ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА РАСПОЗНАВАНИЯ РУКОПИСНЫХ И СТАРОПЕЧАТНЫХ ТЕКСТОВ ИСТОРИЧЕСКИХ ИСТОЧНИКОВ

Айдаров Ю.Р., Волгирева Г.П., Гагарина Д.А., Корниенко С.И., Черепанов Ф.М., Ясницкий Л.Н.

В настоящее время в нашей стране и за рубежом активно ведутся работы, связанные с переводом рукописных и печатных литературных, исторических, технических и другого рода текстов в электронный формат. Актуальность этих работ обусловлена, с одной стороны – необходимостью сохранения историко-культурного и технологического наследия, а с другой – возможностью создания полнотекстовых информационных систем, позволяющих применять современные методы информационного поиска и научного анализа материалов.

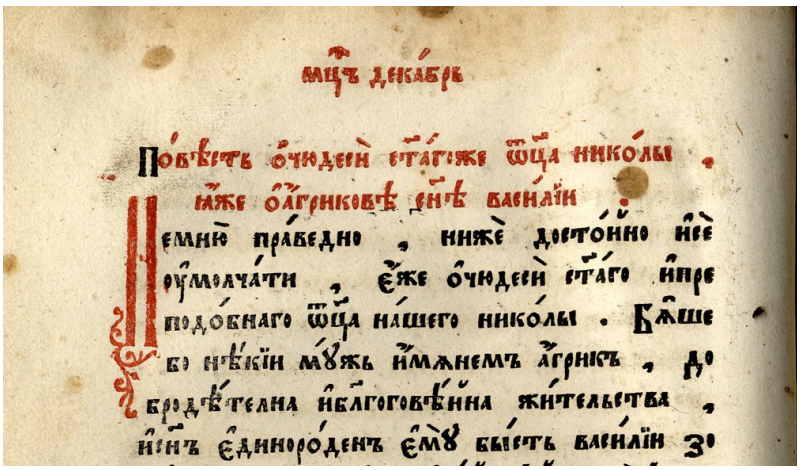

Рис.1. Пример фрагмента старопечатного текста

Среди проблем создания подобных продуктов актуальным является распознавание текстовой информации и представление ее в электронном формате. Традиционно для этих целей используется технология OCR (Optical Character Recognition), включающая шаблонный, структурный и признаковый методы [1], а также их модификации в виде фонтанного преобразования, реализованного в FineReader и самообучающихся алгоритмов CuneiForm. Однако попытки применения этих традиционных методов для распознавания текстов рукописных и старопечатных книг не приносят желаемых результатов. Как видно из примера, приведенного на рис.1, тексты старинных книг весьма сложны для распознавания в техническом отношении. Они написаны на разных диалектах, с использованием разных шрифтов, имеют множество помарок и потертостей.

По-нашему мнению, основанному на опыте Пермской научной школы искусственного интеллекта [2], решение задачи распознавания текстов рукописных и старопечатных книг следует искать в области использования перспективных видов современных информационных технологий, в частности основанных на приемах и методах искусственного интеллекта и параллельных вычислений.

В силу междисциплинарного характера представляемый нами проект[1] осуществляется коллективом исполнителей, в который входят сотрудники историко-политологического и механико-математического факультетов Пермского государственного университета – специалисты в области исторической информатики, источниковедения, палеографии и древнерусского языка, программирования, создания информационных систем, а также искусственного интеллекта.

Ранее нами [4] была предпринята попытка создания демонстрационного прототипа будущей OCR-системы на базе персептрона с сигмоидными активационными функциями. На входное табло, разбитое на $8 \times 10 = 80$ клеточек, поочередно наносились изображения букв старославянской письменности (рис.2, слева), из которых составлялось множество обучающих примеров (рис.2, внизу). Соответственно 80 сигналов, получаемых с каждой буквы, подавались на вход персептрона, обучавшегося методом обратного распространения ошибки. При тестировании нейросети на табло накладывались буквы, отличающиеся от букв обучающего множества помарками, потертостями и небольшими искажениями. Результаты тестирования показали до 80% правильных ответов, что свидетельствует о перспективности выбранного инструментария – нейронной сети персептронного типа.

