«6D075100- Информатика, есептеуіш техника және басқару» мамандығы бойынша философия докторы (PhD) дәрежесін алу үшін ұсынылған Ақанова Ақерке Сапарқызының «Мәтінді семантикалық талдау жасайтын нейрокомпьютерлік жүйе» диссертациялық жұмысының

АҢДАТПАСЫ

Зерттеудің өзектілігі:

Қазақстан - әлемдік білім кеңістігінің деңгейіне сәйкес келуге ұмтылатын қарқынды дамушы елдердің бірі.Сондай-ақ, короновирус пандемиясы әлемдік білім кеңістігіне түзетулер енгізді, ал бұл білім беру қоғамының барлық өкілдерін ақпараттық-коммуникациялық технологияларға жүгінуге мәжбүр етті. Пандемия компьютерлік технологиялар арқылы білім беру қоғамымен байланыс орнатуға мәжбүрледі, осылайша тез ақпарат алмасудың құндылығын көрсетті. Мемлекеттік тілде мәтін түрінде ұсынылған ақпарат ең көп таралған ақпарат түрі болып табылады. Қазақ тілі түркі тілдерінің қыпшақ тобының синтетикалық агглютинативті тілдер типіне жатады.Ондағы берілген сөздер негізден және оған қосылатын бірнеше аффикстерден тұрады, ал бұл бір сөзден 60-қа дейін жаңа сөз алуға мүмкіндік береді.

Демек, семантикалық талдау мақсатында қазақ тіліндегі мәтінді автоматты түрде өңдеу мәтіндер жиынтығынан осы нақты тапсырма үшін ең елеулі, тұжырымдамалық және маңызды сәттерді бөліп көрсетуге мүмкіндік береді және мұғалімнің мәтіндік жұмыстарды тексеруге жұмсайтын шығындарын азайтады. Әрине, қазір компьютерлік лингвистикадағы қазақ тілі үлкен нәтижелерге қол жеткізді.Нейрондық желіні қолдана отырып, қазақ тіліндегі мәтіндерді өңдеуді автоматтандыру саласындағы ғалымдардың зерттеулері назар аударуға тұрарлық, бірақ жеткіліксіз болып көрінеді. Берілген жұмыста зерттелген және баяндалғандай, қазіргі уақытқа дейін морфологиялық талдаусыз, онтолия сөздіктерінсіз қазақ тіліндегі мәтіндерді өңдеу модельдері жоқ.

Осылайша, зерттеу тақырыбының өзектілігі қазақ тіліндегі мәтіндер мазмұнының тақырыптары бойынша жіктеудің жаңа модельдерін құру міндетімен анықталады, ал оның шешімі берілген тақырыпқа мәтін мазмұнының семантикалық жақындығын сапалы анықтауға қол жеткізуге мүмкіндік береді.

Диссертациялық жұмыстың мақсаты

Жұмыстың мақсаты қазақ тіліндегі мәтіндерді семантикалық талдауға арналған нейрокомпьютерлік жүйені құруға ықпал ететін тақырыптық модельдің шығуын терең оқыту арқылы берілген тақырыптағы мәтін мазмұнының семантикалық жақындығын анықтау мәселесін шешу үшін қолданылатын модельдер мен алгоритмдерді құру болып табылады.

Зерттеу аясында берілген мақсатқа жету үшін келесі *міндеттер* орындалады:

Created by Universal Document Converter

1) Мәтінді автоматты өңдеуде қолданылатын мәтінді семантикалық талдаудың әдістері мен модельдерін зерттеу және талдау.

2) Нейрондық желілерді терең оқытуға ықпал ететін технологияны талдау.

3) Қазақ тіліндегі мәтіндерді автоматты өңдеу үшін сөз формаларының сөздігін әзірлеу.

4) Қазақ тіліндегі сөздерге арналған аффикстерді қысқарту алгоритмін әзірлеу.

5) Қазақ тіліндегі құжаттардың тақырыптық моделін әзірлеу.

6) Тақырыптық модельдің нәтижесін оқыту үшін көп қабатты нейрондық желіні жасау.

7) Мәтінге семантикалық талдау жасайтын нейрокомпьютерлік жүйені әзірлеу.

