КЛАССИФИКАЦИЯ СТРОК ТЕКСТА ДОКУМЕНТА ПО ПРИЗНАКУ ЖИРНОСТИ НАЧЕРТАНИЯ

Д. Е. Копылов, А. А. Михайлов

В данной работе рассматривается проблема классификации строк текста документа по признаку жирности начертания. Документы подаются как растровые изображения. Входными данными является изображение страницы документа с координатами строк, полученными после обработки страницы OCR Tesseract [1].

Определение, обладает ли строка свойством жирности является важным для сегментации документа. Конкретно, система Dedoc [2] для того, чтобы пометить строку как заголовок, одним из признаков использует жирность строки.

Сложность применения для данной классификации, например, сверхточных нейронных сетей (CNN) заключается в том, что свойство жирного начертания текста неоднозначно. Текст является жирным по сравнению с другим текстом документа, при этом в другом документе такое же начертание может обозначать не жирный шрифт.

В качестве тестовых данных использовались изображения документов выпускных квалификационных работ. Вручную было размечено 30 изображений, содержащих суммарно 871 строку (из них 97 жирным начертанием). Из 30 изображений 5 являются сложным текстом (3 титульных листа и 2 листа с оглавлением), 20 обычный текст с заголовками, 5 изображений не содержат жирного шрифта.

В результате исследования хорошо себя зарекомендовали три подхода.

Первый подход заключается в подсчете количества пикселей в изображении строки, интенсивность которой меньше некоторого значения. Результат делится на количество пикселей во всей строке. Точность составила 27,36%, а полнота составила 86,6 %.

Второй подход заключается в взятие среднего значения интенсивности в строке за вычетом пробелов. Для данного подхода точность: 81,82%, полнота: 83,51%.

Третий подход состоит в том, что считается частное числа темных пикселей и «периметра» текста. Периметр заменяется числом темных пикселей, полученных при разности исходной строки и ее же, только сдвинутой на один пиксель вправо. Данный подход имеет наилучшие результаты, а именно точность: 85,29% и полноту: 89,69%.

Примечательно, что описанные выше подходы в качестве результата возвращают количественное значения. Полученные значения можно обработать статистически. В предположение, что для документов одного типа среднее количество жирных строк будет составлять конкретный процент от всех строк, можно проводить классификацию по порогу, полученному в попытке удовлетворить заданному соотношению.

В дальнейшем планируется использовать архитектуру нейронных сетей трансформер [3], которая лишена недостатка, описанного выше для CNN. В качестве входных данных также могут подаваться значения, полученные с помощью описанных выше подходов.

Литература

1. Tesseract Open Source OCR Engine (main repository [Электронный ресурс] // GitHub : [сайт]. URL: https://github.com/tesseract-ocr/tesseract (дата обращения: 01.05.2023).

2. Dedoc: система извлечения содержимого и структуры текстовых документов [Электронный ресурс] // Институт системного программирования им. В.П. Иванникова РАН : [сайт]. URL: https://www.ispras.ru/technologies/dedoc/ (дата обращения: 01.05.2023).

3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention Is All You Need. // Advances in neural information processing systems. 2017. Vol. 30. URL: https://arxiv.org/abs/1706.03762v5 (дата обращения: 01.05.2023).