В настоящее время в Пермском госуниверситете проводятся работы по созданию массива рукописных и старопечатных текстов в электронном формате и его кластеризация для выявления типов текстов и шрифтов. Предварительные исследования подтверждают мнение о том, что должны создаваться сразу несколько нейронных сетей, ориентированных на конкретные исторические периоды, в которых были написаны книги. Они должны быть ориентированы также на конкретные российские регионы, отличающиеся диалектами, а результаты нейросетевого распознавания должны корректироваться классическими методами дораспознавания, включающими морфологический, синтаксический, семантический и прагматический виды анализа. Естественно что базы знаний для такого анализа должны создаваться с учетом диалектов, соответствующих пространственным и временным секторам.

[1] Проект поддержан грантом РФФИ № 09-06-00254

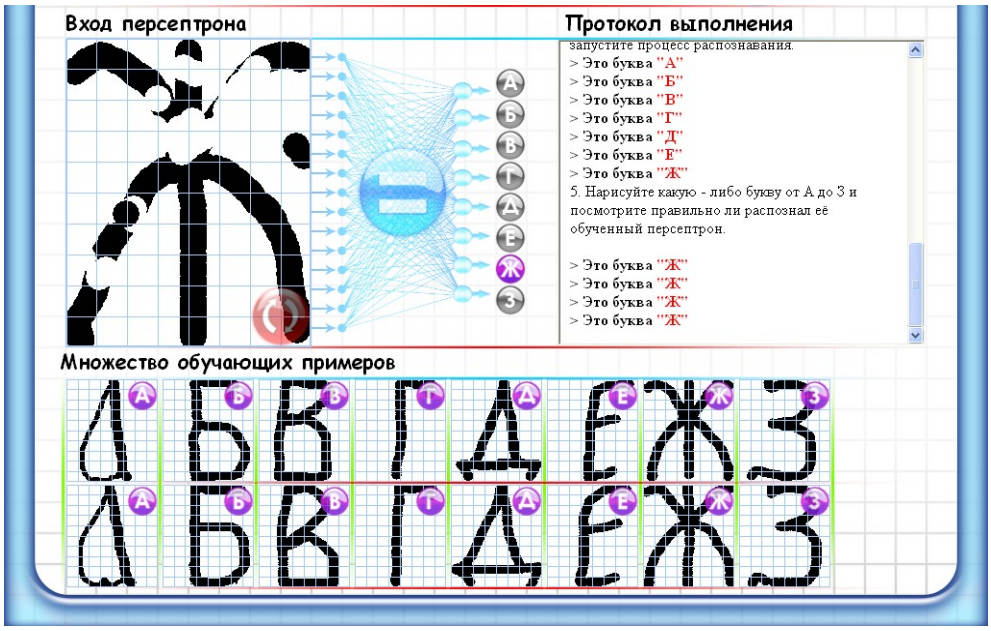

Рис. 2. Рабочее окно демонстрационного прототипа интеллектуальной системы распознавания старопечатных текстов

Литература

1. Ясницкий Л.Н. Введение в искусственный интеллект / Л.Н.Ясницкий. – Издание 3. – М.: Издательский центр «Академия», 2010. – 176с.
2. Ясницкий Л.Н. 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. – 2-е изд. – Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2008. – 75 с
3. Черепанов Ф.М. Симулятор нейронных сетей «Нейросимулятор 1.0». / Ф.М.Черепанов, Л.Н.Ясницкий // Свидетельство об отраслевой регистрации разработки №8756. Зарегистрировано в Отраслевом фонде алгоритмов и программ 12.07.2007.
4. Корниенко С.И. Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам / С.И.Корниенко, Ф.М.Черепанов, Л.Н.Ясницкий : Материалы Междунар. науч. конф. (Казань, 26-30 августа 2008 г.) / отв. ред. В.Д. Соловьев, В.А. Баранов.