Ерболова Асель Серикановнаның
8D06101 (6D075100) – «Информатика, есептеу техникасы және басқару» білім беру бағдарламасы бойынша философия докторы (PhD) дәрежесін алуға ұсынылған «Боттарды идентификациялауда табиғи тіл құрылымдарын зерттеу» тақырыбындағы диссертациялық жұмысының

АҢДАТПАСЫ

Зерттеудің өзектілігі табиғи тілді өңдеу (NLP) және жасанды интеллект (ЖИ) технологияларының қарқынды дамуымен айқындалады. Адамдар мен автоматтандырылған жүйелер (боттар) арқылы қалыптасатын мәтіндер көлемінің ұлғаюы, цифрлық ортада жалған ақпарат таралу қаупінің күшеюі және жасанды интеллект модельдерінің деградациясы бот мәтіндерін анықтаудың тиімді әдістерін әзірлеуді талап етеді. Бұл ақпараттық қауіпсіздікті қамтамасыз ету, цифрлық контентті мониторингтеу және ЖИ жүйелерінің орнықтылығын арттыру міндеттерімен тығыз байланысты. Осыған байланысты мәтіндерді талдау мен жіктеудің заманауи әдістерін қолдану зерттеудің маңыздылығын айқындайды.

Жұмыста ұсынылған әдістер мен тәсілдердің өзектілігі келесі факторлармен негізделеді. Біріншіден, тілдік фракталдық құрылымдардың ішкі өлшемділіктерін семантикалық траекторияларды талдаумен ұштастыру тілдік заңдылықтарды кешенді түрде сипаттауға мүмкіндік береді және тілдік жүйелердің құрылымы мен динамикасын терең түсінуге жағдай жасайды.

Екіншіден, деректердің топологиялық талдау әдістерін қолдану тілдік жүйені зерттеудің жаңа ғылыми құралдарын ұсына отырып, боттарды анықтау дәлдігін арттыруға мүмкіндік береді. Бұл автоматтандырылған жүйелердің қоғамдық пікірге ықпал етуі мен жалған ақпарат тарату тәуекелдерінің артуы жағдайында ерекше өзекті болып табылады. Осылайша, диссертациялық жұмыс табиғи тілдің күрделі құрылымын зерттеуге бағытталған заманауи әдістерді ұсына отырып, ақпараттық жүйелердің тиімділігін арттыруға және цифрлық ортадағы қауіпсіздікті қамтамасыз етуге елеулі үлес қосады.

Зерттеудің мақсаты – тіл құрылымдарын зерттеу әдістерін әзірлеу және оларды адам тарапынан жазылған және боттар арқылы генерацияланған мәтіндерді идентификациялауға бейімдеу.

Қойылған мақсатқа жету үшін келесі негізгі **міндеттер** шешілді:

1. Тілдік объектілердің сипаттамаларын әзірлеу және олардың статистикалық қасиеттерін талдау арқылы адам жазған мәтіндер мен боттар генерациялаған мәтіндердің айырмашылықтарын анықтау.
2. Табиғи тілдегі мәтіндердің семантикалық траекторияларының күрделілігін бағалау үшін уақыт қатарларының хаостық қасиеттерін талдау әдістерін бейімдеу.
3. Күрделі геометриялық объектілердің ішкі өлшемділіктерін бағалау әдістерін тілдік фракталдық құрылымдарды зерттеуге бейімдеу.
4. Табиғи тілдің семантикалық кеңістігін зерттеу үшін деректердің топологиялық талдау әдістерін бейімдеу, әртүрлі n-граммалар (сөздер, биграммалар, триграммалар) үшін «қуыстарды» анықтау, сондай-ақ әртүрлі тілдер аясында табиғи тілдің ірі масштабты құрылымдарының сипаттамаларын салыстырмалы талдау.
5. Анықталған тілдік белгілер негізінде боттар жасаған мәтіндерді идентификациялауға арналған классификациялық модельдерді әзірлеу және олардың тиімділігін эксперименттік тұрғыда тексеру.

Зерттеу объектісі – адам тарапынан жазылған мәтіндер мен автоматтандырылған жүйелер (боттар) арқылы генерацияланған мәтіндерді ажыратуда тиімді болып табылатын табиғи тілдің тілдік сипаттамалары мен құрылымы.

Зерттеу пәні – табиғи тілдің ірі масштабты құрылымын, оның геометриялық қасиеттерін зерттеу әдістері, сондай-ақ адам жазған мәтіндер мен автоматтандырылған жүйелер (боттар) генерациялаған мәтіндерді идентификациялау әдістерін әзірлеу.

