УДК: 004.89:004.912:81'33:34 DOI: 10.65599/III7538

МЕТОДЫ И МОДЕЛИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ЮРИДИЧЕСКИХ ТЕКСТОВ НА ТАДЖИКСКОМ ЯЗЫКЕ

Х.А. Худойбердиев, Г.Х. Сафаров

Политехнический институт Таджикского технического университета имени академика М.С. Осими в городе Худжанд

В статье проведён систематический обзор и классификация методов и моделей интеллектуального анализа юридических текстов. Особое внимание уделено перспективам применения рассмотренных методов к таджикскому языку, для которого характерны отсутствие крупных размеченных юридических корпусов и специализированных языковых моделей. Показано, что наиболее обоснованным для построения интеллектуальной правовой системы на таджикском языке является комбинированный подход, сочетающий построение структурированного нормативно-правового корпуса, формирование семантического поиска на базе многоязычных векторных представлений и ограниченную генерацию ответов с обязательной привязкой к нормативным источникам.

Ключевые слова: *интеллектуальный анализ текста; обработка естественного языка; большие языковые модели; дополненная поиском генерация; юридические тексты; низкоресурсные языки; таджикский язык; семантический поиск; корпусная лингвистика права.*

УСУЛҲО ВА АМСИЛАҲОИ ТАҲЛИЛИ ЗЕҲНИИ МАТНҲОИ ҲУҚУҚӢ БА ЗАБОНИ ТОҶИКӢ

Х.А. Худойбердиев, Ғ.Ҳ. Сафаров

Дар мақола баррасии низомнок ва таснифоти усулҳову амсилаҳои таҳлили зеҳнии матнҳои ҳуқуқӣ гузаронида шудааст. Ба дурнамои истифодаи усулҳои баррасишуда барои забони тоҷикӣ, ки бо набудани корпусҳои бузурги ҳуқуқии ишораткардашуда ва моделҳои махсуси забонӣ тавсиф мешавад, диққати махсус дода шудааст. Нишон дода шудааст, ки барои сохтани низоми ҳуқуқии зеҳнӣ ба забони тоҷикӣ муносибати таркибӣ, ки сохтани корпуси сохтории меъёрии ҳуқуқӣ, ташаккули ҷустуҷӯи семантикиро дар асоси ифодаҳои бисёрзабонаи векторӣ ва тавлиди маҳдуди ҷавобҳоро бо такя ба манбаъҳои меъёрӣ муттаҳид мекунад, асоснок мебошад.

Калидвожаҳо: *таҳлили зеҳнии матн; коркарди забони табиӣ; моделҳои калони забонӣ; тавлиди бо ҷустуҷӯ пурраткардашуда; матнҳои ҳуқуқӣ; забонҳои камзахира; забони тоҷикӣ; ҷустуҷӯи семантикӣ; забоншиносии корпусии ҳуқуқ.*

METHODS AND MODELS OF INTELLIGENT ANALYSIS OF LEGAL TEXTS IN THE TAJIK LANGUAGE

Kh.A. Khudoyberdiev, G.H. Safarov

The article presents a systematic review and classification of methods and models for the intelligent analysis of legal texts. Particular attention is paid to the prospects of applying the considered methods to the Tajik language, which is characterized by the absence of large annotated legal corpora and specialized language models. It is shown that the most justified approach for building an intelligent legal system in the Tajik language is a combined one that integrates the construction of a structured normative legal corpus, semantic search based on multilingual vector representations, and constrained answer generation with mandatory grounding in normative sources.

Keywords: *text mining; natural language processing; large language models; retrieval-augmented generation; legal texts; low-resource languages; Tajik language; semantic search; legal corpus linguistics.*

Введение

Цифровизация правовой сферы и стремительный рост объёма нормативно-правовых актов обусловливают потребность в интеллектуальных методах обработки юридических текстов. Традиционные правовые информационные системы, ориентированные на хранение и поиск документов по формальным признакам, не обеспечивают семантического анализа содержания норм и не позволяют формировать ответы на запросы пользователей, выраженные на естественном языке. В этих условиях возрастает значение методов искусственного интеллекта и обработки естественного языка, способных учитывать смысловое содержание правовых текстов.

