УДК 004.896, 004.855.5

ПОСТРОЕНИЕ СИСТЕМЫ ГЕНЕРАЦИИ СТИЛИЗОВАННЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И НЕЙРОННЫХ СЕТЕЙ

Ларькин В. Д. (Московский авиационный институт, МАИ), **Поволоцкий В. А.** (Московский авиационный институт, МАИ)
Научный руководитель – к.ф.-м.н. Пановский В. Н.
(Московский авиационный институт, МАИ)

В решении задач из области обработки естественного языка и, в частности, генерации текста в последние годы хорошо себя зарекомендовали нейронные сети архитектуры Transformer. Обучаясь на больших корпусах, они строят точные модели языка и показывают лучшие на сегодняшний день результаты в задаче генерации правдоподобной речи. В настоящей работе мы показываем, как, дообучив такую модель на сравнительно небольшом корпусе из конкретной предметной области, можно добиться от модели усвоения стилистики текстов данной области и использовать её в дальнейшем для генерации новых текстов.

Введение. В современном мире с ростом уровня образования и увеличением спроса на специалистов высокой квалификации для оптимизации производственных процессов требуется повышать уровень автоматизации предприятий, чтобы разгрузить работников и предоставить им возможность заниматься интеллектуальным трудом. А с увеличением вычислительных мощностей и развитием компьютерных наук всё больше рутинных задач поддаются автоматизации. Задача автоматизации написания разного рода текстов стоит особенно остро, так как она актуальна в самых разных сферах деятельности. Предложенное нами решение позволяет автоматизировать большую часть работы по созданию базовой структуры документа и его частичному заполнению, предоставляя возможность коррекции и дополнения пользователем «на лету». При этом, область применения описанной системы может быть самой разной: от написания отчётов по различным работам до генерации сценариев юмористических телешоу.

Основная часть. Мы предлагаем использовать одну из новейших русскоязычных языковых моделей ruGPT-3 Small от «Сбера». В сравнении с другими моделями, распространяемыми под свободными лицензиями, её преимуществом является то, что она обучалась на русскоязычном корпусе и заточена под генерацию текста, в первую очередь, на русском языке, а малое относительно других моделей семейства ruGPT-3 число параметров позволяет дообучать её, располагая сравнительно небольшими мощностями.
Перед обучением производится предварительная обработка корпуса, заключающаяся в выделении структурных блоков текста специальными синтаксическими конструкциями на естественном языке, формат которых определён заранее и сохраняется неизменным во всём корпусе. Затем каждый текст обучающей выборки разделяется на части такого размера, чтобы:

1) в токенизированном виде их длина была больше длины контекста модели,
2) их длина была не слишком большой, чтобы при обучении иметь возможность подавать их в модель в случайном порядке,
3) каждая часть была самостоятельным осмысленным текстом, принадлежащим исходной предметной области.

После обучения модель встраивается в пользовательский интерфейс, включающий возможность дополнения написанного текста, в том числе, структурных синтаксических конструкций для поддержания требуемого формата вывода.

Выводы. Вышеописанным способом можно строить системы генерации стилизованного текста для неограниченного числа предметных областей путём использования разных наборов данных. Гибкость задания структуры текстов на естественном языке даёт возможность

поддержания требуемого формата выхода с помощью подачи на вход модели отдельных структурных элементов текста, а предопределённость формата данных позволяет автоматически форматировать и корректировать сгенерированный текст.

Ларькин В. Д. (автор)	Подпись
Пановский В. Н. (научный руководитель)	Подпись