

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ И АНАЛИЗА СТРУКТУР БАЗЫ ЗНАНИЙ ИЗ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Цель исследования – повышение эффективности управления корпоративными знаниями посредством разработки интеллектуальной системы автоматического построения баз знаний. Разработанная система объединяет нейросетевые методы обработки естественного языка и символьные методы формального представления знаний для построения верифицируемых структур из электронных документов.

Введение

Актуальность исследования обусловлена необходимостью перехода от простого хранения документов к построению формализованных структур знаний. Это важно для отраслей с высокой ценой ошибки, таких как промышленность, энергетика и медицина, где требуется верифицируемая логика выводов.

Существующие нейросетевые модели эффективно решают задачи извлечения сущностей, но действуют как «чёрные ящики». Они выявляют статистические корреляции, но не формируют формализованные, логически связанные структуры знаний и не предоставляют прозрачной цепочки рассуждений. Генеративные языковые модели также имеют недостаток: они склонны к генерации правдоподобных, но ложных утверждений («галлюцинаций»), что недопустимо для технических и нормативных баз знаний.

С другой стороны, детерминированные методы, основанные на лингвистических правилах и шаблонах, демонстрируют высокую точность при извлечении стандартизированных данных. Однако их основным ограничением является низкая масштабируемость, так как адаптация к новой предметной области требует значительных трудозатрат.

В связи с этим обоснована необходимость применения гибридного нейросимвольного подхода. Данная методология компенсирует непрозрачность нейронных сетей за счет наложения строгих логических ограничений, задаваемых онтологией предметной области, что обеспечивает интерпретируемость результатов.

I. Анализ существующих решений

В рамках исследования был проведен анализ существующих подходов и систем. Например, система OntoRAG реализует автоматическую деривацию онтологии из неструктурированного текста с использованием алгоритмов кластеризации, превосходя традиционные векторные подходы. Тем не менее, существенным ограничением OntoRAG является высокая вычислительная сложность процесса кластеризации графа при обработке масштабных корпусов данных.

Архитектура Contextual Graph Transformer интегрирует графовые нейронные сети для моделирования локальных связей и трансформерные модели для учета глобального контекста при обработке сложной технической документации. Программный комплекс ANDDigest, ориентированный на биомедицинскую область, сочетает словарные методы и алгоритмы обработки естественного языка, но его зависимость от

словарей ограничивает масштабируемость на новые области.

Проведенное исследование выявило дихотомию: системы либо специализируются на нейросетевом извлечении фактов с низкой верифицируемостью, либо на экспертном моделировании с высокой трудоемкостью. Универсальное решение для русскоязычного сегмента, объединяющее оба подхода, на текущий момент отсутствует.

II. Архитектура системы

Разработанная система имеет модульную архитектуру и структурно разделена на логическую часть, базу знаний и внешние элементы взаимодействия (компоненты диалога).

В состав логической части входят модули предобработки документа и текста, извлечения сущностей и связей, проверки логической согласованности, обработки запросов, поиска контекста, генерации ответа и разрешения конфликтов. Модуль предобработки документа извлекает текстовые данные, которые затем нормализуются в модуле предобработки текста для подготовки к дальнейшей работе. Обработанный текст передается в модуль извлечения сущностей и связей, результаты работы которого проверяются на наличие противоречий в модуле логической согласованности. Верифицированные данные используются для пополнения базы знаний.

III. Программная реализация и интерфейс

Логическая часть системы разработана на языке программирования Python 3.11. Для связи логики и пользовательского интерфейса использован высокопроизводительный асинхронный веб-фреймворк FastAPI.

Для извлечения сущностей и связей реализованы две различные стратегии. Первая — детерминированная, основанная на наборе библиотек Natasha (Navec, Slovnnet, Yargy, Corus, Razdel), которая производит анализ на основе синтаксиса и морфологии русского языка. Вторая — нейросетевая, делегирующая задачу локальной большой языковой модели Qwen 2.5 (7 млрд параметров). Преимуществом Qwen является работа со смыслом: модель глубже понимает контекст и способна выделять целевую информацию даже в «шумном» тексте, не требуя строгой предварительной очистки.

