УДК 004.822

ФОРМИРОВАНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ СТРУКТУРЫ НАУЧНЫХ СТАТЕЙ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Минтус Е.А. (ИТМО)
Научный руководитель – Авдюшина А.Е. (ИТМО)

Введение. Современные методы обработки естественного языка значительно расширяют возможности анализа и структурирования научных текстов. Большие языковые модели (LLM) демонстрируют высокую эффективность в автоматическом извлечении сущностей и структурировании научных публикаций, что позволяет автоматизировать обработку и структурирование научных данных. Рост объема научных данных требует разработки интеллектуальных систем для их автоматического анализа, классификации и определения скрытых закономерностей. Для оценки потенциала LLM в этой области необходимо исследовать их применимость к анализу научных текстов и интеграции в аналитические платформы.

Основная часть. Исследование включало создание специализированного корпуса научных текстов, отобранных из открытых источников, с последующим извлечением ключевых сущностей. Автоматизированное извлечение сущностей и основных фактов осуществлялось путем анализа контекста и выделения значимых элементов текста. Выделенные сущности использовались для последующего структурного анализа и выявления ключевых взаимосвязей в научных публикациях. Для выявления логических и тематических взаимосвязей между публикациями анализировались связи между авторами, методологиями и направлениями исследований. Анализ цитируемости позволил определить публикации с наибольшим научным влиянием. Завершающим этапом стало внедрение векторизации и кластеризации данных. Преобразование текстов в векторные представления, позволило реализовать семантический поиск схожих исследований и определить тематические кластеры публикаций.

Выводы. В ходе исследования был разработан алгоритм автоматического извлечения ключевых сущностей из научных публикаций с использованием больших языковых моделей. Анализ показал, что модели эффективно справляются с выделением авторов, методологий, ключевых терминов и направлений исследований. Применение векторизации и кластеризации позволило сгруппировать публикации по тематикам и упростить поиск схожих исследований.

Дальнейшая работа будет направлена на: повышение точности выделения сущностей за счет расширения корпуса обучающих текстов и оптимизацию алгоритмов векторного представления для более точного тематического анализа

Список использованных источников:

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
2. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP).
3. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Research*.