За последние десятилетия сформировалось значительное многообразие подходов к интеллектуальному анализу юридических текстов – от ранних формально-логических и основанных на прецедентах систем до современных больших языковых моделей. Это многообразие порождает потребность в систематизации и сравнении методов, а также в оценке их применимости к конкретным языковым и правовым условиям. Особую сложность представляет применение указанных методов к низкоресурсным языкам, для которых отсутствуют крупные размеченные корпуса юридических текстов и специализированные языковые модели. К числу таких языков относится таджикский язык, что определяет необходимость отдельного рассмотрения применимости существующих методов к национальной правовой системе Республики Таджикистан.

Целью настоящего исследования является классификация и сравнительный анализ методов и моделей интеллектуального анализа юридических текстов, а также оценка перспектив их применения для таджикского языка. Для достижения поставленной

цели решаются следующие задачи: выделение основных классов методов обработки правовых текстов; сравнение этих классов по критериям точности, проверяемости, масштабируемости и требований к данным; разработка графического и алгоритмического описания комбинированного метода и методики его экспериментальной оценки; обоснование выбора подхода, наиболее пригодного для условий низкоресурсного таджикского языка.

Научная новизна работы заключается в следующем. Во-первых, выполнена систематизация методов интеллектуального анализа юридических текстов под углом их применимости к низкоресурсному таджикскому языку, что ранее в литературе не рассматривалось. Во-вторых, предложен комбинированный метод, в котором структурная сегментация нормативных актов выполняется по иерархии единиц национального законодательства Республики Таджикистан (кодекс –«*боб*» –«*модда*» –«*қисм*» –«*банд*»), а генерация ответа жёстко привязывается к этим структурным единицам, что обеспечивает проверяемость результата в условиях отсутствия размеченных корпусов. В-третьих, для предложенного метода даны его формальная модель и набор метрик оценки, ориентированные на специфику правовых запросов на таджикском языке.

Материалы и методы

Материалом исследования послужили научные работы в области интеллектуального анализа правовых текстов, корпусной лингвистики права, обработки естественного языка и больших языковых моделей, опубликованные в период с 2005 по 2025 год. Методологической основой выступает сравнительный аналитический обзор: рассматриваемые методы группируются по классам и сопоставляются по ряду критериев – интерпретируемость и проверяемость результатов, требования к объёму и качеству обучающих данных, масштабируемость, а также применимость к низкоресурсным языкам. Подобный подход согласуется с принятой в современных обзорных исследованиях практикой систематизации методов по задачам, моделям и сопутствующим ограничениям [6, 7].

Проведённый анализ позволяет выделить четыре основных класса методов интеллектуального анализа юридических текстов, последовательно сменявших друг друга по мере развития искусственного интеллекта и обработки естественного языка. Рассмотрим каждый класс с точки зрения принципов работы, преимуществ, ограничений и применимости к таджикскому языку.

Формально-логические и основанные на прецедентах методы

Исторически первые подходы к автоматизации правовых рассуждений опирались на формальное представление знаний и логический вывод. К данному классу относятся экспертные системы, системы аргументации [1] и методы рассуждения на основе прецедентов (РОП). В основе РОП лежит моделирование аналогического рассуждения, которое юрист применяет, сопоставляя текущую ситуацию с ранее решёнными делами и выявляя сходство между ними. Теоретические основы формализации правовых рассуждений и когнитивного подхода к праву подробно изложены в работах по аргументации в искусственном интеллекте [1].

Сильной стороной данного класса методов является прозрачность и интерпретируемость: каждый вывод системы может быть прослежен по цепочке логических правил или явно указанных прецедентов. Однако этим методам присущи существенные ограничения – необходимость ручной формализации правовых норм и знаний экспертами, высокая трудоёмкость сопровождения базы правил, а также слабая устойчивость к естественно-языковой вариативности формулировок. Применительно к таджикскому языку эти методы не требуют больших размеченных корпусов, что является преимуществом в условиях дефицита данных, однако трудозатраты на ручную формализацию таджикского законодательства делают такой подход плохо масштабируемым.

