«6D070400 – Есептеу техникасы және бағдарламалық қамтамасыз ету»
мамандығының PhD докторанты
Мухсина Куралай Женисбекқызының
«Машиналық оқыту негізінде көп тілді мәтіндік ақпаратты талдау жүйесін құру» тақырыбынадағы диссертациялық жұмысына

АҢДАТПА

Жұмыстың өзектілігі. Қазіргі заманғы көптілді және мультимәдени әлемде тілдердің тоғысу мәселесі, қоғамдарды біріктіру бойынша тілдер саласындағы тиімді және өмірге бейімді бағдарламаларды іздеу ерекше өзекті болып келеді. Мемлекет басшысы бастамашылық еткен және Қазақстан Республикасында іске асырылып жатқан тілдердің үштұғырлығы жобасы мемлекет ішінде де, сол сияқты одан тыс жерлерде де ақпараттық алмасу үшін база ретінде қызмет етеді. Берілген жоба қазақ тілін мемлекеттік тіл ретінде, орыс тілін ұлтаралық қатынас тілі ретінде және ағылшын тілін жаһандық экономикаға сәтті интеграциялау тілі ретінде қарастыра отырып, мемлекеттің ішіндегі ақпараттық қоғамдастықтың белсенді дамуына және Қазақстанның әлемдік жаһандық ақпараттық қоғамдастыққа интеграциясына жәрдемдеседі. Әлемдік ақпараттық кеңістікте халықаралық өзара іс-қимылды жүзеге асыру, Қазақстан Республикасының өнімін, білімін және коммуникациясын ілгерілету үшін қазақ тілін өңдеу жөніндегі арнайы ақпараттық қосымшаларды да, сонымен қатар оны көптілді өңдеу қосымшаларына қосу мүмкіндіктерін де дамыту қажет.

Бүгінде қазақ тілін өңдейтін қосымшалардың саны жеткілікті. Алайда, онымен қоса, қазақ тілін автоматты өңдеу сапасын арттыру қажеттілігіне байланысты көптеген проблемалар шешілмей қалып отыр.

Қазіргі уақытта қазақ тілін автоматты түрде өңдеу мәселесін шешуге Қалдыбай Бектаев, Шәріпбаев Алтынбек Әміруұлы, Әмірғалиев Еділхан Несіпханұлы, Түкеев Уәлшер Әнуарбекұлы, Хусейн Атакан Варол, Рахимова Диана Рамазанқызы, Мұсабаев Рустам Рафикович, Мансурова Мадина Есімханқызы, Мусиралиева Шынар Женісбекқызы және басқа да көптеген көрнекті ғалымдар елеулі үлес қосты.

Алайда, қолданыстағы зерттеулердің негізгі бөлігі морфология мен синтаксисті автоматтандыруға бағытталған, ал оның семантикалық талдау міндеті әлі шешілмеген болып қала беруде. Компьютерлік лингвистика, интеллектуалдық талдау және жасанды интеллект саласындағы қазақ тілін көп тілді жобаларға енгізумен және білімді дамытумен байланысты ғылыми зерттеулерін талдау осы ғылыми бағыттағы қазіргі шешімдер қазақ, орыс және ағылшын тілдерінің мәтіндерін автоматты өңдеу жүйесін әзірлеу бойынша қазіргі қажеттіліктерді сапалы деңгейде қанағаттандыру үшін жеткіліксіз екенін көрсетеді.

Қолданыстағы NLP көптілді қосымшалары көбінесе мәтінді өңдеудің грамматикалық кезеңін ғана қолданады, ал мәтінді семантикалық талдау немесе табиғи тілдің мағынасын талдау жасанды интеллект теориясы,

сонымен қатар компьютерлік лингвистика сияқты негізгі мәселелерінің бірі болып қала беруде.

