Анна Юрьевна Филиппович

ОСОБЕННОСТИ РАСПОЗНАВАНИЯ ПЕЧАТНЫХ ИСТОЧНИКОВ XVIII – нач. XIX вв.[*]

Описана проблема поиска информации в электронных хранилищах исторических документов по их содержанию. Предложено решение этой проблемы, основанное на выделении группы источников с определенными свойствами (способ печати, средства оформления, шрифты и т.п.) и обработки их OCR-системами и системами индексирования текстов. Описаны проблемы распознавания печатных источников XVIII – нач. XIX вв.: ориентация на старинную лексику и использование старинных шрифтов. Предложено решение путем создания лексического ядра языка XVIII – нач. XIX вв. на базе Словаря Академии Российской 1789-1794 гг.

Сегодня идея переноса исторических документов на новые носители информации получила свое воплощение в создании электронных ресурсов на базе современных библиотек и в Интернет. Фактически, осуществляется факсимильное копирование исторических памятников письменности, что подразумевает хранение их постранично в виде набора цифровых изображений. Копии, полученные таким образом, сопровождаются только библиографическими и археографическими описаниями. Основным недостатком такого типа описаний является неполное и ограниченное раскрытие содержания документа. В итоге, оцифровка источников факсимильным способом принципиально не изменяет способ доступа к информации. По-прежнему исследователь должен просматривать значительное количество источников для поиска нужной информации, последовательно "листая" их. Таким образом, решение проблемы поиска информации в созданных электронных хранилищах документов по их содержанию является актуальной задачей современности.

Решение этой проблемы может быть основано на выделении группы источников с определенными свойствами (в числе которых способ печати, используемые средства оформления, шрифтовые гарнитуры и т.п.) и обработки их автоматизированными системами распознавания (OCR-системами) и индексирования текстов. Рассмотрим особенности использования первых.

Ряд ограничений не позволяет эффективно использовать современные OCR-системы для распознавания старинных печатных источников, среди которых ориентация на

[*] Работа выполняется при поддержке Гранта Президента РФ МК-3732.2010.9

современную лексику и определенную группу шрифтов. Так, при оценке эффективности распознавания и корректуры текстов XVIII века и большинство ошибок было связано с особенностями графем шрифта и старинной лексикой [Исследование 2005: с. 281; Исследование 2007: с. 110]. Качество печати и шрифтовое оформление источников, изданных ранее второй половины XVIII века, делает это фактически невозможным и требует создания специализированных систем распознавания. Совершенствование печатного процесса, использование ограниченного набора шрифтов в конце XVIII века является предпосылкой для создания технологии распознавания источников этого периода времени, основанной на современных OCR-системах.

Основным компонентом таких систем является лингвистическая БД, ядро которой – лексическая система языка. Согласно статистическим исследованиям текстов XVIII века [Словарь 2008: с. 108] около 40% слов принадлежат к старинной лексике. Данные слова не будут эффективно распознаны из-за отсутствия их в лингвистической БД OCR-системы. Таким образом, для решения задачи распознавания источников необходимо пополнить лексический состав системы распознавания документов. Фактически речь идет о создании лексического ядра языка XVIII – нач. XIX вв.

В качестве основы ядра предполагается использовать Словарь Академии Российской 1789-1794 гг. (САР), содержащий более 200000 лексических единиц, который в 2001-2005 гг. был переиздан с использованием современных информационных технологий. В течение 2006-2008 гг. в рамках проекта РГНФ "Интегрированная инструментальная информационно-программная среда для автоматизации исследований САР" было создано электронное издание САР, содержащее БД объемом более 44 тысяч структурных единиц.

Для оценки эффективности распознавания источников XVIII – нач. XIX вв. OCR-системой необходимо провести сопоставительные исследования лексики коллекции документов рассматриваемого периода и созданного лексического ядра.

Результаты исследований могут быть использованы для переиздания значительного массива конкретных источников XVIII – нач. XIX вв., разработки систем распознавания исторических текстов, решения практических задач электронного и полиграфического издания древних памятников.

Литература

Филиппович А.Ю. Словарь Академии Российской (1789–1794): информационная технология переиздания. Вступительная статья М.И.Чернышевой. - М.: МГУП, 2008.– 304 с.

Филиппович А.Ю. Исследование эффективности автоматизации корректурных процессов с помощью словаря спеллера при подготовке переиздания Словаря Академии Российской 1789–1794 гг. // Проблемы полиграфии и издательского дела. № 4. – М.: Изд-во МГУП, 2007. – С. 102-112.

Филиппович Анна. Исследование эффективности систем оптического распознавания текстов. // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 7 / Сост. и ред. Ю.Н. Филипповича. – М. Изд-во ООО «Эликс+» 2005. – С. 272-297.