К. А. Плешак

МЕТОДЫ И СРЕДСТВА ОСУЩЕСТВЛЕНИЯ АВТОМАТИЧЕСКОГО СЕМАНТИЧЕСКОГО АНАЛИЗА НА МАТЕРИАЛЕ АНГЛОЯЗЫЧНЫХ ТЕКСТОВ

С достижением научно-технического прогресса перед человечеством открылись новые возможности. Однако развитие технологий не только поспособствовало оптимизации человеческой жизни, но и привело к значительному росту объемов информации. В связи с этим возросла потребность в быстрой и точной обработке данных, что подтолкнуло к разработке различных методов и средств осуществления автоматического анализа текстовой информации. Данная работа посвящена определению методов и средств, применимых к обработке текстов на семантическом уровне. Актуальность исследования заключается в выявлении наиболее эффективных методов и средств семантического анализа данных, в результате осуществления которых усовершенствуется сам анализ смысловой составляющей текста и модернизируется обработка естественного языка.

Семантический анализ может реализовываться статистическими методами и методами глубокого обучения на примере трансформерных моделей. Первые алгоритмы хорошо подходят для выделения ключевых терминов и обнаружения скрытых тематических структур. Их преимущества – высокая скорость обработки данных и небольшие запросы к вычислительным ресурсам. Недостатком таких методов считается осуществление лишь поверхностного анализа. Трансформеры, за счет применения двунаправленного кодирования, позволяют разрешить неоднозначность и кореференцию естественного языка, извлечь латентные смысловые связи. Такие модели подходят для проведения глубокого контекстуального анализа и решения широкого спектра задач.

Выбор средств осуществления автоматического семантического анализа зависит от выбора конкретного метода, специфики исходных данных и поставленной задачи. Наиболее эффективные и простые в применении средства – библиотеки NLTK и SpaCy. Для тематического моделирования широко используются библиотеки Gensim и Stanford CoreNLP. Фреймворки TensorFlow и PyTorch, посредством которых осуществляются методы глубокого обучения, обеспечивают выполнение более комплексных задач семантического анализа и работу с большими массивами текстовых данных.

В ходе исследования было выявлено, что для реализации семантического анализа наиболее актуально внедрение трансформеров, т.к. у них числовые показатели метрик приближены к максимальному значению. Однако данные архитектуры имеют ряд ограничивающих факторов, в т.ч. проблематичность обработки текстов с терминологией или культурно-специфической лексикой, высокие вычислительные затраты и сложность

интерпретации, поэтому они продолжают совершенствоваться. Дальнейшая перспектива развития методов глубокого обучения особенно связана с их интеграцией с другими алгоритмами.