Ережеге негізделген әдістер жоғары интеллектуалдық шығындарды талап ететіндіктен, ал компьютерлік процессорлар көбірек күшке ие болғандықтан, машиналық оқыту әдістері кеңінен таралуда. Алайда, көптілді ақпаратты семантикалық және грамматикалық өңдеу кезінде машиналық оқыту әдістерін қолдану үшін әр табиғи тілдің алдын-ала жасалған грамматикалық және семантикалық белгіленген корпустары қажет.

Жоғарыда айтылғандардың бәрі машиналық оқытуға негізделген көптілді мәтіндік ақпаратты талдау мәселесін кешенді зерттеуге және шешуге арналған диссертациялық жұмыстың өзектілігін анықтайды.

Диссертациялық жұмыстың мақсаты машиналық оқытудың интеллектуалдық талдау модельдері мен әдістерін пайдалану есебінен қазақ, орыс және ағылшын тілдерінің мәтіндік ақпаратын автоматты өңдеу жүйелерінің жұмыс сапасын арттыру болып табылады. Қойылған мақсат аясында машиналық оқыту модельдерін құру үшін мәтіндердің негізгі сипаттамаларын анықтау мақсатымен көптілді мәтіндік ақпаратты талдауды жүзеге асыратын модельдерді, әдістерді, алгоритмдерді әзірлеу, көптілді мәтіндерді интеллектуалдық өңдеу процестерін модельдеудің ғылыми міндеті шешіледі. Алынған алгоритмдер жұмысын құрылған белгіленген корпустар негізінде бағалау мүмкіндігімен мәтіндерді тәжірибелік өңдеу жүйелері түрінде өзінің практикалық іске асырылуына қолжеткізуі керек.

Диссертацияда қойылған мақсатқа жету үшін келесі **міндеттер** шешіледі.

- Әлсіз құрылымдалған және құрылымданбаған мәтіндік массивтерден фактілерді алу моделін әзірлеу және оны қазақ, орыс және ағылшын тілдеріне бейімдеу;
-Жасырын Марков моделін қолдана отырып, ықтималды POS-тегинг әдісін түрлендірк;
- VSM қолдануға негізделген көптілді мәтіндік құжаттардың семантикалық жақындығын анықтау әдісін жасау;
- Мәтіндердің семантикалық жақындығын талдау жүйесі жұмысының сапасын сараптамалық бағалау әдістемесін қалыптастыру;
- Әзірленген модельдерді, әдістер мен алгоритмдерді енгізетін бағдарламалық қосымшаны құру.

Диссертациялық жұмыстың ғылыми жаңалығы.

- Айрықша ерекшелігі НММ және тұрақты өрнектермен ұсынылған ережелерді бір мезгілде пайдалану болып табыталытын қазақ тілінің мәтіндік корпустарын автоматты морфологиялық және семантикалық таңбалаудың гибридтік әдісі түрлендірілді; бұл морфологиялық көп мәнділіктің бір бөлігін шешуге және таңбалаудың толықтығы мен дәлдігін арттыруға мүмкіндік берді;
- Көптілді мәтіндердегі фактілерді сәйкестендіретін семантикалық талдаудың логикалық-лингвистикалық моделі әзірленді, бұл қазақ, орыс және ағылшын тілдерінің мәтіндерінен RDF-триплет түрінде анық түрде ұсынылған білімді

алуға және семантикалық белгіленген оқыту корпустарын қалыптастыруға мүмкіндік берді

- VSM негізіндегі көп тілді мәтіндік құжаттардың семантикалық жақындығын анықтау әдісі жетілдірілді, ол мәтіннің тар бағытта мамандандырылған пәндік салаға жататындығын анықтау үшін PPMI салмақтық функциясын қолданумен ерекшеленеді;

- Оқу корпусының құжат векторларының косинустық ұқсастығының орташа мәнін есептеудің ұсынылған әдісіне негізделген, берілген тар шеңберде мамандандырылған тақырыпқа мәтіндердің семантикалық жақындығын анықтаудың ақпараттық технологиясы жасалды.