Здесь необходимо пояснить, что понимается под «большим размеченным корпусом», поскольку именно требование к таким данным определяет применимость методов машинного обучения. Размеченный корпус – это не просто собрание текстов законов, а коллекция, в которой к текстам вручную добавлены метки: границы и тип структурных единиц: статья, часть, пункт, отрасль права, перекрёстные ссылки между нормами, а для задач вопросно-ответного взаимодействия – пары «вопрос – эталонный

ответ со ссылкой на норму». Именно на таких размеченных примерах обучаются и оцениваются модели. Для англоязычной юридической сферы созданы корпуса подобного масштаба – например, система, объединяющий порядка двухсот тысяч размеченных юридических текстов, и платформа, содержащий тысячи аннотаций, выполненных профессиональными юристами. Создание подобных ресурсов является дорогостоящим: разметку должны выполнять специалисты с юридической квалификацией, а трудоёмкость измеряется человеко-годами экспертной работы; стоимость формирования корпуса объёмом в десятки тысяч размеченных примеров оценивается в сотни тысяч долларов [9, 12].

Для таджикского языка размеченные юридические корпуса подобного масштаба отсутствуют, а их создание с нуля в обозримой перспективе требует целевого государственного финансирования. Данное обстоятельство имеет принципиальное значение для выбора подхода: методы, основанные на обучении модели на размеченных данных, оказываются практически нереализуемыми, тогда как дополненная поиском генерация работает на неразмеченном корпусе нормативно-правовых актов, для которого достаточно структурной разбивки текста без дорогостоящей ручной разметки эталонных ответов. Тем самым ограничение по данным превращается в аргумент в пользу комбинированного подхода, рассматриваемого ниже.

Корпусная лингвистика и интеллектуальный анализ текста

Второй класс методов основан на статистическом анализе больших коллекций правовых текстов. Развитие корпусной лингвистики права позволило перейти от интуитивного толкования к эмпирическому анализу языка закона на основе данных [10]. В рамках данного направления применяются методы интеллектуального анализа текста: тематическое моделирование, анализ сетей текстовых ссылок, частотный анализ и кластеризация. Показательным примером является компьютерный анализ корпуса исламского права, в котором применялись анализ повторного использования текста на основе аннотаций и HTML-разметки, а также тематическое моделирование для исследования эволюции правовых текстов на протяжении длительного исторического периода [10].

Ключевым преимуществом данного класса является способность выявлять скрытые закономерности и структуру в больших массивах правовых документов без ручного кодирования правил. Ограничения связаны с необходимостью тщательной предварительной обработки и разметки текстов, а также с зависимостью качества результатов от объёма и репрезентативности корпуса. Опыт применения корпусных методов к неевропейским и исторически удалённым правовым системам [9] представляет особый интерес для таджикского языка, поскольку демонстрирует принципиальную возможность построения структурированного юридического корпуса для языка с ограниченными ресурсами [10-12].

Статистические и нейросетевые методы представления текста

Третий класс связан с переходом от подсчёта частот к распределённым векторным представлениям текста, отражающим его смысловое содержание. Распределённые векторные представления слов и предложений позволили численно выражать семантическую близость языковых единиц. Принципиальным сдвигом стало появление архитектуры «трансформер», основанной на механизме внимания, на базе которой были разработаны предварительно обученные языковые модели, в частности BERT, продемонстрировавшие высокое качество в широком спектре задач понимания текста. Для информационного поиска по правовому корпусу применяются современные методы плотного извлечения фрагментов на основе векторных представлений [4].

Преимуществом нейросетевых представлений является способность улавливать семантику текста и обеспечивать поиск по смыслу, а не по точному совпадению слов. Применительно к низкоресурсным языкам решающее значение приобретают многоязычные и кросс-языковые модели представлений [3], позволяющие переносить семантические представления между языками. Исследования показывают, что для языков с малым объёмом данных качество представлений может быть повышено за счёт переноса из родственных языков, что особенно актуально для таджикского языка, имеющего близкое родство с персидским. Тем не менее качество представлений для

таджикского языка остаётся ограниченным вследствие недостаточного объёма обучающих данных.

Большие языковые модели и дополненная поиском генерация

Четвёртый, наиболее современный класс методов основан на больших языковых моделях (БЯМ), обученных на значительных объёмах текстовых данных и способных решать задачи в режиме обучения на малом числе примеров [2]. Современные обзоры подтверждают активное применение БЯМ в правовой сфере для задач классификации, суммаризации, извлечения информации и вопросно-ответного взаимодействия [6, 7]. Вместе с тем прямое применение БЯМ в праве сопряжено с принципиальным риском генерации правдоподобных, но недостоверных утверждений, не подтверждённых нормативными источниками.

