УДК 004.89

К ВОПРОСУ ОБ ИСПОЛЬЗОВАНИИ СИСТЕМ МНОГОЯЗЫЧНОГО РАСПОЗНАВАНИЯ ТЕКСТА ДЛЯ РАСШИРЕННОГО КИРИЛЛИЧЕСКОГО АЛФАВИТА

А. А. Голубничий, А. Д. Яблонцева

Хакасский государственный университет им. Н. Ф. Катанова,
пр. Ленина, 92/1, 655017, г. Абакан, Россия, artem@golubnichij.ru

В статье раскрывается проблема использования систем многоязычного распознавания текста для языков тюркской группы, построенных на расширенном кириллическом алфавите. Дается оценка точности работы системы на трех языках: хакасском, тувинском и тофаларском.

Ключевые слова: оптическое распознавание символов, малые языки, LSTM, нейронные сети.

TO THE QUESTION ABOUT THE USE OF MULTILINGUAL TEXT RECOGNITION SYSTEMS FOR THE EXTENDED CYRILLIC ALPHABET

A. A. Golubnichiy, A. D. Yablontseva

Katanov Khakass State University, ave. Lenin, 92/1, 655017, Abakan, Russia, artem@golubnichij.ru

The article reveals the problem of using multilingual text recognition systems for the languages of the Turkic group, built on the extended Cyrillic alphabet. An assessment of the accuracy of the system in three languages is given: Khakass, Tuvan and Tofalar.

Keywords: optical character recognition, minor languages, LSTM, neural networks.

Многоязычное распознавание текста (MOCR) представляет интерес по многим причинам: оцифровка исторических книг, содержащих два или более языков, двуязычные книги, словари и книги. Однако реализация систем, построенных по типу MOCR, несет в себя ряд сложностей: использование языков с одинаковыми или похожими шрифтами (например, арабско-персидский, англо-немецкий и др.); один и тот же алфавит в нескольких языках, например, урду в насталик и насх; архаичные и реформированные орфографии, например, XVIII века (фрактур – исторический немецкий и др.).

Одним из наиболее востребованных подходов в решении задачи построения систем по типу MOCR является использование нейронных сетей типа LSTM, реализованных в Tesseract версии 4 и выше [1].

Масштаб и относительное положение символов важны, особенно для различия символов в тюркском алфавите, построенном на основе кириллицы. Также нормализация текстовой строки – важный шаг в применении сетей LSTM к OCR. Словарь токенов, созданный из набора связки текстовых строк, содержит информацию о х-высоте, исходной линии (геометрические элементы) и форме отдельных символов. Затем эти модели используются для нормализации любой текстовой строки.

Рекуррентные нейронные сети (RNN) показали преимущества благодаря архитектуре LSTM. Архитектура LSTM значительно отличается от более ранних архитектур, таких как сети Элмана и сети с эхосигналом. Традиционные RNN, хотя и хорошо справляются с контекстно-зависимой обработкой, не показали конкурентоспособной производительности для OCR, что связано с проблемой исчезающего градиента. Архитектура Long Short Term Memory была разработана для преодоления данной проблемы.

Для распознавания в рамках исследования использована одномерная двунаправленная архитектура LSTM. Обнаружено, что одномерная архитектура превосходит своих двухмерных или более высокоразмерных собратьев.

В работе использовалась модифицированная версия библиотеки LSTM. Эта библиотека предоставляет одномерные и многомерные сети LSTM, а также выравнивание по основанию с использованием прямого-обратного алгоритма. Библиотека также предоставляет механизм эвристического декодирования для отображения покадрового сетевого вывода на последовательность символов.

На этапе обучения случайно выбранные входные изображения текстовых строк представляются в виде одномерных последовательностей для шага прямого распространения через ячейки LSTM, а затем выполняется прямое-обратное выравнивание выходных данных, далее – обратное распространение для обновления весов, и весь процесс повторяется для следующего случайно выбранного изображения текстовой строки.

Во время тестирования модель LSTM обучена тувинскому языку на хакасском и модель, обученная тофаларскому на тувинском и хакасском, специальные символы были исключены из результатов распознавания. Это позволило правильно оценить влияние неиспользования языковой модели. Если эти слова не были удалены, то результирующая ошибка также будет содержать часть ошибок из-за неправильного распознавания символов.

Итак, удалив эти слова со специальными символами, истинная производительность LSTM-сети обучена языку, содержащему малые алфавиты на язык, содержащий больше алфавитов. Следует отметить, что эти результаты были получены без учета помощи на любом этапе постобработки, например, в языковом моделировании, использовании словарей для исправления ошибок распознавания текста и т. д.

Система Tesseract достигла высоких показателей по сравнению с моделями на основе LSTM. Модель Tesseract для тувинского языка принесла 1,33 %, 5,02 %, 5,09 % и 4,82 % узнаваемости. Ошибки применены к тувинскому, хакасскому, тофаларскому языкам и смешанным данным.

Результаты показывают, что отсутствие языкового моделирования или применение различных языковых моделей влияет на точность распознавания. Поскольку модель для смешанных данных недоступна для Tesseract, влияние оценки такой модели на отдельных языках не может быть вычислено.

Библиографический список

1. Яблонцева А. Д. Обзор технологии tesseract 4.0 и типичные проблемы распознавания текстов // Modern Science. 2021. № 7. С. 392–394.