

ТАБИҒИ ТІЛДІ ӨНДЕУ ТЕХНОЛОГИЯЛАРЫНЫҢ ҚАЗІРГІ ЛИНГВИСТИКАДАҒЫ МАҢЫЗЫ

Жанкеева Ж.Д.

6B01701-«Қазақ тілі мен әдебиеті» ББ

3-курс студенті

zhankeevazhanerke27@gmail.com

Қ.Жұбанов атындағы Ақтөбе өңірлік университеті,
Қазақстан, Ақтөбе

Қазіргі ақпараттық қоғам жағдайында цифрлық технологиялардың қарқынды дамуы адам тілін автоматты түрде өңдеу мәселесін ерекше өзекті етті. Осы бағыттағы маңызды ғылыми салалардың бірі - табиғи тілді өңдеу технологиялары (Natural Language Processing – NLP). NLP – компьютердің адам тілін түсінуі, талдауы және өңдеуімен айналысатын жасанды интеллектінің негізгі бағыттарының бірі болып табылады. Бұл сала лингвистика, информатика, математика және когнитивтік ғылымдардың тоғысында қалыптасып, қазіргі таңда ғылыми-зерттеу мен практикалық қолдану аясында кеңінен дамуда.

Табиғи тілді өңдеу технологиялары мәтінді автоматты талдау, машиналық аударма, сөйлеуді тану, чат-боттар, ақпарат іздеу жүйелері және виртуалды көмекшілер сияқты көптеген интеллектуалды жүйелердің негізін құрайды. Соңғы жылдары нейрондық желілер мен үлкен тілдік модельдердің дамуы NLP саласының мүмкіндіктерін едәуір кеңейтіп, тілдік деректерді өңдеудің сапасын жаңа деңгейге көтерді. Нәтижесінде бұл технологиялар тек техникалық салада ғана емес, қазіргі лингвистика ғылымында да маңызды зерттеу нысанына айналды [1; 1-2-бб].

Қазіргі лингвистикада NLP технологиялары тілдің құрылымын, семантикасын және прагматикалық ерекшеліктерін автоматты түрде зерттеуге мүмкіндік береді. Сонымен қатар олар тілдік корпустар құру, мәтіндерді сандық форматта талдау және сирек қолданылатын тілдерді цифрландыру

бағыттарында тиімді құрал ретінде қолданылады. Әсіресе, қазақ тілі сияқты агглютинативті тілдерді автоматты өңдеу мәселесі ғылыми тұрғыдан ерекше қызығушылық тудырады. Себебі қазақ тілінің морфологиялық құрылымы күрделі болғандықтан, оны компьютерлік өңдеу арнайы алгоритмдер мен тілдік модельдерді қажет етеді.

Табиғи тілді өңдеудің негізгі бағыттары бірнеше деңгейден тұрады: морфологиялық талдау, синтаксистік талдау және семантикалық өңдеу. Бұл бағыттардың әрқайсысы тілдің белгілі бір құрылымдық қабатын зерттеуге арналған.

Морфологиялық талдау – сөздердің құрылымын, олардың грамматикалық формаларын және сөз тудыру ерекшеліктерін автоматты түрде анықтау процесі. Бұл бағытта компьютер сөздің түбірін, қосымшаларын, сөз табын және грамматикалық категорияларын айқындайды. Қазақ тілі агглютинативті тілдер қатарына жататындықтан, бір түбірге көптеген қосымшалардың жалғануы сөз формаларының күрделі жүйесін қалыптастырады. Мысалы, «жазушыларымыздан» сөзін талдау барысында жүйе түбірді, тәуелдік, көптік және септік жалғауларын ажыратуы қажет. Сондықтан қазақ тіліндегі NLP жүйелерінде морфологиялық анализаторлардың рөлі ерекше.

Синтаксистік талдау табиғи тілдегі сөйлемдердің грамматикалық құрылымын анықтауға бағытталған. Бұл кезеңде сөйлем мүшелері арасындағы байланыс, сөздердің орын тәртібі және синтаксистік қатынастар зерттеледі. Компьютерлік синтаксистік талдау сөйлемдегі бастауыш, баяндауыш, толықтауыш және анықтауыш сияқты құрылымдарды автоматты түрде анықтауға мүмкіндік береді. Қазіргі NLP жүйелерінде синтаксистік талдау мәтінді машиналық аударуда, сұрақ-жауап жүйелерінде және мәтін мазмұнын

түсінуде маңызды қызмет атқарады. Синтаксистік құрылымды дұрыс анықтау мәтіннің мағынасын дәл интерпретациялауға негіз болады.