Решением данной проблемы выступает дополненная поиском генерация (ДПГ), при которой релевантные фрагменты извлекаются из внешнего хранилища документов и используются в качестве контекста при формировании ответа [5]. Такой подход обеспечивает привязку ответов к проверяемым источникам и существенно снижает риск галлюцинаций. Применение ДПГ к юридическим задачам продемонстрировало возможность построения интерпретируемых систем вопросно-ответного взаимодействия со ссылками на правовые нормы [7]. Особый интерес представляет подход РОП-ДПГ, объединяющий рассуждение на основе прецедентов с дополненной поиском генерацией [8]; тем самым замыкается логическая связь между первым и четвёртым классами методов, и классическая идея аналогического рассуждения получает новое воплощение в архитектуре современных систем.

Преимуществами данного класса являются высокое качество генерации естественно-языковых ответов, масштабируемость, где обновление корпуса не требует переобучения модели и возможность привязки результатов к источникам. Основными ограничениями для таджикского языка остаются отсутствие специализированных языковых моделей и размеченных юридических данных, а также зависимость качества от точности семантического поиска, ограниченной языковыми ресурсами [10].

Сравнительный анализ методов

Обобщение рассмотренных классов методов по ключевым критериям представлено в таблице 1. В качестве критериев сравнения приняты: интерпретируемость результатов, требования к объёму обучающих данных, масштабируемость и применимость к таджикскому языку в его текущем состоянии.

Таблица 1- Сравнение классов методов интеллектуального анализа юридических текстов

Класс методов	Интерпретируемость	Требования к данным	Масштабируемость	Применимость к таджикскому языку
Формально-логические и РОП	Высокая	Низкие (ручная формализация)	Низкая	Ограниченная (трудоёмкость формализации)
Корпусные методы и text mining	Средняя	Средние	Средняя	Перспективная (при построении корпуса)
Нейросетевые представления и поиск	Низкая	Высокие	Высокая	Условная (через многоязычные модели)
БЯМ и дополненная поиском генерация	Средняя (при ДПГ)	Высокие	Высокая	Перспективная (ДПГ + многоязычные модели)

Примечание: Исследование авторов.

Как следует из таблицы 1, ни один из классов методов не является в чистом виде оптимальным для условий низкоресурсного таджикского языка. Формально-логические методы трудоёмки в сопровождении, чистые нейросетевые и генеративные подходы

предъявляют высокие требования к данным и подвержены риску недостоверной генерации. Наиболее обоснованным представляется комбинированный подход, сочетающий построение структурированного нормативно-правового корпуса, семантический поиск на основе многоязычных представлений с нейросетевыми методами и ограниченную генерацию ответов со ссылками на источники.

Сравнительный анализ генеративных моделей

Помимо сравнения классов методов, для практической реализации системы необходимо сопоставить конкретные генеративные большие языковые модели, пригодные для локального развёртывания. В качестве критериев сравнения приняты: поддержка таджикского языка и многоязычность, размер модели и требования к вычислительным ресурсам, возможность полностью локального использования и характер лицензии. Результаты сопоставления приведены в таблице 2. Следует отметить, что ни одна из рассмотренных моделей не заявляет официальной поддержки таджикского языка; однако модели Qwen 2.5 и Aya Expanse поддерживают персидский язык, что вследствие близкого языкового родства обеспечивает частичную применимость к таджикскому.

Таблица 2 - Сравнение генеративных больших языковых моделей для локального применения

Модель	Многоязычность / таджикский	Размеры (число параметров)	Локальное использование	Лицензия
Qwen 2.5	Более 29 языков; персидский есть, таджикского нет	0.5–72 млрд (в т.ч. 7, 14, 32)	Да (Ollama/LM Studio)	Apache 2.0 (кроме 3B и 72B)
Aya Expanse	23 языка; персидский есть, таджикского нет	8 и 32 млрд	Да (открытые веса)	CC-BY-NC 4.0 (некоммерческая)
Llama 3.1	8 языков; персидского и таджикского нет	8, 70, 405 млрд	Да (открытые веса)	Llama Community License
GPT-4o (OpenAI)	Многоязычная; таджикский ограниченно	Не раскрывается	Нет (только облако)	Проприетарная (API)

Примечание: Исследование авторов.

