УДК 004.85

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ ОБЪЕКТОВ ПОВЕСТВОВАНИЯ ИЗ ТЕКСТА НА РУССКОМ ЯЗЫКЕ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Шаламов В.В. (Университет ИТМО), **Ефимова В.А.** (Университет ИТМО)
Научный руководитель – к.ф.-м.н., Фильченков А.А. (Университет ИТМО)

В данной работе рассматривается метод извлечения из короткого текста (до 5000 символов) на русском языке ключевых объектов, их описаний и положения в пространстве. Это нужно для построения сцены на основе текста.

Введение. Понимание компьютером текста на естественном языке используется во многих областях, оно применяется для автоматизированного поиска информации, ответа на вопросы, автоматического перевода и поддержания диалога. Задачу разбора текста исследуют давно, существует множество библиотек для обработки естественного языка, которые позволяют размечать текст на части речи, извлекать семантические связи между словами, извлекать эмоциональную окраску и именованные сущности [1].

Основная часть. Целью данной работы является извлечение из текста ключевых объектов, их описаний и положения в пространстве. Для решения этой задачи требуется выделить все объекты (существительные и местоимения), извлечь зависящие от них слова, и оценить значимость этих объектов для понимания смысла исходного текста.

В этой работе предлагается подход к разбору текста в указанном формате на основе правил с использованием библиотеки spaCy [2] и оценки их значимости с помощью модифицированного алгоритма EmbedRank [3]. Данный алгоритм предназначен для извлечения ключевых слов из текстов на английской языке, он использует векторное представление текста word2vec, doc2vec, text2vec [4] и библиотеку Stanford CoreNLP [5]. Мы применим его к тексту на русском языке, использовав другие векторное представления BERT [6], ELMO [7], и сравнив полученные варианты между собой. Сравнение предложенных вариантов будет производиться на собранном и размеченном вручную наборе данных из 3000 кратких содержаний книг на русском языке.

Лучший алгоритм будет использован в системе генерации изображения по текстовому описанию для эффективного поиска в базе изображений и модификации найденного изображения для лучшего соответствия исходному описанию.

Выводы. Данная технология может быть использована широким кругом людей для анализа текстов на русском языке, автоматического понимания их содержания и генерации изображения по тексту.

Список литературы.

1. Bird S., Klein E., Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. – " O'Reilly Media, Inc.", 2009.
2. Explosion A. I. spaCy-Industrial-strength Natural Language Processing in Python //URL https://spacy. io. – 2017.
3. Bennani-Smires K. et al. Simple unsupervised keyphrase extraction using sentence embeddings //arXiv preprint arXiv:1801.04470. – 2018.
4. Pagliardini M., Gupta P., Jaggi M. Unsupervised learning of sentence embeddings using compositional n-gram features //arXiv preprint arXiv:1703.02507. – 2017.
5. Manning C. D. et al. The Stanford CoreNLP natural language processing toolkit //Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. – 2014. – С. 55-60.

6. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.
7. Peters M. E. et al. Deep contextualized word representations //arXiv preprint arXiv:1802.05365. – 2018.

Шаламов В.В. (автор) Подпись

Ефимова В.А. (автор) Подпись

Фильченков А.А. (научный руководитель) Подпись