УДК 004.912

ИССЛЕДОВАНИЕ МЕТОДОВ ГЕНЕРАЦИИ СЛУЧАЙНОГО ТЕКСТА НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Трушин И. Е. (Университет ИТМО)
Научный руководитель – к.т.н., доцент Шиков А. Н. (Университет ИТМО)

Аннотация

Обработка естественного языка (Natural Language Processing) используется во многих сферах нашей жизни. Одной из главных прикладных задач NLP является генерация текста. В этом докладе будут рассмотрены различные методы и способы создания текста из существующей выборки.

Введение

Влияние NLP в повседневной жизни для человека неоспоримо велико, хоть и заметить его довольно сложно. Это направление в области искусственного интеллекта используется во многих сферах современности: машинный перевод (Яндекс Переводчик, Google Translate), классификация текста (определение спама и категории писем, определение семантики текста), суммаризация (генерация краткого пересказа текста), вопросно-ответные и диалоговые платформы (различные чат боты, Яндекс Алиса, Google Assistant). Также одной из главных сфер NLP является генерация текста.

Основная часть

В данном докладе будут рассмотрены некоторые способы и метод генерации текста на существующей выборке. Первым из них являются цепи Маркова. Цепь Маркова – инструмент из теории случайных процессов, состоящий из последовательности конечного количества состояний. Связи между значениями цепочки при этом создаются, только если состояния стоят строго рядом друг с другом. Иначе говоря, мы можем определить закон распределения случайной величины и вычислить с какой вероятностью одно из слов появится в тексте после текущего. Другим методом генерации текста является метод N-грамм. N-граммы – непрерывные последовательности n-элементов в предложении. При использовании данного алгоритма выделяют определённые N-граммы – биграммы и триграммы. Отличие последних от биграмм в том, что условная вероятность слова определяется не по одному предшествующему слову в предложении, а по двум. Здесь же обозначено отличие данного подхода от цепей Маркова – условная вероятность определяется по предшествующим словам, а не по следующему. Следующим методом генерации текста являются рекуррентные нейронные сети (RNN). RNN – это подкласс нейронных сетей, являющийся мощным инструментом для моделирования последовательных данных, например, таких как временные ряды или естественный язык. Это достигается благодаря сохранению внутреннего состояния на различных временных шагах последовательности. Иными словами, у RNN имеется специальная ячейка памяти, которая запоминает прошлые шаги.

Выводы

В результате исследования проведен анализ методов генерации текста: цепи Маркова, метода N-грамм и рекуррентные нейронные сети. Продемонстрированы практические результаты методов, по результатам которых сделан вывод о наиболее подходящем методе для выполнения поставленной задачи, а именно, генерация текста на основе существующей выборки.