Из таблицы 2 следует, что проприетарные облачные модели (например, GPT-4o) неприменимы в силу требования полностью локальной и автономной работы системы. Модель Aya Expanse обладает сильной многоязычной подготовкой, однако её некоммерческая лицензия (CC-BY-NC 4.0) ограничивает дальнейшее практическое внедрение. Модель Llama 3.1 поддерживает ограниченный набор языков, не включающий персидско-таджикскую группу. Наиболее сбалансированным выбором является Qwen 2.5 в варианте 14 млрд параметров: он сочетает широкую многоязычную поддержку (включая персидский язык, родственный таджикскому), разумные требования к ресурсам, возможность полностью локального развёртывания и разрешительную лицензию Apache 2.0. При этом окончательный выбор между Qwen 2.5 и Aya Expanse должен определяться эмпирически – по результатам экспериментальной оценки качества ответов на таджикском языке.

Графическое и алгоритмическое описание комбинированного метода

Предлагаемый комбинированный метод объединяет сильные стороны трёх классов и состоит из двух этапов – подготовки корпуса, где однократно выполняется и обработки запроса, где выполняется при каждом обращении пользователя. Общая схема метода приведена на рисунке 1.

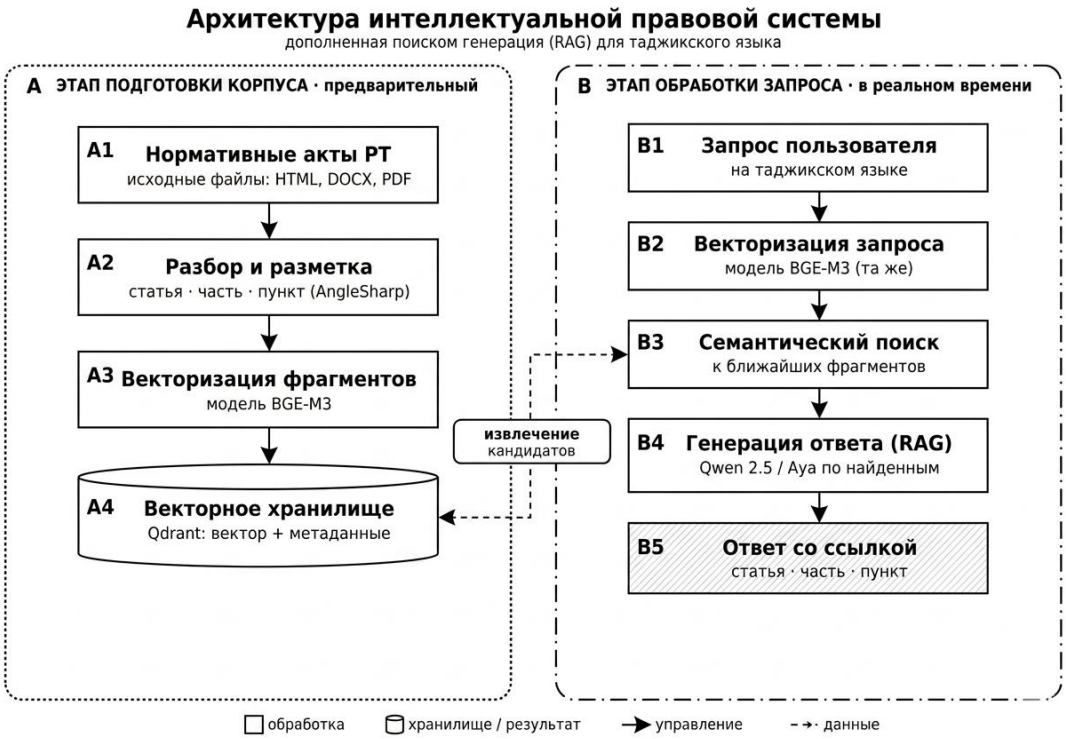

Рисунок 1- Схема комбинированного метода: этап подготовки корпуса и этап обработки запроса

Алгоритмически метод может быть представлен следующей последовательностью шагов.

