УДК: 004.9

АНАЛИЗ ПОДХОДОВ И МОДЕЛЕЙ ГЕНЕРАЦИИ СЕМАНТИЧЕСКИ ОБОСНОВАННЫХ ТЕКСТОВ

Близнюк А.В.

НИУ БелГУ – Белгородский национальный исследовательский университет, Россия, Белгород, e-mail: 1144611@bsu.edu.ru

В данной работе рассматриваются основные проблемы генерации текста, проводится сравнение существующих подходов и моделей генерации семантически обоснованного текста.
Ключевые слова: автоматическая генерация текста, рекуррентные нейронные сети, семантика, машинное обучение, онтология.

ANALYSIS OF APPROACHES AND MODELS FOR GENERATING SEMANTICALLY BASED TEXTS

Bliznyuk A. V.

BelGU – Belgorod State Research University, Belgorod, e-mail: 1144611@bsu.edu.ru

The article considers the main problems of text generation and compares existing approaches and models for generating semantically based text.
Keywords: automatic text generation, recurrent neural networks, semantics, machine learning, ontology.

Одной из основных характеристик современных программных систем различного назначения является возможность представления результатов их работы в соответствии с требованиями, определяемыми не только предпочтениями конкретного пользователя или групп пользователей, но, прежде всего, назначением программной системы и особенностями предметной области.

Существуют различные виды представления выходной информации – таблицы, графики, диаграммы. Особое место занимает представление выходных данных в виде текстов. При генерации таких текстов приходится решать три основные проблемы: каким образом задавать способ изложения текста; как породить текст, соответствующий этому способу изложения на основе выходных данных прикладной программы; как разместить этот текст, чтобы обеспечить возможность его дальнейшего использования.

Целью исследования является анализ подходов и моделей генерации текста.

Материалы и методы.

Основополагающим для предпринимаемого исследования является теоретический метод исследования, включающий анализ, дедукцию, аналогию.

Основная часть.

В настоящее время рынок программных систем не предлагает универсальных средств генерации текстов для произвольных предметных областей.

Представление результатов в виде текста возможно с помощью генераторов отчетов. Они, как правило, ориентированы на представление отчетов в виде таблиц, графиков, диаграмм, поэтому с их помощью можно сгенерировать только тексты простой структуры. Генераторы отчетов не имеют специальных средств анализа полученных результатов, а снабжены только функциями их фильтрации и сортировки. Технология формирования отчетов с помощью таких средств основана на создании шаблона отчета, который в большинстве случаев встраивается в приложение, что делает невозможным любые изменения самого приложения; многие генераторы отчетов формируют отчет в собственном формате без возможности его редактирования и изменения [1].

Существует метод автоматической генерации текстов произвольных структуры и содержания, управляемой онтологией на основе результатов работы прикладной программы, представленных в виде неупорядоченного множества кортежей отношений.

Синонимайзеры и генерация фраз по шаблонам. Часто генераторы текстов совмещены с программами-синонимайзерами, которые автоматически меняют слова на синонимы, в целях рерайта и придания уникальности фразам. Слова, которые надо заменять в шаблоне на синонимы, заменяются макросами [2].

Чем длиннее текст, тем заметнее неестественность в автоподставленных синонимах. Поэтому в текстах «сделанных для людей» (*СДЛ*) синонимайзеры могут применяться только для создания уникальных коротких текстов: заголовков и анкоров с ключевыми словами, комментариев и абзацев. Синонимайзеры более успешно применяются в английском языке, который, в отличие от русского языка, имеет простую морфологию.

Ряд компаний развивает более сложную технологию. Создаются синтаксические структуры по частям речи и членам в предложениях, слова в словарях категоризируются по семантике, с дальнейшей автоподстановкой их в предложения. Данным кругом задач занимается общее направление искусственного интеллекта и математической лингвистики — обработка естественного языка (*Natural Language Processing, NLP*). Оно изучает проблемы компьютерного анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез — генерацию грамотного текста. Решение этих проблем будет означать создание более удобной формы взаимодействия компьютера и человека [3].

Эффективным решением задачи генерации текста является применение рекуррентных нейронных сетей. Преимущество рекуррентных нейронных сетей — в возможности использовать неограниченно длинный контекст. Вместе с каждым словом, поступающим на

вход рекуррентной ячейки, в неё приходит вектор, представляющий всю предыдущую историю — все обработанные к данному моменту слова [2].

Возможность использования контекста неограниченной длины, конечно, только условная. На практике классические RNN страдают от затухания градиента — по сути, отсутствия возможности помнить контекст дальше, чем на несколько слов. Для борьбы с этим придуманы специальные ячейки с памятью. Самыми популярными являются LSTM и GRU [5].

Заключение

В ходе анализа были рассмотрены следующие подходы и модели генерации текста: генераторы отчетов, метод автоматической генерации текстов произвольных структуры и содержания, управляемой онтологией, синонимайзеры и генерация фраз по шаблонам, а также применение рекуррентных нейронных сетей. Были выявлены преимущества и недостатки рассмотренных методов и принято решение – в ходе дальнейшего исследования разработать новую модель генерации семантически обоснованного текста на основе рекуррентных нейронных сетей.

Список литературы:

1. Разработка систем анализа и генерации текстов [Электронный ресурс] /Электрон.дан.URL:https://habr.com/ru/company/meanotek/blog/259355/ (дата обращения 10.10.2019)
2. Нейронные сети. Statistica Neural Networks. Методология и технологии современного анализа данных[Текст]. Горячая Линия – Телеком -2017. - 392 с.
3. Ilya Sutskever, James Martens, Geoffrey Hinton. Generating Text with Recurrent Neural Networks [Текст] // University of Toronto, 6 King’s College Rd., Toronto, ON M5S 3G4 CANADA
4. Большакова Е.И Автоматическая обработка текстов на естественном языке и компьютерная лингвистика.[Текст]/ Е.И. Большакова, Э.С. Клышинский. – М.: Изд-во НИУ ВШЭ, 2017. – 269 с.
5. Generating Logical Forms from Graph Representations of Text and Entities [Электронный ресурс] /Электрон.дан. – URL: https://research.google/pubs/ (дата обращения 15.10.2019).