Қазақстан Республикасы Ғылым және жоғары білім министрлігі

әл-Фараби атындағы Қазақ ұлттық университеті

Матанов Н.Б

МАШИНАЛЫҚ ОҚЫТУ НЕГІЗІНДЕ ҚАЗАҚ ТІЛІНІҢ СИНОНИМДЕРІН ҚҰРАСТЫРУ МОДЕЛЬДЕРІН ЗЕРТТЕУ ЖӘНЕ ӘЗІРЛЕУ

ДИПЛОМДЫҚ ЖҰМЫС

Мамандығы 7М0101 – «Есептеуіш лингвистика»

Қазақстан Республикасы Ғылым және жоғары білім министрлігі

әл-Фараби атындағы Қазақ ұлттық университеті

Ақпараттық технологиялар факультеті

Ақпараттық жүйелер кафедрасы

МАГИСТРЛІК ДИССЕРТАЦИЯ

тақырыбы: «Машиналық оқыту негізінде қазақ тілінің синонимдерін құрастыру модельдерін зерттеу және әзірлеу»

7М06101 – «Есептеуіш лингвистика»

Орындаған _____ (қолы) Матанов Н.Б

Ғылыми жетекші _____ (қолы) Рахимова Д.Р
Phd, аға оқытушы

Қорғауға жіберілді:

Хаттама № 10 , «31» 05 2024ж.

Кафедра меңгерушісі _____ (қолы және мөрі) Мусиралиева Ш.Ж.

Норма бақылаушы _____ (қолы) Мухитова А.А.

Алматы, 2024

ТҮЙІНДЕМЕ

Жұмыстың өзектілігі: Қазақ синонимдерін құру үшін машиналық оқыту мүмкіндіктерін зерделеу қазақ тілі үшін табиғи тілді өңдеу саласындағы зерттеулердің жеткіліксіздігін ескере отырып, маңызды және уақтылы міндет болып табылады. Кластерлеу және машинамен оқыту алгоритмдерін пайдалана отырып, синонимдік компиляторды құру тілдік ресурстарды байытады, тілде сөйлеушілер арасындағы коммуникацияны жақсартады, лингвистикалық зерттеулерді қолдайды және маңызды технологиялық сын-қатерлерді шешеді. Мұндай құрал өзара қарым-қатынас пен түсіністікті едәуір жақсартады, сондай-ақ тілтанушыларға сандық дәуірде мәдени және тілдік бірегейлікті сақтауға және дамытуға ықпал ете отырып, сөздік қорын кеңейтуге көмектеседі.

Жұмыстың мақсаты: Бұл жұмыстың мақсаты қазақ тілінің синонимдік компиляторын жасау үшін машиналық оқыту негізінде алгоритм әзірлеу болып табылады. Бұл қазақ сөздері мен олардың мағыналарының деректер жиынтығын жинау мен өңдеуді, ұқсас сөздерді санатқа топтастыру үшін кластерлеу алгоритмдерін қолдануды және әрбір сөз үшін синонимдер тізімін жасауды қамтиды. Негізгі міндет деректер жинағының ауытқуына және оңтайлы алгоритмді таңдауға байланысты проблемаларды еңсеру болып табылады, бұл тілдік ресурстарды жақсартуға, лингвистикалық зерттеулерді қолдауға және қазақ тілінде сөйлеушілер арасында коммуникацияны жақсартуға мүмкіндік береді.

Зерттеу міндеттері: Машиналық оқыту негізінде қазақ тілінің синонимдерін құрастыру жүйесін ұсыну; Қазақ тілін оқытудың жаңа бағыт беру; Әдістер нәтижелерінің ерекшеліктері мен мүмкіндіктерін талдау; Плагиатка тексеру жүйесін жетілдіру.

Зерттеу пәні: Машиналық аударма технологиялары мен табиғи тілдер.

Зерттеу нысаны: Зерттеу теориялық талдау, алгоритмдерді әзірлеу, оларды іске асыру және тестілеу кезеңдерін, сондай-ақ алынған нәтижелерді бағалауды және оларды нақты жағдайларда қолдануды қоса алғанда, эксперименттік-талдамалық нысанда жүргізіледі.

Зерттеу әдістері: Машиналық оқыту әдістері және нейрондық желілер

Зерттеудің маңыздылығы: Зерттеу қазақ тілінің тілдік ресурстарын жақсарту және оны цифрлық дәуірде қолдануды қолдау үшін маңызды мәнге ие. Машиналық оқыту негізінде синонимдік компиляторды құру сөздік қорын кеңейтуге, тілде сөйлеушілер арасындағы коммуникация мен өзара түсіністікті жақсартуға ықпал етеді, сондай-ақ мәдени және тілдік ұқсастықты сақтау үшін маңызды болып табылатын лингвистикалық зерттеулерді қолдайды.

Алғыс. Бұл зерттеу жұмысы Қазақстан Республикасы Ғылым және жоғары білім министрлігінің **AP19677835** «Қазақстан Республикасының заңнамасы саласындағы мемлекеттік тілге арналған семантикалық тәсілдерге негізделген интеллектуалды сұрақ-жауап жүйесінің үлгілерін зерттеу және әзірлеу» ғылыми гранттық жобасы аясында жүргізілді.

РЕФЕРАТ

Актуальность работы: Изучение возможностей машинного обучения для создания казахских синонимов является важной и своевременной задачей, учитывая отсутствие исследований в области обработки естественного языка для казахского языка. Создание компилятора синонимов с использованием алгоритмов кластеризации и машинного обучения обогащает лингвистические ресурсы, улучшает общение между носителями языка, поддерживает лингвистические исследования и решает важные технологические проблемы. Такой инструмент значительно улучшает общение и понимание, а также помогает лингвистам расширить словарный запас, способствуя сохранению и развитию культурной и языковой идентичности в эпоху цифровых технологий.

Цель работы: Целью данной работы является разработка алгоритма на основе машинного обучения для создания компилятора синонимов казахского языка. Это включает в себя сбор и обработку наборов данных казахских слов и их значений, использование алгоритмов кластеризации для группировки похожих слов по категориям и создание списка синонимов для каждого слова. Основная задача – преодоление проблем, связанных с отклонением набора данных, и выбор оптимального алгоритма, позволяющего улучшить языковые ресурсы, поддержать лингвистические исследования и улучшить общение среди носителей казахского языка.

Задачи исследования: предложить систему составления синонимов казахского языка на основе машинного обучения; Обеспечение нового направления преподавания казахского языка; Анализ особенностей и возможностей результатов методов; Совершенствование системы проверки на плагиат.

Предмет исследования: Технологии машинного перевода и естественные языки.

Объект исследования: Исследование проводится в экспериментально-аналитической форме, включая теоретический анализ, разработку алгоритмов, этапы их реализации и тестирования, а также оценку полученных результатов и их применение в реальных ситуациях.

Методы исследования: методы машинного обучения и нейронные сети.

Значимость исследования: Исследование важно для улучшения языковых ресурсов казахского языка и поддержки его использования в эпоху цифровых технологий. Создание компилятора синонимов на основе машинного обучения способствует расширению словарного запаса, улучшению общения и взаимопонимания между носителями языка, а также поддерживает лингвистические исследования, что важно для сохранения культурного и языкового сходства.

Благодарность. Данное исследование проводилось в рамке научного грантового проекта **AP19677835** «Исследование моделей и разработка интеллектуальной вопросно-ответной системы на основе семантических подходов для государственного языка в сфере законодательства Республики Казахстан» Министерства науки и высшего образования Республики Казахстан.

ABSTRACT

Relevance of the work: Studying the possibilities of machine learning for creating Kazakh synonyms is an important and timely task, given the lack of research in the field of natural language processing for the Kazakh language. Building a synonym compiler using clustering and machine learning algorithms enriches linguistic resources, improves communication between native speakers, supports linguistic research, and solves important technological problems. Such a tool greatly improves communication and understanding, and helps linguists expand their vocabulary, helping to maintain and develop cultural and linguistic identity in the digital age.

The purpose of the work: The purpose of this work is to develop an algorithm based on machine learning to create a compiler of synonyms for the Kazakh language. This involves collecting and processing datasets of Kazakh words and their meanings, using clustering algorithms to group similar words into categories, and creating a list of synonyms for each word. The main objective is to overcome the problems associated with dataset rejection and select the optimal algorithm to improve language resources, support linguistic research and improve communication among Kazakh speakers.

Research objectives: to propose a system for compiling synonyms for the Kazakh language based on machine learning; Providing a new direction for teaching the Kazakh language; Analysis of the features and capabilities of the results of the methods; Improving the plagiarism checking system.

The subject of the research: Machine translation technologies and natural languages.

The object of the research: The research is carried out in an experimental-analytical form, including theoretical analysis, development of algorithms, stages of their implementation and testing, as well as evaluation of the results obtained and their application in real situations.

Research methods: machine learning methods and neural networks.

Significance of the research: The study is important for improving the linguistic resources of the Kazakh language and supporting its use in the digital age. Creating a synonym compiler based on machine learning helps expand vocabulary, improve communication and understanding between native speakers, and also supports linguistic research, which is important for preserving cultural and linguistic similarities.

Acknowledgments. This study was conducted within the framework of the scientific grant project **AP19677835** “Research of models and development of an intelligent question-answer system based on semantic approaches for the state language in the field of legislation of the Republic of Kazakhstan” of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР

NLP
UML
ML
KZ
LSTM
BERT
NLLB
BRNN
RNN

МАЗМҰНЫ

КІРІСПЕ

Natural Language Processing (NLP) адам тілдерін автоматтандырылған өңдеуге мүмкіндік беру арқылы лингвистикалық зерттеулерде төңкеріс жасады. Бұл жұмыс қазақ тіліндегі синонимдерді құрастыру аясында NLP саласына тереңірек үңіледі. Аз зерттелген тіл ретінде қазақ тілі өзінің ерекше тілдік ерекшеліктеріне байланысты NLP үшін ерекше қиындықтар туғызады. Қазақ тілін цифрландыру және талдау жұмыстары үздіксіз жүргізілуде, бірақ синонимдерді жинақтау күрделі мәселе болып қала береді. NLP жүйесіндегі бар синонимдерді құрастыру әдістерін қарастыру және қазақша синонимдерді құрастыру үшін арнайы әзірленген NLP үлгілеріне салыстырмалы талдау жүргізу арқылы бұл зерттеу осы саладағы әртүрлі әдіснамалардың тиімділігіне жарық түсіруді мақсат етеді. Бұл зерттеудің салдары NLP-ді қазақ тілі сияқты аз зерттелген тілдер үшін қалай оңтайландыруға болатыны туралы түсініктерді ұсынатын NLP-тің кең саласына таралады. Қазақ тіліндегі құрастырылған синонимдік дерекқорлардың әлеуетті қолданбаларын зерттей отырып, бұл зерттеу NLP зерттеулері мен қазақ тілдік ландшафтындағы қолданбалардағы болашақ жетістіктерге жол ашады.

Бұл мақала ана тілінде сөйлейтіндер мен тіл үйренушілер үшін қарым-қатынасты жақсарту және сөздік қорын кеңейту мақсатында қазақ тілінде синонимдерді жасау үшін машиналық оқыту саласын зерттейді. Машиналық оқыту лингвистикада кеңірек тараған сайын, оның ағымдағы қолданбаларын, қиындықтарын және ықтимал үлестерін түсіну өте маңызды. Қазақ тілінің бірегей тілдік ерекшеліктері машиналық оқытуды қолдануда белгілі бір кедергілер мен мүмкіндіктер туғызып, жеке тәсілдердің қажеттілігін көрсетеді. Қазақ тілінің бар лингвистикалық ресурстарын зерттей отырып және синонимдерді құрастыру үлгілерін жасаудың әртүрлі әдістемелерін зерттей отырып, бұл зерттеу қазақ тілін өңдеуді жақсартуда машиналық оқытудың толық әлеуетін ашуға бағытталған. Осы аспектілерді жан-жақты талдау арқылы бұл зерттеу қазақ тіліндегі синонимдік өндіріске арналған машиналық оқыту үлгілерінің дамуы мен маңызын ашып көрсетуге бағытталған, бұл сайып келгенде тілдік қабілеттерді арттыруға және мәдениетаралық түсінушілікке жол ашады.

Жұмыстың өзектілігі: Қазақ синонимдерін құру үшін машиналық оқыту мүмкіндіктерін зерделеу қазақ тілі үшін табиғи тілді өңдеу саласындағы зерттеулердің жеткіліксіздігін ескере отырып, маңызды және уақтылы міндет болып табылады. Кластерлеу және машинамен оқыту алгоритмдерін пайдалана отырып, синонимдік компиляторды құру тілдік ресурстарды байытады, тілде сөйлеушілер арасындағы коммуникацияны жақсартады, лингвистикалық зерттеулерді қолдайды және маңызды технологиялық сын-қатерлерді шешеді. Мұндай құрал өзара қарым-қатынас пен түсіністікті едәуір жақсартады, сондай-ақ тілтанушыларға сандық

дәуірде мәдени және тілдік бірегейлікті сақтауға және дамытуға ықпал ете отырып, сөздік қорын кеңейтуге көмектеседі.

Жұмыстың мақсаты: Бұл жұмыстың мақсаты қазақ тілінің синонимдік компиляторын жасау үшін машиналық оқыту негізінде алгоритм әзірлеу болып табылады. Бұл қазақ сөздері мен олардың мағыналарының деректер жиынтығын жинау мен өңдеуді, ұқсас сөздерді санатқа топтастыру үшін кластерлеу алгоритмдерін қолдануды және әрбір сөз үшін синонимдер тізімін жасауды қамтиды. Негізгі міндет деректер жинағының ауытқуына және оңтайлы алгоритмді таңдауға байланысты проблемаларды еңсеру болып табылады, бұл тілдік ресурстарды жақсартуға, лингвистикалық зерттеулерді қолдауға және қазақ тілінде сөйлеушілер арасында коммуникацияны жақсартуға мүмкіндік береді.

Зерттеу міндеттері: Машиналық оқыту негізінде қазақ тілінің синонимдерін құрастыру жүйесін ұсыну; Қазақ тілін оқытудың жаңа бағыт беру; Әдістер нәтижелерінің ерекшеліктері мен мүмкіндіктерін талдау; Плагиатка тексеру жүйесін жетілдіру.

Зерттеу пәні: Машиналық аударма технологиялары мен табиғи тілдер.

Зерттеу нысаны: Зерттеу теориялық талдау, алгоритмдерді әзірлеу, оларды іске асыру және тестілеу кезеңдерін, сондай-ақ алынған нәтижелерді бағалауды және оларды нақты жағдайларда қолдануды қоса алғанда, эксперименттік-талдамалық нысанда жүргізіледі.

Зерттеу әдістері: Машиналық оқыту әдістері және нейрондық желілер

Зерттеудің маңыздылығы: Зерттеу қазақ тілінің тілдік ресурстарын жақсарту және оны цифрлық дәуірде қолдануды қолдау үшін маңызды мәнге ие. Машиналық оқыту негізінде синонимдік компиляторды құру сөздік қорын кеңейтуге, тілде сөйлеушілер арасындағы коммуникация мен өзара түсіністікті жақсартуға ықпал етеді, сондай-ақ мәдени және тілдік ұқсастықты сақтау үшін маңызды болып табылатын лингвистикалық зерттеулерді қолдайды.

Жұмыстың күтілетін нәтижелері:

1.Машиналық оқыту және жақындықты анықтау бойынша әдебиеттерді талдай отырып қазақ тілінде парафрейз жүйесін жасау жәнеде керекті модельдерін жасап яғни оқытып олардың қазақ тілінің сөздеріне тиімді түрде синонимдер ұсына алатынын талдау.

2.Синонимдердің дәлдігін және әртүрлілігін арттыру, үлкен мәтіндік корпустарды пайдалану және алгоритмдерді жетілдіру арқылы.

3. Сөйлемнен жаңа сөйлем жасайтын парафрейз жасайтын модельді құру.

4. Қазақша ресурстарды Python тілі арқылы жинау.

5. Қазақ тілін пайдалана отырып сөздердің жақындығын есептейтін модельдерді зерттеу.

Диссертациялық жұмыстың құрылымы: Зерттеу жұмысының құрылымы кіріспеден, 3 тараудан құралған негізгі бөлімнен, қорытынды және пайдаланылған әдебиеттер тізімінен тұрады.

Кіріспеде: жұмыстың мақсаты, міндеттері, өзектілігі, зерттеу жұмысының әдістері, мазмұны, пәні, нысаны, практикалық және теориялық маңыздылығы туралы ғылыми мәліметтер берілген.

Негізгі бөлімде зерттеу жұмысының тақырыбына сәйкес еңбектермен әдебиеттерге шолу жасалынды, синонимдарды табу мәселелеріне талдау жасалды. Сөздердің жақындығын және қаншалықты синоним екендігін анықтайтын алгоритмдер сипатталды, қазақ тілінде парафрейз жасайтын модель жобаланып құрылды. Тілдік ресурстарды жинақтау жолдары қарастырылып, бағдарлама көмегімен жүргізілген тәжірибелердің нәтижелері аналитикалық тұрғыдан сараланды.

Қорытынды бөлімде магистрлік диссертацияны орындау кезінде орындалған ғылыми-зерттеу жұмыстарының тәжірибесі мен нәтижелері қамтылады.

1 ЗЕРТТЕУ ЖҰМЫСЫ БОЙЫНША ӘДЕБИЕТТЕРГЕ ШОЛУ

1.1 Лингвистикадағы машиналық оқытуға шолу

Лингвистика саласы машиналық оқыту әдістерін біріктіру арқылы айтарлықтай прогреске қол жеткізді. Бұл зерттеу мақаласы қазақ тіліндегі синонимдерді жасау үшін машиналық оқыту үлгілерін әзірлеуге бағытталған және лингвистикалық талдауды есептеу алгоритмдерімен біріктіретін жаңа тәсілді ұсынады. Машиналық оқыту әртүрлі салаларда төңкеріс жасауды жалғастыруда, оны лингвистикада қолдану тілдерді өңдеу мүмкіндіктерін жақсарту үшін жаңа мүмкіндіктер ашады. Дегенмен, машиналық оқытуды лингвистикалық тапсырмаларға, әсіресе қазақ тілі сияқты бірегей ерекшеліктері бар тілдерге бейімдеуде қиындықтар әлі де бар. Өзінің ерекше белгілерімен танымал қазақ тілі тілдік ресурстардың шектеулі болуына байланысты машиналық оқыту қолданбаларына кедергі жасайды. Қазақ тіліне арнайы әзірленген машиналық оқыту үлгілерінің дамуын зерттей отырып, бұл зерттеу осы мәселелерді шешуге және лингвистикалық ресурстарды дамытуға үлес қосуға бағытталған. Қолданыстағы ресурстарды, деректерді жинау әдістерін және осы үлгілердің тілді өңдеуге күтілетін әсерін зерттей отырып, бұл мақала қазақ тіліндегі контекстте тілге қатысты тиімдірек және тиімді қолданбаларға жол ашуға бағытталған.

Табиғи тілді өңдеу (NLP) информатика, жасанды интеллект және лингвистиканың қиылысында орналасқан, адам қарым-қатынасы мен компьютерді түсіну арасындағы алшақтықты жоюға бағытталған. Есептеу әдістерін қолдану арқылы NLP адам тілін компьютерге қолжетімді түрде талдауға, түсінуге және түсіндіруге тырысады [1]. Тіл біліміндегі негізін ескере отырып, NLP адамдар сөйлесу үшін қолданатын табиғи тілді түсінумен ерекше айналысады, ауызша және жазбаша формаларды қамтиды [2]. Бұл процесс компьютерлік бағдарламаларға адам тілін дәл түсіндіруге мүмкіндік беру үшін статистиканы, машиналық оқытуды және терең оқытуды қоса алғанда, әртүрлі модельдер мен тәсілдерді қамтиды. Негізгі мақсат - компьютерлер тілде берілген нақты мазмұнды түсініп қана қоймай, сонымен қатар ауызша немесе жазбаша сөздің астарындағы ниеттер мен эмоцияларды анықтау, осылайша адамдар мен машиналар арасындағы интуитивті өзара әрекеттесуді жеңілдету [1].