Этап предварительной подготовки корпуса:

1) загрузка исходных нормативно-правовых актов в форматах файлов и HTML;

2) синтаксический разбор и структурная сегментация документов с выделением иерархии: кодекс → глава («*боб*») → статья («*модда*») → часть («*қисм*») → пункт («*банд*»);

3) формирование фрагментов, чанков на уровне статьи или части с сохранением структурных метаданных: название документа, редакция, номера главы, статьи, части, пункта;

4) векторизация каждого фрагмента с помощью многоязычной модели;

5) сохранение векторов и метаданных во векторном хранилище.

Этап обработки запроса в реальном времени:

6) приём запроса пользователя на таджикском языке;

7) векторизация запроса с использованием модели пункта 4;

8) семантический поиск k наиболее близких фрагментов в векторном хранилище;

9) формирование контекста из найденных фрагментов и передача его языковой модели вместе с запросом;

10) генерация ответа с обязательным указанием точной ссылки на норму («*модда*»/«*қисм*»/«*банд*») на основе метаданных извлечённых фрагментов.

Математическая модель решения проблемы

Формализуем ключевые этапы предложенного метода. Пусть правовая база данных (корпус нормативно-правовых актов) состоит из множества документов или их фрагментов, статей и пунктов:

$$D = \{d_1, d_2, \dots, d_N\}$$

При этом, входящий запрос пользователя на таджикском языке обозначается **q**. Для преодоления языкового барьера используется многоязычное семантическое пространство размерности **d**, отображение текста в вектор выполняется предварительно обученной многоязычной векторной моделью, на пример в данной исследовании использованы модели BGE-M3, mBERT, XLM-R, LaBSE:

$$f: T \to \mathbb{R}^d$$

где,
T – множество всех текстовых строк;
v = f(q) – запрос в виде векторной модели;
v_i = f(d_i) – векторы документов.

Модель семантического поиска и фильтрации релевантных источников основан на максимизации косинусного сходства между вектором запроса и векторами документов. На основе полученных исходных данных функция сходства получает следующий вид:

$$\mathrm{Sim}(q, d_i) = \frac{\mathbf{v}_q \cdot \mathbf{v}_{d_i}}{\| \mathbf{v}_q \| \| \mathbf{v}_{d_i} \|} = \frac{\sum_{j=1}^{d} v_{q,j}\ v_{d_i,j}}{\sqrt{\sum_{j=1}^{d} v_{q,j}^2} \cdot \sqrt{\sum_{j=1}^{d} v_{d_i,j}^2}}$$

В этом этапе формируется подмножество кандидатов **R ⊂ D**, который состоит из **k** документов с наивысшими оценками сходства:

$$R = \arg\max_{d_i \subset D}{}^{(k)} \mathrm{Sim}(q, d_i)$$

Для решения задачи *ограниченной генерации ответов* **A** на таджикском языке формулируется поиск последовательности токенов **A = (a_1, a_2, …, a^L)**, максимизирующей условную вероятность при заданном запросе **q** и извлечённом контексте **C**. Используя автоматическую регрессионную функцию генеративной большой языковой модели:

$$P(A \mid q, C) = \prod_{t=1}^{L} P\left(a_t \mid a_1, \ldots, a_{t-1}, q, C\right)$$

Для выполнения условия обязательной привязки к источникам вводится штрафная функция с ограничениями на генерацию. Оптимальный ответ **A*** определяется как:

$$A^* = \arg\max_{A} \log P(A \mid q, C) - \lambda \cdot \mathbb{I}(A \nrightarrow C)$$

где, $\mathbb{I}(A \nrightarrow C)$ – функция определения, получаем 1, если сгенерированный текст содержит утверждения, получаем 0, если текст полностью верифицируем;
λ – жёсткий коэффициент, который регулирует свободную генерацию без опоры на контекст.

Следует отметить, что в предложенной математической модели генерация ответов ограничиваются извлечёнными нормативно-правовыми документами, что обеспечивают достоверность результатов.

