ҒТАМР 50.13.13 **https://doi.org/10.58805/kazutb.v.1.30-1415**

ОНТОЛОГИЯЛЫҚ ТЕЗАУРУСТЫ ҚОЛДАНУ ӘДІСТЕМЕСІ СҰРАҚ-ЖАУАП ЖҮЙЕЛЕРІНЕ АРНАЛҒАН БАҒДАРЛАМАЛЫҚ МОДУЛЬ ҚҰРУ

[1]М.Ж. Ергеш, [1]Б.Ж. Ергеш, [1]К.М. Максутова, [2]А.Д. Тулегулов

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
[2]Қ. Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан

✉ *Корреспондент-авторлар: shymkent90@mail.ru, b.yergesh@gmail.com*

Қазақ тілі үшін цифрлық ресурстар тапшылығы жағдайында Зияткерлік бағдарламаланатын жүйелерді қолдау үшін Қазақстан тарихы мектебінің курсының пәндік білімін ресімдеу ерекше өзекті болып отыр. Бұл жұмыста 5-10 сынып оқулықтарының толық корпусы негізінде құрылған Қазақстан тарихы бойынша зияткерлік бағдарламаланатын жүйелерді зерттеу нәтижелері ұсынылады және алынған білім графигінің QA-міндеттері үшін контекст іздеудің сапасына әсері талданады.

Зерттеудің мақсаты-терминдерді теру алгоритмін әзірлеу (person, period, place, source, event, unknown). Бастапқы ”астық” белгілерін алу үшін жоғары дәлдіктегі үлгілер пайдаланылады, содан кейін оқыту логистикалық регрессияны пайдалана отырып, XLM-R трансформаторының эмбеддингтерінде жүргізіледі. Seed-белгілеу бойынша валидацияда негізгі сыныптар (person, period, place) үшін 0,89 және F1-ден 0,92-ге дейінгі дәлдікке қол жеткізіледі.

Зерттеудің ғылыми жаңалығы онтологияның практикалық пайдалылығын бағалау әдісін әзірлеу болып табылады QA-661 мысалдың жиынтығы (528 пойыз /133 тест) кепілдендірілген дұрыс жауап спандарымен. Ұсынылған әдіс бірнеше семантикалық іздеу конфигурацияларын салыстырады: оқулық бөлімдері бойынша негізгі іздеу, онтологиялық ”құжаттар” бойынша іздеу және тезаурус пен онтология арқылы сұраныстар мен контексттерді кеңейту нұсқалары. Нәтижелер онтологиялық құжаттарды қолдану p@1-ді шамамен 0,56-дан 0,86-ға дейін және ndcg@10-ны ≈0,77-ден ≈0,92-ге дейін арттыратынын көрсетеді; біріктірілген ”онтология + кеңейту” схемасы ndcg-нің жоғары деңгейін және сәйкес контексттерді толық қамтуды сақтайды.

Осылайша, ұсынылған онтологиялық бағдарланған тезаурус пен Қазақстан тарихы бойынша білім графы 5–10 сыныптарға арналған оқу материалын құрылымдап қана қоймай, қазақ тіліндегі сұрақ-жауап жүйелерінде контекст іздеу сапасын айтарлықтай жақсартады және доменге бейімделген QA-модельдерді оқыту мен бағалауға арналған берік негіз қалыптастырады.

Түйін сөздер: цифрландыру, алгоритм, интеллектуалды бағдарламаланатын жүйелер, тезаурус, XLM-R, терминдерді теру, RDF/OWL, контекст іздеу.

МЕТОДОЛОГИЯ ПРИМЕНЕНИЯ ОНТОЛОГИЧЕСКОГО ТЕЗАУРУСА ПОСТРОЕНИЯ ПРОГРАММНОГО МОДУЛЯ ДЛЯ ВОПРОСНО-ОТВЕТНЫХ СИСТЕМ

[1]М.Ж. Ергеш, [1]Б.Ж. Ергеш, [1]К.М. Максутова, [2]А.Д. Тулегулов

[1]Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан,
[2]Казахский университет технологии и бизнеса им. К. Кулажанова, Астана, Казахстан,
e-mail: shymkent90@mail.ru, b.yergesh@gmail.com

В условиях дефицита цифровых ресурсов для казахского языка особую актуальность приобретает формализация предметных знаний школьного курса истории Казахстана для поддержки интеллектуальных программируемых систем. В данной работе представляются результаты исследования интеллектуальных программируемых систем по истории Казахстана, построенных на основе полного корпуса учебников 5-10 классов, и анализируется влияние полученного графа знаний на качество поиска контекста для QA-задач.

Цель исследования заключается в разработке алгоритма типизация терминов (person, period, place, source, event, unknown). Для получения начальных «зёрновых» меток используются высокоточные шаблоны, после чего обучение проводится на эмбеддингах трансформера XLM-R с использованием логистической регрессии. На валидации по seed-разметке достигается точность порядка 0,89 и F1 до 0,92 для основных классов (person, period, place).

Научная новизна исследования заключается в разработке метода оценки практической полезности онтологии, формируемой из QA-набор из 661 примера (528 train /133 test) с гарантированно корректными спанами ответов. В предложенном методе сравниваются несколько конфигураций семантического поиска: базовый поиск по разделам учебников, поиск по онтологическим «документам», а также варианты с расширением запросов и контекстов за счёт тезауруса и онтологии. Полученные результаты показывают, что использование онтологических документов повышает P@1 примерно с 0,56 до 0,86 и nDCG@10 с ≈0,77 до ≈0,92; комбинированная схема «онтология + расширение» сохраняет высокий уровень nDCG и почти полное покрытие релевантных контекстов.

Таким образом, предложенный нами метод не только структурируют учебный материал для 5-10 классов, но и демонстрируют существенное улучшение качества извлечения контекста в вопросно-ответных системах на казахском языке, формируя основу для последующих работ по обучению и оценке доменно-специализированных QA-моделей.

Ключевые слова: цифровизация, алгоритм, интеллектуально-программируемые системы, тезаурус, XLM-R, типизация терминов, RDF/OWL, поиск контекста.

METHODOLOGY OF APPLYING ONTOLOGICAL THESAURUS BUILDING A PROGRAM MODULE FOR QUESTION-ANSWER SYSTEMS

[1]M.Zh. Yergesh✉, [1]B.Zh. Yergesh✉, [1]K.M. Maksutova, [2]A.D. Tulegulov

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
[2]K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan,
e-mail: shymkent90@mail.ru, b.yergesh@gmail.com

In the context of a shortage of digital resources for the Kazakh language, the formalization of subject knowledge from the school history course of Kazakhstan is of particular relevance to support intelligent programmable systems. This paper presents the results of a study of intelligent programmable systems on the history of Kazakhstan, built on the basis of a full corpus of textbooks for grades 5-10, and analyzes the impact of the obtained knowledge graph on the quality of context search for QA tasks.

The purpose of the study is to develop an algorithm for term typing (person, period, place, source, event, unknown). To obtain initial ”grain” labels, high-precision templates are used, after which training is carried out on XLM-R transformer embeddings using logistic regression. On validation by seed-labeling, accuracy of the order of 0.89 and F1 up to 0.92 for the main classes (person, period, place) are achieved.

The scientific novelty of the study lies in the development of a method for assessing the practical usefulness of an ontology formed from a QA set of 661 examples (528 train /133 test) with guaranteed correct answer spans. The proposed method compares several semantic search configurations: basic search across textbook sections, search across ontological ”documents,” and variants that expand queries and contexts using a thesaurus and ontology. The results show that using ontology documents increases P@1 from approximately 0.56 to 0.86 and nDCG@10 from ≈0.77 to ≈0.92; the combined ontology + expansion scheme maintains a high level of nDCG and almost complete coverage of relevant contexts.

Overall, the proposed ontology-oriented thesaurus and knowledge graph for the History of Kazakhstan not only structure the educational content for grades 5–10 but also demonstrate a substantial improvement

in context retrieval quality for Kazakh-language QA systems, forming a foundation for future research on training and evaluating domain-specialized QA models.

Keywords: digitalization, algorithm, intelligent programmable systems, thesaurus, XLM-R, term typing, RDF/OWL, and context search.

Кіріспе. Интеллектуалды білім беру жүйелерін дамыту оқу мазмұнын құрылымдалған білім түрінде ұсынуды талап етеді. Қазақ тілі үшін бұл міндет цифрлық білім беру ресурстарының, әсіресе мектеп пәндері саласында, шектеулі болуына байланысты күрделене түседі.5-10 сыныптарға арналған Қазақстан тарихы курсы тарихи оқиғалар, тұлғалар, кезеңдер мен географиялық объектілерге бай болғанымен, дәстүрлі мәтіндік форматта берілуі оны сұрақ-жауап жүйелерінде тиімді пайдалануды қиындатады [1].