NLP-тің пайда болған сәттен бастап қазіргі күйіне дейінгі дамуы, әсіресе оның әртүрлі тілдерде қалай қолданылғаны туралы қызықты эволюцияны көрсетеді. Бастапқыда бұл өріс негізінен ережелерге негізделген жүйелерге сүйенді, бұл үрдіс 1950-1990 жылдар аралығында басым болды. Тіл мамандары мұқият әзірлеген бұл жүйелер алдын ала анықталған ережелер жиынтығы арқылы бірнеше тілді түсіну және өңдеу үшін негіз қалады. NLP-ті тілдерде қолданудағы ең алғашқы кезеңдердің бірі 1954 жылы Джорджтаун-IBM эксперименті болды, ол 60 орыс сөйлемін

автоматты түрде ағылшын тіліне аудару арқылы машиналық аударманың әлеуетін сәтті көрсетті. Бұл эксперимент машиналық аударманың орындылығын көрсетіп қана қоймай, сонымен қатар тіл кедергілерін жоюдағы NLP маңыздылығына баса назар аудара отырып, осы саладағы одан әрі зерттеулер мен әзірлемелерді жандандырды [2]. Қазіргі уақытта Google Translate сияқты NLP қолданбалары айтарлықтай жетілдіріліп, көптеген тілдерде аудармаларды қамтамасыз ету үшін күрделі NLP құралдарын қолдана отырып, жаһандық коммуникацияны бұрынғыдан да қолжетімді етеді [3]. Негізгі ережелерге негізделген жүйелерден әртүрлі тілдердің нюанстарын өңдеуге қабілетті жетілдірілген алгоритмдерге дейінгі бұл эволюция NLP-тің адам тілін түсіну және аударуда жасаған маңызды қадамдарын көрсетеді.

NLP-тің қазақ тілі үшін маңыздылығы тілдік күрделілік пен ресурс тапшылығын қарастыра отырып, негізгі тілді өңдеуден айтарлықтай асып түседі. Қазақ тілі өзінің күрделі синтаксисі мен бай дауыстылығымен ерекше қиындықтар туғызады, оны NLP әдістері шешуге өте ыңғайлы. Мысалы, зерттеушілер қазақ тіліндегі мәтіндерді талдауды түбір сөздерге, тілдің күрделілігін өңдеуді жеңілдететін стратегияға, әсіресе жұрнақтардың мағынаны күрт өзгерте алатынын ескере отырып, жеңілдете алады. Бұл тәсіл өте маңызды, өйткені қазақ тілі әр түрлі мағынаны беру үшін жұрнақтарға тәуелділігімен сипатталады, дәл лингвистикалық үлгілерді жасауды қиындатады [25]. Сонымен қатар, бағдарламалық қосымшалардағы қазақ тіліне арналған лингвистикалық ресурстар мен құралдардың тапшылығы қазақ тілінде сөйлейтін аймақтардағы технологиялық прогреске айтарлықтай кедергі болып табылады. NLP технологиялары Интернетте қолжетімді ақпараттың үлкен көлемін өңдеу және түсіну үшін маңызды болып табылатын Уикипедиядан қазақ тіліндегі атаулар үшін сөйлем векторларын алу сияқты тілдерді өңдеу құралдарын әзірлеуді жеңілдету арқылы осы қиындықтардың шешімін ұсынады. Бұл мүмкіндік лингвистикалық ресурстары шектеулі деп есептелетін қазақ тілі үшін өте маңызды болып табылады, мұндай технологиялық жетістіктер тілдік және мәдени сақтауды ынталандыру және қазақтілді халықтың ана тілінде ақпарат алуға және шығаруына мүмкіндік беру үшін маңызды [25].

Қазіргі уақытта лингвистикада машиналық оқытудың қолданылуы

Жасанды интеллект саласы болып табылатын машиналық оқыту лингвистика саласында бұрыннан келе жатқан мәселелердің инновациялық шешімдерін ұсынатын маңызды орын алды. Табиғи тілді өңдеу (NLP) мәселелеріне заманауи машиналық оқыту әдістерін қолдану арқылы зерттеушілер мен технологтар бұрын қиын немесе қол жетімсіз болған тәсілдермен адам тілін талдап, түсініп және жасай алады. Бұл қолданба

лингвистикадағы классикалық мәселелерді шешуде машиналық оқытудың әмбебаптығы мен күшін бейнелейтін сөйлеуді тану мен машиналық аудармадан бастап сезімді талдау мен мәтінді қорытындылауға дейінгі тапсырмалардың кең ауқымын қамтиды. Машиналық оқыту мен лингвистиканың тоғысуы компьютерлердің адам тілін түсіну және өзара әрекеттесу мүмкіндігін кеңейтіп қана қоймайды, сонымен қатар лингвистерге тіл заңдылықтары мен құрылымдарын талдаудың жаңа құралдарын ұсынады, сол арқылы лингвистикалық зерттеулерде мүмкін болатын нәрселердің шекарасын кеңейтеді.

Қазақ тілінің күрделі морфологиясының қиындықтарын ескере отырып, қазіргі NLP әдістерінің шектеулері айқынырақ болады. Мысалы, токенизация, мәтінді мағыналы элементтерге немесе лексемаларға бөлетін негізгі NLP процесі, жұрнақтар сөздердің мағынасында маңызды рөл атқаратын қазақ тілі сияқты тілдер үшін өте маңызды. Дегенмен, лексемалаудың тиімділігіне тілдің морфологиялық күрделілігі кедергі келтіруі мүмкін, өйткені бұл процесс сөзжасамдық және туынды мағыналық реңктерді дәл түсіре алмауы мүмкін. Сөздерді түбір түріне келтіруге бағытталған септеу және лемматизация, әдістер туралы сөз болғанда бұл мәселе одан әрі күрделене түседі [4]. Қазақ тілінің бай морфологиялық құрылымы сөз формаларының елеулі түрленуіне мүмкіндік береді, бұл әдістерге айтарлықтай қиындық туғызады. Олар тілді тым жеңілдетуі мүмкін, бұл сыни мағынаны жоғалтуға әкеледі, немесе оны жеткілікті түрде жеңілдете алмайды, нәтижесінде талдаудың тиімсіз болуы мүмкін. Сонымен қатар, мәтінді жіктеу және аталған нысанды тану сияқты NLP әдістерін қолдану да әсер етеді [5]. Қазақ тілінің синтаксистік және семантикалық реңктері мәтінді жіктеуде немесе мәтін ішіндегі жалқы есімдерді анықтауда дәлсіздіктерге әкелуі мүмкін, есептеу ресурстары шектеулі және күрделі тілдік ерекшеліктері бар тілдерге қолданылған кезде қазіргі NLP әдістерінің шектеулерін көрсетеді.

Машиналық оқытудың лингвистикалық қосымшалары қандай қиындықтарға тап болады?

Мәтіндегі эмоцияларды автоматты түрде тануды жеңілдететін және миллиондаған параметрлерді қарастыру арқылы күрделі мәселелерді шешетін машиналық оқытудағы (ML) жетістіктерге қарамастан, лингвистикалық салаларда, әсіресе табиғи тілді өңдеуде (NLP) ML қолдану бірнеше маңызды қиындықтарға тап болады. Маңызды мәселелердің бірі – шағын тілдерді талдау және өңдеу, оларда тиімді ML үлгілерін үйрету үшін қажет үлкен деректер жиыны жиі жетіспейді [6]. Бұл жағдай адам тілін түсінуге және манипуляциялауға тән қиындықтармен толықтырылады - NLP компьютерлерге адам тілін түсіндіруге, түсінуге және жасауға мүмкіндік беру арқылы шешуге тырысады [7]. Сонымен қатар, 90-шы

жылдардың басынан бастап машиналық оқыту әдістерінің эволюциясы статистикалық лингвистикадағы жаңалықтармен бірге лингвистикалық қосымшалардағы ML-дің жылдам өсуі мен әлеуетін көрсетеді. Дегенмен, ол табиғи тілдің нюанстарын тиімді өңдеу және түсіну үшін бұл әдістерді бейімдеудің қиындығын да көрсетеді [8]. Бұл күрделілік лингвистикадағы ML қосымшаларын жақсарту бойынша ағымдағы қиындықтар мен күш-жігерді көрсете отырып, машиналардың адам тілін қалай түсінетінін жан-жақты шолулар мен үздіксіз жаңартулардың қажеттілігінен көрінеді [9].

NLP тапсырмаларында қазақ тілінен туындайтын ерекше қиындықтарды түбегейлі түсінуге сүйене отырып, машиналық оқыту әдістері перспективалы шешімдерді ұсынады, бірақ олардың артықшылықтары мен кемшіліктері бар. Маңызды артықшылықтардың бірі - мәтіндік реңктерді тану үшін пайымдау процедуралары мен табиғи тілді өңдеу әдістерін қолданатын жүйелердің дамуы, бұл жұрнақтар мағынаны айтарлықтай өзгерте алатын қазақ тілінде өте маңызды. Сонымен қатар, NLP-де ортақ міндет болып табылатын сезімдік талдауды қолданудың қазақ тілінің күрделі мәселелерін шешуде машиналық оқыту әдістерінің икемділігін көрсете отырып, айтылған ойларды шығару және түсіндіруде үлкен құндылық бар. Дегенмен, бұл әдістердің тиімділігі қиындықтарсыз емес. Қазақ тіліндегі қате сөздерді машиналық оқыту арқылы анықтау мұндай тәсілдердің мүмкіндіктері мен ерекшеліктерін көрсетеді, дегенмен тілдік үлгілер мен технологиялар арқылы ерекшеленетін төмен корреляция және үлкен өнімділік айырмашылықтары бұл әдістерді кең ауқымда біркелкі қолданудағы тән қиындықтарды көрсетеді. Бұл дихотомия қазақша NLP-де машиналық оқытуды қолданудың динамикалық пейзажын бейнелейді, мұнда кеңейтілген талдау мен сезімді интерпретациялаудың артықшылықтары тілдің бірегей лингвистикалық сипаттамаларын тиімді өңдеу үшін талап етілетін өнімділіктің өзгермелілігі мен нюансты түсінуге қатысты мұқият таразылануы керек [10].

Машиналық оқыту лингвистикалық ресурстарды дамытуға үлес қосуы

Тіл білімінде машиналық оқытуды қолдану негізінде бұл технологиялар лингвистикалық ресурстарды дамытуға нақты қалай ықпал ететінін түсіну маңызды. Машиналық оқыту дәстүрлі түрде қолмен және уақытты қажет ететін процестерді автоматтандыру және жақсартуда бірегей артықшылық береді. Мысалы, мәтіндегі эмоцияларды автоматты түрде тану арқылы машиналық оқыту мәтіндік деректер туралы толық ақпарат бере алады, лингвистикалық талдау және қолданбалы қолдану аясын кеңейтеді [11]. Бұл мүмкіндік сезімді талдау ресурстарын әзірлеу үшін ғана маңызды емес, сонымен қатар лингвистикалық дерекқорлардың жалпы сапасы мен тереңдігін жақсартады. Сонымен қатар, машиналық оқытудың

сөйлемдердің қолайлылығын автоматты түрде бағалауды қамтамасыз ету мүмкіндігі анағұрлым күрделі және контекстік сәйкес лингвистикалық құралдарды жасау үшін жаңа мүмкіндіктер ашады 11]. Мұндай құралдар сөйлем құрылымы мен қабылдаудың нюанстарын түсіну өте маңызды болып табылатын тіл үйрену қолданбаларында үлкен көмек көрсете алады. Бұл міндеттердің күрделілігі тілдің күрделі табиғатын және оны тиімді талдау үшін қажет күрделі аналитикалық әдістерді көрсететін ондаған миллион параметрлерді қарастыратын заманауи машиналық оқыту үлгілерінің қажеттілігімен ерекшеленеді. Дегенмен, бұл күрделілік интеллектуалды дизайнмен және тілдік белгілердің және олардың мағыналарының нәзіктіктеріне бейімделген машиналық оқыту үлгілерін қолдану арқылы жүзеге асырылады [12].

Қазақ тілі және оның тілдік ерекшеліктері

Қазақ халқының мәдени-әлеуметтік құрылымымен терең тамырлас байланысы бар қазақ тілі оның мұрасы мен болмысына өмірлік дәнекер қызметін атқарады. Тіл байлығы қазақ қоғамындағы адамгершілік-этикалық тәрбиеде маңызды рөл атқаратын мақал-мәтелдер, нақыл сөздер, әдеби контекстер арқылы айқындалады [13]. Тілдің бұл элементтері жай ғана тілдік өрнектер емес, қазақ болудың мәнін қамтитын ұрпақтан-ұрпаққа жалғасып келе жатқан даналық пен құндылықтармен сусындаған. Оның үстіне тілдің өзі халықтың жалпы жадысы мен тәжірибесінің жанды қоймасы қызметін атқара отырып, ұлттың шығу тегін, тарихын, әдет-ғұрпының көрінісі. Қазақ халқының өткен тарихын, салт-дәстүрін, болмысын сақтап, ұрпаққа жеткізу – қазақ тілі арқылы. Тілдің Отанмен тығыз байланысы, мақтаныш, ұлтжандылық оның маңыздылығын одан әрі айшықтай түседі. Ол жай ғана қарым-қатынас құралы емес, ұлттық бірегейліктің белгісі және қазақ этносының арасындағы бірлік пен мақтанаштың қайнар көзі [13]. Тіл, мәдениет және ұлттық болмыс арасындағы бұл ішкі байланыс қазақ тілінің қайталанбас ерекшеліктерін және оның қазақ мұрасын мәңгілік етудегі орталық рөлін көрсетеді.

Қазақ тіліне тән ерекшеліктердің бірі – сөздің негізіне немесе түбіріне қосымша, қосымша, жалғаулар жалғану арқылы жаңа сөздердің жасалуын жеңілдететін агглютинативті қасиеті. Бұл құрылымдық аспект тілдің икемділігі мен даму мүмкіндігін ғана емес, оның басқа тілдердің элементтерін енгізу қабілетін де атап көрсетеді. Соңғы жылдары қазақ тіліне әсіресе парсы және араб тілдерінен енген көптеген шетелдік және қабылданатын сөздер органикалық түрде сіңісіп кетті[14]. Бұл қарыздар қазақтардың сөздік қорын байытады, қазақтар мен басқа халықтар арасындағы тарихи қарым-қатынастар мен мәдени алмасулар туралы түсінік береді. Сонымен қатар, тілдің дамуы жалғасуда, қазақ қоғамының және оның құндылықтарының динамикалық және дамушы сипатын көрсететін

жаңа сөздер, сөз формалары, сөз тіркестері мен фразеологиялық бірліктер үздіксіз пайда болады. Бұл үздіксіз эволюция тілдің тұрақтылығы мен бейімделгіштігіне баса назар аударып, оны қазақ халқының мұрасының тірі қоймасына және қазіргі өмірінің айнасына айналдырады.

Машиналық оқыту қолданбаларының қиындықтары

Адамның когнитивтік процестерінің күрделілігі, әсіресе ақпаратты жылдам және тиімді өңдеуді қажет ететін салаларда машиналық оқытуды қолдануда ерекше қиындықтар туғызады. Мидың ақпараттың ұлғаюына біртіндеп бейімделу қабілеті көбінесе үлкен көлемдегі деректерді бірден өңдеуге арналған машиналық оқыту қолданбаларынан айтарлықтай айырмашылықты көрсетеді [14]. Бұл сәйкессіздік адам миының біртіндеп үйрену және бейімделу процесін имитациялайтын машиналық оқыту үлгілерін әзірлеуді талап етеді, осылайша олардың нақты әлемдегі сценарийлерде қолданылуын жақсартады. Бұған қоса, мәселе шамадан тыс жүктелген кезде мидың ақпаратты есте сақтау қиындығымен қиындайды, бұл машиналық оқыту қолданбалары да кездесетін құбылыс [14]. Бұл параллель ақпаратты өңдеуді тиімді басқара және басымдық бере алатын машиналық оқыту жүйелерін әзірлеу қажеттілігін көрсетеді, бұл жүйелердің деректермен шамадан тыс жүктелуін қамтамасыз етеді, бұл олардың өнімділігін төмендетеді. Тіл білімінде, әсіресе табиғи тілді өңдеу мәселелерін шешуде машиналық оқытуды пайдалануға алдыңғы назар аудару адамның когнитивтік процестерін дәл имитациялауда машиналық оқыту қолданбаларының тиімділігі мен тиімділігін арттыру үшін осы мәселелерді шешудің маңыздылығын көрсетеді.
Қазақ тілінің тілдік сипаттамалары NLP үшін, әсіресе қазақ әліпбиінің кириллицадан латын графикасына жалғасып жатқан көшуіне байланысты ерекше қиындықтар туғызады. Бұл ауысу таңбалардың өзгеруі ғана емес, тілдің адамдармен де, компьютерлік алгоритмдермен де бейнеленуі мен өңделуіндегі терең өзгерісті білдіреді. Бұл ауысудан туындайтын мәселелер көп қырлы, орфографияға, орфографияға, тілдік реңктердің сақталуына әсер етеді. Ең басты мәселе қазақ тіліндегі дыбыстарды жазудағы кириллица мен латын таңбаларының арасындағы елеулі айырмашылықтар болып табылады, бұл емледегі сәйкессіздіктерге әкелуі мүмкін, демек, NLP жүйелеріндегі сауаттылық пен тілді өңдеуге әсер етеді [15]. Мемлекет бұл ауысуды принциптік мәселе ретінде қарастыра отырып, оның ұлттық бірегейлікті және жаһандық жүйелерге интеграциялануын қамтитын тілдік артықшылықтан тыс маңыздылығын атап көрсетеді. Дегенмен, көшпелі кезеңде латын әліпбиінің құрылымдық ерекшеліктерін ескере отырып, қазақ тілінің төл дыбыстарының сақталуын қамтамасыз етуде көптеген

қиындықтар бар. Бұл қазақ тіліндегі кирилл әліпбиінің тамыры тереңде жатқанын және жаһандық үйлесімділікке қарай ілгерілеу кезінде осы тілдік сипаттарды құрметтейтін шешімдердің қажеттілігін мойындай отырып, көшуге мұқият, кезең-кезеңімен қарауды қажет етеді [15].

Қазақ тілін цифрландыруға, талдауға қандай күш-жігер жұмсалды?

Табиғи тілдерді өңдеу (NLP) технологияларындағы жетістіктер аясында Қазақстан, Қытай, Моңғолия, Өзбекстан және Тәжікстандағы кең тараған сөйлейтіндер үшін қазақ тілін цифрландыру және талдау бойынша айтарлықтай күш-жігер жұмсалды [16]. Жаһандану процесі түркітілдес халықтардың, соның ішінде қазақ тілінде сөйлейтін халықтардың әлеуметтік-экономикалық және мәдени біртұтастықты нығайту үшін бірігуін қажет етеді [14]. Бұл тілдің өзіндік интонациялық ерекшеліктері мен фонетикалық ерекшеліктерін сақтап қана қоймай, оны ұлттық мақтаныш пен патриотизмнің қайнар көзі ретінде дәріптеуді мақсат еткен қазақ тіл білімінің бағдарлы дамуына әкелді [13]. Бұл күш-жігерге қазақтар басым көпшілігін құрайтын Қазақстанның демографиялық құрамы тірек болып, елдің ұлттық болмысындағы тілдің маңыздылығын атап көрсетеді. Қазақ тілін заманауи цифрлық платформаларға және лингвистикалық зерттеулерге интеграциялау басқа түркі тілдерін цифрландыру және талдау үшін прецедент құра отырып, тілді сақтау мен технологиялық инновациялардағы жаһандық үрдістерге сәйкес келеді.

Лингвистикалық ресурстарды әзірлеу кезінде машиналық оқытудың теориялық негіздерінен алшақтай отырып, қазақ тілі сияқты нақты тілдер үшін қолжетімді нақты ресурстарды анықтау өте маңызды. Тіл үйренуге арналған цифрлық платформаларды қолданудың жарқын мысалы ретінде жаңадан бастаушыларға да, тілді қарапайым түсінігі бар адамдарға да қолайлы танымдық бейнероликтерді ұсына отырып, кең аудиторияға қызмет көрсететін «Kazakh Language KZ» YouTube арнасы табылады [17]. Бұл арнаның оқу бағдарламасы мен тәсілін Қазақстан Республикасының Үкіметі бекіткен, бұл оның сенімділігі мен ұлттық лингвистикалық стандарттарға сәйкестігін баса көрсетеді. Бейне ресурстардан басқа, қазақ тілінің экожүйесі құрылымдық тестілеу арқылы адамның қазақ тілін меңгеру деңгейін бағалауға арналған «Қазтест» онлайн платформасымен толықтырылды [17]. Үкімет бекіткен бейнеконтентті онлайн бағалау құралдарымен үйлестіретін тіл үйренудің бұл көп қырлы тәсілі лингвистикалық мұраны сақтау және ілгерілетудегі цифрлық платформалардың әлеуетін көрсетеді.

