Терехова А.Д., студент, **Терехов Г.В.**, ст. преп.,
email: gvterechov@ya.ru
ВолгГТУ, г. Волгоград, Российская Федерация

СОВРЕМЕННЫЕ МЕТОДЫ АВТОМАТИЗИРОВАННОЙ РЕКОМЕНДАЦИИ ТЕГОВ ДЛЯ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

В статье рассмотрены основные современные методы глубокого обучения, применяемые для автоматизации выполнения задач по выделению ключевых особенностей текста. Для каждого метода описаны его особенности работы и в соответствии с ними сделаны выводы для какого класса задач оптимален метод.

Ключевые слова: *Тегирование текста, глубокое обучение, анализ данных, автоматизированные системы, классификация текста.*

MODERN METHODS OF AUTOMATED TAG RECOMMENDATION FOR NATURAL LANGUAGE TEXT

The article discusses the basic modern methods of deep learning that are used for automation of tasks for highlighting the key features of the text. For each method are described its operation features and in accordance with them are made the conclusions for which class of tasks the method is optimal.

Keywords: *Text tagging, deep learning, data analysis, automated systems, text classification.*

В период глобализации, непрерывно растет число активных пользователей Интернета, которые имеют возможность свободно публиковать данные, а также потреблять их со все большего количества сайтов. В связи с этим возникает проблема информационной перегрузки. Более того, недавние исследования показали, что среди других текстовых единиц, таких как заголовок, описание и комментарии пользователей, теги являются наиболее эффективными при выполнении задач, связанных с поиском и классификацией. Чтобы привести контент к более организованному по ключевым характеристикам виду, применяются различные методы автоматизированного подбора тегов.

Для решения задачи автоматизированной рекомендации тегов наиболее часто применяются системы машинного обучения. В данной работе рассмотрено четыре метода, основанных на глубоком обучении, широко используемых для решения задач по классификации текста**:** TagCNN, TagRNN, TagHAN и TagRCNN [1].

Метод TagCNN основан на CNN (convolutional neural network) - технологии, которая оказалась успешной в различных областях классификации текстов. Так как подзадачей автоматической рекомендации

тегов является классификация текста, то можно предположить, что данный метод будет иметь хорошую производительность для автоматизированной рекомендации тегов. Для CNN естественны такие задачи классификации, как анализ эмоций, обнаружение спама или категоризация тем. TagCNN может извлекать локальную семантику из различных позиций текста с помощью n-мерных фильтров, но именно фильтры фиксированных размеров становятся причиной потери семантики в длинных текстах [2].

TagRNN - метод, основанный на технологии RNN (recurrent neural network). Рекуррентные нейронные сети - одна из самых популярных архитектур, используемых при NLP (обработке естественного языка). Преимущество данного метода в улучшенной по сравнении с другими методами способности извлекать контекстную информацию, что может быть выгодным при анализе семантики длинных текстов. Алгоритм данного метода заключается в придании наибольшего веса самой поздней информации, следовательно такой подход будет не эффективным, если ключевая информация находится в начале текста [3].

Метод TagHAN, основанный на моделях HAN (Hierarchical attention networks), способен качественно отбирать формирующие слова и предложения для задач NLP. Так как данный метод основан на иерархической структуре текста, то он является не эффективным при классификации коротких текстов [4].

Наконец, рассмотрим метод TagRCNN, который основан на RCNN (recurrent convolutional neural network). RCNN комбинирует RNN и CNN архитектуры и наследует их преимущества: способность извлечения контекстных особенностей RNN и способность выделения наибольшего числа потенциальных характеристик CNN [5]. Однако, метод TagRCNN требует больше времени на обучение, чем остальные методы.

Вывод. В данной статье были рассмотрены четыре метода для автоматизированной рекомендации тегов, основанных на глубоком обучении: TagCNN, TagRNN, TagHAN и TagRCNN. Методы TagRNN и TagHAN наиболее эффективны при анализе длинных текстов, TagCNN - коротких. Отдельно стоит отметить метод TagRCNN, который способен извлекать контекстные особенности и выделять наибольшее число потенциальных характеристик по сравнению с прочими методами. Результаты проведения комплексных экспериментов показывают, что методы глубокого обучения TagCNN и TagRCNN превосходят по эффективности TagRNN и TagHAN. Сравнив CNN и RCNN подходы на задаче по классификации отношений SemEval-2010, модель RCNN в результате показала F1-меру равную 83,7%, что превосходит все вышесравниваемые подходы. Таким образом, для задач по автоматизации выделения ключевых особенностей текста наиболее подходит метод глубокого обучения TagRCNN.

Список использованных источников

1. Pingyi Zhou, Jin Liu, Xiao Liu, Zijiang Yang, John Grundy. Is deep learning better than traditional approaches in tag recommendation for software information sites?/ Pingyi Zhou, Jin Liu, Xiao Liu, Zijiang Yang, John Grundy// Information and software technology 109 (2019): 1-13.

2. Can Li, Ling Xu, Meng Yan, Yan Lei. TagDC: A tag recommendation method for software information sites with a combination of deep learning and collaborative filtering/ Can Li, Ling Xu, Meng Yan, Yan Lei// The Journal of Systems & Software 170 (2020): 9-11.

3. Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning/ Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang// IJCAI'16: Proceedings of the Twenty-Fifth International Joint Conference on Artificial IntelligenceJuly (2016): 2873–2879.

4. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, Eduard Hovy. Hierarchical Attention Networks for Document Classification/ Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, Eduard Hovy// NAACL-HLT (2016): 1480–1489.

5. Xiaobin Zhang, Fucai Chen, Ruiyang Huang. A Combination of RNN and CNN for Attention-based Relation Classification/ Xiaobin Zhang, Fucai Chen, Ruiyang Huang// Procedia Computer Science 131 (2018): 911–917.

Кулевич А.П., студент, **Сибирный Н.Д.**, студент,
Литвиненко В.В., студент,
Розалиев В.Л., к.т.н., email: yulia.orlova@gmail.com
ВолгГТУ, г. Волгоград, Российская Федерация

РАЗРАБОТКА ПРОГРАММНЫХ СРЕДСТВ АВТОМАТИЗАЦИИ ПРОЦЕССА НАБОРА ДАТАСЕТА АНГЛИЙСКОЙ ДАКТИЛЬНОЙ АЗБУКИ

Авторами статьи разработана методика набора датасета для обучения нейросети, распознающей английскую дактильную азбуку на основе данных, получаемых с инфракрасной камеры.

Ключевые слова: программное средство, автоматизация, датасет, азбука жестов, дактильная азбука, система захвата движения.

SOFTWARE TOOL FOR AUTOMATION OF THE PROCESS OF COLLECTING DATASET OF THE ENGLISH DACTYLES

The authors of the article have developed a method for collecting a dataset for training a neural network that recognizes the English dactyl alphabet based on data received from an infrared camera.

Keywords: software, automation, dataset, gesture alphabet, dactyl alphabet, motion capture system.