Взаимодействие с базой знаний (модуль kb_interaction) реализовано на основе интерфейсов sc-машины, являющейся частью технологии смыслового представления информации OSTIS (Open Semantic Technology for Intelligent Systems).

Пользовательский интерфейс реализован на базе TypeScript в сочетании с React-фреймворком Next.js. Рабочее пространство разделено на три зоны: модуль управления контекстами диалога, рабочую область интеллектуального агента для интерактивного диалога и панель инженерии знаний для визуализации семантического графа и загрузки документов. Безопасная авторизация реализована при помощи GitHub OAuth-приложения.

Заключение

Разработанный программный комплекс представляет собой современное решение для автоматизации извлечения структурированных знаний из неформализованных электронных документов. Научная новизна работы заключается в комбинировании методов глубокого обучения (для гибкости извлечения) и жестких онтологических ограничений (для логической корректности), что позволяет строить объяснимые структуры знаний из неструктурированных русскоязычных текстов.

Практическая значимость подтверждается созданием действующего прототипа, снижающего трудозатраты на первичную обработку технической документации с сохранением возможности верификации каждого факта ссылкой на первоисточник.

Перспективы развития системы включают внедрение более тесных схем взаимодействия между нейросетевыми и символьными модулями, расширение функционала для обработки таблиц, схем и изображений, а также усовершенствование алгоритмов автоматического разрешения смысловых конфликтов на основе анализа авторитетности источников.

Ерофеев Артём Сергеевич, учащийся УО «Национальный детский технопарк».

Кондратенко Владислав Александрович, учащийся УО «Национальный детский технопарк».

Научный руководитель: Зотов Никита Владимирович, ассистент кафедр ИИТ БГУИР.

Научный руководитель: Сальников Даниил Андреевич, ассистент кафедр ИИТ БГУИР.

Список литературы

1. Design and Construction of a NLP Based Knowledge Extraction Methodology in the Medical Domain Applied to Clinical Information – PMC / Electronic resource. – Mode of access: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6230532/>. – Date of access: 09.02.2026.
2. Построение графов знаний нормативной документации на основе семантического моделирования и автоматического извлечения терминов / Электронный ресурс. – Режим доступа: <https://cyberleninka.ru/article/n/postroenie-grafov-znaniy-normativnoy-dokumentatsii-na-osnove-semanticheskogo-modelirovaniya-i-avtomaticheskogo-izvlecheniya>. – Дата доступа: 09.02.2026.
3. Об автоматизации извлечения знаний из информационных ресурсов на основе онтологии предметной области / Электронный ресурс. – Режим доступа: <https://nasplib.isoftware.kiev.ua/items/eeab3435-368d-44cf-975c-d294dd467daa>. – Дата доступа: 09.02.2026.
4. A Semi-Automatic Ontology Development Framework for Knowledge Transformation of Construction Safety Requirements / Electronic resource. – Mode of access: <https://www.mdpi.com/2075-5309/15/4/569>. – Date of access: 09.02.2026.
5. Using AI and NLP for Tacit Knowledge Conversion in Knowledge Management Systems: A Comparative Analysis / Electronic resource. – Mode of access: <https://www.mdpi.com/2227-7080/13/2/87>. – Date of access: 09.02.2026.
6. OntoRAG: Enhancing Question-Answering through Automated Ontology Derivation from Unstructured Knowledge Bases / Electronic resource. – Mode of access: <https://arxiv.org/abs/2506.00664>. – Date of access: 09.02.2026.
7. Contextual Graph Transformer: A Small Language Model for Enhanced Engineering Document Information Extraction / Electronic resource/ – Mode of access: <https://arxiv.org/abs/2508.02532>. – Date of access: 09.02.2026.
8. Использование диаграмм переходов состояний для автоматизированного создания баз знаний / Электронный ресурс. – Режим доступа: <https://cyberleninka.ru/article/n/ispolzovanie-diagramm-perehodov-sostoyaniy-dlya-avtomatizirovannogo-sozdaniya-baz-znaniy>. – Дата доступа: 09.02.2026.