Сұрақ-жауап жүйелерінің тұрақты жұмыс істеуі үшін тек мәтіндік корпус емес, пәндік ұғымдар мен олардың өзара байланыстарын қамтитын құрылымдалған модель қажет. Терминдер мен қатынастар айқындалмаған жағдайда, іздеу жолдық сәйкестікке сүйенеді, бұл қазақ тілінің морфологиялық және терминологиялық ерекшеліктері жағдайында жеткіліксіз [1]. Осыған байланысты тезаурус пен пәндік онтологияға негізделген білім графын қалыптастыру қажеттілігі туындайды.

Ұсынылған тәсіл оқу мәтіндерінің корпусы, терминологиялық қабат және онтологиялық графты біріктіруге негізделеді. Әлсіз бақыланатын машиналық оқыту әдістерін қолдану терминдерді типизациялауға және алынған білім графын сұрақ-жауап жүйесінің семантикалық іздеу контурына кіріктіруге мүмкіндік береді. Онтологиялық модельдің тиімділігі релевантты контекстіні іздеу сапасын бағалау арқылы эмпирикалық түрде айқындалады [2].

Білім беру саласының пәндік облыстары үшін онтологиялар мен тезаурустарды құрастыру мәселесі интеллектуалды оқыту жүйелері, адаптивті электрондық оқулықтар және компьютерлік оқытуды қолдау жүйелері контекстінде кеңінен зерттеледі. Онтологиялар пәндік білімді ұсынудың негізгі тетігі ретінде терминологиялық келісімділікті қамтамасыз етуге, ұғымдар арасындағы қатынастарды айқындауға және білімді әртүрлі білім беру қосымшаларында қайта пайдалануға мүмкіндік береді. Сұрақ-жауап жүйелері мен семантикалық іздеу саласындағы заманауи зерттеулерде нейрондық модельдер мен білім графтарын біріктіретін гибридті архитектураларға ерекше назар аударылады, бұл сұрауларды дәл интерпретациялауға және релевантты мәтіндік фрагменттерді тиімді таңдауға жағдай жасайды. Ресурсы шектеулі тілдер, соның ішінде түркі тілдері үшін, арнайы құрылған домендік корпустар мен онтологиялардың маңыздылығы бірқатар еңбектерде атап өтіледі [3].

Тарих пәні саласында тарихи онтологиялар мен білім графтарына арналған зерттеулер бар болғанымен, олар негізінен ірі еуропалық тілдерге бағытталған. Қазақ тілі үшін, әсіресе Қазақстан тарихының мектеп курсы аясында, жүйелі түрде сипатталған онтологиялар мен тезаурустар іс жүзінде жоқ, бұл пәндік білімді формализациялау модельдерінің және онтологиялардың сұрақ-жауап жүйелеріндегі тиімділігін бағалайтын эмпирикалық зерттеулердің тапшылығын көрсетеді.

Осы зерттеуде эмпирикалық бастапқы материал ретінде 5-10 сыныптарға арналған Қазақстан тарихы мектеп оқулықтарының толық мәтіндері пайдаланылды. Алдын ала белгілеу нәтижесінде әр секциясы оқулық, бөлім және сынып идентификаторларымен сипатталатын 208 мәтіндік секциядан тұратын бірыңғай корпус қалыптастырылды [4]. Корпус мәтіндері регистрді, бос орындарды және тыныс белгілерін нормализациялау рәсімдерінен өтіп, қазақ орфографиясының ерекшеліктеріне бейімделген токенизация негізінде өңделді [5].

Сонымен қатар оқулық мәтіндерінен автоматты түрде алынған онтологиялық триплеттер негізінде 2 973 түйін мен 2 318 қырдан тұратын бағытталған білім графы құрылды, мұнда

қатынастар негізінен хронологиялық, кеңістіктік және себеп-салдарлық сипатта беріледі [6]. Онтологияның контексті іздеу сапасына әсерін бағалау үшін Қазақстан тарихы бойынша ашық типті сұрақтар мен жауаптардан тұратын QA-датасет пайдаланылды: 889 бастапқы жазбаның ішінен жауаптың мәтіндегі орны бірмәнді анықталатын 661 мысал іріктеліп, олар оқу және тест жиынтықтарына бөлінді [7].

Нормаланған корпус пен онтологиялық триплеттер негізінде корпустық термин-кандидаттар мен бастапқы онтология терминдерін біріктіру арқылы 8 345 нормаланған терминнен тұратын пәндік тезаурус құрастырылды. Бұл тезаурус кейінгі терминдерді типизациялау және онтологиялық бағдарланған сұрақ-жауап жүйелерін құру кезеңдері үшін негізгі кіріс дерек ретінде пайдаланылды [8].

Материалдар мен әдістер. Зерттеудің эмпирикалық негізі ретінде жалпы білім беретін мектептің 5-10 сыныптарына арналған Қазақстан тарихы оқулықтарының корпусы пайдаланылды. Корпусқа «Қазақстан тарихы» сериясының ресми бекітілген алты басылымы енгізіліп, олар логикалық тұрғыдан аяқталған фрагменттерге бөлінді. Нәтижесінде корпус 208 мәтіндік секциядан тұрады: 5-сынып - 46, 6-сынып - 54, 7-сынып - 29, 8-сынып - 28, 9-сынып - 29 және 10-сынып - 22 секция. Әр секция оқулық идентификаторларымен, сынып деңгейімен, бастапқы және нормаланған мәтінмен, сондай-ақ токендер санымен сипатталады, бұл лексиканы корпус деңгейінде талдауға және терминдер мен онтологиялық фактілерді нақты мәтіндік фрагменттермен байланыстыруға мүмкіндік береді [9].

Бастапқы PDF-құжаттар алдын ала өңдеуден өтті: UTF-8 форматына түрлендіру, қызметтік элементтерді жою, регистр мен тыныс белгілерін нормализациялау және сөз тасымалдарын қалпына келтіру жүзеге асырылды. Кейін мәтіндер бос орындар мен базалық тыныс белгілері бойынша токенизацияланды.

Нормаланған корпус негізінде термин-кандидаттар іріктелді. Минималды жиілік шегіне сай келетін униграммалар қарастырылып, стоп-сөздер мен қызметтік лексиканы алып тастағаннан кейін 5 692 нормаланған токен-кандидат алынды [10]. Әр кандидат үшін нормаланған форма, жиілік бағасы, жылдық мәндерді көрсететін белгі және дерек көзі тіркелді.

Корпус пен онтология терминдерін біріктіру нәтижесінде тарихи терминдер тезаурусы қалыптастырылды. Әрбір жазба терминнің бастапқы және нормаланған формасын, жиілік бағасын, шығу көзін, термин типін және (бар болған жағдайда) нормаланған синонимдер тізімін қамтиды. Қорытынды тезаурус 8 345 бірегей нормаланған терминнен тұрады және корпус деректері мен онтологиялық құрылымдар арасындағы байланысты қамтамасыз етеді.

Онтологиялық триплеттерді шығару және интеграциялау. Әрбір оқулық үшін фактографиялық тұжырымдарды триплеттер түрінде сақтайтын жартылай автоматты түрде құрылған *ontology_triples.csv* файлы пайдаланылды. Триплеттер келесі үлгіде беріледі:

$$\langle \text{subject, relation_kk, object} \rangle ,$$

оларға қоса book_id, section_id, pages, evidence (негіздеме ретінде алынған мәтін фрагменті) және confidence (триплетті алудың сенімділік бағасы) өрістері тіркеледі.

Алты файлды біріктіру нәтижесінде көлемі 2318 триплеттен тұратын «біріздендірілген онтология» алынды. Осы триплеттер негізінде $G = (V, E)$, түріндегі бағытталған білім графы құрылған, мұнда төбелер жиыны

$$V = \{v : v = \text{subject} \vee v = \text{object}\}$$

2973 бірегей мәндіден (сущность) тұрады, ал қырлар жиыны

$$E = \{(v_i, r, v_j)\},$$

олардың арасындағы қатынастармен таңбаланған доғаларды сипаттайды.

Графтың статистикалық талдауы орташа дәреженің шамамен 1.56, ал максималды дәреженің 27-ге тең екенін көрсетті; визуализация үшін

ең жоғары байланыстылығы бар түйіндердің локалды подграфтары да, дәрежесі екіден кем емес түйіндерден тұратын подграф та пайдаланылды. Триплеттердің оқулықтар бойынша бөлінісі (бір кітапқа 94-тен 776-ға дейін) онтологияның оқу материалын қаншалықты қамтитынын талдау және болашақ эксперименттердегі ықтимал стратификация (қабаттарға бөлу) үшін жеке сақталады.

Онтология екі түрлі түрде сақталады: кестелік ұсынылым (DataFrame) ретінде және графтық құрылым (бағытталған мультиграф) ретінде. Бұл онтологияны бір мезгілде қосымша құжаттар (анықтамалар мен түсіндірмелер) көзі ретінде де, сұрауларды семантикалық кеңейтуге негіз болатын білім графы ретінде де пайдалануға мүмкіндік береді.