Машиналық оқыту алгоритмдері қазақ тіліндегі синонимдерді анықтауы

Қазақ тіліндегі синонимдердің ортақ мәселесін шешуде талдау табиғи тілді өңдеу (NLP) тапсырмаларының тиімділігіне әсер ететін маңызды лингвистикалық мәселені көрсетеді [17]. Қазақ тіліндегі синонимдерді анықтаудың күрделілігі тек лингвистикалық мәселе емес, сонымен қатар есептеу мәселесі болып табылады, өйткені мағыналары жақын сөздерді дәл ажырату күрделі алгоритмдік тәсілдерді қажет етеді. Бұл жерде машиналық оқыту алгоритмдерінің әлеуеті ерекше көрінеді. Үлгіні тану, статистикалық талдау және есептеу лингвистикасын қолдана отырып, бұл алгоритмдер қазақ тіліндегі синонимдік дилемманың перспективалы шешімін ұсынады. Машиналық оқытуды қолдану арқылы адам аналитиктері назардан тыс қалдыруы мүмкін сөздер арасындағы нюанстық үлгілер мен қарым-қатынастарды анықтай отырып, қазақ тіліндегі мәтіннің үлкен деректер жиынтығын жүйелі түрде талдау мүмкін болады. Бұл есептеу тәсілі синонимдерді анықтаудың дәлдігін арттырып қана қоймай, сонымен қатар процесті айтарлықтай жылдамдатады, осылайша қазақ контекстіндегі NLP қолданбаларының жалпы тиімділігін арттырады [17].

Табиғи тілді өңдеу (NLP) зерттеулері саласында синонимдерді құрастыру әртүрлі күрделі әдістер арқылы жүзеге асырылады, олардың әрқайсысы нәтижелердің дәлдігі мен жан-жақты болуына ерекше ықпал етеді. Лексика-синтаксистік үлгілер синонимдік терминдерді анықтау үшін тілдегі құрылымдық және синтаксистік қатынастарды қолдана отырып, іргелі стратегия ретінде анықталды [18]. Бұл тәсіл ұқсас контексте кездесетін сөздер ұқсас мағынаға ие болады деп тұжырымдайтын дистрибуциялық семантиканы қолданумен толықтырылады, осылайша үлкен мәтіндік корпуста сөздердің таралу үлгілерін талдау арқылы синонимдерді анықтауға көмектеседі. Сонымен қатар, графикаға негізделген модельдер синонимдерді құрастыруға инновациялық өлшемді енгізеді. Шеттері семантикалық қатынастарды білдіретін сөздердің желілерін құру арқылы бұл модельдер осы сілтемелердің күші мен жақындығын өлшейтін алгоритмдер арқылы тілдік байланыстарды зерттеуге мүмкіндік береді, осылайша синонимдерді шығаруды жеңілдетеді [18]. NLP жүйесіндегі синонимдерді жинақтаудың бұл көп қырлы тәсілі синонимдерді алудың дәлдігі мен еске түсірілуін арттыру үшін өрістің әртүрлі әдіснамаларға тәуелділігін көрсетеді, бұл біздің табиғи тілді түсіну мен өңдеуді байыту үшін құрылымдық, контекстік және қатынастық деректерді біріктіретін кешенді стратегияны көрсетеді.

Табиғи тілді өңдеу және машиналық оқыту саласында сөз векторлары мен сөз векторларын зерттеу сөздер арасындағы семантикалық қатынастарды түсінуде шешуші рөл атқарады[19]. Сөзді ендіру сөздерді берілген тілдегі мағынасы мен контекстін түсіріп, жоғары өлшемді кеңістіктегі сандық векторлар ретінде көрсетеді. Зерттеушілер қайталанатын нейрондық желілер негізінде қазақ тілінің тілдік модельдерін құруды зерттеп, тілді өңдеу мәселелерін шешуде нейрондық желілердің

маңыздылығын атап көрсетті. Қазақ-ағылшын және ағылшын-қазақ деректерінің синтетикалық параллель корпусында дайындалған бұл тілдік модельдер қазақ тіліне арналған машиналық оқыту қосымшаларының одан әрі дамуының негізін қалады[20].

Қазақ тіліндегі синонимдерді генерациялауға келгенде зерттеушілер ең тиімді тәсілді анықтау үшін машиналық оқытудың әртүрлі үлгілерін салыстыруға тереңірек үңілді. Қайталанатын нейрондық желілер, LSTM қабаттары және BERT модельдері сияқты әртүрлі модельдердің өнімділігін талдау арқылы зерттеушілер қазақ тіліндегі синонимдік қатынастардың нюанстарын жақсырақ түсіре алатын модельді анықтауды мақсат етіп отыр. Қазақ тіліне арнайы бейімделген BERT үлгісіне негізделген сұрақ-жауап жүйесін зерттеу және әзірлеу тіл мәселелерін шешу үшін машиналық оқытудың озық әдістерін пайдалану бойынша жүргізіліп жатқан әрекеттерді көрсетеді[21].

Бағалау өлшемдері қазақ тіліндегі синонимдерді шығару үшін машиналық оқыту үлгілерінің өнімділігі мен тиімділігін бағалау үшін қолданылатын маңызды құрал болып табылады [22]. Зерттеушілер модельдердің дәлдігін, дәлдігін және F1 рейтингін өлшеу үшін олардың мүмкіндіктері мен шектеулері туралы құнды түсініктерді қамтамасыз ету үшін әртүрлі көрсеткіштерді зерттеді. Бағалаудың қатаң критерийлерін қолдана отырып, зерттеушілер өз үлгілерін нақтылай алады, олардың өнімділігін оңтайландырады және сайып келгенде, қазақ тіліне синонимдерді қалыптастыру құралдарының сапасын жақсарта алады. Бағалау метрикасына бұл назар аудару қазақ тілінің спецификалық тілдік сипаттамалары мен талаптарына бейімделген сенімді және сенімді машиналық оқыту шешімдерін әзірлеуге деген ұмтылысты көрсетеді [21].

NLP мақсаты мен әдістерінің іргелі түсінігіне сүйене отырып, оны Қазақстанда қолдану білім, байланыс және технология секторларында бірегей жетістіктер мен артықшылықтарды ұсынады. Кішігірім деректер жиынынан туындаған қиындықтарға қарамастан, Қазақстандағы NLP үлгілері мемлекеттік мекемелерден, ұлттық компаниялардан және квазимемлекеттік органдардан алынған деректерге үйретілген кезде перспективті нәтижелер көрсетті, бұл қазақша-ағылшынша аудармалар үшін мағыналы тілді өңдеудің орындылығын көрсетті. Бұл прогресс қазақша жұрнақтардың толық жиынтығын қосу арқылы NLP үлгілеріне қазақ↔орыс және қазақ↔ағылшын тілдерінің жұптары үшін 0,14-тен 0,16-ға дейінгі BLEU ұпайларына қол жеткізуге мүмкіндік беретін синтетикалық корпустың құрылысы арқылы одан әрі күшейтілді [26]. Мұндай жетістіктер NLP-тің Қазақстанның академиялық және технологиялық ландшафттарында тіл мен қарым-қатынас тәсілін түбегейлі өзгерту мүмкіндігін көрсетеді. Сонымен қатар, Қазақстанның ресми екі тілді үкіметтік интернет көздерінен бастапқы KZ–EN параллель корпусын талдау

үшін Bitextor сияқты құралдарды пайдалану деректер тапшылығын жеңуге және елдегі NLP зерттеу инфрақұрылымын байытуға практикалық көзқарасты көрсетеді [40]. Бұл әзірлемелер бірге деректер жиынтығының шектеулеріне қарамастан NLP әлеуетін пайдалану үшін Қазақстанда қолданылатын бейімделу стратегияларын көрсетеді, осылайша негізгі секторларға айтарлықтай пайда әкеледі.

Синонимдерді құрастыру кезінде деректер жиынтығының ауытқуынан қандай қиындықтар туындайды?

Синонимдерді құрастыру кезіндегі деректер жиынтығының бұрмалану мәселесі лингвистикалық кедергілерден әрі деректерді өңдеу мен алгоритмді жобалау саласына дейін созылады. Атап айтқанда, сингулярлы деректер көзіне тәуелділік мәселені одан әрі ушықтырады, өйткені ол әртүрлі контексттер мен қолданбалардағы синонимдердің көп қырлы сипатын көрсете алмайды [23]. Бұл шектеу тек қамтудың кеңдігі туралы ғана емес, сонымен қатар жақын синонимдер немесе қатысты терминдерден шынайы синонимдерді ажырату үшін қажет түсіну тереңдігіне қатысты. Мысалы, антонимдерді анықтау процесі белгілі бір қиындықты көрсетеді, онда дәстүрлі деректер жиыны жиі оң қауымдастықтарға бұрылады. Бұл тарихи деректердегі олқылықтарды азайту үшін әртүрлі көздерді зерттеуді қажет етеді. Сонымен қатар, деректер жиынындағы тән ауытқу сөздер арасындағы нюансты айырмашылықтарды жасырып, олардың қарым-қатынастарын тым жеңілдетілген немесе дәл емес көрсетуге әкеледі. Осылайша, антонимдердің арасындағы күрделі динамикаға бейімделу үшін «қарым-қатынас» ұғымын қайта анықтау маңызды болып табылады. Сонымен қатар, синонимді құрайтын анықтаманың өзі әртүрлі қолданбаларда айтарлықтай өзгереді, бұл тапсырмаға күрделіліктің тағы бір деңгейін қосады [23]. Бұл қиындықтарды шешу үшін көптеген деректер көздерін пайдаланатын және ойында кеңірек лингвистикалық және контекстік факторларды қарастыратын біртұтас тәсілді қабылдау өте маңызды. Бұл стратегия синонимдерді құрастыру процесін байытып қана қоймайды, сонымен бірге алынған синонимдік жиындардың дәлдігі мен өзектілігін арттырады, осылайша деректер жиынының бұрмалану мәселесіне жан-жақты шешім ұсынады.

Әртүрлі тілдердегі әдістердің өзгермелілігі мәдениет пен контексттің тілге әсерімен, сондай-ақ тілдік құрылымдардағы тән айырмашылықтармен байланысты болуы мүмкін. Мәтінмән тілдердегі сөйлемдердің мағынасын өзгертуде шешуші рөл атқарып қана қоймайды, сонымен қатар тіл мен мәдениет арасындағы терең байланысты көрсетеді, бұл тілдерді өңдеудің әртүрлі есептеу әдістеріне әкеледі [24]. Бұл вариация бір мәдениетте жалпы білім деп саналатын нәрсе басқа мәдениетте мүлдем түсініксіз болуы мүмкін, осылайша бұл тілдерді өңдеу тәсіліне әсер ететіндігімен одан әрі күрделенеді. Мысалы, Word2Vec және FastText сияқты тіл үлгілерінің дамуы бұл әдістердің тілдік әртүрлілікті өңдеуге қалай бейімделетінін көрсетеді. Word2Vec сөздердің тығыз векторлық көріністерін жасау үшін нейрондық желіні пайдалана отырып, сөйлем ішіндегі сөздің контекстін түсіруге назар аударады, ал FastText бұл тәсілді сөздерді n-gram таңбаларының жинағы ретінде көрсету арқылы кеңейтеді, бұл әсіресе бай тілдер үшін тиімді. морфологиясы немесе сөздік қордан тыс сөздерді

өңдеуге арналға. Бұл бейімделу әртүрлі тілдерді өңдеудегі техникалық өзгерістерді көрсетіп қана қоймайды, сонымен қатар NLP артықшылықтарын анағұрлым инклюзивті және жаһандық қол жетімді ету үшін кеңірек міндеттемені көрсетеді [24].

Алгоритмді таңдау синонимді құрастырудың дәлдігіне әсері

Алгоритмді таңдаудың күрделі процесі синонимдерді құрастыруда шешуші рөл атқарады, әсіресе ол енгізуі мүмкін тән ауытқуларға байланысты. Мысалы, алгоритмдер «банк» сияқты екіұшты терминдердің ең көп таралған мағыналарына басымдық бергенде, синонимдік жинақтарды осы танымал интерпретацияларға бұрмалау қаупі бар. Бұл қиғаштыққа ұзынырақ мәтіндерді талдау кезінде біршама қарсы тұруға болады, мұнда қоршаған сөздермен қамтамасыз етілген кең контекст сәйкес мағынаны анықтауға көмектеседі, осылайша алгоритмді таңдау бұрмалануының дәлдікке әсерін азайтады. Дегенмен, бұл жеңілдету стратегиясы қысқа сөйлемдерді өңдеуде жеткіліксіз, мұнда контекстік белгілердің болмауы танымал мағыналарға бірдей бейімділіктен туындаған кірістіру кеңістігіндегі қатені түзете алмайды [25]. Бұл синонимдерді құрастыруда алгоритмді таңдаудың күрделі мәселесін көрсетеді, мұнда мәтіннің ұзындығы мен алгоритмнің контекске сезімталдығы нәтиженің дәлдігін анықтауда маңызды рөл атқарады.

Кешенді зерттеуде бірнеше NLP үлгілері олардың тиімділігі мен әртүрлі машиналық аударма тапсырмалары үшін жарамдылығын анықтау үшін қатаң салыстырудан өтті. Атап айтқанда, зерттеу NLLB үлгісін кестеге келтіретін аударма дәлдігіндегі жетістіктерді баса көрсете отырып, NLLB моделін базалық модельмен қатар қойды. Сонымен қатар, парк моделі аударманың жоғары дәлдігін талап ететін тапсырмаларға сәйкестігін қамтамасыз ету үшін дәл бапталған, бұл жоғары нәтижелерге қол жеткізуде үлгіні теңшеудің маңыздылығын атап өтті [26]. Салыстыру NMT үлгілеріндегі Seq2Seq, RNN, BRNN және Transformer сияқты архитектуралардың кең спектріне кеңейтілді, олар аударма тапсырмаларын өңдеудегі сәйкес тиімділігін анықтау үшін бағаланды. Бұл бағалау RNN, BRNN және Transformer архитектураларын пайдалана отырып, жетілдірілген MT үлгілерін оқыту үшін OpenNMT құрылымын қолданды, осылайша олардың өнімділігі мен күрделі лингвистикалық қиындықтарға бейімделуін мұқият зерттеуді жеңілдетеді [39]. Зерттеудің тереңдігі NLP үлгісінің негізгі және тазартылған нұсқаларын қарама-қарсы қою арқылы одан әрі байытылды, мұнда тазартылған модель базалық үлгіге қарағанда шекті, бірақ тұрақты артықшылық көрсетті, бұл 0,01 BLEU ұпайының жақсаруы арқылы дәлелденді [26]. Hugging Face репозиторийінен алынған модельдермен расталған бұл мұқият салыстыру Ha, Niehues & Waibel (2016), Эдунов және т.б. (2018) және Sennrich, Haddow & Birch (2016), олар

көптілді NMT сценарийлерін, кері аударманың әсерін және машиналық аударма сапасын жақсарту үшін біртілді деректерді пайдалануды зерттеді [39]. Осы әртүрлі, бірақ бағытталған сараптама арқылы зерттеу әртүрлі NLP модельдерінің машиналық аударманың дамып келе жатқан талаптарына қалай бейімделетінін түсінуге айтарлықтай үлес қосады.

Қазақ тілінің синонимдері үшін деректер жиынтығын дайындау

Қазақ тіліндегі сөздер мен мағыналарға арналған деректер жиынын дайындау – оның жан-жақтылығы мен тілді өңдеу тапсырмалары үшін пайдалылығын қамтамасыз ету үшін әртүрлі дереккөздер мен әдістемелерді қамтитын мұқият процесс. Бастапқыда деректер жинағы қазақ тіліндегі өңделмеген мәтінді құрастырады, «.kz» файл пішімінде сақталады, мұнда әр сөйлем анықтық пен ұйымшылдықты сақтау үшін мұқият бөлек жолға орналастырылады. Бұл өңделмеген мәтіндік деректер қазақ тіліндегі жалпы саны 121138 сөйлемнен тұратын айтарлықтай көлемді құрайды, бұл деректер жиынтығының тілге қатысты ауқымдылығын көрсетеді. Деректер жиынтығының машиналық аударма үлгілері үшін пайдалылығын жақсарту үшін, әсіресе қазақ тілін түркі тіліндегі өзбек тіліндегі әріптесімен байланыстыру кезінде деректер жиынтығы параллельді корпусты қамтиды. Бұл параллельді корпус қазақ тіліндегі сөйлемдерді өзбек тіліндегі сөйлемдермен теңестіру арқылы жеңілдетілген, әрбір теңестірілген сөйлемдер жиынына бірегей сәйкестендіру нөмірі берілген және «.csv» файл пішімінде ұсынылған. Бұл теңестіру деректер жиынтығын дайындаудағы тыңғылықты күш-жігерді атап қана қоймайды, сонымен қатар қазақ тіліндегі әрбір сөйлемнің өзбек тіліндегі сөйлемге дәл сәйкес келуін қамтамасыз етеді, жан-жақты екі тілді ресурсты дамытады. Қазақ тіліндегі сөйлемдері өзбек тіліндегі аналогтарына қарағанда орта есеппен сәл ұзағырақ екі тілдің қосылуы осы түркі тілдерінің лингвистикалық сипаттамаларына нюансты түсінік береді, бұл деректер жиынтығын машиналық аудармамен және тілді өңдеумен айналысатын зерттеушілер мен практиктер үшін баға жетпес ресурс [27].

NLP-тің қазақ тілін түсіну және генерациялау үшін қолдану контекстінде синонимдерді құрастырудағы үлгілердің орындалуы айтарлықтай вариацияларды көрсетеді. Варол және т.б. (2019) ағылшын-орыс, қазақ-орыс және ағылшын-түрік тілдерін пайдаланатын трансферттік оқыту әдістерінің үйлесімін пайдалана отырып, қазақ тілінен ағылшын тіліне аудару үшін BLEU ұпайы 0,24 және керісінше үшін chrF ұпайы 0,48 болатын елеулі жетістік көрсетті [26]. Бұл қазақ тіліне қатысты синонимдерді анықтау және құрастыру мүмкіндігінің күшті екендігін көрсетеді, осылайша қазақ тіліне қатысты NLP тапсырмаларын түсіну мен генерациялау аспектілерін жақсартады. Екінші жағынан, Бриакоу мен Карпуат (2019) қазақ-ағылшын және түрік-ағылшын деректерін пайдалана

отырып, BLEU 0,1 ұпайы туралы хабарлады, бұл бір доменде әлдеқайда шектеулі табысты көрсетеді [26]. Бұл нәтижелер арасындағы қарама-қайшылық әдістеме мен деректер алуан түрлілігінің тілге қатысты мазмұнды өңдеу және жасаудағы NLP үлгілерінің тиімділігіне, әсіресе NLP зерттеулерінде аз назар аударылған қазақ сияқты тілдерге әсерін атап көрсетеді. Өнімділіктегі бұл теңсіздік сонымен қатар тілдер арасындағы ұқсастықтарды пайдалана отырып, лингвистикалық ресурстардағы олқылықтарды жою үшін оқытуды трансферттеу әлеуетін көрсетеді.

Қазақ тілінде синонимдер жасау модельдерін әзірлеу табиғи тілді өңдеу және жасанды оқыту саласында маңызды бағыт болып табылады. Зерттеушілер қазақ тілінің лингвистикалық ерекшеліктері мен нюанстарына бейімделген жасанды оқыту алгоритмдерін әзірлеуге күш салды. Қазақ тіліне тән морфологиялық күрделіліктер мен синтаксистік құрылымдарды ескере отырып, зерттеушілер нақтырақ және контекстке сәйкес келетін синонимдер жасау модельдерін жасай алады. Бұл сала бойынша зерттеулер дамыған сайын, қазақ тілінде синонимдерді жинақтауға арналған сенімді модельдерді әзірлеу табиғи тілді өңдеу мүмкіндіктерінің дамуына кең перспективалар ашады.