Семантикалық өңдеу - тілдік бірліктердің мағынасын анықтаумен байланысты күрделі бағыттардың бірі. Бұл деңгейде компьютер тек сөздердің грамматикалық құрылымын ғана емес, олардың мағыналық байланысын да талдайды. Семантикалық өңдеудің негізгі мақсаты - мәтіндегі контексті түсіну және ақпараттың нақты мазмұнын анықтау. Мысалы, бір сөз түрлі контексте түрлі мағына беруі мүмкін. Осындай көпмағыналылықты шешу NLP жүйелерінің басты мәселелерінің бірі болып саналады. Қазіргі таңда семантикалық талдау ақпарат іздеу жүйелерінде, автоматты реферат жасауда, мәтіндік сараптамада және чат-боттарда кеңінен қолданылады.

Табиғи тілді өңдеу технологиялары бірнеше тарихи кезеңдер арқылы қалыптасты. Әрбір кезеңде қолданылған әдістер мен модельдер ғылыми-техникалық прогреске байланысты өзгеріп отырды. NLP технологияларының алғашқы кезеңі ережеге негізделген жүйелермен байланысты. Бұл бағыт XX ғасырдың ортасында қалыптаса бастады. Алғашқы компьютерлік лингвистикалық жүйелер адамның тілдік ережелерін формальды түрде сипаттауға негізделді. Зерттеушілер сөздердің грамматикалық құрылымын, синтаксистік байланыстарын және тілдік заңдылықтарды арнайы алгоритмдер арқылы бағдарламалауға тырысты. Мұндай жүйелерде әрбір тілдік құбылыс алдын ала жазылған ережелер көмегімен өңделді.

Ережеге негізделген модельдердің басты артықшылығы - олардың логикалық анықтығы мен тілдік құрылымды нақты сипаттай алуында болды. Алайда бұл тәсілдің бірқатар кемшіліктері де байқалды. Табиғи тілдің күрделілігі, көпмағыналылығы және контекстік ерекшеліктері ережелер санын шексіз көбейтуді талап етті. Нәтижесінде мұндай жүйелер үлкен көлемдегі мәтіндермен жұмыс істеуде тиімсіз бола бастады.

Статистикалық NLP технологиялары машиналық аударма жүйелерінің дамуына үлкен ықпал етті. Мысалы, алғашқы Google Translate жүйелері статистикалық аударма модельдеріне негізделді. Сонымен қатар Hidden Markov Model (HMM), Naive Bayes және n-gram модельдері мәтінді автоматты классификациялау мен сөйлеуді тану саласында кеңінен қолданылды. Бұл тәсілдер табиғи тілді өңдеу сапасын едәуір жақсартқанымен, мәтіннің терең мағынасын түсінуде белгілі бір шектеулерге ие болды [2; 42-б].

Соңғы онжылдықта NLP технологияларының дамуында нейрондық желілер ерекше орын алды. Терең оқыту (Deep Learning) әдістерінің пайда болуы табиғи тілді өңдеу сапасын жаңа деңгейге көтерді. Нейрондық желілер үлкен көлемдегі деректерді өздігінен үйрену арқылы тілдік құрылымдарды автоматты түрде анықтауға мүмкіндік берді. Әсіресе рекуррентті нейрондық желілер (RNN), ұзақ қысқа мерзімді жад желілері (LSTM) және attention механизмдері NLP саласында маңызды рөл атқарды. Қазіргі кезеңде NLP технологиялары тек техникалық құрал ретінде ғана емес, лингвистика ғылымының жаңа зерттеу әдісіне айналды. Әсіресе аз зерттелген тілдерді цифрландыру, ұлттық тілдік модельдер жасау және электрондық тіл корпустарын дамыту бағытында NLP технологияларының ғылыми маңызы артып келеді.