Архитектура программной реализации

Предложенный метод реализуется в виде полностью локальной программной системы, не зависящей при работе от внешних сетевых сервисов, что обусловлено требованиями к автономности и контролю над обработкой правовых данных. Архитектура построена на разделении ответственности между двумя языковыми платформами. Управляющий слой, реализованный на платформе .NET с использованием языка программирования C#, отвечает за синтаксический разбор и структурную сегментацию нормативных документов с использованием библиотеки AngleSharp для обработки документов в формате HTML. Разрабатываемые API-модули позволяют за верификацию этапов обработки запроса и за достоверное взаимодействие с векторным хранилищем. Модули машинного обучения, будут реализованы на языке программирования Python, которые изолированы и отвечают исключительно за формирования векторной модели входящих данных и генерации языковой модели в виде локальных HTTP-элементов. Векторное хранилище реализуется средствами СУБД Qdrant, модели векторизация BGE-M3, а модель генерации в среде Ollama и/или LM Studio. Использование современных технологий и средств искусственного интеллекта обеспечивают технологическую гибкость

между сгенерированной языковой модели и векторов без наглядной нагрузки на управляющую логику.

Характеристики корпуса и критерии отбора

- Источником данных служат нормативно-правовые акты Республики Таджикистан, доступные в форматах файлов и HTML. Формирование корпуса отвечает следующим критериям:
- достоверная редакция нормативного акта на момент включения в корпус;
- чёткая иерархической структуры кодекс→глава→статья→часть→пункт;
- однозначная структура классификации и сегментации документов;
- кластеризация по отраслям права на базе характерной устойчивой терминологии.

Единицей хранения корпуса выступает фрагмент уровня статьи или её части, снабжённый структурными метаданными. Количественные характеристики корпуса такие как, общий объём, число документов и распределение по отраслям права определяются на этапе полномасштабной апробации метода и составляют предмет отдельного исследования.

Методика экспериментальной оценки

Для проверки работоспособности предложенного метода разработана методика экспериментальной оценки на ограниченной выборке из 5-7 отрывков нормативно-правовых документов Республики Таджикистан. Методика предусматривает следующие действия: из одного из кодексов отбираются 5-7 статей («*модда*»), которые проходят полный цикл подготовки корпуса по шагам 1-5; для каждой статьи формулируется содержательный вопрос на таджикском языке, ответ на который однозначно определяется выбранной статьёй; запросы последовательно обрабатываются по шагам 6-10. Результат каждого запроса оценивается по трём показателям: релевантность извлечённого фрагмента, соответствует ли найденная статья ожидаемой, корректность ссылки на норму верно ли указаны номера («*модда*»/«*қисм*»/«*банд*») и содержательная полнота ответа. Указанная методика обеспечивает воспроизводимость оценки и закладывает основу для последующей апробации метода на полном корпусе нормативно-правовых актов, которая планируется в качестве самостоятельного этапа исследования.
Таким образом, комбинированный подход обеспечивает одновременно достоверность, проверяемость и масштабируемость, что отвечает специфике правовой сферы и языковым условиям Республики Таджикистан.

Заключение

В настоящей работе выполнен систематический обзор и классификация методов и моделей интеллектуального анализа юридических текстов, а также проведена оценка их применимости к таджикскому языку. Выделены четыре основных класса методов: формально-логические и основанные на прецедентах подходы; корпусная лингвистика и интеллектуальный анализ текста; статистические и нейросетевые методы представления текста; большие языковые модели с дополненной поиском генерацией. Рассмотрение этих классов в их историческом развитии показывает закономерный переход от ручной формализации правовых знаний к подходам, основанным на данных, и далее – к гибридным архитектурам, сочетающим информационный поиск с генерацией.

На основании проведённого анализа обосновано, что наиболее целесообразным для построения интеллектуальной правовой системы на таджикском языке является комбинированный подход, сочетающий построение структурированного нормативно-правового корпуса, семантический поиск на основе многоязычных векторных представлений и ограниченную генерацию ответов с обязательной привязкой к нормативным источникам. Такое сочетание обеспечивает достоверность, проверяемость и масштабируемость формируемых результатов, что соответствует повышенным требованиям правовой сферы и языковым условиям Республики Таджикистан.

Рецензент: Максудов Х.П. – к.ф.-м.н., доцент кафедры цифровой экономики ПИПТУ имени академика М.С. Осими в г. Худжанд.