Деректер жиынтығының сапасын үшін қолданылатын критерийлер

Қазақ тіліндегі синонимдерді анықтау негізінде қаланған іргетасқа сүйене отырып, қазақ тілін өңдеуге арналған деректер жиынтығының сапасын алдыңғы қатарлы әдістер мен ресурстарды енгізу арқылы айтарлықтай арттыруға болады. Мысалы, ақпаратты іздеу өнімділігін жақсарту үшін семантикалық ұқсастық шараларын қолдану перспективалы бағыт болып табылады, өйткені ол семантикалық жақын сөздерді анықтауға бағытталған, осылайша бұрын көрсетілген синонимдік мәселені тікелей шешуге бағытталған [28]. Сонымен қатар, сөйлеуді синтездеуге арналған ҚазақТТС және аталған нысанды тану үшін KazNERD [29] сияқты мамандандырылған деректер жиынын құру деректер жинақтары тек саны жағынан ғана емес, сонымен қатар ерекшелігі мен қолдану аясы бойынша да өсіп келе жатқан ландшафтты көрсетеді. Бұл деректер жинақтары сөйлеу синтезі және атаулы нысанды тану сияқты тауашаларды қамтамасыз ете отырып, қазақ тілін өңдеудегі машиналық оқытудың пайдалылығын жай синонимдерді анықтаудан тыс кеңейтеді. Сонымен қатар, бұл деректер жинақтарын әзірлеу тиімді үлгілерді дайындау үшін маңызды мақсатты, сапалы ресурстардың маңыздылығын көрсетеді, осылайша машиналық оқыту мен табиғи тілді өңдеудегі жетістіктерді қазақ тілі үшін толығымен пайдалануға болатынын қамтамасыз етеді.

Жасанды оқыту әдістері табиғи тілді өңдеу саласында төңкеріс жасап, синонимдерді жасау сияқты міндеттер үшін жаңашыл шешімдер ұсынды.

Атап айтқанда, BERT (трансформаторлардан біржақты кодердің ұсыныстары) сияқты жасанды оқыту модельдерін қолдану, тілге қатысты әр түрлі міндеттерде, соның ішінде сұрақ-жауап жүйелері мен тіл модельдеуінде үлкен мүмкіндіктерді көрсетті. Зерттеушілер қазақ тілінде синонимдерді және басқа да тілдік өңдеу міндеттерін жақсарту үшін BERT негізіндегі модельдерді белсенді түрде әзірлеп, жетілдіруде. Рекуррентті нейрондық желілер және LSTM (ұзақ мерзімді қысқа мерзімді жады) модельдері сияқты алдыңғы қатарлы жасанды оқыту әдістерін пайдалана отырып, зерттеушілер қазақ тіліндегі синонимдерді жасауға қатысты ерекше мәселелерді тиімді шеше алады.

Табиғи тілді өңдеудегі үлгі өнімділігін бағалау, әсіресе машиналық аударма (MT) және тәуелділікті талдау (DEP) тапсырмаларында дәлдік пен сенімділікті қамтамасыз ету үшін метрика мен критерийлердің күрделі жиынын қамтиды. BLEU, WER және TER көрсеткіштері MT үлгілерінің сапасын бағалау үшін кеңінен қолданылады, өйткені бұл көрсеткіштер адам сапасының рейтингтерімен жоғары корреляцияны көрсетіп, олардың саладағы сенімділігін нығайтады [39]. Атап айтқанда, MT нәтижесі кәсіби адам аудармасына неғұрлым жақын болса, соғұрлым ол жақсырақ болады деп есептейтін BLEU метрикасы аударма сапасының негізгі көрсеткіші ретінде қызмет етеді. Сол сияқты, TER және WER көрсеткіштері, мұнда төменгі ұпайлар жоғары аударма сапасын көрсетеді, 0%-ға жақын TER ұпайы және төмен WER ұпайы жоғары аударма дәлдігін білдіретінін түсініп қолданылады [39]. Дегенмен, бағалау критерийлері DEP тапсырмаларын, мысалы, екіжақты назар үлгілерін қамтитын тапсырмаларды және mBERT сияқты үлгілерді пайдаланатындар сияқты POS белгілеу тапсырмаларын қарастыру кезінде әр түрлі болады. Бұлар үшін өнімділікті бағалау тек аударма дәлдігі көрсеткіштеріне ғана тәуелді емес, сонымен қатар алдын ала дайындалған контекстік ендірулердің тиімділігін және осы үлгілер бастапқы үлгілерге қарағанда ұсынатын статистикалық маңызды жақсартуларды қарастырады [30]. Бұл тәсіл аударма сапасын өлшеу үшін BLEU, WER және TER сияқты дәстүрлі өлшемдер өте қажет болғанымен, DEP және POS тегтері сияқты тапсырмалардың күрделілігі мен нақты талаптары өнімділік жетістіктерін дәл өлшеу үшін кеңірек және икемді бағалау жүйесін қажет ететінін нақты түсінуге баса назар аударады.

Синонимдерді топтастырудың кластерлік алгоритмдері

Синонимдерді анықтауға арналған ең тиімді кластерлеу алгоритмдерін анықтау үшін осы алгоритмдердің негізгі принциптері мен қолданбаларын түсіну өте маңызды. Кластерлеу процесс ретінде объектілерді ұқсастықтар негізінде топтастыруды қамтиды, бұл таңбаланбаған деректер жиынындағы үлгілерді анықтау үшін машиналық оқытуда және деректерді талдауда

әсіресе пайдалы [31]. Сансыз алгоритмдердің ішінде DBSCAN деректер нүктелерінің тығыздығына назар аудару арқылы кластерлеуге бірегей тәсілімен ерекшеленеді. Бұл әдіс оған ұқсастығы жоғары сөздерді немесе сөз тіркестерін тиімді топтастыруға мүмкіндік береді, бұл оны синонимдерді анықтаудың күшті құралына айналдырады [32]. Сонымен қатар, оның қарапайымдылығы мен тиімділігі үшін жоғары бағаланған k-орташа алгоритмі деректерді кластерлердің белгілі бір санына, 'k'ге бөлу және осы кластерлерді итеративті түрде нақтылау арқылы жұмыс істейді. Бұл тәсіл әсіресе мәтіннің үлкен жинақтары арасындағы синонимдерді анықтау қажет болатын үлкен деректер жиынын өңдеу үшін тиімді [33]. Бұл алгоритмдер бірігіп синонимдерді сәйкестендіру мәселесін шешуге арналған бірқатар әдістерді көрсетеді, олардың әрқайсысы лингвистикалық деректерді өңдеу және талдаудағы күшті жақтары бар.

Түрлі алгоритмдер қазақ тілінің нюанстарын қалай өңдейді?

Қазақ тіліндегі синонимдерді анықтау негізіне сүйене отырып, алгоритмдерді, атап айтқанда, k-орталарын кластерлеу қазақ тілінің тілдік реңктерін одан әрі ашудың перспективалы шешімін ұсынады. Кездейсоқ k құрал жасау арқылы сөздерді топтарға біріктіретін және әрбір сөзді ең жақын ортаға тағайындайтын бұл әдіс ұқсас сөздерді біріктіру арқылы тіл деректерінің күрделілігін өңдеуге шебер [34]. Кластерлеу алгоритмінің бір түрі болып табылатын k-орталар алгоритмі өзінің тиімділігі мен қарапайымдылығымен танылып, оны қазақ тіліндегі нәзік тілдік айырмашылықтарды талдаудың тиімді құралына айналдырады. Деректер нүктелерін – бұл жағдайда қазақ тіліндегі сөздерді немесе сөз тіркестерін – кластерлердің алдын ала белгіленген санына бөлу арқылы k-құралдары тіл үлгілерін тереңірек түсінуге көмектеседі, бұл қазақ тілі сияқты морфологиялық құрылымы бай тілдерді өңдеу үшін өте маңызды [33] . Бұл кластерлік тәсіл объектілердің немесе сөздердің арасындағы ұқсастықтарға назар аудара отырып, тілді нюансты талдауға мүмкіндік береді, бұл тек синонимдерді ғана емес, сонымен қатар осы синонимдердің басқаша қолданылуы мүмкін контекстті анықтауға мүмкіндік береді [34]. Бұл әдіс арқылы қазақ тіліндегі сөздердің жіңішке қолданыстарын ажыратудың күрделі тапсырмасы лингвистикалық қолданбаларда машиналық оқыту алгоритмдерінің бейімделу сипатын көрсете отырып, басқарылатын болады.

Синонимдік топтастырудағы кластерлердің дәлдігі

Синонимдерді құрастыруда ерекшеленген қиындықтарды шешу үшін, әсіресе деректер жиынтығының ауытқуын және қолданбалардағы синонимдердің әртүрлі анықтамаларын жеңу үшін кластерлерді

құрылымдау шешуші рөл атқарады. Кластерлерді басқа кластерлердегі объектілерден ерекшеленетін ұқсас объектілер топтары ретінде қалыптастыру арқылы [35], антонимдерден немесе байланыссыз терминдерден айқын дифференциацияны сақтай отырып, бір топтағы синонимдердің біртектілігін қамтамасыз етуге бағытталған іргелі қадам жасалады. Бұл дифференциация синонимдерді топтастырудың дәлдігін нақтылау үшін өте маңызды, өйткені ол әр кластердегі ұқсастық шекарасын анықтау арқылы әртүрлі анықтамалар мәселесін тікелей шешеді. Сонымен қатар, осы гетерогенді деректер жиынында ішкі топтарды құру жеке кластерлерде біртектіліктің жоғары деңгейін ашуы мүмкін. Бұл нюансты тәсіл кеңірек санаттағы нәзік вариацияларды ескере отырып, синонимдерді дәлірек өңдеуге мүмкіндік береді. Топтастырудың дәлдігін бекіту үшін кластерлеу қадамдарын қайталау сыни әдіс ретінде пайда болады [36]. Бұл итерациялық процесс бастапқы қорытындыларды растап қана қоймайды, сонымен қатар топтастыруларды деректер жиынының тән ауытқуларына және синонимдік анықтамалардағы вариацияларға қарсы үздіксіз реттеу және тексеру арқылы кластерлерді нақтылайды. Осы мұқият процесс арқылы синонимдерді топтастырудағы кластерлердің дәлдігі айтарлықтай артады, бұл бұрын талқыланған деректер жиынының ауытқуы және анықтамалық өзгергіштік мәселелеріне сенімді шешімді қамтамасыз етеді.

Синонимдік жинақта табиғи тілді өңдеу

Табиғи тілді өңдеу және машиналық оқыту саласында сөз векторлары мен сөз векторларын зерттеу сөздер арасындағы семантикалық қатынастарды түсінуде шешуші рөл атқарады. Сөзді ендіру сөздерді берілген тілдегі мағынасы мен контекстін түсіріп, жоғары өлшемді кеңістіктегі сандық векторлар ретінде көрсетеді. Зерттеушілер қайталанатын нейрондық желілер негізінде қазақ тілінің тілдік модельдерін құруды зерттеп, тілді өңдеу мәселелерін шешуде нейрондық желілердің маңыздылығын атап көрсетті. Қазақ-ағылшын және ағылшын-қазақ деректерінің синтетикалық параллель корпусында дайындалған бұл тілдік модельдер қазақ тіліне арналған машиналық оқыту қосымшаларының одан әрі дамуының негізін қалады.

Табиғи тіл өңдеу (NLP) арқылы қазақша синонимдерді құрастыруға ұмтылуда бірнеше жетілдірілген әдістемелер мен концепциялар қолданылады. Ең алдымен, компьютерлер мен адам (табиғи) тілдердің өзара әрекеттесуімен айналысатын NLP пәні тілдің табиғи түрде жазылған және айтылуын түсінуге және басқаруға қабілетті машиналарды құруға ұмтылу арқылы осы талпыныстың негізін қалайды [36]. Бұл түсіну синонимдерді анықтау және топтау үшін өте маңызды, өйткені ол әртүрлі контексттерде бірдей немесе ұқсас мағыналарды білдіретін сөздерді тануға мүмкіндік береді. Сонымен қатар, компилятор дизайны, NLP сыни аспектісі, тілдік

модельдердің тиімділігі мен дәлдігін айтарлықтай арттырады, осылайша лингвистикалық деректердің үлкен көлемін талдау және өңдеу арқылы синонимдерді дәл анықтауға мүмкіндік береді [37]. Бұл контекст пен мәдени реңктер мағынаға қатты әсер ететін қазақ тілінде әсіресе маңызды. Сонымен қатар, сөз векторларын талдайтын әдістерді қолдану арқылы, негізінен, көп өлшемді кеңістіктегі сөздердің орны - берілген мақсатты сөзге контекстік жағынан ұқсас сөздерді табуға болады, осылайша әлеуетті синонимдерді анықтауға болады. Бұл әдіс қазақ тіліндегі синонимдерді іздеу мен жіктеуді одан әрі нақтылай отырып, нақты өмірлік жағдайларда сөздердің қалай қолданылатынын түсіну үшін қажетті тілдік контекстті қамтамасыз ететін ауқымды корпусты немесе мәтіндер жинақтарын пайдаланады [38]. Осы NLP әдістері мен құралдары бірге қазақ тіліндегі синонимдердің кең және нюансты тізімін құрудың күрделі тәсілін ұсынады, бұл жақсы қарым-қатынас пен түсінуді жеңілдетеді.

Қазақ тіліндегі синонимдерді анықтау мәселесін шешу үшін машиналық оқыту алгоритмдерінің әлеуетіне сүйене отырып, осы үдерістегі контексттің таптырмас рөлін атап өту өте маңызды. Мәтінмән тек лексикалық алмастырғыштар мен нақты дискурстағы мақсатты мағынаны сақтайтын дәл синонимдерді ажыратуда тірек болады. Дәл синонимдерді анықтаудағы контексттің маңыздылығын асыра айту мүмкін емес. Сөздің және оның ұсынылған синонимдерінің әртүрлі абзацтарда қолданылуын талдау арқылы әртүрлі жағдайларда берілген мағынадағы нәзік реңктерді байқауға болады. Бұл аналитикалық тәсіл машиналық оқыту алгоритмдері арқылы анықталған синонимдердің тек лингвистикалық тұрғыдан ғана емес, контекстік жағынан да сәйкестігін қамтамасыз етеді. Осылайша, контекстке басымдық беру арқылы біз қазақ тіліндегі синонимдерді анықтау процесін нақтылай аламыз, бұл ұсынылған синонимдердің олар қолданылатын нақты контексттерде шынайы синонимдік болуын қамтамасыз ете аламыз.

Табиғи тілді өңдеудегі (NLP) полисемия мен омонимияның күрделі мәселелерін тиімді шешу үшін зерттеушілер күрделі сөздерді енгізу және терең оқыту үлгілерін әзірледі [39]. Бұл модельдер әртүрлі контекстердегі сөздердің реңкті мағыналарын түсінуге қабілетті, осылайша көп мағыналы сөздер (көп мағыналы) немесе дыбыстары бірдей, бірақ мағыналары әртүрлі сөздер (омонимия) ұсынатын қиындықтарды жеңеді. Мәтінмәнге сезімтал демистификация әдістерін қолдана отырып, NLP әдістері қоршаған мәтінге негізделген сөздердің болжамды мағынасын шеше алады, осылайша тілді түсіну мен өңдеудің дәлдігін айтарлықтай жақсартады [39]. Бұл прогресс NLP жүйелерінің көп мағыналы немесе омонимдік терминдердің болуына қарамастан олардың енгізгенін дәл түсіндіру арқылы пайдаланушылармен сөйлесу мүмкіндігін арттырып қана қоймайды, сонымен қатар белгілі бір сөздер қолданылатын контекстті түсіну арқылы мәтінді дәлірек санаттарға

жіктеуге мүмкіндік береді [37]. Сонымен қатар, қолмен аннотация немесе лексикон құру және автоматты әдістер сияқты қол күштерін біріктіретін гибридті тәсілдерді пайдалану NLP жүйелерінің адам тілінің күрделілігін өңдеудегі сенімділігін арттырады және олардың коммуникация мен ақпаратты өңдеудің тиімді құралдары болып қалуын қамтамасыз етеді. қолданудың кең ауқымы [40].

Қазақ тілін үйренуге машиналық оқыту негізіндегі синонимді құрастырушылардың әсері

Көмекші іздеу жүйесі сияқты синонимдік құрастырушыларға машиналық оқытуды біріктіру пайдаланушының іздеу ауқымын кеңейту үшін семантикалық жақын сөздерді пайдалану арқылы тіл үйренуді айтарлықтай жақсартады. Семантикалық байланысы бар синонимдерді ұсына отырып, бұл құралдар оқушыларға тілдік нюанстар мен түсініктердің кең ауқымын зерттеуге мүмкіндік береді, бұл олардың сөздік қорын және түсіну дағдыларын кеңейту үшін өте маңызды. Дегенмен, бұл тәсіл шектеусіз емес, өйткені тым жалпы синонимдер маңызды емес іздеу нәтижелеріне әкелуі мүмкін, әсіресе тілдерді жан-жақты үйрену үшін маңызды болып табылатын Wikipedia мақалалары сияқты білім беру ресурстарына қол жеткізу кезінде. Осы қиындықтарға қарамастан, бұл мағыналық жағынан жақын сөздерді анықтау дәлдігі, әсіресе NLPL репозиторийінің жетілдірілген есептеу үлгілері мен полиглоттық сөздерді ендіру арқылы қазақ тілі сияқты тілдерде іздеу нәтижелерінің дәлдігі мен өзектілігін айтарлықтай жақсартуды көрсетеді. Семантикалық жағынан жақын сөздерді анықтаудағы жоғары дәлдік көрсеткіштерімен дәлелденетін бұл дәлдік оқушылардың жоғары сапалы, өзекті материалға қол жеткізуін қамтамасыз ету арқылы тіл үйрену құралдарының тиімділігіне тікелей ықпал етеді [29].

Қазақ тіліндегі синонимдерді анықтау негізіне сүйене отырып, технология алдыңғы қатарлы машиналық аударма (MT) machine translation (MT) әдістерін қолдана отырып, қазақ тілінде сөйлейтіндер арасындағы байланысты жақсартудың негізгі құралы ретінде қызмет етеді. Қазақ-ағылшын тілінде арнайы дайындалған Neural Machine Translation (NMT) үлгілерін қабылдау аударма сапасының айтарлықтай жақсарғанын көрсетеді, бұл тілдер арасындағы мағынаны тереңірек түсінуге және жеткізуге мүмкіндік береді [22]. Сонымен қатар, толық жиынтық жалғауларды (CSE) complete set of endings пайдаланатын морфологиялық сегментацияны біріктіру тілдің күрделі морфологиялық ерекшеліктерін қарастыра отырып, қазақ тіліне машиналық аударманың тиімділігін арттырады. Бұл әдістемелік ілгерілеу қазақ тіліне бейімделген машиналық аударма жүйелерін бірге жетілдіретін ережеге негізделген және нейрондық желі тәсілдерінің үйлесімімен толықтырылады. Бұл технологиялық

инновациялар мәтіндерді дәл аударуды жеңілдетіп қана қоймайды, сонымен қатар қазақ тілінде сөйлейтіндердің жоғары сапалы қазақ тілінде ғылыми мақалалар мен мемлекеттік құжаттарды қоса алғанда кең ауқымды ақпаратқа қол жеткізуін қамтамасыз етеді. Қазақ тілінің лингвистикалық реңктерін дәл көрсететін құралдарды ұсына отырып, технология қарым-қатынас алшақтықтарын өтейді, дүние жүзіндегі қазақ тілінде сөйлейтіндер арасында тереңірек түсіністік пен байланыс орнатады [40].

Зерттеу нәтижелерінің қазақ тіліндегі NLP зерттеулерінің салдары көп қырлы және лингвистикалық ресурстарды әртараптандыру мен сапасын арттырудың маңызды қажеттілігін көрсетеді. Біріншіден, зерттеудің әртүрлі көздерден кең ауқымды параллель корпус құруға ұмтылысы, атап өтілгендей, мұндай деректерден әзірленген NLP үлгілерінің кеңеюі мен қолданылуын шектейтін біржақтылықтарды қамтуы мүмкін үкіметтік мәтіндерге шектен тыс тәуелділікке қарсы тұрады. Бұл сенім зерттеу ландшафтын бұрмалап қана қоймайды, сонымен қатар қазіргі тәжірибелердің еңбекті көп қажет ететін сипатын одан әрі баса көрсете отырып, деректер сапасы мен теңестіруді қамтамасыз ету үшін айтарлықтай қолмен араласуды қажет етеді. Сонымен қатар, зерттеудің қол жеткізген \acbleu ұпайы мен \acwer арқылы дәлелденген сенімді \acrbmt жүйесін дамытуға бағытталғандығы қазақ тіліне машиналық аударманың нақты қиындықтары мен ілгерілеуін атап көрсетеді, әсіресе қазақ тілі және тілдері үшін параллельді оқыту деректерінің сыни тапшылығын көрсетеді. түрік. Бұл жағдай қазақстандық NLP зерттеулеріндегі кеңірек мәселені білдіреді - соңғы жетістіктерге қарамастан, әлі де ресурстар мен зерттеу тереңдігіндегі шектеулермен күресіп келе жатқан, әрі қарай дамыту мен барлау үшін бірлескен күш-жігерді қажет ететін сала [26].