Қазіргі таңда табиғи тілді өңдеу (NLP) технологияларының ең кең таралған бағыттарының бірі — машиналық аударма жүйелері. Машиналық аударма - бір тілдегі мәтінді автоматты түрде екінші тілге аударуға арналған интеллектуалды технология. Бұл жүйелер халықаралық коммуникацияны жеңілдетіп, ақпарат алмасу үдерісін жеделдетуде маңызды рөл атқарады. Әсіресе жаһандану жағдайында көптілді ақпараттық кеңістіктің қалыптасуы машиналық аударма жүйелерінің маңызын арттырды.

Машиналық аударманың ең танымал платформаларының бірі - Google Translate жүйесі. Google компаниясы алғашында статистикалық модельдерді пайдаланғанымен, кейін Transformer архитектурасына негізделген нейрондық желілерге көшті. Қазіргі уақытта Google Translate жүзден астам тілді қолдайды және мәтіндік, дыбыстық, тіпті визуалды аудармаларды жүзеге асыра алады. Жүйе үлкен көлемдегі көптілді корпустар арқылы үйретілгендіктен, контекстік сәйкестікті анықтауда жоғары нәтижелер көрсетеді. Машиналық аударма жүйелері барлық тілдер үшін бірдей сапалы нәтиже бере бермейді. Әсіресе қазақ тілі сияқты морфологиялық құрылымы күрделі тілдерде аударма мәселелері жиі кездеседі. Қазақ тілі агглютинативті тіл болғандықтан, бір түбірге бірнеше қосымшаның жалғануы аударма жүйелеріне қиындық туғызады. Мысалы, сөздердің септелуі, жіктелуі және тәуелденуі контекстік мағынаның өзгеруіне әсер етеді. Сонымен қатар қазақ тіліндегі сапалы цифрлық корпустардың аздығы да машиналық аударманың даму қарқынын тежейді. Нейрондық модельдердің тиімді жұмыс істеуі үшін үлкен көлемдегі параллель мәтіндік базалар қажет. Алайда қазақ тіліндегі мұндай ресурстардың шектеулі болуы аударма жүйелерінің қателік деңгейін арттырады. Кейбір жағдайда қазақ тіліндегі сөйлем құрылымы ағылшын немесе орыс тілдерінің синтаксисіне бейімделіп, табиғи емес аудармалар пайда болады.

Ғалымдардың пікірінше, қазақ тіліндегі машиналық аударманы жетілдіру үшін ұлттық тілдік корпустарды дамыту, морфологиялық анализаторларды жетілдіру және арнайы қазақ тілдік модельдерін құру қажет. Бұл бағыттағы зерттеулер қазіргі таңда Қазақстандағы цифрлық лингвистика саласының маңызды міндеттерінің бірі болып саналады.

Қазіргі кезеңде мәтінді классификациялау үшін Support Vector Machine (SVM), Naive Bayes және нейрондық желілер сияқты алгоритмдер кеңінен

қолданылады. Терең оқыту модельдерінің дамуы мәтіндік деректерді өңдеу сапасын едәуір жақсартты. Әсіресе BERT және GPT секілді трансформер модельдері мәтіннің контекстік ерекшеліктерін ескеріп, классификация дәлдігін арттыруға мүмкіндік берді. Ақпарат іздеу жүйелері де NLP технологияларының маңызды қолданылу саласына жатады. Google, Yandex және Bing сияқты іздеу жүйелері пайдаланушы сұранысын түсіну үшін табиғи тілді өңдеу алгоритмдерін қолданады. Қазіргі іздеу жүйелері тек кілт сөздерді сәйкестендіріп қана қоймай, сұраныстың семантикалық мағынасын да талдай алады. Мысалы, пайдаланушы «Қазақстандағы ең ірі университет» деп сұраныс енгізсе, жүйе сөздердің жай ғана сәйкестігін емес, олардың мағыналық байланысын да есепке алады. Бұл NLP технологияларының семантикалық өңдеу деңгейінің дамығанын көрсетеді. Мұндай технологиялар ғылыми ақпарат іздеу, электрондық кітапханалар және цифрлық архивтерде де белсенді түрде енгізілуде.