Зерттеу әдістері интеллект теориясын, жүйенің жалпы теориясын, жүйелік талдауды, түпкілікті предикаттар алгебрасын және машиналық оқыту әдістерін кешенді қолдануға негізделген. Түпкілікті предикаттар алгебрасы кейін оқу корпусын қалыптастырумен табиғи сөйлемдермен берілетін семантикалық ақпаратты формализациялау үшін қолданылады. Машиналық оқыту әдістері тар пәндік саладағы мәтіндердің тиесілігін анықтау модельдерін және көптілді корпустарды семантикалық белгілеу алгоритмін құру үшін қолданылады.

Зерттеу нысаны. Қазақ, орыс және ағылшын тілдеріндегі мәтіндік ақпаратты автоматты өңдеу жүйелері.

Зерттеу пәні көптілді мәтіндік ақпаратты интеллектуалдық семантикалық талдаудың модельдері мен алгоритмдері болып табылады.

Жұмыстың практикалық маңыздылығы қазақ, орыс және ағылшын тілдеріндегі мәтіндердің көптілді корпустарын автоматты түрде семантикалық белгілеу жүзеге асыруға мүмкіндік беретін бағдарламалық қосымшаның және талданатын мәтіннің ықтимал криминалды құрауышын анықтауға мүмкіндік беретін қосымшаның қорғауға шығарылатын ережелері негізінде әзірлеумен негізделеді. Сондай-ақ қазақ, орыс және ағылшын тілдеріндегі криминалды боялған мәтіндердің семантикалық белгіленген корпустарын әзірлеуде негізделеді.

Жұмыс нәтижелерінің қолданбалы құндылығы кез-келген мүдделі мемлекеттік органдардың компьютерлік желілердегі криминалды боялған мәтіндерін анықтау мүмкіндігімен айқындалады.

Қорғауға шығарылатын ережелер. Зерттеу нәтижелері бойынша төменде келтірілген міндеттер шешілді:

- Қазақ, орыс және ағылшын тілдеріне бейімделген әлсіз құрылымдалған және құрылымданбаған мәтіндік массивтерден фактілерді алу моделі әзірленді. Табиғи тіл сөйлемдерінің семантикасын модельдеу үшін түпкілікті предикаттар алгебрасының математикалық аппаратын таңдау негізделген.

-Жасырын Марков моделін қолдана отырып, ықтималды POS-тегинг әдісі түрлендірілді.

- VSM қолдануға негізделген көп тілді мәтіндік құжаттардың семантикалық жақындығын анықтау әдісі жасалды;

- Мәтіндердің семантикалық жақындығын талдау жүйесі жұмысының сапасын сараптамалық бағалау әдістемесі қалыптасты;

- Келіп түскен мәтіндерде криминалды мағынаның болуын анықтауға мүмкіндік беретін және семантикалық белгілеуді жүзеге асыратын бағдарламалық кешен құрылды.

Тақырыптың ғылыми-зерттеу бағдарламаларының жоспарларымен байланысы

Диссертациялық жұмыс ҚР БҒМ мен Ғылым комитеті Ақпараттық және есептеу технологиялары институтының «Құрылымданбаған және әлсіз құрылымдалған мәтіндік массивтердегі криминалды маңызды ақпаратты іздеу және талдау әдістері мен модельдері» ғылыми-зерттеу гранттық жұмыстарының күнтізбелік жоспарына сәйкес орындалды.

Жарияланымдар. Диссертация тақырыбы бойынша жүргізілген зерттеулердің негізгі нәтижелері 17 жарияланымда ұсынылған, оның ішінде 4 – ҚР БҒМ ҒК ұсынатын ғылыми басылымдарда, 6 – Scopus және Web of Science дерекқорына кіретін халықаралық ғылыми басылымдарда, 7 – халықаралық ғылыми практикалық конференциялар материалдарында.

