УДК 004.89

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ СОЗДАНИЯ АЛГОРИТМА ОПРЕДЕЛЕНИЯ НАБОРА СЛОВ ДЛЯ ИЗУЧАЮЩИХ АНГЛИЙСКИЙ ЯЗЫК

Насыров Н.Ф. (Университет ИТМО),
Научный руководитель - кандидат технических наук, доцент Федоров Д.А.
(Университет ИТМО)

Введение. Большие языковые модели уже достигли уровня, который позволяет генерировать последовательность шагов алгоритмов, их параметров, генерировать программный код на различных языках программирования. Разными компаниями разрабатываются программные решения, позволяющее анализировать программный код, определять возможные уязвимости повышать эффективность алгоритмов. Метрики бенчмарков учитывают качество предлагаемых решений по решению алгоритмических задач и генерации кода. Однако большие языковые модели не всегда эффективны в вопросах описания алгоритмов решения сложно формализуемых задач. В исследовании рассматривается задача создания алгоритма для веб-приложения, позволяющего расширять словарный запас изучающих английский язык.

Основная часть. Каждый эксперт (в роли которого может выступать, например, преподаватель-лингвист, репетитор или непосредственно сам обучающийся) формирует пул следующих для изучения слов, опираясь на свой опыт, экспертную оценку знаний лексики обучающегося, используемую методику изучения слов и так далее. Также следует учитывать проблему субъективности понятия сложности английских слов [1]. Очевидно, что для разработки программного решения для расширения словарного запаса изучающих английский язык, затруднительно представить модель, достоверно интерпретирующую экспертную оценку.

В связи с этим целесообразно исследовать возможности больших языковых моделей для систематизации и формализации данных, которые могут использоваться в решении поставленной задачи. Безусловно, можно описать некоторые параметры, которые следует учитывать в вопросе определения следующих для изучения слов. К таким параметрам можно отнести частотность слов по различным словарям [2], соответствие слов, а также уровень владения английским языком по уровням Common European Framework of Reference for Languages (CEFR) [3], список изученных ранее слов.

Большие языковые модели могут помочь определить дополнительные параметры, не очевидные на первый взгляд, которые также необходимо учитывать для определения следующих для изучения слов. В работе использовались как отечественные, так и зарубежные большие языковые модели. Анализ их ответов позволил выделить дополнительные параметры, которые, возможно, следует учитывать при реализации алгоритмов. Стоит отметить, что любое предложение больших языковых моделей, безусловно, должно быть подвергнуто экспертной оценке. Тем не менее, ряд параметров, предложенных большим языковыми моделями, подлежит дальнейшему исследованию для определения целесообразности их использования в реализации алгоритма описанной задачи. В частности, были дополнительно выявлены

такие параметры, длина слова, вхождение слов во фразовые глаголы. Также было сформулировано предложение использовать алгоритмы машинного обучения для мультиклассификации слов на основе перечисленных выше параметров (класс соответствует уровню сложности слова).

Выводы. В ходе исследования были определены параметры, которые могут оказать влияние на точность определения набора следующих для изучения слов английского языка. Результаты исследования имеют практическую значимость и будут использоваться при разработке веб-приложения.

Список использованных источников:

1. Сосчитать незримое: достоверно определяем словарный запас [сайт]. — URL: https://habr.com/ru/companies/skyeng/articles/301214/
2. Frequency List. Explore the top 5000 words in English: [сайт]. — URL: https://frequencylist.com/ (дата обращения: 15.02.2025).
3. The CEFR Levels [сайт]. — URL: https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions (дата обращения: 15.02.2025).

Насыров Н.Ф. (автор)

Федоров Д.А. (научный руководитель)