Трансформерлерді пайдалана отырып терминдерді әлсіз бақыланатын типизациялау. Терминдерді ықшам пәндік сыныптар жүйесіне келтіру үшін әлсіз бақыланатын оқытудың екі кезеңді сұлбасы іске асырылды:

1. эвристикаға негізделген seed-таңбаларды тағайындау,

2. трансформерлік ұсынуларға сүйенген көпсыныпты классификация.

Бірінші кезеңде әрбір терминге τ домендік эвристикалар негізінде анықталатын «тұқымдық» (seed) таңба

$$y^{seed}(\tau) \in \{person, period, place, event, source, unknown\}$$

сәйкестендірілді. Бұл үшін персоналияларға, топонимдерге, тарихи кезеңдерге және дереккөздерге тән индикаторлық жұрнақтар мен лексемалар тізімдері, сондай-ақ жылдар мен жылдық интервалдарды анықтайтын үлгілер (регулярлы өрнектер) пайдаланылды.

Алынған seed-жиын айқын теңгерімсіз болды: 8345 терминнің ішінен сенімді таңба 860 терминге ғана берілді (оның ішінде 411 - «person», 363 - «period», 86 - «place», ал «event» және «source» сыныптары үшін тек жекелеген жағдайлар тіркелді); қалған терминдер «unknown» сыныбына жатқызылды. Машиналық оқыту моделін оқыту үшін ең көп толтырылған үш сынып: *person, period* және *place* қана пайдаланылды.

Екінші кезеңде терминдер көптілді *xlm-roberta-base* трансформерлік моделінің көмегімен векторлық кеңістікке бейнеленді. Әрбір термин үшін

$$h_\tau \in \mathbb{R}^{768},$$

өлшемді векторлық ұсыну есептелді; ол модельдің соңғы жасырын қабатындағы [CLS] токенінің эмбеддигі ретінде алынды.

Осы эмбеддингтер негізінде көпсыныпты логистикалық регрессия моделі келесі ықтималдық функциясымен оқытылды:

$$p(y = c \mid h_\tau) = \frac{\exp(w_c^\top h_\tau + b_c)}{\sum_{c' \in C} \exp(w_{c'}^\top h_\tau + b_{c'})},$$

Мұнда $C = \{$"person", "period", "place"$\}$, ал параметрлер жиыны $\{w_c, b_c\}$seed-таңбалар бойынша кросс-энтропия шығынын минимизациялау арқылы бағаланды:

$$L = -\sum_{\tau \in T_{\text{seed}}} \log p(y^{\text{seed}}(\tau) \mid h_\tau).$$

Модельдің орнықтылығын бағалау үшін таңдаманы (выборка) 80/20 қағидаты бойынша қабатталған бөліну (стратификация) арқылы оқыту және валидация жиынтықтарына бөлдік. Валидация жиынтығында дәлдік (accuracy) шамамен 0.89, ал макро-F1 шамамен 0.84 деңгейінде болды, бұл трансформерлік ұсынулардың seed-эвристикалармен жеткілікті дәрежеде келісімді екенін көрсетеді.

Трансформер негізіндегі модель шығарған нәтижелік сыныптық болжам

$$y^{\text{ML}}(\tau)$$

seed-таңбамен жоғары дәлдік қағидасы бойынша біріктірілді:

$$y(\tau) = \begin{cases} y^{\text{seed}}(\tau), & \text{егер } y^{\text{seed}}(\tau) \neq \text{unknown}, \\ y^{\text{ML}}(\tau), & \text{онда.} \end{cases}$$

Осы гибридті сұлбаның нәтижесінде терминдердің пәндік сыныптар бойынша таралуы төмендегідей болды:

- 1158 термин - «person»,
- 585 термин - «period»,
- 238 термин - «place»,
- 35 термин - «source»,
- 14 термин - «event»,
- 6315 термин - «unknown».

Алынған сыныптық таңбалар әрі қарай пәндік білім графының құрылымын талдауда, сондайақ сұрауларды мақсатты түрде семантикалық кеңейтуде (әсіресе персоналиялар мен тарихи кезеңдерді басымдықпен қосу кезінде) қолданылды.

Сұрақ-жауап жүйесі үшін контекстіні іздеу моделі. Онтология мен тезаурустың релевантты контекстіні іздеуге әсерін бағалау үшін арнайы QA-жиынтық құрылды. Бастапқы 889 аннотацияланған жазбаның ішінен оқулық фрагменттері мен мәтіндік дәлелдермен байланыстырылған 661 сұрақ іріктеліп, олар 528 оқыту және 133 тест мысалына бөлінді.

Контекстіні іздеу әрбір сұрақ үшін құжаттарды (оқулық секциялары мен онтологиялық анықтамаларды) ранжирлеу міндеті ретінде қарастырылды. Берілген кандидаттар жиыны үшін базалық модель сұрақ пен құжат арасындағы мәтіндік ұқсастыққа (косинустық ұқсастық немесе Okapi BM25) негізделген ранжирленген тізім құрады.

Экспериментте төрт конфигурация қолданылды. Plain нұсқасында индексе тек оқулық мәтіндері енгізіліп, сұрау ретінде сұрақ мәтіні ғана пайдаланылды. OntologyDocs конфигурациясында индексе онтологиялық evidence және subject–relation–object өрістерінен алынған анықтамалық құжаттар қосылды. Plain+Expansion нұсқасында индекс өзгеріссіз қалып, сұраулар тезаурус негізінде семантикалық түрде кеңейтілді. OntologyDocs+Expansion конфигурациясы кеңейтілген индекстеу мен сұрауларды онтология және тезаурус арқылы кеңейтуді бір мезгілде іске асырады.

Ранжирлеу сапасы ақпараттық іздеу саласындағы стандартты метрикалар бойынша бағаланды: алғашқы k құжат бойынша дәлдік $P@k$, толықтық $P@k$ және нормаланған дисконтталған жинақталған пайдалық DCG $nDCG@k$. Бинарлық релеванттік жағдайында (құжатта жауаптың дұрыс фрагменті бар/жоқ) жекелеген сұрақ qүшін, ранжирлеу π_qберілген кезде метрикалар келесідей анықталды:

$$P@k(q) = \frac{1}{k}\sum_{i=1}^{k} rel(\pi_q(i)), \quad R@k(q) = \frac{1}{|\,R_q\,|}\sum_{i=1}^{k} rel(\pi_q(i)),$$

мұнда $\text{rel}(d) \in \{0,1\}$- құжат dүшін релеванттілік индикаторы, ал R_q- сұрақ q үшін барлық релевантты құжаттар жиыны.

DCG метрикасы k нүктесіндегі шектеуде келесі формула бойынша есептелді:

$$DCG@k(q) = \sum_{i=1}^{k} \frac{2^{rel(\pi_q(i))} - 1}{\log_2(i+1)},$$

ал нормаланған нұсқасы:

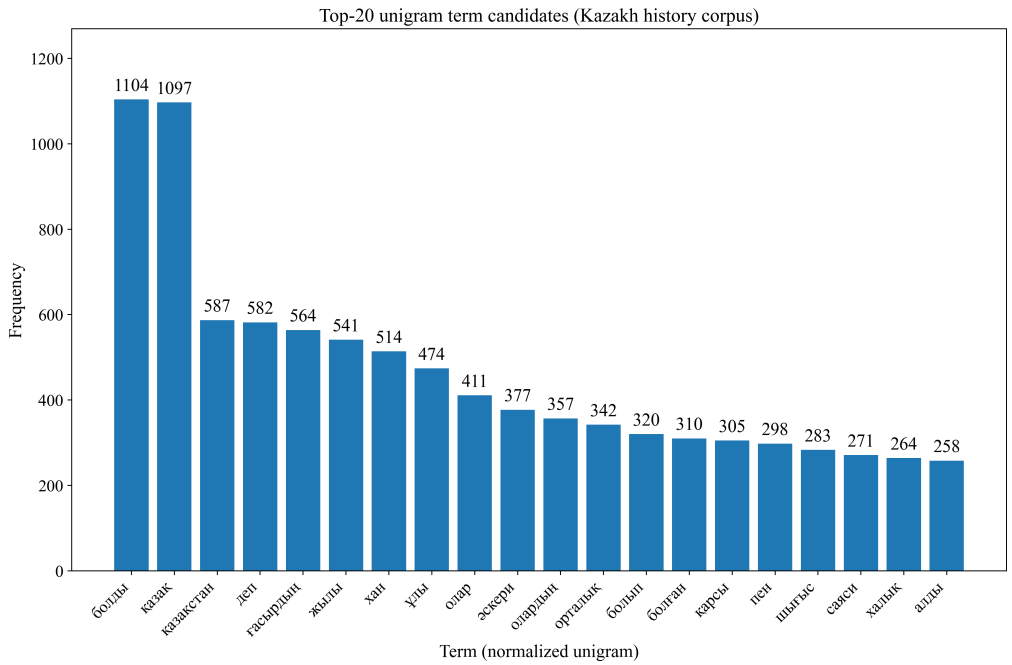
$$nDCG@k(q) = \frac{DCG@k(q)}{IDCG@k(q)},$$

мұнда $IDCG@k(q)$- идеал ранжирлеу жағдайындағы DCG-дің ең жоғарғы мүмкін мәні (яғни, релевантты құжаттардың барлығы тізімнің басында тұрғандағы мән).

