УДК 004.89

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ КОНТЕНТА ДЛЯ ГЕНЕРАЦИИ ПРЕЗЕНТАЦИЙ

Власенко Н. А.
(Университет ИТМО)
Научный руководитель – к. т. н. Русак А. В.
(Университет ИТМО)

В данной работе проводится исследование алгоритмов языковых моделей для задачи суммаризации объемных документов на русском языке и их сравнение на подготовленных наборах данных. Также была предложена архитектура приложения для генерации презентаций.

Введение. В современном информационном обществе все больше важности приобретает умение представлять информацию в понятной и увлекательной форме. Актуальность работы обусловлена тем, что несмотря на то, что исследования в области обработки естественного языка прогрессируют с каждым годом, все ещё существуют проблемы в обработке русского языка ввиду меньшего количества исследований. Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 623106 «Исследование методов интеллектуальной обработки контента для генерации эффективных презентаций».

Основная часть. Создание презентаций является творческим процессом, требующим задействовать аналитические навыки для выделения главных мыслей и передачи их в сжатой и визуально приятной форме. Рассмотрим шаги, связанные с созданием презентационных слайдов на основе документа: определить тематику презентации, основные разделы, структуру контента и его расположение на слайде, суммаризировать текст, извлечь формулы, извлечь рисунки, расположить все на слайде в соответствии со структурой, сгенерировать уникальный стиль. Во многих исследованиях используются алгоритмы целочисленного линейного программирования (ILP) [1], RoBERTa для кодировки предложений и bidirectional GRU для извлечения эмбеддингов текста, для извлечения эмбеддингов изображений используется ResNet-152 [2], а также для парсинга документов используется Stanford parser и разделение предложений на характерные для английского языка словосочетания с глаголами или с существительными[3][4]. Для проведения исследований были выбраны несколько наборов данных: собран корпус дипломных работ студентов ИТМО за 2022 и 2023 года, стандартный корпус с автоматическим обобщением новостей GusevGazeta, тексты научных статей на английском языке TitleAbstractSummaryDatasets и этот же набор, переведенный на русский язык. С помощью методов суммаризации и подготовки данных, таких как TextRank, который позволяет построить граф по тексту и извлечь ключевые слова и связи, извлечения эмбеддингов с помощью Summa и Navec из проекта Natasha, а также жадная oracle суммаризация в тандеме с рекуррентной нейронной сетью SummaRuNNer [5], трансформер Bart, предобученный на русском датасете и новая генеративная модель Gigachat от Сбера, основанная на ruGPT-3.5.

Выводы. В данной работе были реализованы алгоритмы и подходы к суммаризации объемных русских текстов из разных датасетов, акцент делался на научные статьи и дипломные работы на разные темы. Лучших результатов удалось достигнуть с помощью жадного oracle summary и рекуррентной модели SummaRuNNer для экстрактивной суммаризации документов на основе GRU по метрикам качества BLEU и ROUGE-1, ROUGE-2 и ROUGE-L. Также была разработана архитектура ПО, описаны классы и методы для приложения, которое позволит по загруженным PDF файлам генерировать презентации в формате PPTX.

Список использованных источников:

1. Y. Hu and X. Wan. PPSGen: Learning-Based Presentation Slides Generation for Academic Papers // IEEE Transactions on Knowledge and Data Engineering. – 2015. №27(4). 1085-1097. DOI: 10.1109/TKDE.2014.2359652.
2. Fu, T.-J., Wang, W. Y., McDuff, D., & Song, Y. DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents //Proceedings of the AAAI Conference on Artificial Intelligence. – 2022. №36(1). 634-642. DOI: 10.1609/aaai.v36i1.19943
3. Wang, S., Wan, X., & Du, S. Phrase-Based Presentation Slides Generation for Academic Papers //Proceedings of the AAAI Conference on Artificial Intelligence. - 2017. №31(1). DOI: 10.1609/aaai.v31i1.10481
4. Sefid, A., Wu, J., Mitra, P., & Giles, C.L. Automatic Slide Generation for Scientific Papers // CEUR Workshop Proceedings. - 2019. №2526. (11-16).
5. Nallapati, R., Zhai, F., & Zhou, B. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. // AAAI Conference on Artificial Intelligence. 2016. DOI: 10.1609/aaai.v31i1.10958