Диссертацияның құрылымына кіріспе, 4 бөлім, қорытынды, пайдаланылған дереккөздер тізімі және бес қосымша кіреді. Диссертацияның жалпы көлемі 121 бет, 26 сурет, 6 қосымшаны құрайды. Әдебиеттер тізімі 145 дереккөзден тұрады.

Кіріспеде диссертациялық жұмыстың таңдалған тақырыбының өзектілігіне негіздеме келтірілген. Ғылыми-зерттеу жұмысының мақсаты, нысаны, пәні және міндеттері тұжырымдалған. Өткізілген зерттеудің нәтижелері сипатталған, олардың ғылыми жаңалығы мен практикалық маңыздылығы көрсетілген. Диссертациялық жұмыстың негізгі нәтижелерінің апробациясы туралы мәліметтер келтірілген.

Диссертациялық жұмыстың **бірінші бөлімінде** қазіргі заманғы лингвистикалық ресурстар мен қазақ тілі мәтіндерін автоматты өңдеу жүйелеріне шолу жасалынды, қазақ тілі мәтіндерін автоматты өңдеуді формализациялау мен алгоритмдеу мәселелеріне талдау жүргізілді. Бөлімде мәтіндік ақпаратты өңдеу кезінде қолданылатын машиналық оқытудың заманауи әдістеріне талдау жасалды және көп тілді құрылымданбаған мәтіндерден ақпарат алуға мүмкіндік беретін Open IE тәсілдемелері бөлініп көрсетілді. Жүргізілген талдау негізінде зерттеу міндеттерін белгілеу жүзеге асырылды.

Екінші бөлімде көптілді мәтіндік ақпаратты интеллектуалдық өңдеу процестерін модельдеуге арналған түпкілікті предикаттар алгебрасының математикалық аппаратын таңдау негізделген. Сөйлемдегі іс-әрекеттің қатысушылары арасындағы қатынасты анықтайтын табиғи тілдердің құрылымын формализациялау үшін қолдануға қатысты предикаттар алгебрасының құралдары мен предикаттар операцияларының негіздері қарастырылады. Бөлімде берілген тілдің сөйлем сөздерінің грамматикалық және семантикалық сипаттамаларының қатынасы арқылы сөйлем партиципанттарының семантикалық функцияларын сипаттайтын әзірленген Open IE логикалық-лингвистикалық моделі келтірілген. Қазақ, орыс және ағылшын тілдерінің мәтіндерінен құрылымдалған машинамен оқылатын

ақпаратты автоматты генерациялау үшін берілген модельдің бейімделуі көрсетілген. Бөлімде көптілді мәтіндік ақпараттан фактілерді алудың әзірленген математикалық моделі негізінде алынған ағылшын тіліндегі сөйлемдердегі әрекетке шақыру фактісін өзгертіп жазу алгоритмі келтірілген

Үшінші бөлім машиналық оқыту модельдері негізінде көптілді мәтіндерді морфологиялық және семантикалық талдаудың әдістері мен алгоритмдерін әзірлеуге арналған. Бөлімде жасырын Марков моделін (HMM) қолданатын ықтималды морфологиялық және семантикалық белгілеу әдісі көрсетілген. Тегтер тізбегінің ықтималдығын бағалау әдісінде қолданылатын функция екі ықтималдыққа байланысты: тегтер тізбегінің шартты ықтималдығы және осы тегпен токеннің белгіленуінің шартты ықтималдығы. Қазақ тілінің мәтіндерін семантикалық белгілеу алгоритмі сипатталған. Оқыту жүзеге асырылатын қазақ тілінің корпусын алғашқы белгілеу жұрнақтар мен лингвистикалық қағидалар тізімін пайдалануға негізделеді. Үшінші бөлімде VSM құжаттарының екі векторының арасындағы косинустық ұқсастықты есептеуге негізделген көп тілді мәтіндік құжаттардың семантикалық жақындығын анықтау әдісі көрсетілген, ол векторлардың координаттары ретінде салмақтық функцияны білдіретін PPMI өлшемін қолданады.