Жүйе үшін финалдық метрика мәндері тест жиынтығындағы барлық сұрақтар бойынша усреднение (орташа мәнін алу) арқылы есептелді.

Нәтижелер мен талқылау. Корпус пен лексиканың сипаттамалары. Қазақстан тарихы бойынша 5-10 сыныптарға арналған мектеп оқулықтары негізінде қалыптастырылған корпус 208 мәтіндік секциядан тұрады және оқу бағдарламасының негізгі тақырыптық бағыттарын бірқалыпты қамтиды. Секциялар көлемі қысқа кіріспе фрагменттерден бастап мыңнан астам токені бар ірі параграфтарға дейін ауытқиды, бұл контекст іздеу эксперименттері үшін реалистік жағдайды қамтамасыз етеді [11].

1-суретте 5-10 сынып оқулықтарының 208 секциясынан алынған ең жиі кездесетін 20 нормаланған униграмманың жиілік үлестірімі көрсетілген. Ең жоғары жиіліктер жалпы қолданыстағы және тарихи дискурсқа тән лексемаларға тиесілі. Үлестірім айқын оң жақты асимметриямен сипатталады: жоғары жиілікті токендердің аз бөлігі жалпы кездесулердің елеулі үлесін құрайды, ал төмен жиілікті бірліктер үлестірімнің «құйрығын» қалыптастырады.



1-сурет. Қазақстан тарихы- мектеп оқулықтары корпусындағы термин-кандидаттардың ТОП-20 униграммасы

Бұл нәтиже терминологиялық кандидаттарды кейінгі кезеңде қосымша сүзгіден өткізудің қажеттілігін көрсетеді, өйткені тек жиілікке негізделген іріктеу бір жағынан әлеуетті маңызды тарихи ұғымдарды, екінші жағынан пәндік термин болып табылмайтын жалпы қолданыстағы грамматикалық және қызметтік бірліктерді қатар қамтиды [12] .

1-кесте онтологиялық бағдарланған тезаурус элементтерін ұсыну форматын иллюстрациялайды: әрбір термин үшін оның бастапқы мәтіндік формасы, нормаланған жазылуы, корпустағы бағаланған жиілігі, дереккөзі (корпус немесе онтология) және бірлік типі (бірсөзді немесе көпсөзді термин) көрсетіледі.

Келтірілген фрагмент ресурстың хронологиялық кезеңдерді, антропонимдер мен топонимдерді, сондай-ақ абстрактілі тарихи ұғымдарды қамтитынын көрсетеді, бұл білім графын құру және сұрақ-жауап жүйелерін қолдау үшін маңызды.

«Қазақстан тарихы» пәндік облысы терминдерін типизациялау үшін эвристикалық ережелер мен терминдердің үлестірілген векторлық ұсынуларына негізделген қадағаланатын оқытуды біріктіретін әлсіз бақыланатын гибридті сұл-

ба қолданылды. Бұл тәсіл тек шаблондық ережелерге сүйенуден бас тартып, контексті ескеретін, жалпылай алатын және белгісіздік деңгейін басқаратын модель арқылы тезаурус терминдерінің негізгі бөлігін автоматты түрде типизациялауға мүмкіндік береді [13].

Эвристикалық seed-таңбалар. Бірінші қадамда әрбір нормаланған термин $\tilde{t}_i$үшін ережеге негізделген функция қолданылды:

$$s(\tilde{t}_i) \in \{\text{”event”}, \text{”place”}, \text{”period”}, \text{”source”}, \text{”person”}, \text{”unknown”}\},$$

1-кесте. Қазақстан тарихы мектеп курсының тезаурусынан фрагмент (терминдер және олардың нормаланған формаларының үлгілері)

№	**Term**	**term_norm**	**freq_est**	**source**	**term_type**
1	1953–1964 жылдар	1953-1964 жылдар	81	ontology	multiword
2	Әйелдер	Әйелдер	19	ontology	single
3	б.з.д. 3 мыңжылдық	б з д 3 мыңжылдық	11	ontology	multiword
4	арал теңізі жанынан	арал теңізі жанынан	58	ontology	multiword
5	мәдениеттің өркендеу кезеңі	мәдениеттің өркендеу кезеңі	32	ontology	multiword
6	сырдария бойы қалалары	сырдария бойы қалалары	267	ontology	multiword
7	жеке тұлғаларға	жеке тұлғаларға	78	ontology	multiword
8	Найза	Найза	16	ontology	single
9	зияд ибн салих	зияд ибн салих	27	ontology	multiword
10	Уақыт	Уақыт	78	ontology	single
11	кіші обалардағы мәйіттер	кіші обалардағы мәйіттер	45	ontology	multiword
12	сопақша ойықтар	сопақша ойықтар	0	ontology	multiword
13	еуропалықтардың қыпшақ тілі үйренуі	еуропалықтардың қыпшақ тілі үйренуі	194	ontology	multiword
14	даяң хан	даяң хан	514	ontology	multiword
15	қаратау жотасы	қаратау жотасы	13	ontology	multiword
16	қимақ тайпасының басшысы	қимақ тайпасының басшысы	122	ontology	multiword
17	қалалық мәдениеттің қалыптасуы	қалалық мәдениеттің қалыптасуы	113	ontology	multiword

яғни терминге «оқиға» (event), «орын» (place), «кезең» (period), «дереккөз» (source), «тұлға» (person) немесе «белгісіз» (unknown) сыныптық таңбаларының бірі беріледі.

Бұл функция қазақ және орыс тілдеріндегі суффикстер мен негізгі сөздерге (хан, би, сұлтан, қаласы, ғасыр, жылдары және т.б.) негізделген эвристикалық «көрсеткіштерді», сондай-ақ жылдық интервалдарды анықтайтын детекторды пайдаланады [14].

Жалпы тезауруста 8345 бірегей нормаланған термин бар. Seed-таңбалардың терминдер бойынша таралуы 2-кестеде келтірілген.

2-кесте. Тезаурустағы эвристикалық seed-таңбалардың таралуы

Класс	Терминдер саны	8345-тен үлесі, %
Unknown	7436	89,1
Person	411	4,9
Period	363	4,3
Place	86	1,0
Source	35	0,4
Event	14	0,2

3-кесте. Сыныптар бойынша оқыту жиынтығының көлемі (seed-таңбалар бойынша белгілеу)

Класс	Seed-үлгілер саны	860 ішіндегі үлесі, %
Person	411	47,8
Period	363	42,2
Place	86	10,0

Кесте деректері қарапайым ережелердің тұлғалар, тарихи кезеңдер және географиялық объектілердің шектеулі бөлігін ғана жоғары дәлдікпен айқындайтынын, ал терминдердің басым бөлігі «unknown» санатында қалатынын көрсетеді. Бұл машиналық оқыту модельдерін қолдану қажеттілігін айқын дәлелдейді.

Векторлық термин ұсынулары. Екінші қадамда әрбір термин t_iкөптілді XLM-RoBERTa трансформерлік моделі көмегімен бекітілген өлшемді векторға кодталды:

$$h_i = f_\theta(t_i) \in \mathbb{R}^{768},$$

мұнда $f_\theta(\cdot)$- параметрлері θболатын XLM-R моделі, ал h_iвекторы модельдің соңғы жасырын қабатындағы [CLS] токенінің ұсынуынан алынады. Осылайша 8345 терминнің барлығы үшін эмбеддингтер матрицасы құрылды:

$$H \in \mathbb{R}^{8345\times768}.$$

Классификаторды seed-деректерде оқыту. Қадағаланатын оқыту (обучение с учителем) үшін тек жақсы ұсынылған (жеткілікті көлемдегі) seed-таңбалар пайдаланылды. Бұл мақсатта машиналық оқыту классификаторы келесі сыныптар жиыны үшін оқытылды:

$$C_{\text{ML}} = \{\text{”person”}, \text{”period”}, \text{”place”}\}.$$

event және source сыныптарына жатқызылған терминдер санының аздығына (тиісінше 14 және 35 бірлік) байланысты олар оқу жиынтығына енгізілмеді және кейінгі кезеңдерде тек ереже-негізді (rule-based) тәсілдер арқылы өңделді [15].

Нәтижесінде қадағаланатын оқыту үшін пайдаланылған терминдер саны 860 бірлікті құрады (3-кесте).