Зерттеу әдістері

Зерттеу барысында қойылған міндеттерді орындау үшін түрлі әдістер қолданылды.Семантикалық талдау жүргізу үшін мәтіндерді автоматты өңдеу әдістері, аналитикалық зерттеулер мен математикалық статистика әдістері, соның ішінде студенттің таралуына негізделген гипотезаларды (статистикалық өлшемдер) статистикалық тексеру әдістері кеңінен қолданылды.Терең нейрондық желілерді оқытудың математикалық моделін құру үшін кері қатені тарату әдісі, сонымен қатар компьютерлік модельдеу құралдары қолданылды.Жұмыстың істәжірибелік бөлігін орындау барысында математикалық есептеулерді автоматтандыру құралдары және «Python» бағдарламалау тілі негізінде нәтижелердің сұлбаларын көрсету құралдары қолданылды.

Берілген диссертацияның **зерттеу нысаны** қазақ тіліндегі мәтіндерді автоматты түрде өңдеу болып табылады.

Берілген диссертацияның **зерттеу пәні** қазақ тіліндегі мәтіндердің мазмұнын тақырып бойынша жіктеу болып табылады.

Осы зерттеудің **ғылыми жаңалығы** - қазақ тіліндегі мәтіндерді жіктеу мәселелерін шешу кезінде сөз формаларының сөздігін жасау үшін олардың аффикстерін қысқарту алгоритміне негізделген пайдаланылатын мәтін семантикасының дәлдігін анықтауды талап ететін компьютерлік лингвистиканы зерттеуде нейрондық желі моделін әзірлеуден тұрады.

Сенімділік және негіздеме. Нәтижелердің негізділігі мен сенімділігі зерттеу тақырыбы бойынша әдеби дереккөздерді жан-жақты талдау және нейрондық желіні оқыту тәжірибесімен расталады. Компьютерлік лингвистика саласындағы танымал шетелдік және отандық ғалымдардың еңбектері сәйкес зерттеуде нейрондық желілер негізінде мәтінді өңдеуді автоматтандыру үшін модельдер мен әдістер таңдалды және қолданылды. Құжаттағы тақырыптар мен олардың сөздердің таралу ықтималдығын есептеу барысында барлық факторлар, нейрондық желіні оқытудағы қателіктер функциясы, оқыту оптимизаторы және метрика ескеріледі. Атқарылған жұмыстардың дұрыстығы зерттеу қорытындыларының рецензияланған ғылыми журналда жариялануына және оқу процесіне

Created by Universal Document Converter

нейрокомпьютерлік жүйенің ендірілуіне негізделеді.

Теориялық маңыздылығы қазақ тіліндегі мәтінді өңдеумен байланысты зерттеулерде стемминг алгоритмі мен нейрондық желілік модельді қолдану болып табылады.

Практикалық маңыздылығы:

– тақырыптық модельдеу (LDA) барысында сөздік қорын құру үшін қолданылатын қазақ тіліндегі сөздер үшін аффикстерді қысқарту алгоритмі;

– LDA моделінің шығыс деректерін терең оқыту үшін 4 қабатты нейрондық желі модельдері үшін:

а) нейрондық желінің оңтайлы алгоритмдері (қабаттары) таңдалды;

б) нейрондық желіні оқытудың нәтижесіне әсер ететін нейрондық желіні құрастырудың оңтайлы параметрлері анықталды.

– тақырыптық модель және нейрондық желінің құрастырылған моделі негізінде қазақ тіліндегі мәтіндік мәліметтерді семантикалық талдауға арналған нейрокомпьютерлік жүйе.

Қазақ тіліндегі сөздердің аффикстерін қысқарту алгоритмі және оқытылған нейрондық желі қазақ тіліндегі мәтіндік жұмыстарды семантикалық талдау үшін қолданылатын нейрокомпьютерлік жүйеде өзінің практикалық іске асырылуына ие болды.

Қазақ тіліндегі сөздердің аффикстерін қысқарту алгоритмі және оқытылған нейрондық желі қазақ тіліндегі мәтіндік жұмыстарды семантикалық талдау үшін қолданылатын нейрокомпьютерлік жүйеде өзінің практикалық іске асырылуына ие болды.