Сөйлеуді тану технологиялары (Speech Recognition) - адамның ауызша сөзін автоматты түрде мәтінге айналдыруға арналған NLP бағыттарының бірі. Бұл технологиялар қазіргі цифрлық қоғамда кең таралып, адам мен компьютер арасындағы өзара әрекеттесудің жаңа формасын қалыптастырды. Сөйлеуді тану жүйелерінің алғашқы үлгілері шектеулі сөздік қорымен ғана жұмыс істесе, қазіргі нейрондық желілер негізіндегі модельдер күрделі сөйлеу үлгілерін жоғары дәлдікпен өңдей алады. Deep Learning технологиялары дыбыстық сигналдарды талдап, сөздерді автоматты түрде ажыратуға мүмкіндік берді. Қазіргі таңда Siri, Google Assistant, Alexa және Алиса сияқты дауыс көмекшілері NLP технологияларының ең танымал мысалдары болып табылады. Бұл жүйелер пайдаланушының дауыстық командасын түсініп, қажетті әрекеттерді орындай алады. Мысалы, ауа райын айту, хабарлама жіберу, музыка қосу немесе интернеттен ақпарат іздеу сияқты қызметтер

табиғи тіл арқылы жүзеге асырылады. Дауыс көмекшілерінің тиімді жұмыс істеуі үшін бірнеше технология қатар қолданылады: сөйлеуді тану, мәтінді түсіну, семантикалық талдау және жауап генерациясы. Яғни жүйе тек сөздерді танумен шектелмей, пайдаланушының ниетін де анықтауы қажет. Сөйлеуді мәтінге айналдыру технологиялары медицина, білім беру және медиа салаларында да кеңінен қолданылады. Мысалы, дәрігерлердің дауыстық жазбаларын автоматты түрде мәтінге айналдыру уақытты үнемдеуге мүмкіндік береді. Сонымен қатар онлайн білім беру платформаларында лекцияларды автоматты транскрипциялау студенттердің ақпаратты қабылдауын жеңілдетеді.

Сөйлеуді тану – қазіргі кезде жылдам дамып келе жатқан ақпараттық технологиялардың бірі. Сөйлеуді тану технологиясы соңғы жылдары машиналық оқыту алгоритмдерінің қарқынды дамуының арқасында айтарлықтай алға жылжыды [6; 73-б].

Қазақ тіліндегі сөйлеуді тану технологиялары әлі де бастапқы даму кезеңінде. Оның негізгі себептері - қазақ тіліндегі дыбыстық корпустардың аздығы және сөйлеу ерекшеліктерінің күрделілігі. Қазақ тіліндегі диалектілік айырмашылықтар, фонетикалық өзгерістер және еркін сөз тәртібі сөйлеуді тану жүйелеріне қосымша қиындықтар туғызады [5; 29-б]. Сөйлеуді автоматты тану жай ғана машинаның адам сөзін түсінетінін және адам дауысының тікелей компьютерге командалар беретінін, ал компьютердің танылған және өңделген дауысқа сәйкес командаларды орындайтынын, осылайша адам мен компьютер арасындағы интеллектуалды әрекетті жүзеге асыратынын білдіреді [5-6].

Соған қарамастан, соңғы жылдары қазақ тіліндегі дауыс технологияларын дамыту бағытында маңызды зерттеулер жүргізілуде. Ұлттық IT-компаниялар мен ғылыми орталықтар қазақ тілін танитын интеллектуалды

жүйелерді әзірлеуге кірісіп отыр. Бұл зерттеулер қазақ тілін цифрлық кеңістікте дамытуға және ұлттық технологиялық инфрақұрылымды нығайтуға үлкен мүмкіндік береді. Қазіргі цифрлық қоғам жағдайында ұлттық тілдердің ақпараттық кеңістікте толыққанды қызмет атқаруы табиғи тілді өңдеу (NLP) технологияларының даму деңгейімен тікелей байланысты. Осы тұрғыдан алғанда қазақ тілін цифрландыру мәселесі қазіргі тіл біліміндегі маңызды ғылыми бағыттардың біріне айналды. Қазақ тіліндегі NLP жүйелерін дамыту үшін ең алдымен сапалы тілдік корпустар мен цифрлық ресурстардың болуы аса маңызды.