Зерттеу әдістері: статистикалық әдістер (айырмашылықтарды анықтау тесттері, үлестірімдерді талдау); табиғи тілдердегі семантикалық траекториялардың хаостық қасиеттерін талдау әдістері; ішкі өлшемділікті сандық бағалау әдістері (n-граммалар негізіндегі геометриялық және фракталдық тәсілдер); деректердің топологиялық талдау әдістері (персистентті гомология); боттарды идентификациялау әдістері (классификациялық модельдер, таңдама жиынтығын бөлу).

Диссертациялық зерттеудің **ғылыми жаңалығы**:

1. Табиғи тілдер өзін-өзі ұйымдасқан-критикалық жүйелер ретінде көптеген тілдер жиынтығында талданды. Эсперанто тілінің табиғи тілдерден айырмашылығы ретінде гаусстық үлестірімді көрсететін бірегейлігі анықталды. Дәрежелік үлестірімдердің статистикалық сипаттамалары, оның ішінде дәрежелік заң параметрлері тілдік жүйелерді талдау мен жіктеудің сенімді метрикалары болып табылатыны көрсетілді.
2. Табиғи тілдердегі семантикалық траекторияларды талдау нәтижелері ұсынылып, олардың хаостық табиғаты айқындалды. Кластерлік талдау энтропия және күрделілік көрсеткіштерімен корреляцияланатын типологиялық кластерлерді анықтап, тілдік динамикадағы хаостың маңыздылығын көрсетті.
3. Табиғи тілдердің ішкі өлшемділігін бағалау әдістері әзірленіп, олардың мультифракталдық табиғаты анықталды, сондай-ақ өлшемділік сипаттамаларының векторлық көріністерді алу әдістерінің өзгерістеріне қатысты инварианттылығы дәлелденді.
4. Деректердің топологиялық талдау әдістері негізінде тілдегі «қуыстарды» анықтау әдісі әзірленді. Бірінші ретті персистентті гомологияларды есептеу табиғи тілдің ірі масштабты құрылымдық ерекшеліктерін анықтауға және таңдама артефактілерін жоюды қамтамасыз етті.
5. Адамдар жасаған және боттар генерациялаған мәтіндерді ажыратуға арналған жоғары тиімді классификациялау алгоритмдері ұсынылып, көптеген тілдер үшін 96%-дан жоғары дәлдікке қол жеткізілді.

Зерттеу нәтижелерінің практикалық маңыздылығы адам жазған мәтіндер мен автоматтандырылған жүйелер (боттар) генерациялаған мәтіндерді тиімді ажыратуға арналған боттарды идентификациялау әдістерінің әзірленуімен айқындалады. Энтропия және семантикалық траекториялардың күрделілігі сияқты тілдік сипаттамаларды талдау әдістері автоматты түрде боттарды анықтау жүйелерінің дәлдігі мен тиімділігін едәуір арттыруға мүмкіндік берді, бұл контентті автоматты генерациялау мен жалған ақпарат таралуы күшейген жағдайда ерекше маңызды.

Ұсынылған деректердің топологиялық талдау әдістері және персистентті гомология мәтіндердің құрылымдық ерекшеліктерін зерттеуде қолданылуы мүмкін, бұл әртүрлі салаларда талдау мен жіктеу мүмкіндіктерін кеңейтеді. Табиғи тілдерді салыстырмалы талдау нәтижелері олардың ірі масштабты құрылымдарын ескере отырып, тілдерді жіктеу тәсілдерін қайта қарастыруға ықпал етеді.

Қорғауға шығарылатын негізгі нәтижелер:

1. Табиғи тілдерді өзін-өзі ұйымдасқан-критикалық жүйелер ретінде сипаттайтын модель ұсынылып, оның тілдердің кең ауқымын салыстырмалы талдауда тиімді екендігі негізделді.
2. Тілдің хаостық табиғатын сипаттайтын семантикалық траекторияларды талдау моделі әзірленіп, оның тілдерді салыстырмалы талдауда қолданылу мүмкіндігі дәлелденді.
3. Тілдік фракталдық құрылымдарды (ішкі өлшемділіктер мен «қуыстарды») талдау моделі әзірленіп, оның мультифракталдық қасиеттерді анықтауда және тілдердің туыстық гипотезаларын негіздеуде тиімді екендігі көрсетілді.
4. Адамдар жазған және боттар генерациялаған мәтіндерді идентификациялауға арналған классификациялық модель әзірленіп, оның жоғары тиімділігі және практикалық маңыздылығы эксперименттік тұрғыда дәлелденді.