Литература

1. Bench-Capon T., Dunne P.E. Argumentation in artificial intelligence // Artificial Intelligence. 2007. Vol. 171. N 10–15. P. 619–641. https://doi.org/10.1016/j.artint.2007.05.001.

2. Karpukhin V., Oguz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W. Dense passage retrieval for open-domain question answering // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 6769–6781.

3. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Kuttler H., Lewis M., Yih W., Rocktaschel T., Riedel S., Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474.

4. Lai J., Gan W., Wu J., Qi Z., Yu P.S. Large language models in law: A survey // AI Open. 2024. Vol. 5. P. 181–196. https://doi.org/10.1016/j.aiopen.2024.09.002.

5. Ariai F., Mackenzie J., Demartini G. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges // ACM Computing Surveys. 2025. Vol. 58. N 6. Article 163. https://doi.org/10.1145/3777009.

6. Louis A., van Dijck G., Spanakis G. Interpretable long-form legal question answering with retrieval-augmented large language models // Proceedings of the AAAI Conference on Artificial Intelligence. 2024. Vol. 38. N 20. P. 22266–22275. https://doi.org/10.1609/aaai.v38i20.30232.

7. Vogel F., Hamann H., Gauer I. Computer-assisted legal linguistics: Corpus analysis as a new tool for legal studies // Law and Social Inquiry. 2018. Vol. 43. N 4. P. 1340–1363. https://doi.org/10.1111/lsi.12305.

8. Lange C., Latif M.A., Celik Y., Lyklema A.M., van Kuppevelt D.E., van der Zwaan J. Text mining Islamic law // Islamic Law and Society. 2021. Vol. 28. N 3. P. 234–268. https://doi.org/10.1163/15685195-bja10009.

9. Krasadakis P., Sakkopoulos E., Verykios V.S. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages // Electronics. 2024. Vol. 13. N 3. Article 648. https://doi.org/10.3390/electronics13030648.

10. Худойбердиев, Х. А. Об алгоритме проверки орфографии на примере таджикского языка / Х. А. Худойбердиев // Политехнический вестник. Серия: Интеллект. Инновации. Инвестиции. – 2021. – № 3(55). – С. 58-63. – EDN XDGRTM.

11. Худойбердиев Х.А., Назаров А.А. Корпуси параллелии забони тоҷикӣ-русӣ: коркард ва тавсифи он // Вестник ПИТТУ имени академика М.С. Осими. 2019. N 1(10). С. 7–12.

12. Усманов З.Д., Худойбердиев Х.А. Алгоритм безударного озвучивания таджикского текста // Доклады Академии наук Республики Таджикистан. 2007. Т. 50. N 4. С. 316–319.

МАЪЛУМОТ ДАР БОРАИ МУАЛЛИФОН - СВЕДЕНИЯ ОБ АВТОРАХ - INFORMATION ABOUT AUTHORS

TJ	RU	EN
Худойбердиев Хуршед Атохонович	Худойбердиев Хуршед Атохонович	Khudoyberdiev Khurshed Atokhonovich
Доктори илмҳои техникӣ, дотсенти кафедраи «БваНИ»	Доктор технических наук, доцент кафедры «ПиИС»	Doctor of Technical Sciences, Associate Professor of the Department of "PandIS"
Донишкадаи политехникии ДТТ ба номи академик М.С. Осимӣ дар шаҳри Хуҷанд	Политехнический институт ТТУ имени академика М.С. Осими в городе Худжанде	Khujand Polytechnic Institute of TTU named after academician M.S. Osimi
E-mail: tajlingvo@gmail.com		
TJ	RU	EN
Сафаров Ғаниҷон Ҳабибуллоевич	Сафаров Ганиджон Хабибуллоевич	Safarov Ghanijon Habibulloevich
Унвончӯи дараҷаи доктори фалсафа (PhD), кафедраи «БваНИ»	Соискатель учёной степени доктора философии (PhD), кафедра «ПиИС»	PhD applicant, Department of "PandIS"
Донишкадаи политехникии ДТТ ба номи академик М.С. Осимӣ дар шаҳри Хуҷанд	Политехнический институт ТТУ имени академика М.С. Осими в городе Худжанде	Khujand Polytechnic Institute of TTU named after academician M.S. Osimi
E-mail: safarov.ghanijon@gmail.com		