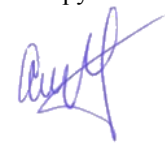
На правах рукописи



Смирнов Сергей Владимирович

ТЕХНОЛОГИЯ И СИСТЕМА АВТОМАТИЧЕСКОЙ КОРРЕКТИРОВКИ РЕЗУЛЬТАТОВ ПРИ РАСПОЗНАВАНИИ АРХИВНЫХ ДОКУМЕНТОВ

Специальность 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2015

Работа выполнена в Федеральном государственном бюджетном учреждении науки Санкт-Петербургском институте информатики и автоматизации Российской академии наук (СПИИРАН).

Научный руководитель: **Кулешов Сергей Викторович**, доктор технических наук

Официальные оппоненты: **Городецкий Андрей Емельянович**, доктор технических наук, заслуженный деятель науки РФ, профессор, заведующий лабораторией "Интеллектуальных электромеханических систем", Федеральное государственное бюджетное учреждение науки Институт проблем машиноведения Российской академии наук (ИПМаш РАН)

Пиотровская Ксения Раймондовна, кандидат технических наук, профессор кафедры методики обучения математике и информатике, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Российский государственный педагогический университет им. А.И. Герцена» (РГПУ им. А.И. Герцена)

Ведущая организация: **Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики» (Университет ИТМО)**

Защита состоится «____» _____2015 г. в ____ часов на заседании диссертационного совета Д.002.199.01 при Федеральном государственном бюджетном учреждении науки Санкт-Петербургском институте информатики и автоматизации Российской академии наук по адресу: 199178, Санкт-Петербург, В.О., 14 линия, 39.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук

Автореферат разослан «____» _____2015 г.

Ученый секретарь
диссертационного совета Д.002.199.01 Фаткиева Роза
кандидат технических наук, доцент Равильевна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертации. Достоверность результатов оптического распознавания сильно зависит от качества исходного изображения, лексикона, используемого при написании текста, особенностей шрифтов, наличия сторонних объектов, шумов и многих других факторов. Высокая точность достигается в случае распознавания изображений, где текст размещен на монотонно ровном фоне с хорошей контрастностью; тезаурус, используемый при написании текста, соответствует встроенному словарю системы распознавания и не содержит редких слов и словоформ; начертание букв и слов позволяет однозначно произвести сопоставление с шаблоном.

Существующие коммерческие системы распознавания текста («Abbyy Finereader», «Nuance OmniPage» и др.), а также системы с открытыми исходными кодами («Cuneiform», «Tesseract» и др.) достигают высокой точности результатов при обработке современных качественных печатных документов. В случае же распознавания архивных документов, происхождение которых датируется десятками лет назад, количество допущенных ошибок в результатах распознавания значительно возрастает и эффективность применения средств автоматизации снижается. Результаты, получаемые на выходе систем распознавания необходимо подвергать последующей корректировке.

Методы автоматической корректировки ошибок распознавания во многом основываются на адаптации известных подходов корректировки орфографических ошибок, использующих скрытые марковские модели, нейронные сети, n-граммы слов и символов, конечные автоматы. Также применяются методы, объединяющие результаты нескольких систем распознавания, использующие дополнительную информацию о контексте и эвристические алгоритмы. Большой вклад в теорию и практику корректировки ошибок в текстах внесли Philips L., Brill E., Kolak O., Mays E., Fossati D., Kukich K., Reynaert M. и другие зарубежные ученые. Среди отечественных авторов в области автоматической обработки текстов можно выделить труды Арлазарова В.Л., Шоломова Д.Л., Постникова В.В., Захарова В.П. и других.

Во многих случаях существующие методы требуют привлечения ручного труда, предназначены для обработки современных текстов и не пригодны для обработки результатов распознавания архивных документов, отличающихся обилием узкоспециализированных терминов и нестабильным уровнем качества.

Решению описанных проблем и разработке системы распознавания архивных документов с применением методов автоматической корректировки и посвящена данная диссертационная работа.

Объектом исследования является процесс распознавания архивных документов.

Предметом исследования являются методы и технология автоматической корректировки результатов распознавания архивных документов.

Цель работы и задачи исследования. Основной целью диссертационной работы является разработка технологии и системы распознавания архивных документов с автоматическим обнаружением и корректировкой допущенных

ошибок.

