Н. И. Липлянина

ИССЛЕДОВАНИЕ МЕТОДОВ И АЛГОРИТМОВ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ НАУЧНЫХ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Автоматическое построение текста на естественном языке представляет собой одну из самых сложных задач в области искусственного интеллекта и компьютерной лингвистики. Оно имеет широкие применения в различных сферах, включая генерацию новостей, создание статей для сайтов, составление отчетов и выполнение других задач, требующих быстрого создания текстов.

В ходе проведенного нами исследования были выделены 3 основных метода автоматического построения текстов на естественном языке: метод генерации текстов по шаблонам, лингвистически мотивированный метод (метод генерации на основе лингвистических правил и знаний) и метод на базе машинного обучения, предполагающий применение как нейросетевых моделей, так и классических алгоритмов машинного обучения.

Проведенный анализ существующих подходов к автоматическому генерированию текстов на русском, английском и китайском языках на материале корпусов текстов научного стиля продемонстрировал, что каждый из представленных методов имеет свои сильные и слабые стороны. Метод генерации текстов по шаблонам, хоть и позволяет создавать относительно связные и грамотные тексты, сильно ограничен в плане гибкости и креативности. Подход на основе лингвистических правил, в свою очередь, требует значительных трудозатрат на создание обширных баз знаний и синтаксических правил, что может препятствовать его широкому применению. В противовес этим более традиционным методам, технологии на базе рекурентных нейронных сетей, а также сетей архитектуры «трансформер» демонстрируют значительно более высокий потенциал в области генерации текстов естественного языка. Ключевыми этапами непосредственно процесса генерации новых текстов являются предварительная подготовка текстовых данных, включая их векторизацию для представления в численном виде, определение оптимальной архитектуры и параметров модели, а также ее обучение на представительном корпусе текстовых данных, что обеспечивает возможность генерировать оригинальные связные тексты на различных языках. Основные ограничения подхода связаны, с качеством и репрезентативностью используемых для обучения модели данных, а также с выбором подходящей архитектуры нейронной сети и параметров используемой модели, что влияет на обеспечение семантической и логической согласованности достаточно длинных и сложных текстовых конструкций в синтезируемых текстах.