Қорғауға шығарылатын ережелер:

– қазақ тіліндегі сөздердің аффикстерін қысқарту алгоритмі;

– қазақ тіліндегі құжаттардың тақырыптық моделі.

– қазақ тіліндегі мәтіндердің семантикалық талдау векторларын оқытуға арналған нейрондық желі моделі;

–нейрондық желіні терең оқыту;

– қазақ тіліндегі мәтіндерді талдауға арналған нейрокомпьютерлік жүйе және берілген тақырыптың семантикалық жақындығын ықтималды анықтау.

Жұмысты апробациялау. Жұмыстың апробациясы С. Торайғыров атындағы Павлодар мемлекеттік университетінің базасында жүзеге асырылды және енгізу актісі алынды.Негізгі жарияланымдар халықаралық ғылыми конференцияларда баяндалды, ғылыми журналдарда жарияланды, әзірленген электрондық өнімдер авторлық құқықпен қорғалатын объектілерге құқықтардың мемлекеттік тізілімінде тіркелді:

- Профессор Болат Әбдікәрімұлының 80 жылдығына арналған «кәсіби білім берудегі инновациялар: мәселелері мен болашағы» атты республикалық ғылыми-практикалық конференция материалдары, ақпан, 2019 ж., Астана.

- Technology Audit and Production Reserves. DOI: 10.15587/2706-5448.2020.217613

Created by Universal Document Converter

- 2020 жылғы 20 ақпандағы № 8300 және 2021 жылғы 27 мамырдағы №18050 авторлық құқықпен қорғалатын объектілерге құқықтардың мемлекеттік тізіліміне мәліметтерді енгізу туралы куәліктер.

Диссертация нәтижелері 10 жұмыста жарияланды. Оның ішінде 1 монография, ҚР БҒМ Білім және ғылым саласындағы бақылау комитеті ұсынған журналдарда 3 мақала, отандық ғылыми басылымда 1 мақала, Scopus деректер базасына енген халықаралық ғылыми басылымда 1 мақала, шетелдік ғылыми басылымда 1 мақала, халықаралық және республикалық конференциялар материалдарында 1 жұмыс, 2 авторлық құқық туралы куәлік.

Автордың жеке қосқан үлесі.

Диссертациялық зерттеу барысында алынған негізгі эксперименттік және теориялық нәтижелерді автор өзі алды.Бірлескен авторлар тобы құрамындағы жарияланымдарда ізденуші қол жеткізілген нәтижелерді алу, жалпылау және талдау барысында негізгі идеяларға ие негізгі автор болып табылады.Диссертацияның құрылымы келесі бөлімдерді құрайды: кіріспе бөлім, негізгі бөлім (үш тарау), қорытынды, пайдаланылған әдебиет көздері тізімі және қосымшалар.Жұмыс компьютерлік мәтіннің 113 бетінде баяндалған, 30 сурет, 15 кесте және 199 библиографиялық дереккөздерден тұрады.

Зерттеудің негізгі нәтижелері.

Ғылыми дереккөздерді зерттеу нәтижесінде мәтінді автоматты түрде өңдеуде қолданылатын **әдістердің жіктелуі** келтірілген (4-сурет).

Қазақ тіліндегі мәтінді семантикалық талдаудың нейрокомпьютерлік жүйесін жасауда қолданылған терең нейрондық оқытуда **қатені тарату әдісі** сипатталған.

Нейрондық желілер негізінде объектілерді тану технологияларына салыстырмалы талдау жүргізілді, ондаApache Singa, Caffe, Deeplearning4j, Keras, Microsoft Cognitive Toolkit, MXNet, Neural Designer, OpenNN, Theano, Torch, Tensor Flow, NLTK, Gensim секілді 13 технологияның салыстырмалы сипаттамалары көрсетілген.Жоғарыда келтірілген талдаудан мәтіндік деректерді жіктеуде Gensim технологиясын қолдану анағұрлым сенімді.

Анықтамалық сөздік қорын құру үшін стеммер Портер негізінде **аффикстерді жою алгоритмі құрастырылды**.

Сөз формаларын жасауды әртүрлі әдістермен жасауға болады, бірақ ең көп таралғаны стеммингболып табылады.Стемминг- сөздер мен жұрнақтарды қысқарту арқылы сөз негізін алу процесі.Python бағдарламалау тілінде қазақ тіліндегі мәтіндерге арналған аффикстерді қысқарту бойынша стеммердің жұмысы іске асырылды.

