпусын ұсынады. Сонымен қатар, чех, неміс және орыс тілдері үшін тәжірибе жүргізіп, синтетикалық параллель корпусты қолдану барысында нейрондық машиналық аударма үлгісі аталған мәліметтер жинағы үшін жаңа нәтижелерге қол жеткізуге мүмкіндік беретінін көрсетеді. Авторлар әрбір тіл үшін алдын ала Transformer нейрондық машиналық аударма жүйесін синтетикалық мәліметтерге дайындайды. Құрастырылған жүйенің өнімділігі барлық үш тілде де әр тілдің жеке құрастырылған грамматикалық қателіктерді түзету үлгілерінің өнімділігінен жоғары болып табылады.

[13] жұмыста авторлар Live Journal, ВКонтакте және т.б. сияқты әлеуметтік желілер мен басқа да блогтардағы мәтіндердің орфографиясын түзетуді қарастырады. Мұндай мәтіндердегі қатемен жазылған сөздердің үлесі өте жоғары болып келеді, себебі теру барысында қате кетуі, орфографиялық қателер де кездесуі мүмкін, мұндай қателерді тиімді түрде түзету морфологиялық және синтаксистік талдау сияқты мәтінді әрі қарай өңдеудің қажетті алғышарты болып табылады. Авторлар түзетуге арналған кандидатты таңдау үшін фонетикалық ұқсастықпен қатар қашықтықты өңдеуді қолданған. Аталған кандидаттарды бағалау үшін тілдік үлгі мен қате үлгісін қатар қолданады және оларды қайта қарастыру үшін сызықты классификациялау алгоритмдерін пайдаланады. Содан кейін соңғы кезеңде қате үлгісінің бағасын, кандидат пен түзету арасындағы Левенштейннің өлшенген қашықтығын, тілдік үлгі бағасын және сөздік және сөздіктен тыс сөздердегі түзетулер саны, бас әріптерді пайдалану және т.б. функциялар қолданылады. Аталған жүйе орыс тіліндегі орфографияны тексерудің алғашқы

SpellRuEval сайысына қатысып, барлық көрсеткіштер бойынша жеңіске жеткен және F1-өлшем 75%-ды құраған.

[6] еңбекте авторлар тіл үйренуге арналған "Lang-8" журналына талдау жасау арқылы жапон тілін үйренушілерінің корпусын алу жұмысын жүргізген. Авторлар жапон тілін үйренушілер құрастырған 900 мыңға жуық сөйлем алып, ол жердегі қателерді түзету үшін символдарға негізделген машиналық талдау әдісін қолданған.

[14] жұмыста авторлар оқытуға арналған үлкен көлемдегі аннотациялық мәліметтерді пайдаланбайтын әдістерді көрсетеді. Зерттеу жұмысының нәтижесінде аталған әдістердің морфологиясы бай тілдердегі грамматикалық қателіктерді түзету үшін өте пайдалы екендігін көрсетеді. Сондай-ақ, бірнеше тілге талдау жасап, аталған тілдердің корпустарындағы қате жазылған сөздердің үлесін келтіреді.

Берілген жұмыстың авторлары жоғарыда келтірілген жұмыстағы корпустағы қате сөздер көлеміне қазақ тілді корпус ішіндегі қате сөздердің үлесін қосты. Орфографиялық дұрыс жазылмаған, қазақ тіліне тән әріптер кирилл әріптерімен алмастырылған сөздер қате ретінде танылды (Кесте 1).

Кесте 1. Орыс, ағылшын, араб және қазақ тіліндегі корпустардағы қате сөздер үлесі

Корпус	Қате үлесі (%)
Орыс тілі (RULEC-GEC)	6.3
Ағылшын тілі (FCE)	17.7
Ағылшын тілі (CoNLL-test)	10.8-13.6
Ағылшын тілі (CoNLL-train)	6.6
Ағылшын тілі (JFLEG)	18.5-25.5
Араб тілі	28.7
Қазақ тілі	**13.7**

Жоғарыдағы кестеден қазақ тілді корпустағы сөздердің 13.7%-ның қате жазылғандығын және қате жазылғанын, соның ішінде діни-экстремистік сипаттағы сөздердің үлесі 2%-ды құрайтындығын көруге болады. Қазіргі таңда аталған қателіктерді автоматты түрде анықтау және түзету әдістеріне шолу жасалуда, келешекте қазақ тіліндегі қате сөздерді тиімді түрде анықтайтын әдіс құрастыру жоспарлануда.

