УДК 004.021

АВТОМАТИЧЕСКОЕ СОЗДАНИЕ ТЕКСТОВ СИНОНИМОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Вахрушев К.К., Национальный исследовательский университет ИТМО
Научный руководитель – к.ф.-м.н., доц. Фильченков А. А,
Национальный исследовательский университет ИТМО

В работе рассматриваются существующие способы перефразирования текста. Анализируются проблемы существующих решений и предлагается свой алгоритм создания текстов синонимов, включающий методы машинного обучения.

Введение. Перефразирование текстов – это одна из задач в области обработки естественного языка, которая имеет множество применений в актуальных сферах машинного обучения, таких как: генерация текста, перевод текстов, создание вопросно-ответных систем, генерирование дополнительных обучающих данных. Большое количество времени задача перефразирования решалась с использованием статистических и основанных на правилах подходах. Самые современные подходы основаны на использовании рекуррентных нейронных сетей и Encoder-Decoder архитектур. Большие универсальные языковые модели, такие как GPT-2, обладая представлением о синтаксисе и грамматике, также оказались способными перефразировать текст после дообучения. Но у существующих решений остается много проблем, в их числе: невозможность генерировать перефразирования не только на уровне предложений, но и для более длинных частей текста, без необходимости разбивать текст на более мелкие куски; слабое обогащение перефразированного текста новыми словами; сильная стилистическая связанность перефразированного текста и исходного; неиспользование морфологических особенностей языка. Это вынуждает продолжать поиски новых подходов к решению задачи перефразирования.

Основная часть. Для решения существующих проблем в задаче перефразирования текста необходимо было исследовать различные способы генерации перефразированных текстов. Но для обучения необходимы были большие корпуса тренировочных данных и хорошие метрики для оценки качества перефразированного текста. Так как на сегодняшний момент не существует удовлетворяющих нас корпусов и метрик, был разработан упрощенный язык из алфавита, позволяющего ввести точную метрику и самостоятельно генерировать тренировочные данные. Это позволило быстро тестировать различные алгоритмы и найти необходимый. Построенные на упрощенном языке модели в дальнейшем можно обобщить на любой язык.

Выводы. Рассмотренный подход позволил создать алгоритм частично превосходящий существующие алгоритмы перефразирования текстов. В дальнейшем необходимо проверить, как обогащение корпусов тренировочных данных с помощью разработанного алгоритма может улучшить показатели существующих алгоритмов. Также полученное решение даст возможность создавать перефразирования нужного размера, с использованием выбранной лексики и в определенной стилистике.

Вахрушев К.К. (автор) Подпись

Фильченков А.А (научный руководитель) Подпись