Google Translate сияқты платформалар пайдаланатын табиғи тілді өңдеу (NLP) және табиғи тілді түсіну (NLU) технологияларының дамып келе жатқан саласы қазақ тіліндегі құрастырылған синонимдік дерекқорлардың инновациялық қолданбаларына жағдай жасайды. Түрлі жанрлар бойынша 135 миллионнан астам сөзді қамтитын бай репозиторий болып табылатын Қазақ тілі корпусы (KLC) осы қолданбаларға кең негіз береді [41]. Атап айтқанда, Kazakh Dependency Treebank-тің KLC бөліктерін лексикалық, морфологиялық және синтаксистік аннотациялармен қайта белгілеу әрекеті жақсартылған тіл үлгілері мен қолданбаларына негіз қалайды [41]. Бұл жинақталған синонимдік деректер базалары тілдік нюанстарды ұсына отырып және қазақ тілін тереңірек түсінуге көмектесу арқылы білім беру және ғылыми зерттеулерге айтарлықтай үлес қоса алады. Сонымен қатар, оларды көркем әдебиетте тілді байыту, сол арқылы авторлардың экспрессивтік мүмкіндіктерін кеңейту және Қазақстанның мәдени мұрасына үлес қосу үшін қолдануға болады. Әдебиеттер мен білім беруден басқа, бұл деректер базасы құқықтық құжаттарды талдауда, бұқаралық ақпарат құралдарының контентін құруда, тіпті кеңейтілген сұрау

мүмкіндіктері арқылы іздеу жүйелерінің тиімділігін арттыруда практикалық қолдану үшін уәде береді [26][41]. Бұл дерекқорлардың машиналық аударманы және көптілді іздеу тапсырмаларын жақсартудағы әлеуеті олардың тілдік кедергілерді жоюдағы және қазақ тілінің ресурстарына жаһандық қолжетімділікті ынталандырудағы рөлін ерекше көрсетеді [41]. Құрастырылған синонимдік дерекқорларды NLP және NLU технологияларына біріктіру бар құралдарды толықтырып қана қоймайды, сонымен қатар тілді өңдеу және түсіну мүмкіндіктерін жақсартуға жол ашады, осылайша қазақ тіліндегі коммуникацияны, зерттеуді және цифрлық қолданбаларды жақсартады.

Бағалау өлшемдері қазақ тіліндегі синонимдерді шығару үшін машиналық оқыту үлгілерінің өнімділігі мен тиімділігін бағалау үшін қолданылатын маңызды құрал болып табылады. Зерттеушілер модельдердің дәлдігін, дәлдігін және F1 рейтингін өлшеу үшін олардың мүмкіндіктері мен шектеулері туралы құнды түсініктерді қамтамасыз ету үшін әртүрлі көрсеткіштерді зерттеді. Бағалаудың қатаң критерийлерін қолдана отырып, зерттеушілер өз үлгілерін нақтылай алады, олардың өнімділігін оңтайландырады және сайып келгенде, қазақ тіліне синонимдерді қалыптастыру құралдарының сапасын жақсарта алады. Бағалау метрикасына бұл назар аудару қазақ тілінің спецификалық тілдік сипаттамалары мен талаптарына бейімделген сенімді және сенімді машиналық оқыту шешімдерін әзірлеуге деген ұмтылысты көрсетеді.

Салыстырмалы анализ

	SOZDIKQOR	Sozdik kz	QuilBot	Paraphraser	REF-N-WRITE
Website address	https://sozdikqor.kz/sinonim	https://sozdik.kz/	https://quillbot.com/	https://www.paraphraser.io/	https://www.ref-n-write.com/paraphrasing-tool/
Statistics	Yes	No	Yes	Yes	No
Web keyboard	Yes	Yes	No	No	No
Clear button	Yes	Yes	Yes	Yes	Yes
Modes	No	No	Yes	Yes	No
Copy button	Yes	No	Yes	Yes	No

	SOZDIKQOR	Sozdik kz	QuilBot	Paraphraser	REF-N-WRITE
Main Language	Kazakh	Kazakh Russian	All types of English	English Espanol Dutch French Norwegian	English
JS - framework	jQuery	jQuery	React LazySizes	Core-js jQuery	Isotope Core-js jQuery
Symbols number	10 000	One word	125 words	400 words	200 words

Protocol	Open Graph	Open Graph	Open Graph	Open Graph	Open Graph
Domain created	2022	2003	2017	2020	2014
WEB Server	Ngingx	Ngingx	Ngingx	PHP	PHP
UI framework	Bootstrap	Bootstrap	Bootstrap	Codelgniter	Bootstrap
Load time	896 ms	891 ms	560 ms	920 ms	670 ms

Техникалық сипаттамаларына сәйкес, SOZDIKQOR, Sozdik kz, QuilBot сияқты қызметтер Node.Js бағдарламалау тілінде жазылғанын, олардан айырмашылығы REF-N-WRITE және Paraphraser жүйелерінде PHP қолданылғанын байқауға болады. UI құрылымы үшін барлық жүйелер Bootstrap пайдаланды. SOZDIKQOR және Sozdik kz сияқты жүйелер негізінен қазақ және орыс тілдерін пайдаланады. Қалған жүйелер негізінен ағылшын тілінде жұмыс істегенімен, парафраза жүйесі испан, голланд сияқты басқа тілдерде де жұмыс істей алады. Таңбалар саны бойынша Парафразада 400 сөз болуы мүмкін, ал Sozdik.kz тек 1 сөзден тұрады.

Функционалдық сипаттамалары бойынша оның қаншалықты жан-жақты және пайдаланушылар үшін қолжетімді екенін анықтауға болады. Мысалы, барлық жүйелерде барлығын тазалау сияқты функциялар бар. QuilBot және Paraphraser, REF-N-WRITE жүйелерінде онлайн пернетақта жүйесі жоқ екенін байқадық. Бұл шетелдік пайдаланушылар үшін үлкен кемшілік. Көптеген сөздердің қанша өзгергенін көрсететін статистика бар. Қалған жағдайда, ортақ көп нәрсе бар, өйткені олар ұқсас функцияны қамтамасыз етеді. Қорытындылай келе, қазақстандық жүйелерде синонимизаторлардың түрлері жоқ екенін атап өтуге болады.
Тұтынушы ретінде қорытындылайтын болсақ, QuilBot әлдеқайда жақсы деп айтуға болады, өйткені оның техникалық және функционалдық талаптары бар. Сонымен қатар оның парафразалық машиналармен жұмыс істеу тәжірибесі көбірек.

2 ҚАЗАҚ ТІЛІНІҢ СИНОНИМДІК СӨЗДЕРІН АНЫҚТАЙТЫН МОДЕЛЬДІ ЖОБАЛАУ ЖӘНЕ ӘЗІРЛЕУ

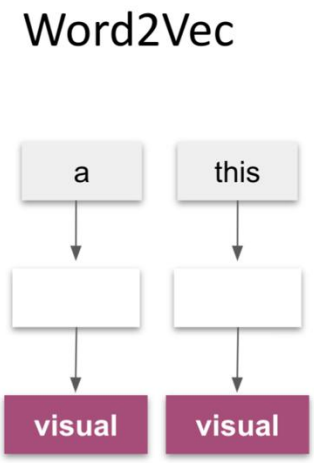
2.1 WORD2VEC

Лингвистикалық зерттеулер мен есептеуіш лингвистика саласында синонимдерді, әсіресе әртүрлі контексттегі сөздерді анықтау және жіктеу бірінші кезектегі міндеттердің бірі болып табылады. Синонимдерді анықтаудың заманауи әдістемелері синонимдерді таңдаудың дәлдігі мен тиімділігін арттыру үшін озық есептеу әдістерін қолдана отырып, айтарлықтай дамыды. Соңғы зерттеулерге сәйкес, сөздер арасындағы синонимдік қатынастарды анықтау үшін лингвистикалық заңдылықтарды, сөздерді қолдануды және контекстті талдайтын алгоритмдерді біріктіретін автоматты синонимдерді таңдау барған сайын жетілдіріліп келеді [42].

Синонимдерді таңдау мәселелерін түбегейлі түсінуге негізделген Word2Vec оны дәстүрлі әдістерден айқын ажырататын жаңа тәсілді ұсынады. Тілдік ортада синонимдерді тану үшін қолмен немесе жартылай автоматтандырылған процестерді жиі қолданатын талқыланған процедуралардан айырмашылығы, Word2Vec семантикалық талдаудың дәлдігі мен тиімділігін арттыру үшін автоматтандырылған өңдеу әдістерін пайдаланады. Мәтінмәндегі сөздерді қолдану үлгілерін анықтау үшін үлкен деректер жиынтығын талдауға негізделген бұл әдістеме қарапайым лексикалық ұқсастықтан шығып, сөздердің мағыналарын толығырақ түсінуге мүмкіндік береді. Айналасындағы сөздерге негізделген сөздің болуын болжау үшін нейрондық желі үлгілерін пайдалану арқылы Word2Vec басқа әдістермен жіберіп алуы мүмкін синонимдер арасындағы нәзік семантикалық айырмашылықтарды анықтай алады. Бұл мүмкіндік мәтінмәндік синонимдерді анықтауға және сол синонимдерді пішімделген мәтін деректерінен автоматты түрде таңдауға байланысты мәселелерді шешу үшін өте маңызды. Сонымен, лингвистикадағы лексикалық синонимияны зерттеу мақсаттарына, жоғарыда айтылғандай, синонимдік шекараларды тиімдірек анықтау үшін теориялық лингвистика мен статистикалық талдауды біріктіретін Word2Vec инновациялық тәсілі ықпал етеді [43].

Word2Vec-тің әртүрлі іске асыруларының арасында векторлық бейнелеуге деген көзқарасына байланысты екі нұсқа ерекшеленеді: Үздіксіз сөмке (CBOW) моделі және Skip-Gram моделі. Word2Vec инновациясының мәні N-өлшемді векторлық кеңістікке сөздерді енгізу арқылы тілдік контекстті неғұрлым нюансты түсінуді ұсынатын дәстүрлі сөздер қаптамасы әдісінен ауытқу болып табылады. Бұл векторлық кеңістік статикалық емес; ол семантикалық жағынан ұқсас сөздерді бір-біріне жақынырақ орналастыратын динамикалық репрезентацияларды жасауға мүмкіндік береді, сол арқылы бұрын назардан тыс қалған мағынаның нәзік

тұстарын түсіреді [44]. CBOW үлгісі мақсатты сөзді қоршаған мәтінмән сөздеріне негізделген болжайды, бір шығыс векторын жасау үшін мәтінмәнді тиімді орташалайды. Керісінше, Skip-Gram моделі мақсатты сөз негізінде қоршаған контекстік сөздерді болжау арқылы бұл тәсілді өзгертеді, әсіресе кішірек деректер жинақтарымен жұмыс істегенде және сөздер арасындағы алыс байланыстарды түсіргенде тиімді болады. Бұл әдістемелер Word2Vec бағдарламасының бейімделгіштігі мен дәлдігін көрсетеді, бұл бұрынғы үлгілердің мүмкіндіктерінен әлдеқайда асатын егжей-тегжейлі семантикалық талдауға мүмкіндік береді.

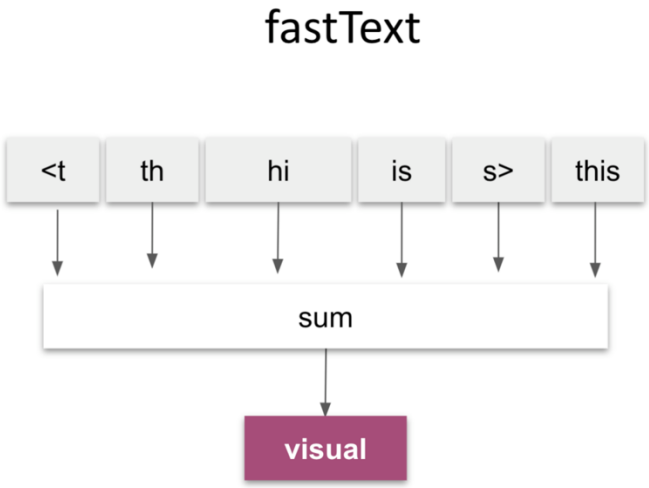


2.2 FASTTEXT

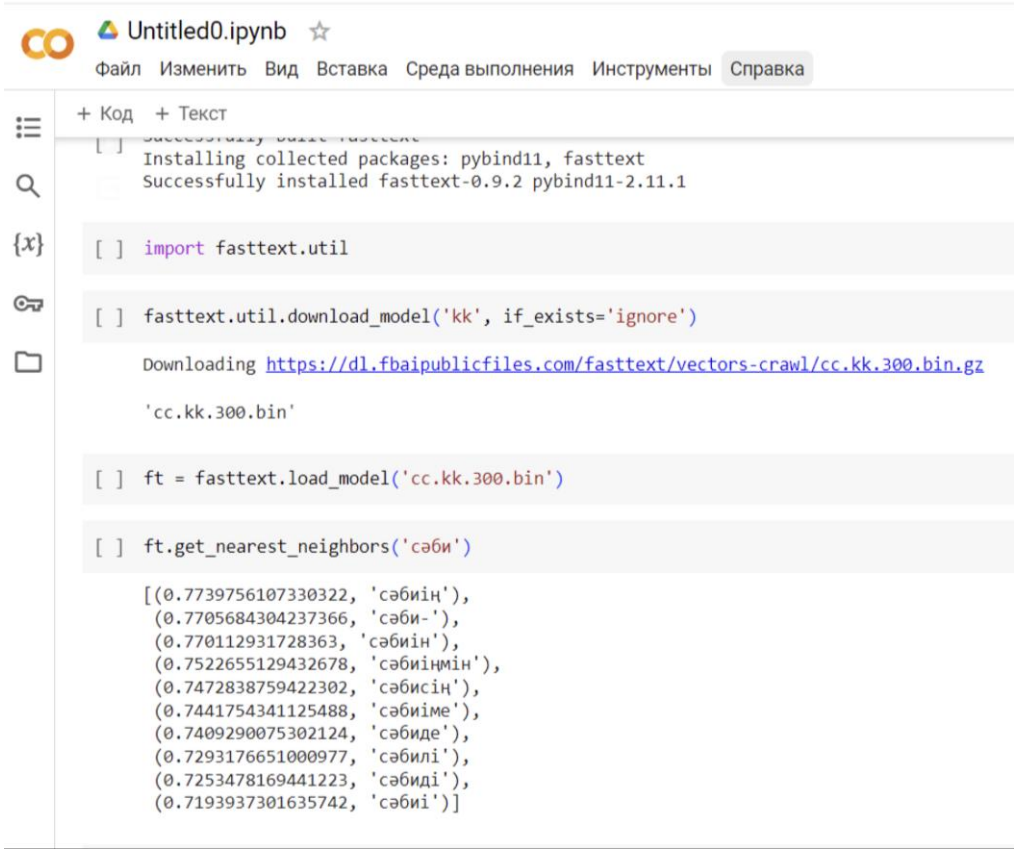
Табиғи тілді өңдеудің қуатты моделіне айналған FastText мәтіндік деректерді түсінуге және өңдеуге жаңа тәсілді енгізеді. Сөздерді өңдеудің ең кіші бірлігі ретінде қарастыратын дәстүрлі үлгілерден айырмашылығы, FastText сөздерді n-grams таңбалары деп аталатын кішірек бірліктерге бөлу арқылы бұл тұжырымдаманы өзгертеді. Бұл әдіс сөздердің морфологиясы мен құрылымын бұрынғы үлгілер жасай алмаған түрде түсіріп, тілді толығырақ талдауға мүмкіндік береді. Осы n-граммдарды талдай отырып, FastText тілдердің кең ауқымын, соның ішінде бай және күрделі морфологиясы бар тілдерді тиімді өңдей алады, осылайша оның жан-жақтылығы мен әртүрлі табиғи тілді өңдеу тапсырмаларында қолдану мүмкіндігін арттырады [45].

FastText-тің табиғи тілді өңдеуге бірегей тәсілі, әсіресе мәтінді жіктеу объектісі арқылы, оны RNNLM, Word2vec және GloVe сияқты бұрынғы үлгілерден ерекшелендіреді. Табиғи тілді өңдеу (NLP) саласындағы мәтінді жіктеудің маңыздылығын асыра айту мүмкін емес, өйткені ол көңіл-күйді талдаудан бастап спамды анықтауға дейінгі сыни қолданбалардың негізінде

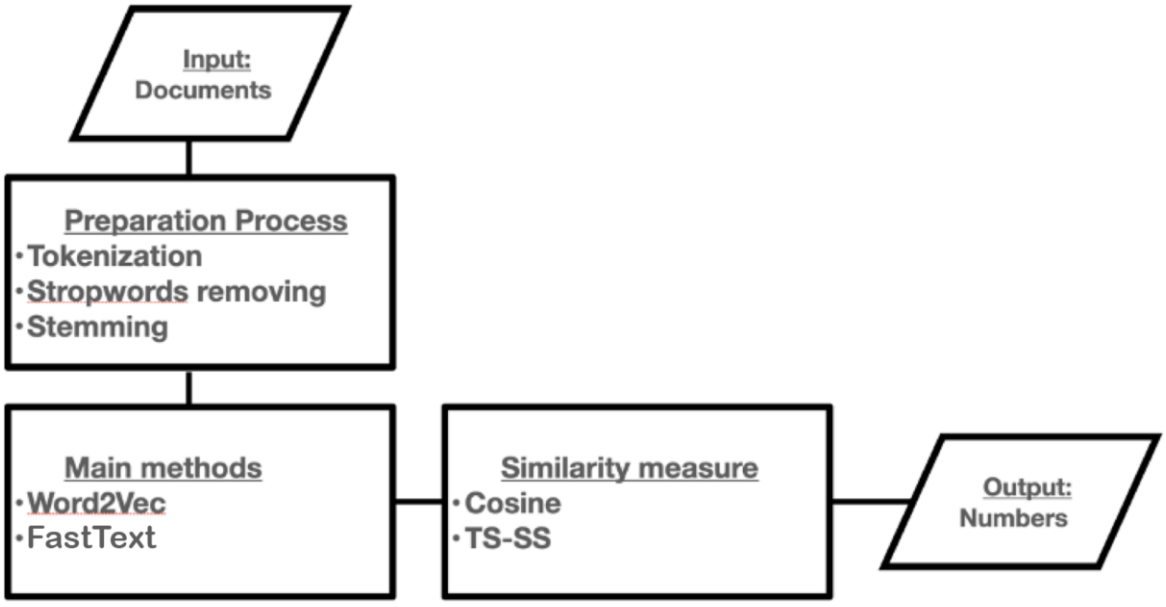
жатыр [46]. Алдыңғы мәтіндерден айырмашылығы, FastText-тің сөздерді кішірек бірліктерге - n-грамм таңбаларға бөлу әдісі - сөздің морфологиясын Word2vec және GloVe сияқты үлгілер жасай алмайтындай етіп түсіріп, тілді неғұрлым егжей-тегжейлі түсінуге мүмкіндік береді. Бұл егжей-тегжейлі тәсіл әсіресе бай морфологиясы бар тілдер үшін пайдалы, мұнда сөздің мағынасы шағын вариациялармен айтарлықтай өзгеруі мүмкін. Сонымен қатар, NLP технологияларының тарихи эволюциясы адам тілінің күрделілігін дәлірек түсіне алатын үлгілерге ауысуды көрсетеді [45]. Мәтінді машинада оқылатын пішінге түрлендіруді қамтитын FastText пайдаланатын векторлау әдістері құжаттарды индекстеу және семантикалық талдау сияқты әртүрлі тапсырмаларды орындау үшін өте маңызды, осылайша FastText-тің әртүрлі NLP тапсырмаларын шешудегі бейімделгіштігі мен тиімділігін көрсетеді [47].



Бұл суретте fasttex моделдің дайын оқытылған жүйесін қолданып көрген кезде, қаншалықты қате мағлұмат шығарып, дұрыс істемейтінін байқасақ болады.



FastText өзінің оқытылған моделі



Енгізу: (Input)

Кезең құжаттарды жүктеуден басталады, олар мәтіндік файлдар, электрондық кестелер, дерекқорлар немесе тіпті Интернет көздерінен мәтін ағындары болуы мүмкін. Тапсырмаға байланысты бұл мәтіндер пішімдеуді жою, кодтауларды түрлендіру немесе кескіндерден мәтін алу сияқты алдын

ала өңдеуді қажет етуі мүмкін.