Для достижения поставленной цели в диссертационной работе поставлены и решены следующие задачи:

1. Сравнение качества существующих систем оптического распознавания, классификация основных видов допускаемых ошибок и анализ существующих подходов к корректировке ошибок распознавания.
2. Разработка метода автоматической корректировки результатов распознавания архивных документов, выполняющего поиск ошибок и генерацию упорядоченного по рангу списка корректировок для их замены.
3. Разработка технологии распознавания архивных документов различных тематических областей и корректировки полученных результатов.
4. Проектирование, разработка и апробация системы распознавания документов архивного фонда, отвечающей требованиям разработанной технологии и реализующей предложенный в работе метод корректировки.

Методы исследования. Для решения поставленных задач в работе используются методы теории множеств, теории вероятности, статистического анализа, корпусной и компьютерной лингвистики. Реализация разработанных алгоритмов произведена в соответствии с объектно-ориентированной методологией разработки программного обеспечения.

Положения, выносимые на защиту. На основе проведенных теоретических работ и их экспериментальной апробации на защиту выносятся следующие положения:

1. Метод автоматической корректировки ошибок распознавания архивных документов на основе рейтинго-ранговой модели текста.
2. Правила ранжирования и выбора наилучших корректировок, основанные на частотных характеристиках и статистической вероятности сочетаемости с предшествующими словами.
3. Технология распознавания архивных документов с последующей корректировкой результатов.
4. Архитектура и компонентная модель системы распознавания и автоматической корректировки результатов, с входящим в ее состав инструментарием настройки конфигурации для обработки архивных документов различных тематических областей.

Научная новизна работы состоит в следующем:

1. Разработан метод автоматической корректировки ошибок распознавания архивных документов на основе рейтинго-ранговой модели текста, основной особенностью которого является способность выявлять и устранять ошибки распознавания документов, содержащих большое количество узкоспециализированной терминологии, за счет автоматического формирования тезаурусов без необходимости предварительного обучения.
2. Разработаны правила ранжирования и выбора наилучших корректировок, основанные на предварительно проведенном n-грамм анализе корпуса результатов распознавания и тематических текстов и учитывающие

статистическую вероятность сочетаемости с предшествующими словами.

3. Разработан инструментарий, позволяющий эксперту ограничивать пространство конфигураций процесса обработки архивных документов для повышения качества распознавания.
4. Разработаны технология и система распознавания архивных документов и автоматической корректировки результатов, позволяющие производить потоковую обработку больших наборов документов с учетом лексикона и специфики их предметной области.

Обоснованность и достоверность научных положений обеспечены аналитическим обзором исследований и разработок в данной области, подтверждаются положительными итогами практического использования результатов диссертации, а также апробацией основных научно-практических положений в печатных трудах и докладах на всероссийских и международных конференциях.

Практическая ценность работы заключается в создании программной системы, реализующей теоретические результаты работы, которая может использоваться в проектах массовой оцифровки и распознавания документов фондов государственных архивов, библиотек, музеев, судов, ЗАГС и других учреждений.

Разработанная в диссертационной работе технология и система автоматического распознавания и корректировки результатов позволяет значительно повысить скорость обработки документов и сократить потребность трудоемкой дорогостоящей ручной работы.

Предложенные в диссертационной работе подходы, методы и алгоритмы автоматического обнаружения и корректировки ошибок оптического распознавания позволяют значительно повысить качество конечных результатов.

Реализация результатов работы. Представленные в работе методы и алгоритмы были реализованы на языке программирования Java в виде программных модулей системы оптического распознавания текста и введены в эксплуатацию в составе государственной информационной системы «Государственные архивы Санкт-Петербурга» (государственный контракт №0172200006113000229_146076 от 24.12.2013)

Апробация результатов работы. Основные положения и результаты диссертационной работы представлялись на конференциях: I Всероссийская электронная научно-практическая конференция-форум молодых ученых и специалистов «Современная российская наука глазами молодых исследователей - 2011»; IV Всероссийская научно-практическая конференция "Научное творчество XXI века" с международным участием (Красноярск, 2011); XVI Международная научно-практическая конференция «Перспективы развития информационных технологий» (Новосибирск, 2013); XXI Международная научно-практическая конференция «Перспективы развития информационных технологий» (Новосибирск, 2014); XIV Санкт-Петербургская международная конференция «Региональная информатика (РИ-2014)» (Санкт-Петербург, 2014); X Всероссийская научно-практическая конференция «Электронные ресурсы

библиотек, музеев, архивов» (Санкт-Петербург, 2014); XVII Всероссийская объединенная научная конференция «Интернет и современное общество» (Санкт-Петербург, 2014).