Осы деректер негізінде көпсыныпты логистикалық классификатор оқытылды:

$$p(c \mid h_i) = \text{softmax}(W h_i + b), c \in C_{\text{ML}},$$

мұнда $W \in \mathbb{R}^{|\mathscr{C}_{\text{ML}}|\times768}$, $b \in \mathbb{R}^{|\mathscr{C}_{\text{ML}}|}$.

Параметрлер (W, b) стандартты кросс-энтропиялық шығын функциясын минимизациялау арқылы бағаланды:

$$L = -\sum_{i \in S} \log p(s(\tilde{t}_i) \mid h_i),$$

мұнда S- C_{ML}сыныптарына жататын seed-таңбамен белгіленген терминдердің индекстер жиыны.

Сыныптар арасындағы дисбаланста өтеу үшін логистикалық регрессияда сыныптар бойынша салмақталған тәсіл *(class_weight = ”balanced”)* қолданылды, ал модель сапасы стратификацияланған бөлу арқылы алынған валидациялық ішкі жиынтықта (seed-деректердің 20 %-ы) бағаланды.

4-кесте. ML-классификатордың seed-валидациядағы сапасы

Класс	Precision	Recall	F1-score	Support
Period	0,984	0,863	0,920	73
Person	0,915	0,915	0,915	82
Place	0,577	0,882	0,698	17

Валидация жиынтығында жалпы дәлдік 0,890 (172 мысал), макроорташа $F_1 = 0,844$, ал салмақталған $F_1 = 0,895$болды. Бұл, салыстырмалы түрде seed-үлгілер саны аз болғанына қарамастан, модельдің тұлғалар (person) мен тарихи кезеңдерді (period) жақсы ажырата алатынын көрсетеді. place сыныбы бойынша дәлдік етілгендей төмендеу, себебі мысалдар ең аз дәл осы класта шоғырланған, дегенмен толықтық (recall) мәні жоғары деңгейде сақталады.

Гибридті түрде бүкіл тезаурусқа сыныптық таңба беру. Келесі қадамда оқытылған классификатор бастапқыда seed-таңбасы unknown болған терминдерді қоса алғанда, тезаурустағы барлық терминдерге қолданылды. Әрбір термин үшін ықтималдықтардың үлестірімі $p(c \mid h_i)$есептеліп, одан кейін сенім шегі $\tau = 0,6$қолданылды. Соған сәйкес аралық сыныптық болжам $\hat{c}_i$мына ережемен анықталды:

$$\hat{c}_i = \begin{cases} \text{argmax}_{c \in C_{ML}} p(c \mid h_i), & \text{егер } \max_c p(c \mid h_i) \geq \tau, \\ \text{”unknown”}, & \text{әйтпесе.} \end{cases}$$

Осыдан кейін термин үшін қорытынды сыныптық таңба y_iкелесі гибридті қағида бойынша тағайындалды:

$$y_i = \begin{cases} s(\tilde{t}_i), & \text{егер } s(\tilde{t}_i) \in \{event, source\}, \\ s(\tilde{t}_i), & \text{егер } s(\tilde{t}_i) \neq unknown, \\ \hat{c}_i, & \text{егер } s(\tilde{t}_i) = unknown. \end{cases}$$

Яғни, сирек, бірақ маңыздылығы жоғары event және source сыныптары үшін эвристикалық таңбалар әрқашан сақталады; ал person, period және place сияқты массалық сыныптарда ML-модель бастапқыда белгісіз (unknown) күйінде қалған терминдердің едәуір бөлігін қосымша типизациялайды [16].

Тезаурус бойынша сыныптардың гибридті өңдеуден кейінгі қорытынды үлестірімі 5-кестеде келтірілген.

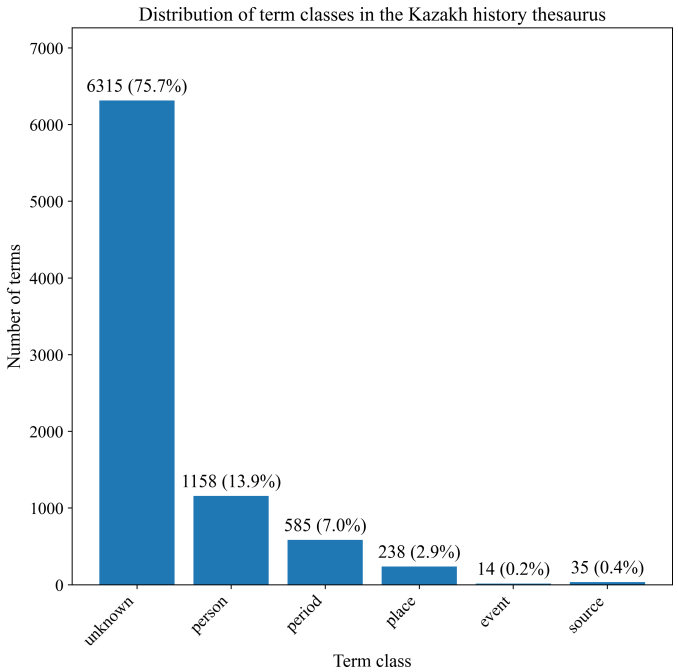
Бастапқы эвристикалық белгілеумен салыстырғанда, типі анықталмаған терминдердің үлесі 89,1 %-дан 75,7 %-ға дейін төмендеп, тұлғалар, тарихи кезеңдер және географиялық объектілер саны екі еседен артық артты. Сирек кездесетін «оқиға» және «дереккөз» сыныптары үшін seed-таңбалар сақталып, онтологиялық графты құру кезеңіне толық енгізілді.

5-кесте. Терминдер сыныптарының қорытынды үлестірімі (гибридті схема)

Класс	Терминдер саны	8345-тен үлесі, %
Unknown	6315	75,7
Person	1158	13,9
Period	585	7,0
Place	238	2,9
Event	14	0,2
Source	35	0,4

Нәтижесінде 8 345 терминнен тұратын типтелген тезаурус қалыптастырылды, оның құрамына 1 158 тарихи тұлға, 585 кезең, 238 топоним, сондай-ақ жекелеген оқиғалар мен дереккөздер кіреді. Бұл типизация онтологиялық триплеттерді генерациялау, білім графын құру және сұрақ-жауап жүйелерінде контекст іздеу сапасын арттыру үшін негізгі тірек компонент болып табылады.

2-суретте терминдердің алты сынып бойынша үлестірімі көрсетілген: person (13,9 %), period (7,0 %), place (2,9 %), event (0,2 %), source (0,4 %) және басым үлеске ие unknown (75,7 %). unknown сыныбының жоғары үлесі қателіктен емес, әлсіз бақыланатын белгілеу кезінде қабылданған әдістемелік жорамалдардың салдары болып табылады.



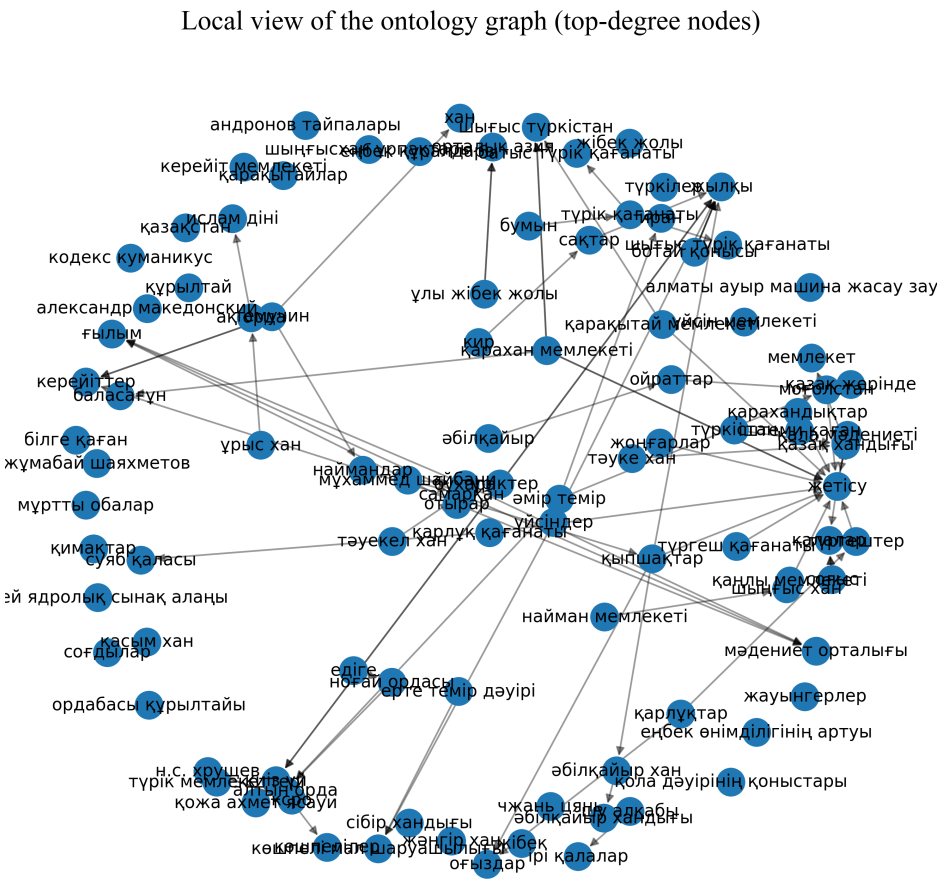
2-сурет. Қазақстан тарихы (5–10 сыныптар) бойынша онтологиялық бағдарланған тезаурустағы термин сыныптарының үлестірімі