Дайындау барысы: (Preparation process)

Токенизация: Бұл мәтінді құраушы элементтерге – лексемаларға бөлу. Токендер әдетте сөздер болып табылады, бірақ тыныс белгілері мен арнайы таңбаларды да қамтуы мүмкін. Токенизация маңызды қадам болып табылады, себебі ол барлық келесі өңдеу қадамдары үшін деректердің түйіршіктілігін анықтайды.

Стоп сөздерді жою: Стоп сөздерді өте жиі кездесетін және талдау контекстінде маңызды мағыналық жүктемені көтермейтін сөздер (мысалы: «ау» «әй», «па», «ей»). Оларды жою талдауға арналған деректер көлемін азайтуға және мәтіннің неғұрлым ақпаратты элементтеріне назар аударуға көмектеседі.

Түйіндеу: Түбірлік жалғаулар мен жұрнақтарды алып тастау арқылы сөздерді түбір түріне келтіреді. Бұл мәтіндегі сөздердің ауыспалылығын азайтуға көмектеседі және бір түбірлі сөздерді сәйкестендіруді жеңілдетеді. Белгілі бір құжат тілімен тиімді жұмыс істейтін стемерді таңдау маңызды, өйткені әртүрлі тілдердің морфологиялық ерекшеліктері әртүрлі.

Негізгі әдістер: (Main methods)

Word2Vec: Бұл әдіс қолданылу контекстіне негізделген сөздердің векторлық көріністерін жасау үшін нейрондық желілерді пайдаланады. Word2Vec оқыту үшін мәтіндік деректердің үлкен көлемін қажет етеді және сөздер арасындағы күрделі семантикалық және синтаксистік қатынастарды анықтауға қабілетті.

FastText: Word2Vec идеяларын кеңейте отырып, FastText әрбір сөзді n-граммдар жиынтығы ретінде қарастырады. Бұл модельге белгісіз сөздерде жақсырақ жұмыс істеуге мүмкіндік береді және қате жазылған немесе қате жазылған сөздер үшін векторлық көріністердің сапасын жақсартады.

Ұқсастық өлшемі: (Similarity measure)

Косинус ұқсастығы: Екі вектор арасындағы ұқсастық дәрежесін анықтау үшін қолданылады (біздің жағдайда мәтіндердің векторлық бейнелері). Бұл көрсеткіш ұқсас құжаттарды кластерлеу және іздеу мәселелерінде жиі қолданылады.

TS-SS: Бұл аббревиатура ұқсастықтың белгілі бір статистикалық өлшеміне сілтеме жасай алады, бірақ оны дәл анықтау үшін қосымша ақпарат қажет. Мұндай көрсеткіштер мәтіндерді салыстыруға арналған әртүрлі статистикалық тәсілдерді қамтуы мүмкін, мысалы, сөздердің немесе сөз тіркестерінің таралуына негізделген.

Шығару: (output):

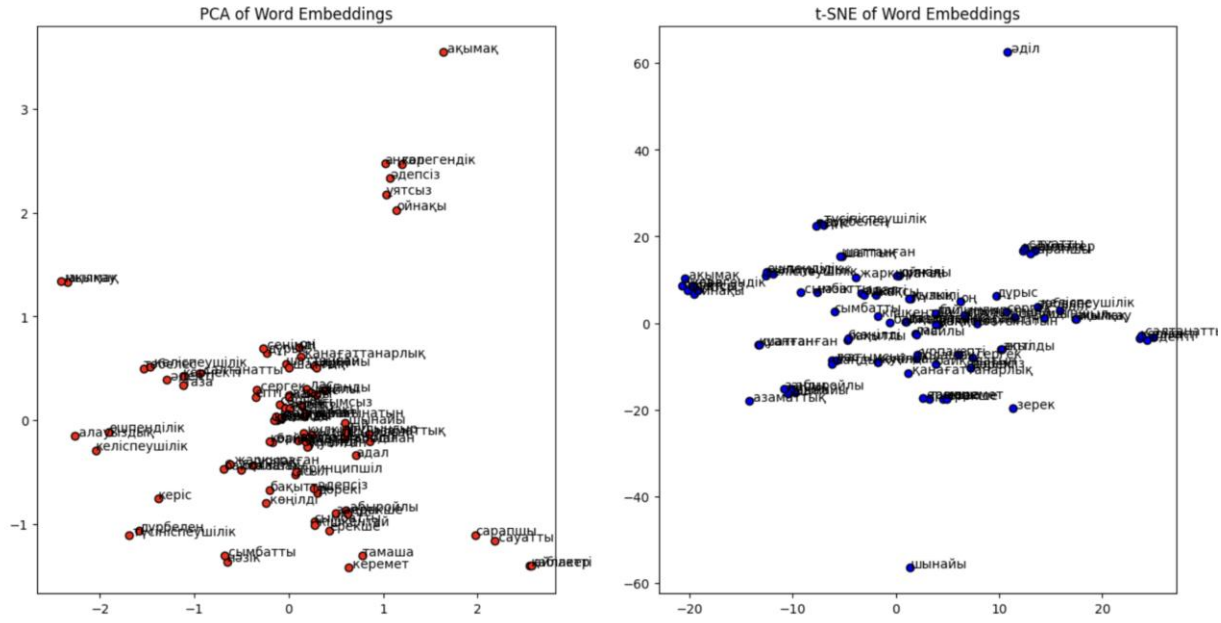
Процестің нәтижесі құжаттар арасындағы ұқсастық дәрежесі, жеке сөздердің маңыздылығы немесе ұсыныс механизмдері, құжаттарды іздеу немесе аналитикалық құралдар сияқты нақты қолданбалар үшін маңызды басқа көрсеткіштер ретінде түсіндірілетін сандық мәндер болып табылады.

PCA

Негізгі құрамдастардың талдауы (PCA) деректердегі бастапқы өзгеру осьтерін анықтау және пайдалану арқылы сөздерді енгізуге тән күрделілікті жеңілдету негізінде жұмыс істейді. Енгізілген сөздер ішінде әрбір өлшемнің қаншалықты дисперсияға ие екендігін бағалау арқылы PCA өлшемдері азырақ ақпараттың негізгі бөлігін түсіретін маңызды мүмкіндіктерді дәл анықтайды. Бұл процесс бастапқы деректерді жүйелі түрде негізгі құрамдас бөліктер немесе меншікті векторлар деп аталатын өлшемдердің жаңа жинағына түрлендіретін PCA қолданатын математикалық алгоритмде терең тамыр жайған [47]. Бұл негізгі құрамдас бөліктер ерікті осьтер ғана емес; олар бірінші негізгі құрамдас ең маңызды дисперсияны, әрбір келесі құрамдас келесі ең маңызды сомаға ие болуын қамтамасыз ететін, дисперсияны барынша арттыратын деректердегі бағыттар болып табылады. Сақталатын негізгі құрамдастардың санын анықтау өте маңызды, өйткені ол PCA түсіндірілген жиынтық дисперсияға тікелей әсер етеді. Бұл жинақталған дисперсия негізгі құрамдастардың түсіндірілген дисперсия коэффициенттерін қосу арқылы есептелетін маңызды өлшем болып табылады, ол өлшемді азайту процесінен кейін бастапқы деректердің өзгергіштігінің қанша бөлігі алынғанын сандық түрде анықтайды. Бастапқы өлшемділігі айтарлықтай жоғары болуы мүмкін, көбінесе жүздеген өлшемдерден асатын сөздерді енгізу контекстінде, бұл қысқарту деректерді жеңілдетіп қана қоймайды, сонымен қатар табиғи тілді өңдеу немесе машина сияқты келесі тапсырмалар үшін оның ең ақпаратты аспектілерін сақтайды. оқыту үлгілері.

Қазақ синонимдері үшін сөздерді енгізуді талдауда негізгі құрамдас талдауды (PCA) пайдалану табиғи тілді өңдеу саласындағы айтарлықтай ілгерілеушілікті көрсетеді. PCA арқылы сөзді ендірудің өлшемділігін тиімді түрде азайту арқылы зерттеушілер деректер ішіндегі негізгі вариация осьтерін анықтай алады, осылайша ақпараттың негізгі бөлігін аз өлшемдермен тасымалдайтын маңызды мүмкіндіктерді түсіреді. Бұл тәсіл лингвистикалық деректерге тән күрделілікті жеңілдетіп қана қоймай, синонимдер арасындағы нәзік семантикалық айырмашылықтарды анықтауға көмектеседі. Сақталатын негізгі құрамдастардың санын анықтауға баса назар аудару өте маңызды, өйткені ол PCA түсіндірілген жиынтық дисперсияға тікелей әсер етеді, осылайша бастапқы деректердің

өзгермелілігінің өлшемділіктен кейінгі қысқаруы сақталатынын сандық түрде анықтайды. Сонымен қатар, кішірейтілген өлшемдік кеңістіктегі косинус ұқсастығы сияқты ұқсастық өлшемдері арқылы сөздерді кірістірулерді салыстыру қазақ тілінің кең ауқымды сөздік қорындағы синонимдерді анықтаудың дәлдігін арттырады. Түпнұсқа деректерді негізгі құрамдас бөліктерге немесе жеке векторларға PCA арқылы жүйелі түрлендіру оның лингвистикалық талдауды жеңілдетудегі тиімділігін көрсетеді.



t-SNE

Қазақ тілінің күрделі морфологиялық құрылымын ескере отырып, дәстүрлі атаулы тұлғаны тану (NER) жүйелері тілдің агглютинативті сипатына байланысты синонимдерді дәл анықтау және байланыстырумен күреседі. Осы Word2Vec ендірулеріне өлшемді азайту әдісі t-SNE қолдану арқылы зерттеушілер кәдімгі талдау әдістері арқылы бірден көрінбейтін синонимдік кластерлер туралы түсініктерді ұсына отырып, сөздер арасындағы байланыстарды көрнекі түрде көрсете алады. Бұл тәсіл қазақ тіліне тән тілдік қиындықтарды жеңіп қана қоймайды, сонымен қатар сөйлеу қабілеті бұзылған балалардың тілдік қажеттіліктеріне бейімделген тиімдірек білім беру және емдік ресурстарды құрудың негізгі құралы болып табылады.

t-SNE-ны қазақ синонимдері үшін сөздерді енгізуді талдау үшін қолдану лингвистикалық зерттеулер мен табиғи тілді өңдеу қолданбаларын кеңейтудегі оның трансформациялық әлеуетін көрсетті. Word2Vec кірістіруіне өлшемді азайту әдістемесі ретінде t-SNE қолдану зерттеушілерге сөздер арасындағы күрделі қарым-қатынастарды, әсіресе қазақ тілінің бірегей сипаттамалары контексінде көрнекі түрде анықтауға мүмкіндік берді. Жоғары өлшемді кірістіру кеңістігінен анағұрлым

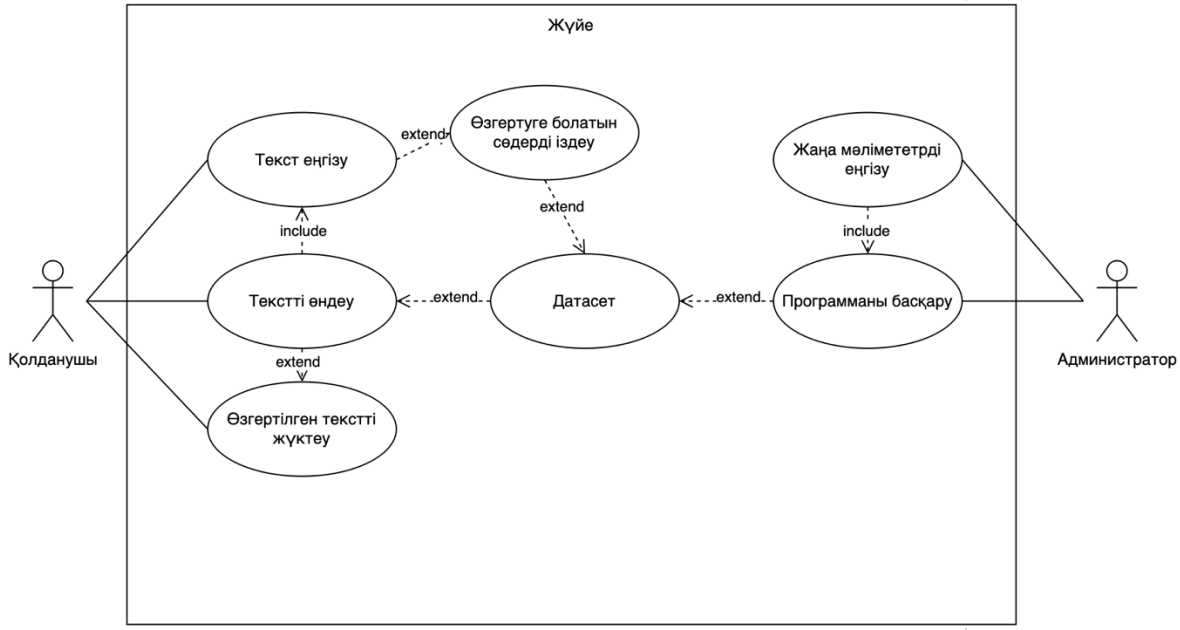
түсіндірілетін 2D визуализациясына ауыса отырып, t-SNE мағыналары ұқсас сөздердің жақындығын тиімді сақтайды, осылайша синонимдік кластерлердің анық көрнекі көрінісін ұсынады және қазақ тілінің семантикалық реңктеріне жарық түсіреді. t-SNE ұсынған визуализация жергілікті сөздердің ұқсастықтарын сақтап қана қоймайды, сонымен қатар қазақ синонимдері арасындағы күрделі қарым-қатынас торын көрсете отырып, ұқсас сөздердің шоғырларын ашады. Қазақ тіліндегі деректерге t-SNE қолданумен байланысты қиындықтарға қарамастан, оның төменгі өлшемді ұсынудағы сөздер арасындағы маңызды семантикалық қарым-қатынастарды сақтау қабілеті оның лингвистикалық талдау құралы ретіндегі құндылығын көрсетеді. Алға қарай келешектегі зерттеулер басқа тілдерге арналған сөздерді ендіруді талдауда t-SNE қолдануын одан әрі оңтайландыру жолдарын зерттеп, туындауы мүмкін кез келген шектеулерді немесе қиғаштықтарды қарастыруы мүмкін. Жалпы алғанда, осы зерттеуде ұсынылған нәтижелер табиғи тілді өңдеу саласындағы білімнің үздіксіз ілгерілеуіне ықпал етеді және t-SNE әлеуетін тілдік құрылымдар мен қарым-қатынастарды зерттеудің құнды құралы ретінде көрсетеді.

3 ЖҮЙЕНІ ЖОБАЛАУ ЖӘНЕ ЖҮЗЕГЕ АСЫРУ

3.1 Use case Қолдану нұсқасы диаграмма

Бағдарламалық қамтамасыз етуді әзірлеу индустриясында Бірыңғай модельдеу тілі (UML) диаграммаларын пайдалану күрделі жүйелерді жобалау мен енгізуді жақсартудың маңызды құралына айналды. UML диаграммалары бағдарламалық жасақтама жүйелерінің құрылымдық және мінез-құлық аспектілерінің көрнекі көрінісі ретінде қызмет етеді, даму процесінің байланыс, талдау және жобалау кезеңдеріне көмектеседі. Жүйе архитектурасын, қарым-қатынастарын және өзара әрекеттесуін визуализациялаудың стандартталған әдісін ұсына отырып, UML диаграммалары әзірлеу процесін жеңілдетеді, оны тиімдірек және қатесіз етеді. Бұл зерттеу мақаласы бағдарламалық жасақтаманы әзірлеудегі UML диаграммаларының мағынасы мен қолданылуын, UML диаграммаларының әртүрлі түрлерін, олардың бағдарламалық жасақтаманы әзірлеу тапсырмаларын жеңілдетудегі артықшылықтарын және бағдарламалық жасақтаманы әзірлеудің өмірлік циклінің әртүрлі кезеңдеріндегі негізгі қолданбаларын зерттейді.

Қолдану нұсқасы диаграммалары бағдарламалық жасақтаманы әзірлеуде маңызды құрал ретінде қызмет етеді, пайдаланушының көзқарасы бойынша жүйелік талаптарды визуализациялауға және түсінуге көмектеседі. Қолдану нұсқасы диаграммаларын түсіну диаграммалық бейнелеудің артындағы іргелі ұғымдарды зерттеуді қамтиды. Біріншіден, нақты мақсаттарға жету үшін қатысушылар (пайдаланушылар немесе сыртқы жүйелер) мен жүйе арасындағы өзара әрекеттесуді сипаттайтын қолдану нұсқасы диаграммасының мәнін түсіну маңызды. Қолдану нұсқасы диаграммаларында актерлер маңызды рөл атқарады, өйткені олар рөлдері және жүйемен өзара әрекеттесу негізінде анықталады. Сонымен қатар, пайдалану жағдайлары диаграммасының құрамдас бөліктері жүйенің функционалдығы мен әрекетін бірге сипаттайтын пайдалану жағдайларын, қатысушыларды, қатынастарды және жүйе шекараларын қамтиды.

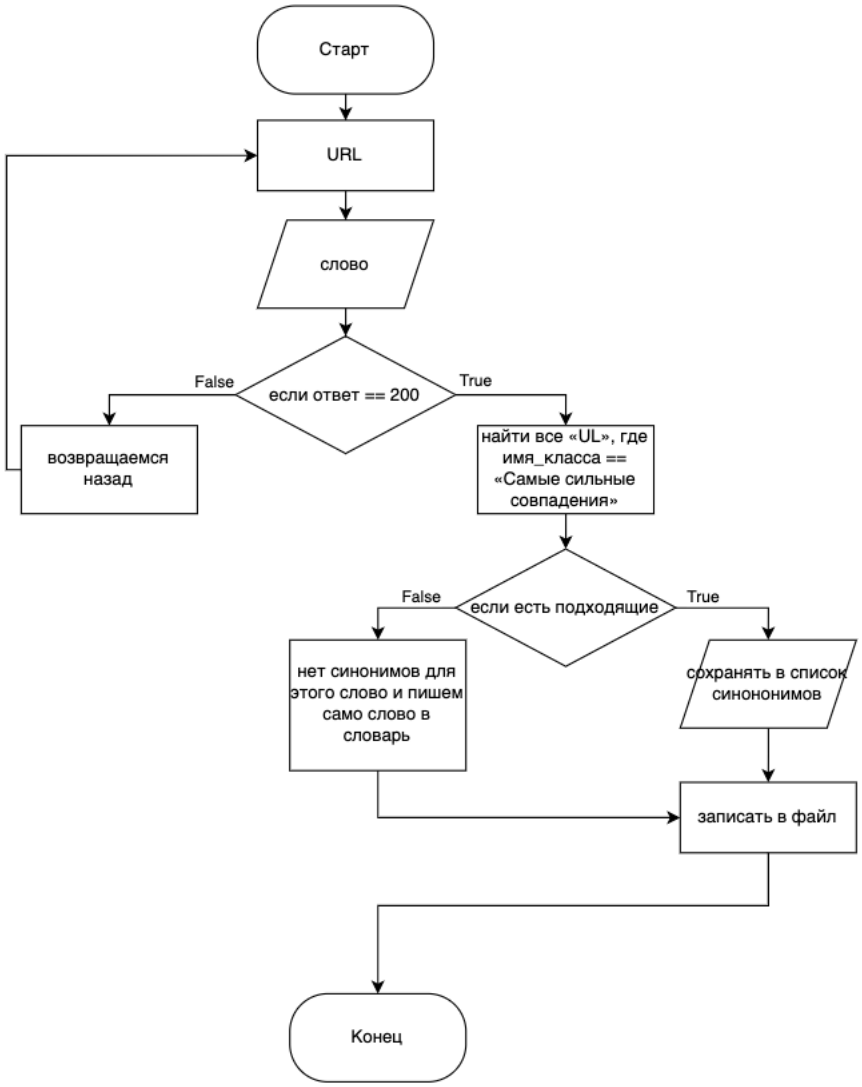


Жүйеде қазақ тілінде парафрейз жасайтын функционал бар. Яғни қолданушы керек мәтінді немесе сөйлемді еңгізіп сол тексттің өзгертілген нұсқасын жүктей алады. Ал администратор деген қолданушы осы жүйенің ұйымдастырушысы болып табылады. Осы жүйенің басты қолданушыларын екіге бөліп қарастырамыз:

