ПРИКЛАДНАЯ ЛИНГВИСТИКА

В. Д. Гурская

МЕТОДЫ И СРЕДСТВА АВТОМАТИЧЕСКОГО ВЫЯВЛЕНИЯ ЗНАНИЙ ОСНОВНЫХ ТИПОВ НА МАТЕРИАЛЕ АНГЛОЯЗЫЧНЫХ ТЕКСТОВ

Современное общество характеризуется стремительным развитием информационных технологий, что приводит к существенному увеличению объема данных и информации. Учитывая постоянный рост потока информации, становится необходимым автоматизировать процессы обработки и анализа текста. Важнейшим компонентом этого начинания является автоматическое определение знаний из текстов, что может значительно улучшить и ускорить наше понимание больших наборов данных. Актуальность исследования обосновывается тем, что в современном мире объем информации продолжает расти ежедневно, что делает автоматическое извлечение знаний из текстов все более важным для эффективного анализа данных. Текущие достижения в области технологий, включая нейронные сети, представляют новые возможности для автоматического выявления, подчеркивая актуальность этой области.

Для автоматического извлечения знаний основных типов на материале англоязычных текстов в качестве примера использовали набор данных «IMDb Movie Reviews» с Kaggle. Платформа, которая представляет собой совокупность датасетов, содержащих научные статьи, рецензии, новостные сводки, а также пользовательские посты из социальных сетей. Основное внимание будет уделено методам, которые позволяют автоматически выявлять и классифицировать информацию из текстовых данных.

Выбор конкретного метода зависит от поставленной задачи и характеристик обрабатываемых данных. Если основная цель – высокая точность классификации, предпочтительнее использовать трансформеры или RNN. Приоритет в скорости обработки может быть отдан SVM или случайному лесу, а для тематического анализа оптимальными окажутся LDA или K-means. Эти результаты подчёркивают важность грамотного выбора метода в зависимости от целей исследования, создавая прочную основу для эффективного извлечения знаний из англоязычных текстов.

Анализ особенностей англоязычных текстов выявил, что они обладают разнообразием лексики, синтаксических конструкций и стилистических особенностей, что требует адаптации методов извлечения знаний. Важные типы знаний, такие как факты, отношения и события, были определены как ключевые для автоматического извлечения, что открывает новые возможности для их применения в различных областях, включая бизнес, науку и технологии.

Оценка эффективности предлагаемых методов с использованием экспериментальных данных показала, что использование современных нейросетевых моделей, в частности трансформеров, значительно повышает качество анализа по сравнению с традиционными методами.