Бұл жұмыста әлеуметтік желі мәтіндеріндегі экстремистік бағыттағы сөздерді анықтау мақсатында семантикалық үлгілерді құру кезіндегі қиындықтардың бірі болып табылатын грамматикалық қателерді түзету мәселесі көтерілді. Ағылшын, жапон, орыс, неміс, чех тілдеріндегі қателерді түзетуге қатысты жұмыстарға шолу жасалды. Құрастырылған қазақ тілді корпустағы қате сөздердің үлесі анықталды. Келешекте кіріс мәтінге морфологиялық және семантикалық талдаудың тиімді орындалуына септігін тигізетін қазақ тіліндегі қателерді автоматты түрде анықтау және түзету үлгісін құру жоспарлануда.

Әдебиеттер

1. Zhongye J., Peilu W., Hai Zh. Grammatical Error Correction as Multiclass Classification with Single Model // Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, 2013. – Pages 74–81.

2. Boyd A. Using wikipedia edits in low resource grammatical error correction // Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, 2018. – Pages 79– 84.

3. Rozovskaya A., Roth D. Grammar error correction in morphologically rich languages: The case of russian. Transactions of the Association for Computational Linguistics, 2019. – Pages 1–17.

4. Šebesta K., Bedˇrichová Z., Šormová K., Štindlová B., Hrdlicka M., Hrdlicková T., Hana J., Petkevic V., Jelínek T., Škodová S., Janeš P., Lundáková K., Skoumalová H., Sládek S., Pierscieniak P., Toufarová D., Straka M., Rosen A., Náplava J., Poláckova M. CzeSL grammatical error correction dataset (CzeSL-GEC) – http://hdl.handle.net/11234/1-2143 (Қаралған күні: 05.01.2020)

5. Yu L., Lee L., Chang L. Overview of grammatical error diagnosis for learning chinese as a foreign language // Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications, 2014. – Pages 42–47.

6. Mizumoto T., Komachi M., Nagata m., Matsumoto Y. Mining revision log of language learning sns for automated japanese error correction of second language learners // Proceedings of 5th International Joint Conference on Natural Language Processing, 2011. – Pages 147–155.

7. Zaghouani w., Mohit B., Habash N., Obeid O., Tomeh N., Rozovskaya A., Farra N., Alkuhlani S., Oflazer K. Large scale arabic error annotation: Guidelines and framework, 2015.

8. Náplava J., Straka M. Grammatical Error Correction in Low-Resource Scenarios // Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, 2019. – Pages 346–356.

9. Bolatbek M., Mussiraliyeva Sh., Tukeyev U. Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language // Journal of Mathematics, Mechanics and Computer Science, №1 (97), 2018. – Pages 134-142.

10. Shalabayev K., Alipbay K., Bolatbek M., Mussiraliyeva Sh. Definition and classification of extremist texts in vkontakte social network // Vestnik KazNRTU, No. 5 (135), 2019. – Pages 80-86.

11. Sanasoft. Онлайн проверка орфографии http://www.sanasoft.kz/c/ru/node/48 (Қаралған күні: 05.01.2020).

12. Golding A., Roth D. A winnow-based approach to context-sensitive spelling correction // Machine learning, 1999. —Vol. 34.—N. 1–3.—P. 107–130.

13. Sorokin A., Shavrina T. Automatic spelling correction for Russian social media texts // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", 2016.

14. Rozovskaya A., Roth D. Grammar Error Correction in Morphologically Rich Languages: The Case of Russian // Transactions of the Association for Computational Linguistics, 2019. —Vol. 7, —P. 1–17.

ISO 9001-2015 ХАЛЫҚАРАЛЫҚ СТАНДАРТЫНДАҒЫ БІРТҰТАС ЖҮЙЕ РЕТІНДЕ ЖАҢАЛЫҚТАР БАҒДАРЛАМАЛАРЫНЫҢ САПА МЕНЕДЖМЕНТІНІҢ МОДЕЛІ

Орақ Б.Б.

e-mail: orakbb@narxoz.kz
Нархоз университеті АҚ, Қазақстан