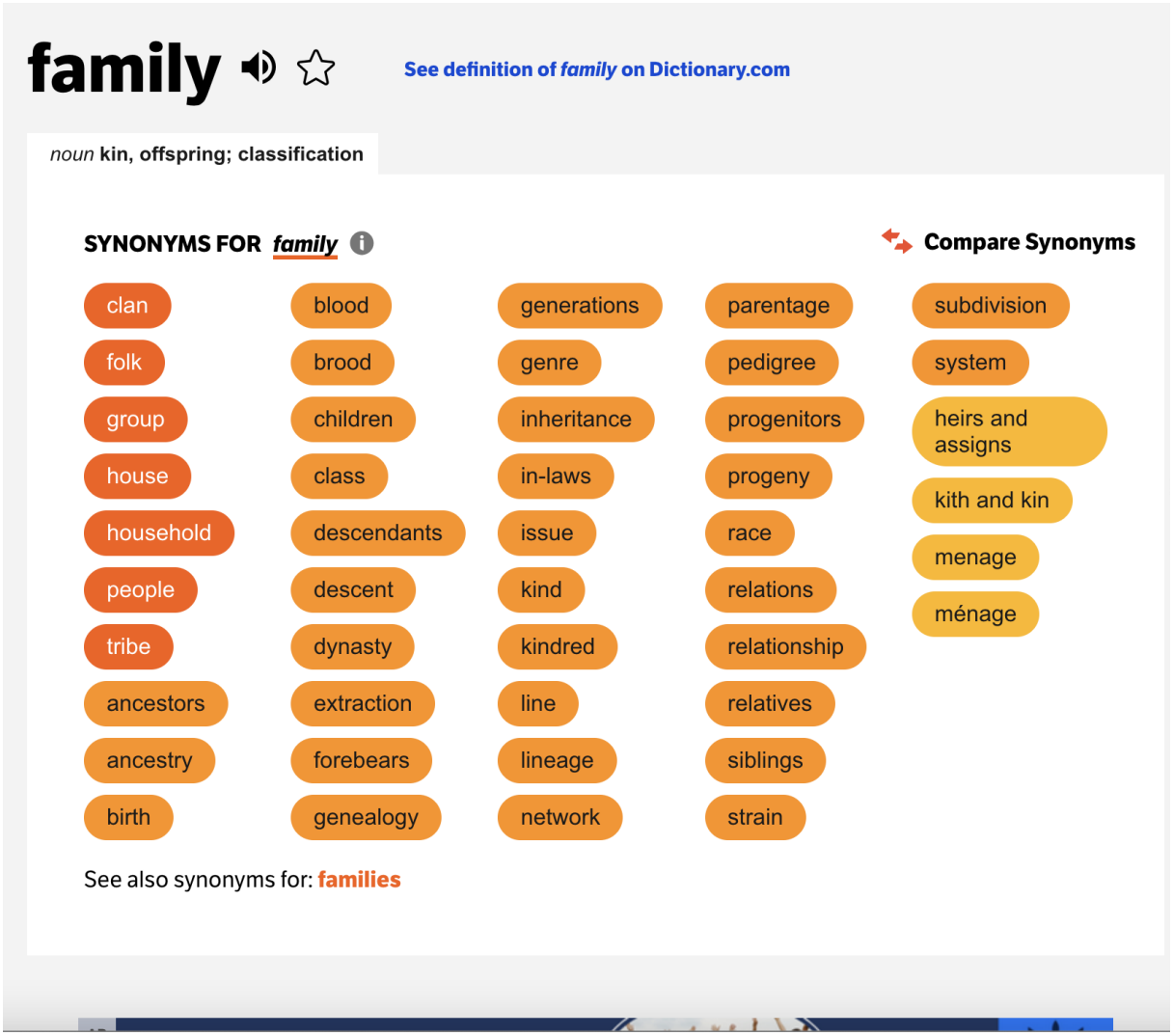
1. Қолданушы – өзі жазған мәтінді және парафрейз жасалған текстті көре алады. Оған қоса шыққан тексттің құрылымы ойдағыдан шықпаса, қолданушы өз бетінше өзгертуге мүмкіндігі бар.
2. Администратор – программист немесе жүйені әзірлеуші. Жүйенің ішіндегі бүкіл функционалдың кодын жәнеде басқаруын қамтамасыз етеді. Яғни осы жүйені өзгертіп жаңа функционал қосып деректермен мағлұматтарды өзгертіп жаңартып отырады.

3.2 Beautiful Soup

Beautiful Soup ең алдымен, деректерді алудың күрделі тапсырмасын басқарылатын және жеңілдетілген операцияға түрлендіру арқылы HTML талдау процесін жеңілдетудің негізгі құралы ретінде ерекшеленеді. Жаңадан бастаушылар үшін BeautifulSoup қолжетімділігі мен қарапайым, интуитивті тәсілі арқылы HTML талдауы өте жеңілдетілген. Бұл сонымен қатар оның күрделі HTML және XML құжаттарын құрылымдық пішінге талдау және түрлендіру, мазмұнды әрі қарай өңдеу үшін қолжетімді ету мүмкіндігімен де көрінеді. Жаңадан бастағандар үшін жай ғана жеңілдетумен қатар, BeautifulSoup HTML талдауын нәтижелі гипермәтіндік белгілерді толыққанды нысандарға түрлендіру мүмкіндігімен тиімдірек ету арқылы тәжірибелі әзірлеушілердің қажеттіліктерін қанағаттандырады. HTML тегтеріне сәйкес атрибуттары бар бұл нысандарды өңдеудің әртүрлі қажеттіліктеріне сәйкес өңдеуге немесе сұрауға болады [48]. Бұл мүмкіндіктің мәнін асыра бағалау мүмкін емес, өйткені ол қолмен қырғышты азайтады және веб-скрепинг пен деректерді өңдеудің озық жобалары үшін берік негіз береді.



Бұл диаграммада Beautiful Soup тың қалай жұмыс істейтіндігі белгіленген. Яғни біріншіден керек сайттың URL алып сол сайтка өзіміздің корпустан сөздерді кезегімен еңгіземіз, егерде сайт бізге бастапқыда 200 деген жауап қайтарған болатын болса, онда келесі кезеңде бізге шыққан синонимдердің ішінен тек қана class_name дері ең жақын ұқсастықта деген сөздерді аламыз. Ал 200 қайтарылмаған жағдай болса кері url жазатын кезеңге қайтарады. Егер сол class_name мен іздегенде ішінде сөздер шықпайтын болса, берілген бастапқы сөздің өзін жазамыз, ал бізге керек синонимар шықса оларды кезек-кезекпен жазып бір файлға сақтаймыз. Ол файлдың аты корпус синонимов болады.



Ағылшын тезаурус сайтының Отбасы сөзіне шыққан үлгісі

Суретте «Family» сөзінің синонимдері көрсетілген. «Family» сөзі экранның жоғарғы жағында көрсетілген. Синонимдер бірнеше бағанға бөлінген және. Неғұрлым дәл синонимдер ашық қанық қызғылт сары түспен ерекшеленеді. Біршама дәлме-дәл ұқсастығы азайған сөздер сары түспен берілген.

Астыңғы суретте веб парақшадан керекті мәліметті алу үшін керек болған пакеттермен кітапханаларды импорттау коды.

```
!pip install beautifulsoup4 requests
```

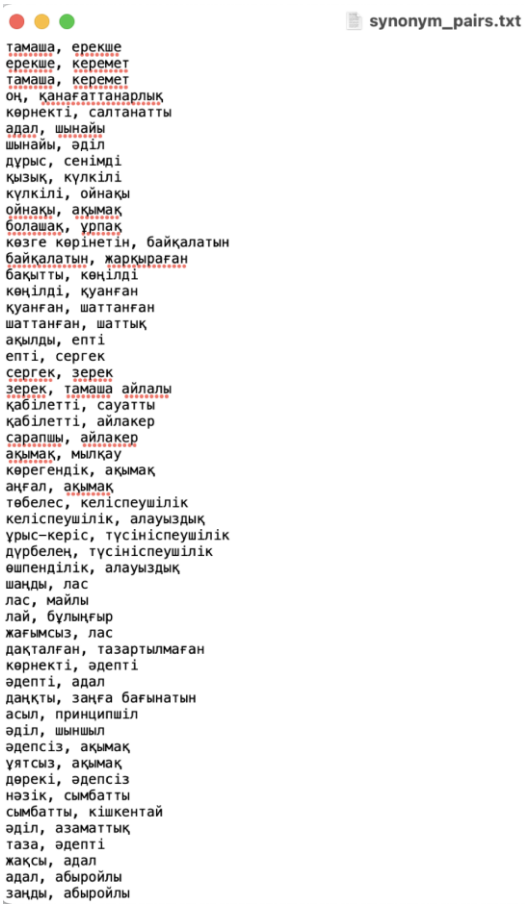
Показать скрытые выходные данные

```
import requests
from bs4 import BeautifulSoup
```

Тезаурыс сайтынан өзіміз берген сөздердің синонимдарының ішінен “Ең күшті сәйкестіктерді” жіктеп аламыз жоғарыдағы суретте “Family” сөзімен мысал келтірілген

```
def scrape_strongest_synonyms(word):
    # Синонимдерді алу үшін URL мекенжайы
    url = f"https://www.thesaurus.com/browse/{word}"
    # Бет мазмұнын алу үшін сұрау жасау
    response = requests.get(url)
    if response.status_code != 200:
        print(f"Failed to retrieve data for word: {word}")
        return []
    # Беттің HTML мазмұнын талдау
    soup = BeautifulSoup(response.text, 'html.parser')
    # Көрсетілген сынып атауы бар ul элементін табу
    ul_class_name = 'LhpJmDktqk95S2vWg1JZ'
    matches_ul = soup.find('ul', class_=ul_class_name
    if not matches_ul:
        print(f"Strongest matches section not found for word: {word}")
        return []
    synonyms = [match.text for match in matches_ul]
    words = []
    # Әрбір сөзді «Ең күшті сәйкестіктер» тізімінен шығарып, List қа жазу
    if matches_ul:
        for li in matches_ul.find_all('li'):
            word = li.text.strip()
            words.append(word)
    else:
      print("Couldn't find the matches list with the specified class.")
    return(words)
```

Ең күшті сәйкестіктерді таңдап алу



Сионимдік жұп корпусы

Бұл суретте бір-біріне ең жақын синоним сөздер жиынтығы көрсетілген. Осы корпус арқылы сөздердің арасындағы жақындықты өзіміз қолмен коя аламыз. Осы жұп корпустағы сөздер басқа корпустағы сөздерге қарағанда жақын синонимдар бірлестігі болып саналуына білікті маман сөз береді.

3.3 Selenium

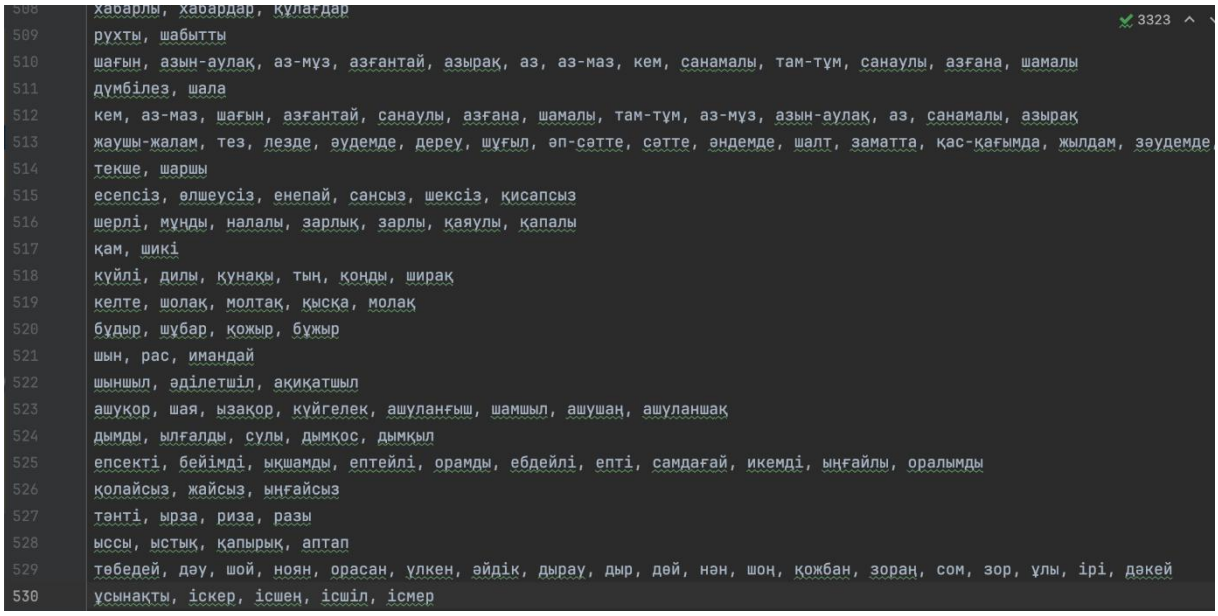
Selenium WebDriver және ChromeDriverManager сияқты басқа құралдарды веб-талдау және автоматтандыруды жылдам дамытады, тиімділік пен точности жою үшін тұрақты емес. Selenium WebDriver, мощный инструмент автоматизации, ойнайды решающую роль в улучшении задач веб-анализа және автоматизации, позволяя пользователям взаимодействовать с веб-элементами, моделировать дествия в веб-элементами, моделирование дествия және веб-сайттарды пайдалану. Сонымен, ChromeDriverManager бағдарламасы ChromeDriver файлын басқаратын автоматтық басқаруды орнату және орнату процестерін жоғарылатады, бұл Selenium WebDriver браузерімен Chrome браузері үшін пайдалы және пайдалы болып табылады. Осыған байланысты, Selenium WebDriver және ChromeDriverManager бағдарламаларын орындауға мүмкіндік береді обслуживание, связанные с их интеграцией. Рассматривая

эти аспекты, данная исследовательская статья призвана предоставить комплексное руководство по использованию Selenium WebDriver және ChromeDriverManager үшін тиімді веб-анализі және автоматизациясы, проливая свет национальные в развитительных веб-сайттар және тестирования.

```
# Файлдан сөздерді оқу
with open('/Users/matanov/Downloads/kazakh_adj.txt', 'r', encoding='utf-8') as file:
    words = file.readlines()
# Нәтижелерді жазу үшін файлды ашу
with open('results.txt', 'w', encoding='utf-8') as result_file:
    # Файлдағы әрбір сөзді қарап шығу
    for word in words:
        word = word.strip()  # Қосымша бос орындар мен жаңа жолдарды жою
        editor = driver.find_element(By.CSS_SELECTOR, value: "div.ql-editor[data-placeholder='Мәтінді енгізіңіз ...'] >
        editor.clear()  # Енгізу өрісін тазалау
        editor.send_keys(word)  # Сөзді енгізу
        editor.send_keys(Keys.ENTER)  # Форманы жіберу
        # Нәтижелерді жүктеп алуға уақыт береміз
        time.sleep(2)
        # Нәтижені аламыз және көрсетеміз
        result = driver.find_element(By.ID, value: "result_text")
        result_text = result.text
        # print(f'"{word}":', result_text)
        result_file.write(f'{result_text}\n')  # Нәтижені файлға жазамыз
        time.sleep(1)  # Келесі енгізу алдында кідірту
```

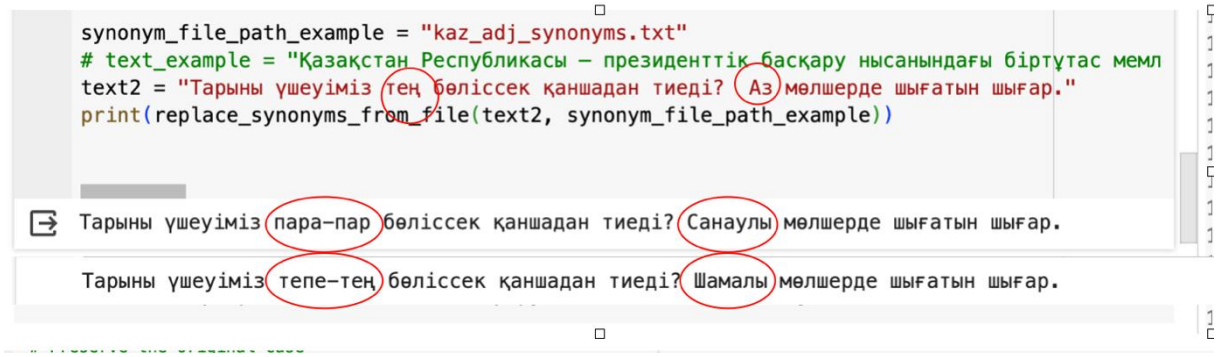
Сөздердің синонимдарын парсинг жасау

Файлды оқу: Скрипт kazakh_adj.txt файлын ашады және барлық жолдарды words тізіміне оқиды. Әрбір жолда, бәлкім, бір сөз бар. Шығыс файлын ашу: Ол results.txt файлын жазу үшін ашады. Бұл файл әрбір сөзге жүргізілген операциялардың нәтижелерін сақтайды. Сөздер бойынша цикл: Әрбір сөз үшін: Ол ақырын жолақты strip() арқылы тазартады. Сценарий веб-беттегі мәтін енгізу өрісін (CSS селекторын пайдаланып) табады және оның барлық мазмұнын тазартады. Ол тазартылған сөзді бұл өріске енгізеді және тапсырманы өңдеуді бастау үшін (мысалы, Enter пернесін басу арқылы) тапсырады. Нәтижелерді алу: Кешіктіруден (2 секунд) кейін, ол result_text ID-імен анықталған элементтен мәтінді ұстайды, бұл тапсырманың нәтижесі/шығарылғаны болуы мүмкін. Нәтижелерді жазу: Сценарий алынған мәтінді results.txt файлына жазады, әрбір жазбадан кейін түсініктілік үшін жаңа жол қосады. Кідіріс: Әрбір цикл итерациясының соңында 1 секундтық кідіріс бар, бұл серверді немесе веб драйверді асыра пайдаланбау үшін жасалған.



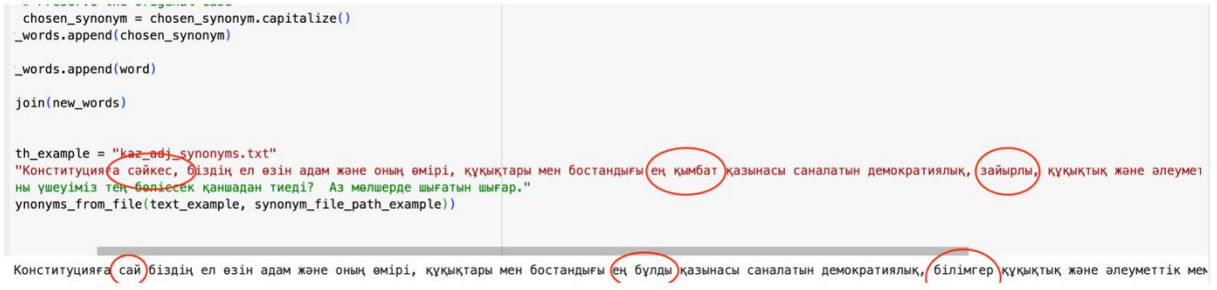
Шыққан синонимдер сөздігі

3.4 Тәжірибе



Бірінші нәтиже

Бірінші нәтижеге келетін болсақ, бұл жерде көріп тұрғандарыңыздай “ТЕҢ” сөзін “ПАРА-ПАР” немесе “ТЕПЕ-ТЕҢ” деген сөздерге ауыстырған. Ал “АЗ” деген сөзді “САНАУЛЫ” немесе “ШАМАЛЫ” деген секілді сөздерге алмастырды.



Екінші нәтиже

Екінші нәтижеде, “СӘЙКЕС” сөзін “САЙ” деп ауыстырып, үтірдіде алыс тастады. Келесі “ЕҢ ҚЫМБАТ” деген сөз тіркесті “ЕҢ БҰЛДЫ” деп ауыстырды, қарасаңыз “ЕҢ” күшейтпелі сөзі орнында қалып екінші сөзі ауысты, бірақ бұлды деген мағына жағынан толықтай келеді деп келісе алмаймын. Ал “ЗАЙЫРЛЫ” деген сөзді “БІЛІМГЕР” деген сөзге алмастырып жіберді.