Біріншіден, эвристикалық ережелер тек жоғары дәлдіктегі бастапқы «seed» таңбаларды қалыптастыру үшін қолданылып, морфологиялық және лексикалық үлгілердің шектеулі жиынын қамтыды. Екіншіден, логистикалық регрессия тек person, period және place сыныптары бойын-

ша оқытылып, модель нәтижелері ықтималдығы p ≥ 0,6 болған жағдайда ғана қабылданды, ал сенімділігі төмен жағдайлар әдейі unknown санатында қалдырылды. Үшіншіден, семантикалық тұрғыдан маңызды, бірақ сирек кездесетін event және source сыныптары тек қатаң эвристикалық ережелер арқылы белгіленіп, модель арқылы жасанды толықтыру жүргізілген жоқ.

Нәтижесінде unknown санатына жалпы қолданыстағы лексемалармен қатар, ағымдағы кезеңде сенімді типизациялау белгілері жеткіліксіз терминдер де жинақталды. Бұл айқын белгіленген сыныптар үшін жоғары дәлдікті қамтамасыз етіп, онтологиялық графты құру барысында қателік тәуекелін төмендетеді және кейінгі қосымша белгілеу үшін әдістемелік тұрғыдан үйлесімді кеңістік қалдырады.

3-суретте 2 973 түйін мен 2 318 бағытталған қырдан тұратын онтологиялық графтың локалды проекциясы көрсетілген. Визуализацияда дәрежесі ең жоғары 100 түйін таңдалып, Қазақстан тарихының мектеп курсына арналған пәндік графтың құрылымдық «өзегін» айқындауға мүмкіндік береді. Түйіндер тезаурус терминдеріне, ал қырлар оқулық мәтіндерінен алынған типтелген семантикалық қатынастарға сәйкес келеді [17].



3-сурет. Қазақстан тарихы бойынша онтологиялық графтың локалды фрагменті (байланыстылығы ең жоғары түйіндер)

Жүйеде unknown сыныбына жататын терминдердің салыстырмалы түрде көп болуы қате немесе әдістемелік қайшылық болып саналмайды. Бұл категорияға саналы түрде екі топ енгізілді: (1) person, period, place, event және source сияқты нысаналы онтологиялық рөлдердің ешқайсысына жатпайтын, мазмұны жағынан жалпы немесе абстрактілі ұғымдарды білдіретін тер-

миндер; (2) трансформерлік модель белгілі бір сыныпқа жатқызу ықтималдығын сенім шегінен төмен (p <0,6) деп бағалаған терминдер. Осылайша, unknown техникалық класс ретінде «өзге домендік концепттер» жиынын білдіреді және мәнді сыныптар үшін типизацияның жоғары дәлдігін сақтауға, күмәнді таңбаларды жасанды түрде тағайындаудан бас тартуға, сондай-ақ онтологиялық графты құрудың әдістемелік корректілігін қамтамасыз етуге мүмкіндік береді.

Бағалау әдістемесі. Екі құжат коллекциясы үшін де (plain-контекст және онтологиялық тұрғыдан байытылған контекст) классикалық BM25 ранжирлеушісі қолданылды. Бағалау 5-10 сыныптарға арналған Қазақстан тарихы оқулықтары негізінде қалыптастырылған 133 «сұрақ–контекст» жұбынан тұратын тест жиынтығында жүргізілді. Әрбір сұрақ үшін жалғыз релевантты құжат ретінде сол сұраққа бастапқыда аннотацияланған сәйкес оқулық фрагменті алынды. BM25 релеванттық функциясы стандартты түрде мына түрде берілді:

$$\mathrm{BM25}(q,d)=\sum_{t\in q}\mathrm{IDF}(t)\,\frac{f(t,d)\,(k_1+1)}{f(t,d)+k_1\left(1-b+b\frac{|d|}{|\bar{d}|}\right)},$$

мұнда $f(t,d)$- dқұжатындағы tтерминының жиілігі, $|\ d\ |$- құжаттың ұзындығы, $|\,\bar{d}\,|$- коллекциядағы құжаттардың орташа ұзындығы, ал гиперпараметрлер $k_1 = 1,5$және $b = 0,75$мәндерінде бекітілді.

Жүйенің сапасы Precision@k, Recall@k және nDCG@k метрикалары бойынша өлшенді. Әр сұрауға бір ғана релевантты құжат сәйкес келгендіктен, q_iсұрағы үшін метрикалар мына түрде есептелді:

$$P@k(q_i)=\frac{1[\mathrm{rank}_i\le k]}{k},\ R@k(q_i)=1[\mathrm{rank}_i\le k],\ nDCG@k(q_i)=\frac{1/\log_2(\mathrm{rank}_i+1)}{1},$$

мұнда rank_i- релевантты құжаттың ранжирленген тізімдегі позициясы, $1[\cdot]$- индикатор-функция (шарт орындалса 1, орындалмаса 0). Әр метриканың қорытынды мәні барлық сұрақтар бойынша арифметикалық орта ретінде есептелді.

Салыстыру үшін төрт конфигурация қарастырылды: Plain - онтологиялық префикссіз және сұрауларды кеңейтусіз тек бастапқы оқулық мәтіндеріне негізделген нұсқа; OntologyDocs - құжаттар онтологиялық анықтамалармен (onto-defs префиксі) толықтырылған конфигурация; Plain+Expansion - құжаттар өзгеріссіз қалып, сұраулар тезаурус синонимдері мен онтологиялық графтағы көрші терминдер есебінен семантикалық түрде кеңейтілген нұсқа; OntologyDocs+Expansion – онтологиялық тұрғыдан байытылған құжаттар мен сұрауларды кеңейту бір уақытта қолданылатын конфигурация. Жиынтық нәтижелер 6-кестеде келтірілген, онда барлық конфигурациялар үшін негізгі метрикалардың (P@1, P@5 және nDCG@10) орташа мәндері көрсетіліп, әрбір бағандағы ең жоғары көрсеткіштер ерекшеленіп белгіленген [18].

Талдау және интерпретация. Кесте 6 деректері онтологиялық тұрғыдан байытылған құжаттық ұсынылымның (*OntologyDocs*) негізгі метрикалардың басым бөлігі бойынша ең жоғары нәтижелер беретінін көрсетеді. Базалық *Plain* конфигурациясымен салыстырғанда, құжат контекстіне онтологиялық анықтамаларды қосу P@1 көрсеткішін 0,556-дан 0,857-ге дейін арттырып, nDCG@10 мәнін 0,774-тен 0,925-ке дейін жақсартты, бұл релевантты фрагменттердің ранжирлеу тізімінің жоғарғы бөлігінде неғұрлым дәл орналасқанын білдіреді.

OntologyDocs+Expansion конфигурациясы ранжирлеу сапасы бойынша екінші нәтиже көрсетіп (nDCG@10 = 0,873), R@10 метрикасы бойынша ең жоғары толықтыққа (0,992) жетеді, яғни релевантты контекст дерлік барлық жағдайда алғашқы он нәтиженің ішінде бола-

ды. Ал тек сұрауларды кеңейтуге негізделген *Plain+Expansion* нұсқасы базалық модельмен салыстырғанда сапаның төмендеуіне алып келеді, бұл онтологиялық құрылымсыз агрессивті кеңейтудің сұрау мәтініндегі шуды арттыруымен түсіндіріледі.

Жалпы алғанда, алынған нәтижелер зерттеудің негізгі гипотезасын растайды: Қазақстан тарихы бойынша онтологиялық ұйымдастырылған білім графын мәтіндік индекс құрамына енгізу қазақ тіліндегі сұрақ-жауап жүйелерінде релевантты контекстіні алу тиімділігін айтарлықтай арттырады. Сұрауларды кеңейту оң әсерді тек онтологиялық тұрғыдан байытылған құжаттармен бірге қолданылған жағдайда ғана береді [19].