Төртінші бөлімде алынған нәтижелерді практикалық іске асыру келтірілген. Бөлімде Ван Ризбергеннің толықтық, дәлдік және өлшем коэффициенттерін қамтитын әзірленген кортеж модельдердің тиімділігінің объективті өлшенетін көрсеткіштері ретінде пайдаланылатын сандық бағалау метрикасын қолдану негізделген. Сондай-ақ орыс, қазақ және ағылшын мәтіндерінің үш корпусында Open IE әзірленген моделін іске асырудың практикалық нәтижелері көрсетілген. Құрылған корпустардың жалпы көлемі: 700 000 сөзден тұратын 6000 мәтін. Ағылшын тілі үшін факт триплетін алу дәлдігі 87,2%, орыс тілі үшін 82,4% және қазақ тілі үшін 71,0% құрайды. Сонымен қатар, бөлімде диссертациялық зерттеуде ұсынылған машиналық оқыту әдістеріне негізделген берілген тар шеңберде мамандандырылған тақырыпқа мәтіндердің семантикалық жақындығын анықтаудың әзірленген ақпараттық технологиясы сипатталған және осы технологияның сапасын бағалау моделі келтірілген .

Қорытындыда осы диссертациялық жұмыстың негізгі нәтижелері мен қорытындылары баяндалған.

Қорғауға шығарылатын ғылыми ережелердің, тұжырымдар мен ұсынымдардың негізділігі математикалық аппаратты пайдаланудың дұрыстығымен; эксперименттердің дұрыс қойылуымен; теориялық зерттеулер мен эксперименттік деректер нәтижелерінің сапалы және сандық сәйкестігімен; зерттеу нәтижелерінің практикалық қолданылуымен расталады.

Жұмыстың апробациясы. Диссертациялық жұмыстың нәтижелері халықаралық ғылыми конференцияларда, Есептеу және ақпараттық технологиялар институтының жыл сайынғы ғылыми конференцияларында, Қазақ ұлттық университетінің жас ғалымдары мен мамандарының ғылыми

конференцияларында, сондай-ақ әл-Фараби атындағы ҚазҰУ «Информатика» кафедрасының ғылыми семинарларында баяндалды.
Авторлық құқық объектісіне құқықтарды мемлекеттік тіркеу туралы куәлік алынды .

Ғылыми жарияланымдар:

1. Мамырбаев О.Ж., Мухсина К.Ж. Мәтін үндесітілігін анықтауға арналған қолданыстағы жүйелерді талдау//«ҚР ҰҒА Хабарлары. Физика-математикалық сериясы», 2017. - №5 (315). – Б.149-155.

2. Мамырбаев О.Ж., Мухсина К.Ж. Анализ текстовых сообщений с применением векторной формы//"Международная научно-практическая конференции «Математический методы и информационные технологии макроэкономического анализа и экономической политики», посвященной празднования 80-летнего юбилея академика НАН РК Абдыкаппара Ашимовича Ашимова", Алматы, 11.04.2017-12.04.2017.-С.136-144.

3.Хайрова Н.Ф., Избасаров Е.Ж., Мамырбаев О.Ж., Мухсина К.Ж. Формальная модель оценивания качества экстракции и идентификации знаний из слабоструктурированной текстовой информации// Материалы научной конференции ИИВТ МОН РК «Современные проблемы информатики и Вычислительных технологий». – 2018. - С.306 – 310.

4.Мамырбаев О.Ж., Хайрова Н. Ф., Мухсина К.Ж. Моделирование грамматических способов выражения семантики факта в английском предложении // III Международной научной конференции «Информатика и прикладная математика», посвященная 80-летнему юбилею профессора Бияшева Р.Г.и 70-летию профессора Айдарханова М.Б. 26-29 сентября 2018 года, Алматы. -С.136-144.

5.Petrasova S., Khairova, N., Lewoniewski W., Mamyrbaev O., Mukhsina K. Similar text fragments extraction for identifying common wikipedia communities// MDPI № 66 от 10.12.2018 https://doi.org/10.3390/data3040066.