Апробациялар және жарияланымдар

Диссертациялық зерттеудің негізгі нәтижелері Манаш Қозыбаев атындағы Солтүстік Қазақстан университетінде (Ақпараттық-коммуникациялық технологиялар кафедрасы), сондай-ақ Ұлттық зерттеу университеті «Жоғары экономика мектебінде» өткен ғылыми семинарларда баяндалды. Диссертация тақырыбы бойынша барлығы 7 ғылыми еңбек жарияланған. Оның ішінде: Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым және жоғары білім саласындағы сапаны қамтамасыз ету комитеті ұсынған басылымдарда – 1 мақала; халықаралық ғылыми конференциялар материалдарында – 1 жарияланым; Scopus деректер базасында индекстелетін ғылыми журналдарда процентилі 39-дан 80-ге дейінгі 5 мақала жарияланған.

- зерттеу нәтижелері бойынша жарияланған мақалалар, соның ішінде Web of Science және Scopus деректер базасында индекстелетін ғылыми журналдарда

1. Gromov V.A., Borodin N.S., and **Yerbolova A.S.**, A Language and Its Dimensions: Intrinsic Dimensions of Language Fractal Structures// *Complexity Volume 2024*, https://doi.org/10.1155/2024/8863360*, (Indexed in Scopus, Q1)*

2. Gromov V.A., Dang Q.N., Kogan A.S., **Yerbolova A.S.**, Spot the bot: the inverse problems of NLP. *PeerJ Computer Science* 10:e2550, 2024, https://doi.org/10.7717/peerj-cs.2550*, (Indexed in Scopus, Q1)*

3. Gromov V.A., Dang Q.N., **Yerbolova A.S.**, Language and Its Holes: the First Order Homologies of the Large-scale Geometrical Structure of a Natural Language// Complexity, 2025 (1) Article ID 9659172, https://doi.org/10.1155/cplx/9659172, *(Indexed in Scopus, Q2)*

4. **Yerbolova A.S.**, Tomashchuk K.K., Kogan A.S., Dang Q.N., Skrynnikova I.V., and Gromov V.A., Relative Chaoticity of Natural Languages// Complexity, Volume 2026, https://doi.org/10.1155/cplx/5519690, *(Indexed in Scopus, Q2)*

5. **Yerbolova A.S.,** Kurmashev I. G., Revealing intrinsic dimensionality patterns in semantic spaces of natural languages using graph algorithms, Eastern-European Journal of Enterprise Technologies, 1/2 (138), 2026, https://doi.org/10.15587/1729-4061.2026.351509 , *(Indexed in Scopus, Q3)*

- ҚР ҒЖБМ ҒЖБССҚЕК ұсынған ғылыми журналдарда жарияланған мақалалар:

6. **Ерболова А.С.,** Громов В.А., Аканова А.С. Жасанды интеллект жүйелері арқылы генерацияланған мәтіндерді анықтаудың семантикалық әдістері // Вестник Алматинского университета энергетики и связи No 1(72), 2026

- халықаралық және республикалық конференциялардағы баяндамалар бойынша тезистер мен мақалалар:

7. Поймай бота: крупномасштабная структура естественного языка // Проектирование будущего. Проблемы цифровой реальности: труды 7-й Международной конференции (15-17 февраля 2024 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2024. — С. 281-312. — https://keldysh.ru/future/2024/6-3.pdf https://doi.org/10.20948/future-2024-6-3

Зерттеу нәтижелерін енгізу.

Диссертациялық зерттеу нәтижелерінің практикалық маңыздылығы олардың ғылыми-зерттеу және білім беру қызметіне енгізілуімен және қолданылуымен расталады.

Зерттеу нәтижелерін практикалық апробациялау және енгізу бірқатар ғылыми және білім беру ұйымдарында жүзеге асырылып, тиісті енгізу актілерімен расталған. Атап айтқанда, диссертациялық зерттеу нәтижелері Қазақстан Республикасы Ұлттық қорғаныс университетінің ғылыми-зерттеу қызметінде қолданылып, мәтіндік ақпаратты талдау, семантикалық құрылымдарды зерттеу және деректерді зияткерлік өңдеу міндеттерін шешуде пайдаланылуда.

Әзірленген әдістер Қазтұтынуодағы Қарағанды университетінің білім беру процесіне енгізіліп, оқу процесінде, оның ішінде деректерді талдау және жасанды интеллект

бағыттары бойынша оқу сабақтарын жүргізуде, сондай-ақ курстық және дипломдық жобалауда қолданылуда, бұл тиісті енгізу актілерімен расталған.

