Распознавание текстов рукописных и старопечатных книг на основе нейросетевых технологий

С. И. Корниенко, Ф. М. Черепанов, Л. Н. Ясницкий
Пермский государственный университет, Пермь, Россия

Среди проблем перевода исторических источников в машиночитаемый формат к наиболее сложным и трудоемким относится создание версий рукописных и старопечатных книг в формате электронного текста. Чтобы получить полноценный электронный текст, как правило, используется его ручной компьютерный набор с помощью специальных редакторов. В значительной мере указанные трудности связаны с тем, что сканирование с последующим распознаванием такого рода текстов не дает удовлетворительных результатов.

Одной из причин этого является нестандартность и разнообразие шрифтов источников-оригиналов, наличие надстрочных знаков, наконец, просто погрешности написания, «затертости», помарки и иные дефекты, затрудняющие распознавание символов.

Представляется, что решение указанной задачи может лежать в плоскости создания специализированных программных продуктов на основе наиболее перспективных видов современных информационных технологий. В настоящее время к таковым относятся технологии, при создании которых используются приемы и методы искусственного интеллекта.

С целью использования технологий искусственного интеллекта для решения проблем перевода в машиночитаемый формат текстов рукописных и старопечатных книг сотрудники Лаборатории исторической и политической информатики историко-политологического факультета и кафедры прикладной математики и информатики Пермского университета начали совместные работы, конечной целью которых может стать создание специализированного программного комплекса.

Пермской научной школой искусственного интеллекта [Пермская 2008] накоплен значительный опыт разработки и внедрения интеллектуальных информационных систем, среди которых немалое место занимают системы, предназначенные для распознавания сложных в техническом отношении

текстов (нечетких, загрязненных, с множеством помарок, написанных с использованием разных шрифтов), таких, как автомобильные номерные знаки, подписи технической документации. В основе таких систем, как правило, лежат нейросетевые технологии, дополненные алгоритмами предварительной обработки изображений (контрастирование, бинаризация, удаление посторонних фрагментов и др.) и средствами дораспознавания, включающими орфографический, синтаксический и семантический виды анализа.

В настоящее время с целью апробации идеи применения методов искусственного интеллекта для распознавания символов текстов рукописных и старопечатных книг создан демонстрационный прототип, в основе которого заложен персептрон слоистой структуры с сигмоидными активационными функциями. В качестве обучающего множества были использованы шрифты старославянской письменности. После обучения на вход персептрона подавались символы из соответствующего текста, причем некоторые из них были искажены помехами, а также выполнены шрифтом, отличающимся от того, который был использован в обучающем множестве.

Как показали эксперименты, нейронная сеть успешно распознавала до 80 % символов древнего текста.

Полученные результаты дают основание говорить о перспективности использования подобного нейросетевого модуля для создания системы распознавания текстов рукописных и старопечатных книг. Существуют реальные возможности повысить процент правильно распознаваемых образов, при этом не только отдельных символов, но и их сочетаний, в том числе с надстрочными знаками, за счет подбора наиболее оптимальных парадигм нейронных сетей, а также дополнения их специализированными алгоритмами пред- и постобработки изображений. В дальнейшем предполагается автоматизировать демонстрационный прототип, доведя его до уровня коммерческой программы.

Список литературы

Пермская 2008 – *Пермская научная школа искусственного интеллекта и ее инновационные проекты* / Л. Н. Ясницкий, В. В. Бондарь, С. Н. Бурдин и др. ; под ред. Л. Н. Ясницкого. – Москва-Ижевск : НИЦ «Регулярная и

хаотическая динамика», 2008. – 75 с.

OCR of manuscripts and early printed books using neural networks

Sergey I. Kornienko, F. M. Cherepanov, Leonid N. Yasnitsky
Perm State University, Perm, Russia

This paper describes the possibilities for using artificial intelligence technologies, in particular the neural networks, for recognition of handwritten and early printed texts. Early results using the technology are described.