

УДК 51-77

Возможности по применению LLM в разрабатываемых методах для анализа и структуризации текстов на естественном языке

Речинский Владислав Алексеевич

rechvlad@mail.ru

Московский политехнический университет, Москва, Россия

Аннотация. Рассмотрена методика для анализа текстов различных предметных областей и представления их в виде концептуальных графов, а также возможность применения нейронных сетей типа LLM для улучшения данной методики. Представлены этапы разрабатываемой методики: анализ текста, систематизация полученных данных и построение графов. Рассмотрены возможности по применению LLM в сферах обучения и анализа текстов. Перечислены потенциальные возможности для интеграции больших языковых моделей (LLM) в разрабатываемую методику для автоматизации этапов, требующих человеческого вмешательства. С учетом специфики свойств LLM и разрабатываемой методики сделано заключение о возможностях автоматизации разрабатываемой методики и ее применения при процессе обучения.

Ключевые слова: анализ текста, методика анализа текста, систематизация данных, программное обеспечение, большие языковые модели

Введение. Тексты представляют собой основу для передачи информации — будь то технические или поэтические тексты, — но тексты могут быть трудны для восприятия из-за обилия узкоспециальной терминологии. Сам по себе текст является языковой системой — совокупностью языковых единиц разных уровней, связанных устойчивыми отношениями. Для применения в сфере обучения наиболее подходящие — это системы, основанные на смысловых отношениях. Их применение облегчает усвоение материала, поскольку способствует более эффективному восприятию информации в учебниках, статьях или лекциях. Визуализация информации помогает выделить наиболее важные ее части. Для этого можно использовать граф, который характеризует выделенные понятия и связи-отношения между ними на основе их дефиниций и дескрипций [1].

Методы и материалы; результаты. Методика построения концептуальных графов текста состоит из трех основных методов:

1. Анализ материала заключается в работе с текстом предметной области и выделении ключевых слов в ней.

2. Систематизация полученных данных заключается в анализе полученной на этапе «Анализа текста» информации и ее систематизации путем поиска дефиниций и дескрипций выделенных ключевых слов, которые будут выступать в качестве смысловых единиц получаемой системы.

3. Процесс построения графов — метод, при котором происходит применение семантических и математических методов для связи ключевых слов,

полученных на этапе «Систематизации данных», между собой и их последующая визуализация.

Сама методика была реализована в виде программного обеспечения, представленного веб-страницей, которая строит два типа понятийных графов [2]:

- 1) граф, основанный на семантической сети;
- 2) граф, основанный на законе Брэдфорда.

В свою очередь, нейронные сети, основанные на LLM (большие языковые модели) — это мощный инструмент, помогающий при анализе естественного языка. Основным их преимуществом является возможность работать непосредственно со смысловой составляющей текстов [3]. Данная особенность позволяет улучшить разработанную методику.

Заключение. Использование закона распределения Брэдфорда для выявления тематического ядра и зон в понятийном графе описания предметной области любого текста, построенном с использованием выделения семантических связей между словами позволяет более эффективно усваивать информацию из статьи, так как наиболее важные ее части выделяются в ядро, а остальные распределяются по зонам, что позволяет начать ее изучение с ключевых частей, а в последующем переходить к менее важным. Включение в методику нейросетей позволит уменьшить количество времени, требуемое на анализ текста и повысить его эффективность.

Список источников

- [1] Маршакова-Шайкевич И.В. *Россия в мировой науке: библиометрический анализ*. Москва, Российская Академия наук, Институт Философии, 2008, 218 с.
- [2] Rechinsky V.A., Philippovich Y.N., Philippovich A.Y. Presentation of scientific documentation in the form of graphs using developed software. Proc. SPIE 13065, Third International Conference on Optics, Computer Applications, and Materials Science (CMSD-III 2023). 2024. <https://doi.org/10.1117/12.3025081>
- [3] Hadi M.U., Al-Tashi Q., Qureshi R., Shah A. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. TechRxiv, 2023. <https://doi.org/10.36227/techrxiv.23589741>