Қазақ тіліндегі цифрлық ресурстар қатарына электрондық сөздіктер, морфологиялық анализаторлар, синтаксистік талдау жүйелері және мәтіндік деректер базалары жатады. Соңғы жылдары қазақ тіліндегі онлайн сөздіктер мен автоматты аударма жүйелері де қарқынды дами бастады. Сонымен қатар ғылыми орталықтар қазақ тіліндегі мәтіндерді таңбалау (annotation) жұмыстарымен айналысып келеді. Алайда қазақ тіліндегі цифрлық ресурстардың көлемі әлі де жеткіліксіз. Әсіресе үлкен көлемдегі таңбаланған мәтіндік корпустардың аз болуы нейрондық модельдерді үйрету процесін қиындатады. Қазіргі трансформер модельдері миллиардтаған сөздерден тұратын деректер қорын қажет етеді. Сондықтан қазақ тіліндегі сапалы цифрлық контенттің тапшылығы NLP технологияларының дамуын баяулататын негізгі факторлардың бірі болып отыр. Сонымен қатар қазақ тіліндегі ғылыми мәтіндердің, терминологиялық базалардың және академиялық ресурстардың жеткіліксіздігі де маңызды мәселе болып саналады. Қазақ тілінің ғылыми стильдегі мәтіндері көбейген сайын NLP жүйелерінің сапасы да арта түседі. Осы себепті қазіргі таңда цифрлық тілдік ресурстарды көбейту мемлекеттік тіл саясатының маңызды бағыттарының бірі ретінде қарастырылуда. Цифрлық кеңістіктегі қазақ тілінің үлесін арттыруда

ең басты кедергілердің бірі – сапалы контент тапшылығы. Әлеуметтік желілерде қазақ тілінде жүргізілетін блогтар мен арналар көбейіп келеді, бірақ олардың басым бөлігі әлі де коммерциялық тұрғыдан тұрақсыз [4; 3-6].

Соңғы жылдары жасанды интеллект технологияларының дамуы қазақ тіліндегі NLP жобаларының болашағына жаңа мүмкіндік ашты. Әлемдік деңгейде ірі тілдік модельдердің (Large Language Models - LLM) дамуы ұлттық тілдерге арналған интеллектуалды жүйелерді құруға жол ашып отыр. Осыған байланысты қазақ тіліндегі цифрлық технологияларды жетілдіру мәселесі стратегиялық маңызға ие болуда. Қазақ тіліндегі NLP жобаларының болашағы ең алдымен ұлттық тілдік модельдерді дамытуға байланысты. Ұлттық тілдік модельдер белгілі бір тілдің грамматикалық, семантикалық және мәдени ерекшеліктерін ескере отырып жасалады. Мұндай модельдер қазақ тіліндегі мәтіндерді түсіну, генерациялау және автоматты аудару сапасын арттыруға мүмкіндік береді. Қазіргі таңда әлемде GPT, BERT, Gemini сияқты ірі тілдік модельдер кеңінен қолданылуда. Алайда бұл модельдердің басым бөлігі ағылшын тіліндегі деректер негізінде оқытылған. Сондықтан қазақ тіліндегі контексті толық түсіну деңгейі әлі де шектеулі болып отыр. Осы себепті қазақ тіліне бейімделген ұлттық нейрондық модельдерді құру өзекті ғылыми міндеттердің бірі саналады. Ұлттық тілдік модельдерді дамыту тек лингвистика үшін ғана емес, білім беру жүйесі үшін де маңызды. Қазіргі таңда электрондық оқыту платформалары, интеллектуалды тест жүйелері және автоматты бағалау технологиялары NLP әдістеріне негізделіп жұмыс істейді. Қазақ тіліндегі сапалы NLP жүйелері білім беру ресурстарын мемлекеттік тілде дамытуға мүмкіндік береді. Сонымен қатар NLP технологиялары ғылым саласында да үлкен рөл атқарады. Ғылыми мәтіндерді автоматты сараптау, академиялық ақпарат іздеу және ғылыми деректерді жүйелеу процестері

қазіргі зерттеулердің тиімділігін арттырады. Қазақ тіліндегі ғылыми контенттің көбеюі ұлттық ғылымның цифрлық кеңістіктегі орнын нығайтады.