6-кесте. Индекс пен сұраудың әртүрлі конфигурациялары үшін контексті алу сапасы

Конфигурация	P@1	P@5	nDCG@10
Plain	0.556	0.180	0.774
OntologyDocs	**0.857**	**0.195**	**0.925**
Plain+Expansion	0.444	0.176	0.705
OntologyDocs+Expansion	0.752	0.191	0.873

7-кесте. BM25-тің төрт конфигурациясы үшін контексті алу сапасын салыстыру (тест жиынтығы, n = 133)

Метрика	Plain	OntologyDocs	Plain+Expansion	OntologyDocs+Expansion
P@1	0.556	0.857	0.444	0.752
R@1	0.556	0.857	0.444	0.752
nDCG@1	0.556	0.857	0.444	0.752
P@3	0.278	0.318	0.258	0.291
R@3	0.835	0.955	0.774	0.872

7-кестеде Қазақстан тарихы бойынша QA-жиынтығы үшін контексті алу сапасын бағалау нәтижелері келтірілген. Тек оқулық мәтіндеріне негізделген Plain конфигурациясы орташа көрсеткіштер береді (P@1 = 0,556; nDCG@10 = 0,774). Онтологиялық тұрғыдан байытылған құжаттарды қосу (OntologyDocs) барлық метрикалардың жүйелі түрде өсуіне алып келеді: бірінші позициядағы дәлдік 0,857-ге дейін, ал nDCG@10 мәні 0,925-ке дейін артады. Бұл онтологиялық анықтамаларды контекстке енгізу релевантты фрагменттің ранжирлеу тізімінің жоғарғы бөлігінде орналасу ықтималдығын едәуір арттыратынын көрсетеді.

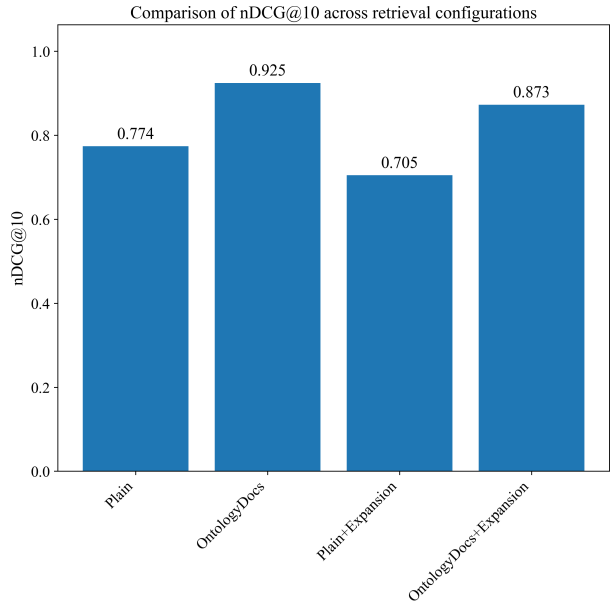
Керісінше, құрылымданған онтологиялық префикссіз тек сұрауларды кеңейтуге негізделген Plain + Expansion конфигурациясы Plain нұсқасымен салыстырғанда сапаның төмендеуіне әкеледі. Бұл эффект сұраудың шамадан тыс кеңейтілуімен және семантикалық құрылыммен қолдау таппаған лексикалық шудың артуымен түсіндіріледі.

Комбинацияланған OntologyDocs+Expansion тәсілі компромисс нәтижені көрсетеді: ол көптеген метрикалар бойынша OntologyDocs конфигурациясынан төмен болғанымен, R@10 = 0,992 және P@10 = 0,099 мәндері бойынша ең жоғары толықтықты қамтамасыз етеді. Бұл релевантты құжаттың дерлік барлық жағдайда топ-10 нәтижелер ішінде табылатынын және сұра-

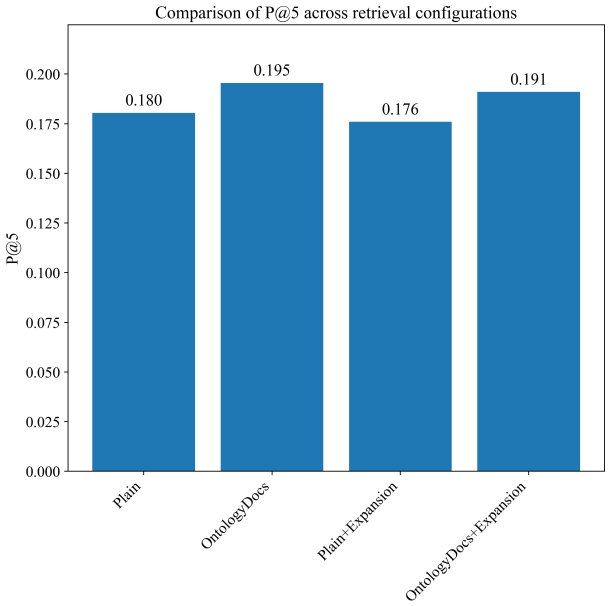
уды кеңейту механизмдерінің ранжирлеудің үлкен тереңдіктерінде қосымша пайда беретінін білдіреді.

Жалпы алғанда, алынған нәтижелер сапаны арттырудың негізгі факторы құжаттарды онтологиялық тұрғыдан байыту екенін, ал сұрауларды кеңейтуді онтологиялық құрылымданған индекспен бірге қолдану орынды екенін растайды [20].

4-суретте іздеу модулінің төрт конфигурациясы үшін nDCG@10 мәндерінің салыстырмалы диаграммасы көрсетілген.



4-сурет - Контексті алу конфигурациялары үшін ранжирлеу сапасын (nDCG@10) салыстыру



5-сурет - Қазақстан тарихы мектеп корпусы бойынша әртүрлі іздеу конфигурациялары үшін P@5 дәлдігін салыстыру

Онтологиялық құжаттарды пайдаланатын OntologyDocs конфигурациясы ранжирлеудің ең жоғары сапасын көрсетеді (nDCG@10 = 0,925), бұл оқу материалын құрылымдауда айқын ұйымдастырылған білім графының тиімділігін дәлелдейді. Базалық Plain нұсқасының көрсеткіші төменірек (nDCG@10 = 0,774), ал онтологиялық корпуссыз тек сұрауды кеңейтуге негізделген Plain+Expansion конфигурациясы лексикалық шудың артуына байланысты сапаның одан әрі төмендеуіне әкеледі (nDCG@10 = 0,705).

Онтологиялық құжаттарға сұрауды кеңейтуді қосатын OntologyDocs+Expansion нұсқасы базалық модельмен салыстырғанда нәтижелерді жақсартқанымен (nDCG@10 = 0,873), тек онтологиялық құрылымданған коллекцияны қолданудан сәл төмен қалады. Бұл семантикалық іздеу міндеттерінде, әсіресе қазақ тіліндегі оқу контенті жағдайында, сапасы мұқият тексерілген онтологияның шешуші рөл атқаратынын көрсетеді.

5-суретте төрт іздеу конфигурациясы үшін P@5 метрикасының салыстырмалы өзгерісі келтірілген: Plain, OntologyDocs, Plain+Expansion және OntologyDocs+Expansion.

P@5 метрикасы бойынша ең жоғары мәнді OntologyDocs конфигурациясы көрсетеді (≈0,195), бұл онтологиялық тұжырымдарды индекске айқын енгізудің пайдалы екенін тағы да дәлелдейді. «Жазық» (құрылымдалмаған) құжаттар үстінен жүргізілетін таза сұрау кеңейтуі дәлдіктің сәл төмендеуіне әкеледі (≈0,176-ға дейін), ал онтологиялық құжаттар мен кеңейтілген сұрауды біріктіретін схема (OntologyDocs+Expansion) дәлдікті ең жақсы мәнге жуық деңгейде ұстап тұрады (≈0,191), топ-5 ішінде неғұрлым мазмұнды және ақпараттық

релевантты фрагменттер жиынтығын қамтамасыз етеді.

Алынған нәтижелер Қазақстан тарихы бойынша оқу материалын онтологиялық бағдарланған түрде құрылымдау сұрақ-жауап жүйелеріндегі контексті іздеу сапасын едәуір арттыратынын көрсетеді. Таза мәтіндік индекстен онтологиялық анықтамалармен байытылған индексқа көшу ранжирлеу метрикаларының тұрақты өсіміне алып келеді. «Субъект–қатынас–объект» триплеттері негізінде қалыптастырылған қысқа анықтамаларды индексе енгізу сұрақ формулировкаларын тиісті оқулық фрагменттерімен дәлірек сәйкестендіреді.

XLM-R эмбеддингтері мен логистикалық регрессияға негізделген әлсіз бақыланатын типизация person, period және place сыныптарын сенімді ажыратуға мүмкіндік береді. unknown сыныбының кең үлесін сақтау әдістемелік тұрғыдан негізделген шешім болып табылады, себебі ол күмәнді терминдерді күшпен типизациялаудан сақтап, білім графындағы қателердің тәуекелін төмендетеді. Бұл тәсіл ресурсты кейінгі кезеңдерде эксперттік белгілеу немесе белсенді оқыту арқылы кеңейтуге қолайлы жағдай жасайды.