Зерттеу нәтижелерінің жекелеген бөліктері Ұлттық зерттеу университеті «Жоғары экономика мектебінің» ғылыми-зерттеу қызметінде пайдаланылып, семантикалық көріністерді талдау және мәтіндік деректерді өңдеу міндеттерін шешуде қолданылуда, бұл да енгізу актілерімен расталған.

Диссертациялық зерттеу нәтижелері бойынша авторлық құқық объектісіне құқықтарды мемлекеттік тіркеу туралы №66441 куәлік (2026 жылғы 19 қаңтар) «Табиғи тілдердің семантикалық кеңістіктерінің ішкі өлшемділігін талдау» алынды (Приложение В).

Ізденушінің қосқан жеке үлесі. Ізденушінің жұмысқа қосқан жеке үлесі - әдістер мен алгоритмдерді әзірлеу, сондай-ақ зерттеу тақырыбы бойынша зерттеу нәтижелерін жариялауға ұсыну.

Диссертацияның көлемі мен құрылымы. Диссертациялық жұмыс кіріспеден, төрт бөлімнен, қорытындыдан, пайдаланылған әдебиеттер тізімінен және қосымшадан тұрады.

Кіріспеде таңдалған диссертация тақырыбының өзектілігі негізделеді, зерттеудің мақсаттары мен негізгі міндеттері тұжырымдалады, жұмыстың жаңалығы мен практикалық маңыздылығы, таңдалған тақырып бойынша шолу жүргізіледі.

Бірінші тарауда табиғи тілді күрделі жүйе ретінде қарастыратын, оның ішінде оның фракталдық құрылымдары мен автоматтандырылған жүйелер (боттар) генерациялаған мәтіндерді идентификациялау әдістеріне арналған ғылыми әдебиеттерге шолу ұсынылған. Тарауда тіл, семантика және хаостық құбылыстар арасындағы өзара байланыс мәселелері жүйеленіп, негізгі ғылыми зерттеулер мен тәсілдерге талдау жүргізілген.

Жаратылыстану-ғылыми парадигма аясында табиғи тіл өзін-өзі ұйымдасқан-критикалық жүйелер ретінде қарастырылады. Бұл бағытта мәтіндердегі дәрежелік заңдылықтар, сондай-ақ олардың қалыптасуына әсер ететін психофизиологиялық және когнитивтік факторлар талданып, негізделеді. Сонымен қатар, мәтіндердің семантикалық траекторияларының хаостығы мен күрделілігі мәселелері қарастырылған. Тілдік процестерді сипаттайтын сандық заңдылықтар мен принциптерге, пермутациялық энтропия негізіндегі хаостықты бағалау әдістеріне және басқа да статистикалық тәсілдерге жүйелі талдау жүргізілген.

Маңызды бағыттардың бірі ретінде тілдік фракталдық құрылымдарды және табиғи тілдердің ішкі өлшемділігін бағалау әдістерін зерттеу қарастырылған. Бұл бөлімде заманауи тәсілдерге, атап айтқанда, страндық аттракторлар, деректер графтары және персистентті гомологияға негізделген әдістерге талдау жүргізілген. Сонымен қатар, автоматты түрде генерацияланған мәтіндерді идентификациялау мәселелері қарастырылған. Қарапайым мәтіндік сипаттамаларды талдаудан бастап семантикалық кеңістіктер мен мәтіндердің эмоционалдық ерекшеліктеріне негізделген күрделі әдістерге дейінгі әртүрлі тәсілдерге талдау жүргізілген.

Екінші тарауда 18 тілдік отбасына жататын 52 тілді қамтитын репрезентативті корпус негізінде табиғи тілдің ірі масштабты құрылымын талдаудың әдістер кешені әзірленген. Табиғи тілдердің өзін-өзі ұйымдасқан-критикалық жүйелер ретінде қарастырылатыны және оларға үлестірімдердің дәрежелік заңдылықтарының тән екендігі негізделген. Мәтінді эмбеддингтер кеңістігіндегі семантикалық траектория түрінде модельдеу тәсілі ұсынылып, энтропия мен күрделілікті талдау негізінде тілдік деректердің хаостық табиғаты айқындалған. Тілдік құрылымдардың ішкі өлшемділігін бағалау әдістері, сондай-ақ семантикалық кеңістікті топологиялық талдау әдістері әзірленіп, олардың фракталдық және топологиялық ерекшеліктерін, соның ішінде тілдің тұрақты гомологиялық құрылымдарын («қуыстарын») анықтау мүмкіндігі дәлелденген. Алынған белгілер жиынтығы негізінде машиналық оқыту алгоритмдерін қолдану арқылы боттар генерациялаған мәтіндерді идентификациялаудың жоғары тиімді әдістері ұсынылып, олардың тиімділігі эксперименттік нәтижелермен расталған.

