УДК 004.89

УПРОЩЕНИЕ ПРОЦЕССА ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Ле Д.В., Университет ИТМО, Санкт-Петербург
Научный руководитель – Осипов Н.А., доцент, кандидат технических наук
Университет ИТМО, Санкт-Петербург

В данной работе рассмотрены основные проблемы обработки естественного языка. Проанализированы основные направления обработки, методы, инструменты и библиотеки, доступные для решения задачи обработки естественного языка и трудности, возникающие при выполнении этих задач. Построены стандартные этапы обычного NLP–конвейера и получены подходы к разработке и направления искусственного интеллекта.

Введение. Обработка естественного языка и синтаксический анализ естественного языка являются важными проблемами искусственного интеллекта, которые изучаются многими учеными по всему миру в течение последних 50 лет. Основные проблемы обработки языка и лингвистики — это изучение лексики и семантики. В этих исследованиях невозможно обойтись без работы со словарями и архивами. Но не всегда существует возможность доступа к этим ресурсам. Основная особенность этих типов данных заключается в том, что они не структурированы или частично структурированы и не могут храниться в фиксированных форматах, таких как таблицы. Помогает компьютерным системам понимать и обрабатывать человеческий язык — это главная цель обработки естественного языка.

Основная часть. Обработка естественного языка — общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез — генерацию грамотного текста. Решение этих проблем будет означать создание более удобной формы взаимодействия компьютера и человека. Но существует много причин, по которым обработка естественного языка является сложной задачей поэтому компьютеры не могут в полной мере понимать живой человеческий язык, однако они на многое способны. В некоторых ограниченных областях NLP может делать по-настоящему волшебные вещи и экономить огромное количество времени. И что еще лучше, последние достижения в NLP легко доступны через библиотеки Python с открытым исходным кодом, такие как spaCy, textacy и neurocoref с помощью моделей one-hot coding и word2vec. С другой стороны сам процесс чтения и понимания текста сложен. Выполнение любой сложной комплексной задачи в машинном обучении обычно достигается путем разбиения задачи на маленькие части и решается по отдельности. Таким способом называется NLP–конвейера.

Выводы. В данной работе были рассмотрены основные проблемы и способы анализа естественного языка. Были ознакомлены с инструментами анализа текста, были построены стандартные этапы обычного NLP–конвейера и получены подходы к разработке модели искусственного интеллекта.

Ле Д. В. (автор) Подпись

Осипов Н.А. (научный руководитель) Подпись