Тарихи оқу дискурсында сипаттамалық және жалпы ұғымдардың басым болуы unknown сыныбының жоғары үлесін табиғи етеді. Мұндай бірліктерді қатаң онтологиялық категорияларға мәжбүрлеп жатқызу онтологияның интерпретациялануын төмендетер еді, сондықтан қолданыстағы граф нақты және тексерілетін сущностік объектілерге шоғырланады, бұл мектеп курсының білім беру мақсаттарымен үйлеседі.

Эксперименттер онтология мен тезаурустың пайдасы құжат деңгейінде айқын көрінетінін, ал құрылымданған индекспен қолдаусыз сұрауларды кеңейту нәтижелерді нашарлататынын көрсетті. Семантикалық кеңейту тек онтологиялық құрылыммен тығыз байланысқан жағдайда ғана тиімді.

Комбинацияланған OntologyDocs+Expansion конфигурациясы интегралдық метрикалар бойынша сәл төмен болғанымен, жоғары тереңдікте дерлік толық толықтықты қамтамасыз етеді, бұл оны күрделі тапсырмалар мен тақырыптық жинақтар құру сценарийлері үшін перспективалы етеді.

Жалпы алғанда, құрылған білім графы мен типтелген тезаурус қазақ тіліндегі пәндік language model-дерді оқыту, тест тапсырмаларын автоматты генерациялау және тарихи контентті цифрлық білім беру экожүйелерінде қайта пайдалану үшін әмбебап инфрақұрылымдық негіз бола алады.

Қорытынды. Бұл жұмыста қазақ тіліндегі Қазақстан тарихы мектеп курсын формализациялауға арналған кешенді онтологиялық тәсіл ұсынылып, іске асырылды. Тәсіл нормаланған оқулықтар корпусын, онтологиялық бағдарланған тезаурусты және білім графын құруды, сондай-ақ оларды сұрақ-жауап жүйелерінде контекст іздеу үдерісіне интеграциялауды қамтиды. Корпус деректері мен бастапқы онтологияны біріктіру нәтижесінде 8 345 терминнен тұратын тезаурус қалыптастырылып, ол әлсіз бақыланатын оқыту арқылы типтелді, бұл тарихи тұлғаларды, кезеңдерді, топонимдерді, оқиғалар мен дереккөздерді сенімді түрде айқындауға мүмкіндік берді.

2 318 триплет негізінде құрылған білім графын іздеу инфрақұрылымына енгізу 661 сұрақтан тұратын QA-жиынтықта контексті алу сапасының едәуір артатынын көрсетті: бірінші позициядағы дәлдік шамамен бір жарым есе өсіп, nDCG@10 мәні 0,92 деңгейіне жетті. Бұл төмен ресурсты қазақ тілі үшін сұрақ-жауап жүйелерінің тиімділігін арттыруда пәндік білімді айқын онтологиялық құрылымдау шешуші фактор екенін растайды.

Алынған нәтижелер онтологияны кеңейту, оны генеративті модельдермен және қайта ранжирлеу тетіктерімен біріктіретін гибридті QA-архитектуралар құру, сондай-ақ ұсынылған әдістемені басқа пәндер мен түркі тілдеріне тарату сияқты келешек зерттеулерге негіз қалайды. Жалпы алғанда, жасалған онтологиялық тезаурус пен білім графы қазақ тіліндегі цифрлық білім беру жүйелерінің сапасын арттыруға бағытталған практикалық маңызы бар инфрақұрылымдық ресурс болып табылады.

Әдебиеттер

1. Chanthati S. R. A centralized approach to reducing burnouts in the IT industry using work pattern monitoring using artificial intelligence (Vector Search, Semantic Search, Python, Large Language Models, MongoDB Atlas)//ESS Open Archive eprints. -2024: 322, 32212020. DOI 0000-0001-5778-8140.

2. Stankevičius L., & Lukoševičius M. Extracting sentence embeddings from pretrained transformer models//Applied Sciences. -2024.-Vol.14(19): 8887. DOI 10.3390/app14198887.

3. Akhter E.AI in the classroom: Evaluating the effectiveness of intelligent tutoring systems for multilingual learners in secondary education//ASRC Procedia: Global Perspectives in Science and Scholarship. -2025.-Vol 1(1). -P.532-563. DOI 10.63125/gcq1qr39

4. Madatov K. A., Sattarova S. Creation of a corpus for determining the intellectual potential of primary school students// In 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM). -2024.-P.2420–2423. DOI 10.1109/EDM61683.2024.10615103.

5. Heilala V., Araya R., Hämäläinen, R. Beyond text-to-text: An overview of multimodal and generative artificial intelligence for education using topic modeling//In Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing. -2025.-P.54-63. DOI 10.1145/3672608.3707764.

6. Götz S., Granger S. Learner corpus research for pedagogical purposes: An overview and research perspectives. //International Journal of Learner Corpus Research. -2024.-Vol.10(1). -P.1-38. DOI 10.1075/ijlcr.00039.got.

7. Fante, C., Ravicchio, F., & Manganello, F. Navigating the evolution of game-based educational approaches in secondary STEM education: A decade of innovations and challenges// Education Sciences. -2024.-Vol. 14(6):662. DOI 10.3390/educsci14060662.

8. Lee J., et al. The life cycle of large language models in education: A framework for understanding sources of bias//British Journal of Educational Technology. -2024.- Vol.55(5). -P.

1982-2002. DOI 10.1111/bjet.13505.

9. Bao K. Comparative analysis of representations of feminism across Chinese social media: A corpus-based study of Weibo and Zhihu//Social Media + Society. -2024.-Vol.10(3). DOI 10.1177/20563051241274688.

10. Chen L. C. An extended TF-IDF method for improving keyword extraction in traditional corpus-based research: A climate change corpus case study//Data & Knowledge Engineering. -2024.-Vol.153:102322. DOI 10.1016/j.datak.2024.102322.

11. San N., et al. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens//arXiv preprint arXiv:2402.02302. DOI 10.48550/arXiv.2402.02302

12. Chataut S., et al. Comparative study of domain-driven term extraction using large language models// arXiv preprint arXiv:2404.02330.-2024. DOI 10.48550/arXiv.2404.02330

13. Kapyshev G., Nurtas M., Altaibek A. Speech recognition for the Kazakh language//Procedia Computer Science.-2024.-Vol.231.-P.369-372.DOI 10.1016/j.procs.2023.12.219.

14. Hsu E., Roberts K. Leveraging large language models for knowledge-free weak supervision in clinical natural language processing//Scientific Reports. -2025.-Vol.15:8241. DOI 10.1038/s41598-024-68168-2.

15. Ghaffari A. Precision seed certification through machine learning//Technology in Agronomy. -2024.-Vol.4: e019. DOI 10.48130/tia-0024-0013.

16. Mosquera C., Ferrer L., Milone D. H., et al. Class imbalance on medical image classification: Towards better evaluation practices for discrimination and calibration performance// European Radiology. -2024.-

Vol.34.-P.7895-7903. DOI 10.1007/s00330-024-10834-0.

17. Cortes C., DeSalvo G., Mohri M. Theory and algorithms for learning with rejection in binary classification//Annals of Mathematics and Artificial Intelligence. -2024. Vol.92.-P.277-315. DOI 10.1007/s10472-023-09899-2.

18. Alaofi M., Arabzadeh N., Clarke C. L. A., Sanderson, M. Generative information retrieval evaluation// Information Access in the Era of Generative AI. -2024.-Vol.51. DOI 10.1007/978-3-031-73147-1_6.

19. Luo Z., Wang Y., Ke W., Qi R., Guo Y., Wang P. Boosting LLMs with ontology-aware prompt for NER data augmentation//IEEE International Conference on Acoustics, Speech and Signal Processing. -P.12361–12365. DOI 10.1109/ICASSP48485.2024.10446860.

20. Xia Y., et al. Knowledge-aware query expansion with large language models for textual and relational retrieval//In Proceedings of NAACL-HLT. -2025.-Vol1.-P.4275-4286. DOI 10.18653/v1/2025.naacl-long.216.

Авторлар туралы мәлімет

Ергеш М. Ж. - PhD докторанты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан, e-mail:Shymkent90@mail.ru;

Ергеш Б. Ж. - PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан, e-mail:b.yergesh@gmail.com;

Максутова К.М. - PhD докторанты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан, e-mail: qunkabai@gmail.com;

Тулегулов А. Д. - физика-математика ғылымдарының кандидаты, қауымдастырылған профессор, Қ. Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан, e-mail: tad62@ya.ru.

Information about the authors

Yergesh M. Zh. - PhD doctoral student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: Shymkent90@mail.ru;

Yergesh B. Zh. - PhD, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: b.yergesh@gmail.com;

Maksutova K. M. - PhD doctoral student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: qunkabai@gmail.com;

Tulegulov A. D. - Candidate of Physical and Mathematical Sciences, Associate Professor, K. Kulazhanov Kazakh University of Technology and Business. e-mail: tad62@ya.ru.