Қазақ тіліндегі мәтінді өңдеу үшін LDA көмегімен тақырыптық модель жасалды.Тақырыптық үлгіні жасау үшін құжаттар жиынтығынан тұратын қазақ тіліндегі мәтіндер корпусы пайдаланылады. Содан кейін мәтіндер корпусы токенизациядан өтеді, ал бұл бізге мәтінді сөздерге бөлуге мүмкіндік береді.Стеммердің көмегімен жұрнақтар мен жалғауларының негіздері кесіліп, сөздік жасалады. Берілген сөздіктен алынған сөздер

Created by Universal Document Converter

сөздердің бір-біріне шамалас белгісін анықтайтын биграммалардың көмегімен алынады, алынған нәтижелерден кейін тақырыптар bag-of-words (BOW) технологиясын қолдана отырып анықталады, ал бұл жұмыста жақын кілт сөздерді табатын doc2bow дайын алгоритмі қолданылды. Содан кейін LDA алгоритмін қолдана отырып, кілт сөздер тақырыптар бойынша және тақырыптар құжаттар бойынша таратылады. LDA алгоритмінің нәтижесінде, біз тақырыптар туралы таразылармен кілт сөздерді аламыз, құжаттар бойынша таразылармен тақырыптарды бөлеміз.

Алынған LDA моделінің нәтижесі құрастырылған нейрондық желі моделінің көмегімен терең оқыту арқылы бағаланады.Осылайша, оқыту үшін 4 қабатты нейрондық желі құрастырылды.Embedding () енгізу қабатыкілт сөздерді таразымен және тақырыптарды векторлық мәліметтерге түрлендіреді. SpatialDropout1D () екінші қабаты желідегі нейрондардың реттелуін тудырады.LSTM үшінші қабаты өз ішінде тағы екі реттегіш қабатты қамтиды: біреуі қарапайым dropout, екіншісі қайталанатын dropout.

Төртінші қабат- Dense Шығыс тығыз қабаты. Әр қабат нейрондық желіде алынған мәліметтерге қатысты өз функциясын орындайды.Әрбір қабат өлшенген мөлшерді есептеу арқылы нейрондарды өңдеу алгоритмін қамтиды. Оқу процесі қатенің кері таралуынан, яғни кірістен кіріске дейін алынған қатені кері бағытта есептеу процесінен басталады.

Қазақ тілінде мәтіндік деректерді өңдеудің нейрокомпьютерлік жүйесі өңделді. Нейрокомпьютерлік жүйенің дамудың келесі кезеңдері келтірілген:

1. Қазақ тіліндегі мәтіндер корпусын құру.
2. Қазақ тіліне арналған аффикстерді қысқарту алгоритмін қолдану арқылы негіздер сөздігін жасау.
3. Таңдалған биграммалармен немесе тақырыптарға қатысты униграммалармен сөздікті жаңарту (кілт сөздер).
4. Тақырыптық модель құру (LDA).
5. Көп қабатты нейрондық желіні құру.
6. Көп қабатты нейрондық желіні оқыту.
7. Тапсырма мен нейрондық желінің оқытылған моделінің байланысы.
8. Адам-машиналық өзара әрекеттесу үшін интерфейс құру.
9. Қазақ тіліндегі мәтіндерге семантикалық талдау жүргізуге арналған нейрокомпьютерлік жүйені апробациялау.

Мәтінді семантикалық талдаудың нейрокомпьютерлік жүйесінің архитектурасы жасалды.Нейрокомпьютерлік жүйені апробациялау қазақ тіліндегі мәтінің мазмұнының өзінің тақырыбына сәйкестігін тексеруден тұрады. Сарапшылардың нәтижелері мен бағдарламаның өзі арасындағы сызықтық байланысты көрсеткен тәжірибелік деңгейде өзара байланыстылық коэффициенті есептелді, ал бұл дамыған модельдер мен оқытылған нейрондық желінің сенімділігін растайды. Бағдарлама Торайғыров университетінде қолданып ендіру актісі алынды.

Created by Universal Document Converter