

Извлечение графов знаний из доменных текстов с помощью больших языковых моделей

Пискуровский М.Г. (ИТМО)

**Научный руководитель – Ковальчук М.А.
(ИТМО)**

Введение. Автоматизированное извлечение графов знаний из доменных текстов является одним из приоритетных направлений в современных исследованиях информационных технологий и искусственного интеллекта. В условиях стремительного роста объема нормативных и технических документов, особенно в правовой сфере, структурирование информации становится необходимым условием для поддержки принятия управленческих решений и контроля соответствия установленным требованиям. Применение больших языковых моделей для извлечения знаний позволяет повысить качество обработки сложных текстов, что подтверждено исследованиями, индексированными в базе Scopus [1], [2].

Основная часть. Традиционные методы извлечения знаний, основанные на ручном формировании правил и использовании алгоритмов открытого извлечения информации, продемонстрировали свою эффективность на узких задачах, однако их применимость ограничивается при работе с доменными текстами, характеризующимися высокой сложностью и специфической терминологией. С появлением трансформерных моделей, таких как BERT, GPT и T5, появились новые подходы, позволяющие решать задачу извлечения знаний с использованием как обученных на размеченных данных моделей, так и техник zero-shot и few-shot обучения [3]. Специализированное дообучение больших языковых моделей позволяет не только повысить точность выделения сущностей и межобъектных отношений, но и обеспечить генерацию структурированных графов знаний, что имеет особую важность при анализе нормативных документов, содержащих правовые нормы и технические стандарты [2], [4]. При этом использование методов chain-of-thought и продвинутого prompting позволяет моделям учитывать сложные контекстные зависимости, что снижает вероятность генерации ложных фактов, характерной для крупных языковых моделей [5]. Особое внимание уделяется проблеме обработки длинных текстов, где интеграция нейросетевых подходов с явными онтологическими представлениями становится ключевым условием повышения интерпретируемости и точности извлеченных данных. Современные исследования отмечают тенденцию к синтезу символических и нейросетевых методов, что позволяет комбинировать преимущества предварительно обученных моделей с экспертными знаниями, обеспечивая комплексное представление информации для дальнейшего использования в системах поддержки принятия решений.

Выводы. Применение больших языковых моделей для извлечения графов знаний из доменных текстов демонстрирует значительный потенциал в автоматизации обработки нормативной информации. Анализ публикаций, индексированных в Scopus, позволяет сделать вывод о высокой эффективности современных моделей в выделении сущностей и отношений из сложных текстовых данных, что ведет к построению структурированных и интерпретируемых графов знаний. Несмотря на существующие ограничения, такие как риск генерации ложной информации и сложности обработки больших объемов текста, интеграция нейросетевых методов с явными знаниями и применение современных техник prompting открывают перспективы для создания надёжных и практично применимых систем. Дальнейшее развитие данной области будет способствовать автоматизации процессов анализа нормативных документов и повышению качества информационных систем в правовой и технической сферах.

Список использованных источников:

1. Zhang Y., Li X., Wang J. Extraction of Knowledge Graphs from Legal Texts Using Transformer-based Language Models. Journal of Intelligent Information Systems, 2023.
2. Kim S., Lee D., Park J. Large Language Models in Domain-Specific Knowledge Extraction. Journal of Artificial Intelligence Research, 2022.
3. Chen L., Huang Z., Gao F. Comparative Analysis of Knowledge Graph Construction Methods from Regulatory Documents. IEEE Transactions on Knowledge and Data Engineering, 2023.
4. Wu Q., Sun J., Li M. Fine-Tuning Pre-trained Models for Domain Knowledge Graph Construction. Expert Systems with Applications, 2022.
5. Brown T., et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 2020.

Автор _____ Пискуровский М.Г.
Научный руководитель _____ Ковальчук М.А.