6.Khairova N., Petrasova S., Lewoniewski W., Mamyrbaev O., Mukhsina K. Automatic extraction of synonymous collocation pairs from a text corpus // Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, FedCSIS 2018, 2018, DOI: 10.15439/2018F186 Номер статьи 8511195, -P. 485-488.

7.Khairova N., Petrasova S., Lewoniewski W., Mamyrbaev O., Mukhsina K. Comparative analysis of the informativeness and encyclopedic style of the popular web information sources// Lecture Notes in Business Information Processing 320, 2018, DOI: 10.1007/978-3-319-93931-5_24 -P. 333-344.

8.Mamyrbaev O., Turdalyuly M., Mekebayev N., Mukhsina K., Keylan A., Bagher B., Nabieva G., Duisenbayeva A., Akhmetov B. Continuous Speech Recognition of Kazakh Language // AMCSE 2018 - International Conference on Applied Mathematics, Computational Science and Systems Engineering. Vol. 24 – 2019.

9.Мамырбаев О.Ж., Мухсина К.Ж., Хайрова Н. Ф., Колесник А.С. Лингвистические инструменты выявления криминально окрашенной

текстовой информации веб-контента // Қазақстан-Британ техникалық университетінің Хабаршысы – 2018. - №3(46). – Б. 112-117.

10.Хайрова Н. Ф., Мамырбаев О.Ж., Мухсина К.Ж., Колесник А.С. Автоматическая генерация структурированной машинно-читаемой информации из мультиязычных текстов // Информатика и прикладная математика: Матер. IV междунар. науч. конф. – Алматы, 2019. – Ч.2. - С. 509 – 519.

11.Мамырбаев О.Ж., Хайрова Н.Ф., Мухсина К.Ж. Қазақ тіліндегі мәтіндердегі қылмыстық мәнді коллакцияларды анықтау // Вестник КазАТК. – 2019. – № 3 (110). – С. 170-175.

12.Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. The Aligned Kazakh-Russian Parallel Corpus Focused on the Criminal Theme // 3rd International Conference on Computational Linguistics and Intelligent Systems, 2019, Volume 2362.

13.Khairova N., Petrasova S., Mamyrbaev O., Mukhsina K. Detecting Collocations Similarity via Logical-Linguistic Model // In Proceedings of the Workshop on meaning relations between phrases and sentences - May 23, 2019, Gothenburg, Sweden, pp. 15-22.

14. Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. The Influence of Various Text Characteristics on the Readability and Content Informativeness // In Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 1: ICEIS, ISBN 978-989-758-372-8, DOI: 10.5220/0007755004620469 - pp. 462-469.

15.Khairova N., Petrasova S., Mamyrbaev O., Mukhsina K. Open Information Extraction as Additional Source for Kazakh Ontology Generation // ACIIDS 2020, LNAI 12033, 2020. https://doi.org/10.1007/978-3-030-41964-6_8 -P. 86–96,

16.Khairova N., Kolesnik A., Mamyrbaev O., Mukhsina K. Logical-linguistic model for multilingual Open Information Extraction // Cogent Engineering (2020), https://doi.org/10.1080/23311916.2020.1714829 00: 1714829.

17. Хайрова Н. Ф., Колесник А.С., Мамырбаев О.Ж., Мухсина К.Ж. Выровненный казахско-русский параллельный корпус, ориентированный на криминальную тематику// Вестник Алматинского университета энергетики и связи № 1 (48) 2020- С. 84-92.

Авторлық құқық объектісіне құқықтарды мемлекеттік тіркеу туралы куәліктер :

Авторлық құқықпен қорғалатын объектілерге құқықтардың мемлекеттік тізіліміне мәліметтерді енгізу туралы 2020 жылғы 8 сәуірдегі № 9180 куәлік, авторлары: Мамырбаев О. Ж., Жұмажанов Б. Ж., Мухсина К. Ж.