Болашақта қазақ тіліндегі NLP технологияларының дамуы мемлекеттік қолдау мен ғылыми инфрақұрылымның сапасына тікелей байланысты болады. Ұлттық тілдік корпустарды кеңейту, ашық деректер платформаларын құру және жасанды интеллект саласындағы ғылыми зерттеулерді қаржыландыру қазақ тілінің цифрлық трансформациясын жеделдетеді. Осылайша, табиғи тілді өңдеу технологиялары қазақ тіл білімінің жаңа ғылыми бағытын қалыптастырып отыр. Бұл технологиялар қазақ тілінің цифрлық кеңістіктегі қолданылу аясын кеңейтіп қана қоймай, ұлттық мәдениеттің заманауи технологиялармен интеграциялануына мүмкіндік береді.

Қазіргі ақпараттық қоғам жағдайында табиғи тілді өңдеу (Natural Language Processing – NLP) технологиялары лингвистика ғылымының маңызды әрі қарқынды дамып келе жатқан бағыттарының біріне айналды. Жасанды интеллект пен компьютерлік лингвистиканың тоғысында қалыптасқан бұл сала адам тілін автоматты түрде өңдеу, талдау және интерпретациялау мүмкіндіктерін кеңейтіп отыр. Әсіресе, үлкен тілдік модельдердің дамуы табиғи тілдің контекстік және семантикалық ерекшеліктерін терең түсінуге мүмкіндік берді. Қазақ тілін оқытуда жасанды интеллектіні (ЖИ) қолдану мақсаты – сандық қоғамдағы мемлекеттік тілді тиімді үйретудің жолы ретінде қарастыру. Жасанды интеллект әдетте адамның интеллектуалды қабілетін қажет ететін және жүйелерді құрумен айналысатын технологиялық құрал. Жасанды интеллект оқытуды, логиканы, жоспарлауды, сөйлеуді тануды, қабылдауды және есептерді шешуді имитациялауға және талдауға тырысады [3; 47-б].

Қазақ тіл біліміндегі NLP технологияларының дамуы ұлттық тілдің цифрлық кеңістіктегі болашағымен тікелей байланысты. Қазақ тілінің

агглютинативті құрылымы мен морфологиялық күрделілігі табиғи тілді автоматты өңдеуде белгілі бір қиындықтар туғызғанымен, қазіргі ғылыми зерттеулер бұл мәселелерді кезең-кезеңімен шешуге мүмкіндік беріп отыр. Қазақ тіліндегі ұлттық корпустарды қалыптастыру, морфологиялық анализаторларды жетілдіру және арнайы тілдік модельдерді әзірлеу бағытындағы жұмыстар мемлекеттік тілдің цифрлық әлеуетін арттыруға ықпал етеді. NLP технологиялары қазақ тілінің ғылыми, білім беру және ақпараттық кеңістіктегі қолданыс аясын кеңейтуге мүмкіндік береді. Сондықтан табиғи тілді өңдеу саласындағы ғылыми зерттеулерді дамыту, цифрлық ресурстарды көбейту және халықаралық тәжірибелерді енгізу қазақ тіл білімінің болашақтағы дамуы үшін аса маңызды бағыт болып табылады.

Пайдаланылған әдебиеттер тізімі

1. NLP дегеніміз не? Бұл қалай жұмыс істейді, артықшылықтар, қиындықтар, мысалдар <https://kk.shaip.com/blog/what-is-nlp-how-it-works-benefits-challenges-examples/> Маусым 29, 2025
2. Оразбек Қ.С., Ернияз Б.З. Қазақ тіліндегі мәтіндерді талдауға арналған машиналық оқыту әдістері. Оңтүстік Қазақстан ғылым Жаршысы, 2025 42-бет
3. Асамбаев А. Ж. Жасанды интеллект негіздері: Оқулық. Алматы, ЖШС РПБК «Дәуір» 2011 ж. – 136 б.
4. «Жетісу-Ақпарат» Қазақ тілінің цифрлық кеңістіктегі орны мен болашағы <https://7-su.kz/janalyqtar/kuenderek/qazaq-tilinin-czifrlыq-kenistiktegi-orny-men-bolashagy-67552/>
5. M.Y. Alva, M. Nachamai, J. Paulose // A comprehensive survey on features and methods for speech emotion detection 2015 IEEE international conference on electrical, computer and communication technologies (ICECCT), IEEE (2015), pp. 1-6.



6. Kalimoldayev M., Mamyrbayev O., Mekebayev N., Kydyrbekova, A. (2020). Algorithms for detection gender using neural networks // International Journal of Circuits, Systems and Signal Processing. 2020, 154-159.