3.5 Метрикалар

```python
from rouge import Rouge

reference = "Тарыны үшеуіміз тең бөліссек қаншадан тиеді?  Аз мөлшерде шығатын шығар."
candidate = "Тарыны үшеуіміз тепе-тең бөліссек қаншадан тиеді? Шамалы мөлшерде шығатын шығар."

rouge = Rouge()

scores = rouge.get_scores(candidate, reference)

print("ROUGE scores:")
for key, value in scores[0].items():
    print(f"{key}: {value['f']:.4f} (F1) {value['p']:.4f} (Precision) {value['r']:.4f} (Recall)")
```

```
ROUGE scores:
rouge-1: 0.8000 (F1) 0.8000 (Precision) 0.8000 (Recall)
rouge-2: 0.5556 (F1) 0.5556 (Precision) 0.5556 (Recall)
rouge-l: 0.8000 (F1) 0.8000 (Precision) 0.8000 (Recall)
```

Rouge метрикасы

```python
from nltk.translate.meteor_score import meteor_score
from nltk import word_tokenize, download
download('punkt')

reference_texts = ["Тарыны үшеуіміз тең бөліссек қаншадан тиеді?  Аз мөлшерде шығатын шығар."]
candidate_text = "Тарыны үшеуіміз тепе-тең бөліссек қаншадан тиеді? Шамалы мөлшерде шығатын шығар."

references = [word_tokenize(ref) for ref in reference_texts]
candidate = word_tokenize(candidate_text)

score = meteor_score(references, candidate)

print(f"METEOR score: {score:.4f}")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
METEOR score: 0.8221
```

Meteor метрикасы

Берілген жәнеде өзгертілген сөйлемдердің ұқсастығын анықтайтын метрикалар қолданылды. Сөйлемнің ұқсастығын бағалау әртүрлі табиғи тілді өңдеу тапсырмаларында, әсіресе қазақ тілі сияқты ресурсы аз тілдер контекстінде шешуші рөл атқарады. N-грамдық қабаттасу және еске түсіру-

дәлдік есептеулері сияқты негізгі құрамдас бөліктерімен танымал ROUGE метрикалары автоматты қорытындылау және машиналық аударма тапсырмаларында кеңінен қолданылған. Екінші жағынан, METEOR сөйлемнің ұқсастығының тиімділігін бағалау үшін семантикалық ұқсастық өлшемдерін енгізу және негізгі және синонимдік ақпаратты пайдалану арқылы басқа тәсілді ұсынады. Бұл көрсеткіштердің қыр-сырын және олардың қазақ тіліндегі көрсеткіштерін түсіну осы тілге бейімделген мәтінді бағалау жүйесін жетілдіру үшін өте маңызды.

ҚОРЫТЫНДЫ

Синонимдер тілдік өңдеуде шешуші рөл атқарады, қатынасудың байлығын және икемділігін арттырып, адамдарға өз ойлары мен идеяларын тиімді жеткізуге мүмкіндік береді. Тілдегі синонимдердің маңыздылығын түсіну арқылы зерттеушілер жасанды оқыту әдістерін қолдана отырып синонимдерді жасау модельдерін зерттеуді зерттейді. Жасанды оқыту синонимдерді контекстік релеванттылық пен семантикалық ұқсастыққа негізделген әдістермен автоматты түрде анықтау және ұйымдастыруға үмітті тәсіл ұсынады. Жасанды оқыту алгоритмдерінің мүмкіндіктерін пайдалана отырып, зерттеушілер тіл ішіндегі синонимдер қатынастарының нюанстарын және күрделіліктерін тиімді бейнелей алатын күрделі модельдерді жасай алады. Бұл модельдер синонимдерді жинақтау процесін жеңілдетіп қана қоймай, табиғи тілді өңдеудің одан әрі жетілдірілген қолданбаларына жол ашады.

Жұмысты орындау кезіндегі жасалған жұмыстар реті:

1. Зерттеу жұмыстары негізінде әдебиеттер мен ақпарат көздеріне анализ және шолу жасау.
2. Синонимдерді анықтап жақындығын табатын әдістері, алгоритмдері мен қалай қолдануын зерттеу.
3. Жұптас синонимдер сөздігін құру және зерттеу барысында алынған сөздерді өңдеу.
4. Қазақ тілінің синоним сөздері бар сөздерді анықтайтын моделді жобалау және әзірлеу.
5. Ғылыми-зерттеу жұмысының бағдарламасын құрастыру және сынау кезінде алынған нәтижелерді салыстырмалы талдау.

Жұмыс бойынша алынған нәтижелер:

1. Көп қолданылатын сөздердің барлық синонимдерін алу.
2. Қазақ тілінің синонимдер сөздерін анықтайтын модельді құру.
3. Сызбалардағы, диаграммалардағы жүйе архитектурасын құрастыру. Жүйе архитектурасы келесі кезеңдерден негізделген: қазақ тілі корпусын дайындау, бір-біріне мағынасы жағынан жақын сөздерді анықтайтын программаны әзірлеу, парафрейз программасын құру.
4. Парафрейз процессін іске асыратын программалық кодын жазу.
5. Қазақ тілінің морфологиясын зерттеу, тілдік қорларын жинау.

Диссертациялық жұмыста қазіргі кездегі компьютерлік лингвистиканың маңызды бағыттарының бірі – табиғи тілді өңдеу саласына ақпараттық технологияларды енгізу мәселесі зерттеліп, қарастырылады. Диссертациялық жұмыс 3 бөлімнен тұрады:

Магистрлік диссертацияның бірінші бөлімі «Зерттеу жұмыстары

бойынша әдебиеттерге шолу» деп аталады және лингвистика мен nlp түсінігі, тілдің мәселенің тарихы, лингвистикалық ресурстар түсінігі және Python бағдарламалау тілі туралы толық ақпаратты қамтиды, оған қоса қазақ тілінің морфологиясы және сөздердің жақындығы тұралы мағлұмат берілген.

Жұмыстың екінші бөлімі «Қазақ тілінің синонимдік сөздерін анықтайтын модельді жобалау және әзірлеу» деп аталып, осы бөлімде синонимдерді, яғни мағынасы жағынан жақын сөздерді анықтау әдістері мен алгоритмдері, парсинг, стемминг, тазалау алгоритдері мен технологиялары, қазақ тілі сөздерінің жақындығын көзге көрсететін технологиялар қарастырылды.

Ал жоғарыда аталған процестерді бейнелеу үшін сөздер арасындағы байланыстарды көрнекі түрде көрсетеді. Бұл бөлімде тілдік, яғни корпустық ресурстарды жинау, бағдарламаны жүзеге асыру және практикалық талдау қарастырылған. Бөлім «Жүйені жобалау және жүзеге асыру» деп аталды.

Зерттеуге сәйкес көптеген веб парақшалардан ақпараттарды алып парсинг секілді алгоритмдерді қолданып қазақ тілі корпусын жиналды. Қазақ тілі морфологиялық және лексикалық формасы мықты агглютинативті тіл болғандықтан парафрейз жүйесі талдау жасалып жобаланды. Корпус екіге болінді: Бірінші ол тек қана синоним сөздердің жиынтығы, ал екінші корпус үлкен қазақша тексттер болып келеді.

Жұмыстың ғылыми үлесі дамып келе жатқан қазақ тілінің синонимдік корпусын жинақтау жәнеде парафрейз жүйесін іске қосу болып саналады.

Бұл жобамен, жасалған зерттеуіммен 2022-2024 жылдары бірнеше хаалықаралық жәнеде басқа конференцияларға қатысу бұйырды. Осы зерттеуден алынған нәтижелермен қорытындыларға сәйкес ғылыми мақалалар мен баяндамалар жарық көрді:

1. 2023 жылы ICCCI (Computational Collective Intelligence 2023) конференциясында Development of a Dictionary for Preschool Children with Weak Speech Skills Based on the Word2Vec Method атты мақала жарық көрді.

2. 2023 жылы «Шет тілдер» кафедрасы ұйымдастырған «Globally Interconnected: New Opportunities and Challenges for Sustainable Development» атты IV халықаралық ғылыми және оқу әдістемелік жинақта мақалам шықты.

3. «Фараби әлемі» атты студенттер мен жас ғалымдардың халықаралық ғылыми конференциясына қатысып мақала жарияланды.

4. «ǴYLYM JÁNE BILIM - 2024» XIX Халықаралық ғылыми конференциясы, Астана қаласында баяндама жасалып мақала жинаққа басылды.

Болашақта қазақ тілінің парафрейздік жүйесін жетілдіру үшін

корпусты кеңейтіп, толықтыру жұмыстарын жүзеге асырамын.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1. “Natural Language Processing - Overview.” 2021. GeeksforGeeks. July 14, 2021. https://www.geeksforgeeks.org/natural-language-processing-overview/.
2. “What Is Natural Language Processing (NLP)?” 2024. Www.techtarget.com. January 5, 2024. https://www.techtarget.com.
3. Harrison, Conrad J., and Chris J. Sidey-Gibbons. 2021. “Machine Learning in Medicine: A Practical Introduction to Natural Language Processing.” *BMC Medical Research Methodology* 21 (1). https://doi.org/10.1186/s12874-021-01347-1.
4. PROJECTPRO. 2024. “10 NLP Techniques Every Data Scientist Should Know.” ProjectPro. April 11, 2024. https://www.projectpro.io/article/10-nlp-techniques-every-data-scientist-should-know/415.
5. Wolff, Rachel. 2021. “Natural Language Processing (NLP): 7 Key Techniques.” MonkeyLearn Blog. October 19, 2021. https://monkeylearn.com/blog/natural-language-processing-techniques/.
6. “Машинное обучение в лингвистике — все самое интересное на ПостНауке.” n.d. Postnauka.org. https://postnauka.org/longreads/82189.
7. “Что такое обработка естественного языка? – Объяснение NLP – AWS.” n.d. Amazon Web Services, Inc. https://aws.amazon.com/ru/what-is/nlp/.
8. Велихов, Павел . 2016. “Машинное обучение для понимания естественного языка.” Издательство “Открытые системы.” February 25, 2016. https://www.osp.ru/os/2016/01/13048649.
9. Straub, Jakob. 2021. “Машинное обучение: что такое обработка естественного языка?” Lingoda. April 29, 2021. https://blog.lingoda.com/ru/obrabotka-yestestvennogo-yazyka/.
10. Sawicki, Jan, Maria Ganzha, and Marcin Paprzycki. 2023. “The State of the Art of Natural Language Processing - a Systematic Automated Review of NLP Literature Using NLP Techniques.” Data Intelligence, July, 1–47. https://doi.org/10.1162/dint_a_00213.
11. Некоммерческий Фонд развития науки и образования ‘Интеллект.’” n.d. Intellect-Foundation.ru. https://intellect-foundation.ru/possibilities/studentu/electives/metodyi-mashinnogo-obucheniya-dlya-resheniya-proektnyix-zadach-kompyut/.Зуйкова, Ася. 2023. “Машинное обучение: типы, задачи, примеры.”

12. РБК Тренды. January 27, 2023. https://trends.rbc.ru/trends/industry/60c85c599a7947f5776ad409.
13. Gulzhana Jenalayeva, Gaukhar Niyar, and Moldir Zhubanyshbayeva. 2021. “Conceptualization of the Kazakh Language in the Linguistic Consciousness of the Kazakhs.” Journal of Humanities and Social Sciences Studies 3 (4): 67–71. https://doi.org/10.32996/jhsss.2021.3.4.8.
14. Житникова, Оксана. 2023. “Сложно ли выучить казахский язык?” Zakon.kz. July 26, 2023. https://www.zakon.kz/sovety/6401262-slozhno-li-vyuchit-kazakhskiy-yazyk.html.
15. Yermekova T.N, Odanova S.A, Zhangabayeva N.A., Esenbayeva G.Zh, and Akylbayeva T.Zh. 2018. “ACTUAL PROBLEMS of the KAZAKH LANGUAGE during the TRANSITION to the LATIN SCRIPT - European Journal of Natural History.” World-Science.ru. European Journal of Natural History. September 14, 2018. https://world-science.ru/en/article/view?id=33917
16. Kevin Martens Wong. 2014. “Kazakh: Language of the Steppes | Unravel Magazine.” Unravel. 2014. https://unravellingmag.com/articles/langprofile-kazakh/.
17. Rakhimova, Diana, Yerkin Suleimenov, and Dinara Makulbek. 2023. “Analysis of Kazakh Language Abbreviations Based on Machine Learning Approach.” 2023 8th International Conference on Computer Science and Engineering (UBMK) (September). https://doi.org/10.1109/ubmk59864.2023.10286751.
18. Henriksson, Aron, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, and Martin Duneld. 2014. “Synonym Extraction and Abbreviation Expansion with Ensembles of Semantic Spaces.” Journal of Biomedical Semantics 5 (1): 6. https://doi.org/10.1186/2041-1480-5-6.
19. Kudubayeva, Saule, Botagoz Zhusupova, and Meruyert Salkenova. 2022. “EasyChair Preprint Machine Learning for Syntactic and Morphological Analysis of Text in the Kazakh Language.” https://easychair.org/publications/preprint_open/dNc2
20. Kuanyshbay, Darkhan & Shoiynbek, Aisultan & Amirgaliyev, Yedilkhan. (2019). Построение языкой модели для казахского языка на основе рекуррентных нейронных сетей.
21. D. Rakhimova, D. Kassymova, and D. Isabaeva. 2021. “RESEARCH and DEVELOPMENT of a QUESTION and ANSWER SYSTEM BASED on the BERT MODEL for the KAZAKH LANGUAGE.” Habaršy. Fizika-Matematika Seriâsy 76 (4): 119–27. https://doi.org/10.51889/2021-4.1728-7901.16.
22. “ИССЛЕДОВАНИЕ СОВРЕМЕННЫХ АНАЛИЗАТОРОВ ТЕКСТОВОЙ ИНФОРМАЦИИ.” n.d. Cyberleninka.ru. Accessed May 20, 2024. https://cyberleninka.ru/article/n/issledovanie-sovremennyh-analizatorov-tekstovoy-informatsii/viewer.

23. Dandekar, Nikhil. 2016. “How to Use Machine Learning to Find Synonyms.” Medium. May 10, 2016. https://medium.com/@nikhilbd/how-to-use-machine-learning-to-find-synonyms-6380c0c6106b.
24. Otten, Neri Van. 2023. “Semantic Analysis in NLP Made Easy, Top 10 Best Tools & Future Trends.” Spot Intelligence. October 16, 2023. https://spotintelligence.com/2023/10/16/semantic-analysis-in-nlp/.
25. “Data Augmentation for Text Classification.” 2019. Stack Overflow. February 2019. https://stackoverflow.com/questions/54797225/data-augmentation-for-text-classification.
26. Yeshpanov, Rustem, Alina Polonskaya, and Huseyin Atakan Varol. 2024. “KazParC: Kazakh Parallel Corpus for Machine Translation.” Arxiv.org. April 24, 2024. https://arxiv.org/html/2403.19399v3.
27. Bobur Allaberdiev, Gayrat Matlatipov, Elmurod Kuriyozov, and Zafar Rakhmonov. 2024. “Parallel Texts Dataset for Uzbek-Kazakh Machine Translation.” Data in Brief 53 (April): 110194–94. https://doi.org/10.1016/j.dib.2024.110194.
28. Dauren Ayazbayev, Andrey Bogdanchikov, Kamila Orynbekova, and Iraklis Varlamis. 2023. “Defining Semantically Close Words of Kazakh Language with Distributed System Apache Spark.” Big Data and Cognitive Computing 7 (4): 160–60. https://doi.org/10.3390/bdcc7040160.
29. Yeshpanov Rustem, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. “Proceedings of the Thirteenth Language Resources and Evaluation Conference.” GitHub. European Language Resources Association. June 2022. https://aclanthology.org/2022.lrec-1.44.
30. Şahin, Gözde Gül. 2021. “To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP.” Computational Linguistics, December, 1–38. https://doi.org/10.1162/coli_a_00425.
31. CR, Aravind. 2021. “Exploring Clustering Algorithms: Explanation and Use Cases.” Neptune.ai. September 17, 2021. https://neptune.ai/blog/clustering-algorithms.
32. Shin, Terence. 2022. “Top Three Clustering Algorithms You Should Know instead of K-Means Clustering.” Medium. December 12, 2022. https://terenceshin.medium.com/top-five-clustering-algorithms-you-should-know-instead-of-k-means-clustering-b22f25e5bfb4.
33. Akalin, Altuna. n.d. 4.1 Clustering: Grouping Samples Based on Their Similarity | Computational Genomics with R. Compgenomr.github.io. https://compgenomr.github.io/book/clustering-grouping-samples-based-on-their-similarity.html.
34. McGregor, Milecia. 2020. “8 Clustering Algorithms in Machine Learning That All Data Scientists Should Know.” FreeCodeCamp.org. September

21, 2020. https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/.
35. Prasad, Sunit. 2020. “Types of Clustering Algorithms in Machine Learning with Examples.” Blogs & Updates on Data Science, Business Analytics, AI Machine Learning. July 5, 2020. https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/.
36. “Natural Language Processing (NLP) - a Complete Guide.” 2023. Www.deeplearning.ai. January 11, 2023. https://www.deeplearning.ai/resources/natural-language-processing/.
37. Mittal, Deepti. 2023. “Exploring the Intersection of Compiler Design and Natural Language Processing.” Medium. January 21, 2023. https://medium.com/@deeptimittal29/exploring-the-intersection-of-compiler-design-and-natural-language-processing-237e2db584c7.
38. Mayo, Matthew. 2022. “Natural Language Processing Key Terms, Explained.” KDnuggets. May 16, 2022. https://www.kdnuggets.com/2017/02/natural-language-processing-key-terms-explained.html.
39. Bakrey, Mohamed. 2023. “All about Lexicons in NLP.” Medium. March 25, 2023. https://mohamedbakrey094.medium.com/all-about-lexicons-in-nlp.
40. Vladislav Karyukin, Diana Rakhimova, Aidana Karibayeva, Aliya Turganbayeva, and Asem Turarbek. 2023. “The Neural Machine Translation Models for the Low-Resource Kazakh–English Language Pair.” PeerJ. Computer Science 9 (February): e1224–24. https://doi.org/10.7717/peerj-cs.1224.
41. Bogdanchikov, Andrey, Dauren Ayazbayev, and Iraklis Varlamis. 2022. “Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text.” Big Data and Cognitive Computing 6 (4): 123. https://doi.org/10.3390/bdcc6040123.
42. Полицына, Екатерина Валерьевна, Сергей Александрович Полицын, Александр Сергеевич Поречный, and Екатерина Евгеньевна Милованова. 2020. “Анализ подходов к автоматическому выделению контекстных синонимов из текстов на русском языке.” Вестник ВГУ. Серия: Системный анализ и информационные технологии, no. 3 (September): 120–32. https://doi.org/10.17308/sait.2020.3/3046.
43. IBM. 2021. “Об исследовании текста.” Www.ibm.com. August 18, 2021. https://www.ibm.com/docs/ru/spss-modeler/saas?topic=analytics-about-text-mining.
44. Edward. 2018. “Блог про HR-аналитику: Анализ тональности текста с использованием Word2vec и реализацией в Pipeline Python.” Блог про HR-аналитику. 2018. https://edwvb.blogspot.com/2018/07/analiz-tonalnosti-teksta-s-ispolzovaniem-word2vec-i-realizaciej-v-pipeline-python.html.

45. “Обзор четырёх популярных NLP-моделей.” 2020. Библиотека программиста. April 21, 2020. https://proglib.io/p/obzor-chetyreh-populyarnyh-nlp-modeley-2020-04-21.
46. “🤖 Практическое руководство по NLP: изучаем классификацию текстов с помощью библиотеки FastText.” 2021. Библиотека программиста. August 21, 2021. https://proglib.io/p/prakticheskoe-rukovodstvo-po-nlp-izuchaem-klassifikaciyu-tekstov-s-pomoshchyu-biblioteki-fasttext-2021-08-28.
47. Кощенко, Екатерина. 2019. “Представление отношений между словами естественного языка в виде разложения на симметрический и кососимметрический линейный оператор – Выпускные квалификационные работы студентов НИУ ВШЭ – Национальный исследовательский университет ‘Высшая школа экономики.’” Www.hse.ru. 2019. https://www.hse.ru/edu/vkr/296298068.
48. “Эффективный парсинг веб-страниц с помощью BeautifulSoup Python — легкое извлечение данных.” 2024. Fineproxy.org. February 18, 2024. https://fineproxy.org/ru/beautifulsoup-python-web-scraping/.