УДК 004.81

ГЕНЕРАЦИЯ ТЕКСТОВ, ЭКВИВАЛЕНТНЫХ ПРЕДОСТАВЛЕННОМУ ОБРАЗЦУ

Е. Д. ГУМЕННИКОВ
Научный руководитель И. А. МУРАШКО, д-р техн. наук, доц.
Гомельский государственный технический университет им. П. О. Сухого
Гомель, Беларусь

Автоматическая генерация эквивалентных текстов заключается в автоматическом написании текста, семантически идентичного предоставленному образцу. Система, способная качественно решать данную задачу, может быть применена в таких отраслях, как построение чатботов, разработка голосовых интерфейсов, а также для автоматизации работы «копирайтеров».

Исходными данными для такой программы служит фрагмент текста – образец. Выходными же данными будет текст, эквивалентный предоставленному, но содержащий другие формулировки.

Решение описанной задачи с помощью применения простой рекуррентной нейронной сети имеет некоторые недостатки, обусловленные тем, что такая архитектура не имеет механизма «памяти», что накладывает серьезные ограничения на способность данной системы оценивать контекст встречающихся в тексте слов. Однако *LSTM*-сеть обладает такой памятью. Предложенное решение базируется на нейронной сети данной архитектуры.

В основе предложенной нейросетевой модели генератора эквивалентного текста лежит модель *seq2seq.* Данная модель базируется на архитектуре *LSTM.* Оригинальный текст по слову подается на вход *LSTM* нейронной сети, играющей роль кодировщика. Выход этой сети есть состояние ячейки, полученное при обработке текущего элемента исходного текста и предыдущего состояния ячейки. Полученный вектор подается в качестве входных данных для второй *LSTM*-сети, которую называют декодировщиком. Ее предназначение состоит в генерации очередного слова-эквивалента. Данный вектор в описанной модели называется вектором промежуточного представления. Промежуточное представление используется в популярных нейросетевых моделях, предназначенных для решения задач автоматического перевода, и, как правило, представляет граф, кодирующий входной текст. Система-переводчик генерирует выходной текст на основе данной промежуточной структуры. Подобная модель может быть успешно применена для решения задачи генерации эквивалентных текстов.

LSTM-сети являются наилучшей архитектурой для решения задач обработки последовательностей и решения задач, связанных с обработкой текстов. Однако качество выполнения тех или иных задач связано не только с архитектурой, лежавшей в основе нейронной сети. Огромную роль также играет качество и объем данных, предоставленных сети для обучения.