Үшінші тарауда әзірленген әдістерді табиғи тілдерді талдауға қолдану нәтижелері келтірілген. 18 тілдік отбасына жататын 52 тілді қамтитын зерттеу негізінде табиғи тілдердің үлестірімдерінің дәрежелік заңдылықтарды көрсететіні анықталып, бұл олардың өзін-өзі ұйымдасқан-критикалық жүйелер ретіндегі табиғатын растайды, ал эсперанто тілі гаусстық үлестірімді көрсетуі арқылы ерекшелік ретінде негізделген. Мәтіндердің семантикалық траекторияларын талдау нәтижелері тілдердің хаостық табиғатын растады, әрі олардың басым көпшілігінің хаостық қасиеттерге ие екендігі дәлелденген. Кластерлік талдау энтропия және күрделілік параметрлерімен корреляцияланатын типологиялық топтарды анықтап, сондай-ақ ареалдық және типологиялық заңдылықтарды қоса алғанда, бірқатар лингвистикалық гипотезаларды растады. Ішкі өлшемділіктің алынған бағалары тілдердің мультифракталдық табиғатын және қолданылатын әдістерге қатысты олардың тұрақтылығын көрсетті, бұл ішкі өлшемділікті тілдің инвариантты сипаттамасы ретінде қарастырудың негізділігін дәлелдейді.

Топологиялық талдау әдістері семантикалық кеңістіктің тұрақты гомологиялық құрылымдарын («тілдің қуыстарын») анықтауды қамтамасыз етіп, олардың тілдік жүйелердің терең семантикалық шектеулерін бейнелейтіні негізделген. Осы құрылымдарға қатысты тілдік бірліктердің қашықтықтары негізінде боттар генерациялаған мәтіндерді идентификациялау тәсілі ұсынылып, оның тиімділігі негізделген. Эксперименттік нәтижелер адам жазған мәтіндер мен боттар генерациялаған мәтіндер арасында статистикалық тұрғыдан мәнді айырмашылықтардың бар екенін көрсетіп, классификацияның жоғары тиімділігін растады.

Төртінші тарауда әзірленген әдістер негізінде боттар генерациялаған мәтіндерді идентификациялау міндетін шешу нәтижелері қарастырылған. Кең ауқымды есептеу эксперименттері барысында классификация тиімділігінің эмбеддинг параметрлерін және семантикалық траекторияларды сипаттамаларын таңдауға едәуір тәуелді екендігі анықталып, оңтайлы параметрлердің тілге байланысты өзгеретіні негізделген. Энтропия мен күрделілікті талдауға, сондай-ақ n-граммаларды кластерлеуге негізделген белгілер адам жазған мәтіндер мен автоматтандырылған жүйелер генерациялаған мәтіндерді тиімді ажыратуға мүмкіндік беретіні көрсетілген.

Ең жоғары нәтижелер әртүрлі тілдер үшін белгілер мен классификаторлардың әртүрлі комбинацияларын қолдану кезінде қол жеткізілетіні анықталып, бұл әмбебап модельдің жоқтығын дәлелдейді. Сонымен қатар, классификацияның жоғары дәлдігіне қол жеткізілген (бірқатар тілдер үшін F1-көрсеткіші 0,9-дан жоғары), соның ішінде тесттік таңдамадағы мәтіндер оқыту кезеңінде пайдаланылмаған модельдермен генерацияланған жағдайларда да. Ұсынылған тәсілдің боттар генерациялаған мәтіндерді тұрақты идентификациялауды қамтамасыз ететіні және жалпылау қабілетіне ие екендігі көрсетіліп, оның цифрлық контентті талдау және сүзгіден өткізу міндеттерінде қолданылу мүмкіндігі негізделген.

Қорытындыда жүргізілген зерттеудің негізгі нәтижелері тұжырымдалып, жұмыстың ғылыми және практикалық маңыздылығы туралы қорытындылар жасалған.

Қосымшада зерттеудің практикалық материалдары берілген.

Автор ғылыми жетекші – техника ғылымдарының кандидаты, қауымдастырылған профессор Ильдар Гусманович Курмашевке, сондай-ақ шетелдік кеңесші – физика-математика ғылымдарының докторы, профессор Василий Александрович Громовқа зерттеу жүргізу барысында көрсеткен бағалы көмегі, кеңестері мен қолдауы үшін шынайы алғыс білдіреді.