Разработанное программное обеспечение было апробировано на документах фондов центральных государственных архивов Санкт-Петербурга в составе государственной информационной системы «Государственные архивы Санкт-Петербурга», свидетельство о регистрации информационной системы в Реестре государственных информационных систем Санкт-Петербурга №2053/14/08 подписано 21.11.2014 г.

Публикации. Основные результаты по материалам диссертационной работы опубликованы в 13 печатных работах, среди них 6 работ в рецензируемых изданиях из перечня ВАК, получено 2 свидетельства о государственной регистрации программы для ЭВМ.

Структура и объем работы. Диссертационная работа включает введение, четыре главы, заключение, список использованных источников (122 наименования) и три приложения. Объем работы – 130 страниц машинописного текста, включая 34 рисунка и 16 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована важность и актуальность темы диссертации, сформулированы цели диссертационной работы и решаемые задачи, определяется научная новизна работы, а также ее практическая значимость, приводится краткое содержание работы по главам.

В первой главе приводится аналитический обзор предметной области и существующих систем оптического распознавания, определяется степень их пригодности к распознаванию архивных документов, выявляется необходимость корректировки допускаемых ошибок распознавания, приводится классификация ошибок по видам и анализ существующих подходов к их корректировке, уточняются требования к разрабатываемой системе.

Сфера деятельности государственных архивов включает в себя широкий спектр задач, связанных с комплектованием, учетом, использованием и обеспечением сохранности документов. Эффективность выполнения каждой задачи архива имеет сильную зависимость от скорости нахождения и получения доступа к нужным документам. Поиск документов является своего рода «узким» местом во всех рабочих процессах архива (рисунок 1) и накладывает серьезное ограничение на скорость выполнения ежедневных задач.

Снижение влияния данного ограничения может быть достигнуто за счет автоматизации процессов пополнения поисковой базы и развития поисковых механизмов, использующихся в архивах.

Разрабатываемая в данной работе система пакетного распознавания архивных документов, является тем самым инструментом, с помощью которого возможно существенно увеличить скорость пополнения и объем поисковой базы, путем добавления в нее результатов распознавания, удовлетворяющих критериям качества. Причем, для достижения поставленной задачи при

распознавании не требуется построения полной электронной копии документа. Пользователю результаты поиска будут отображаться в виде подсвеченных областей текста на электронном образе документа.

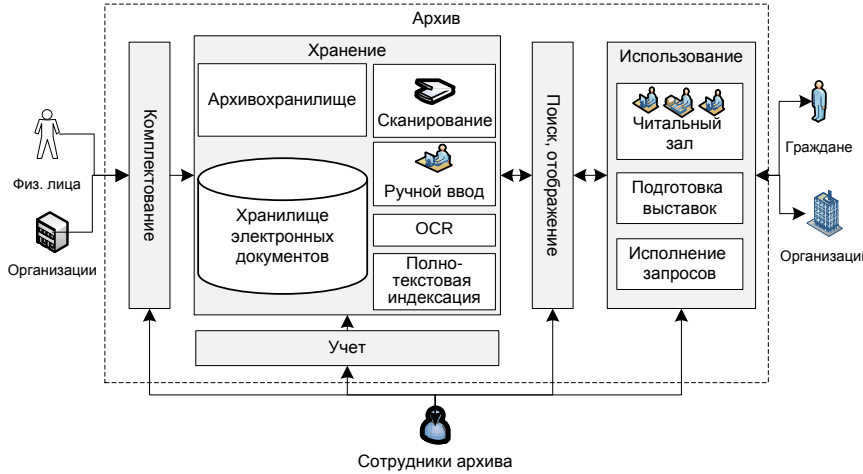


Рис. 1. Схема типовых рабочих процессов государственного архива

На сегодняшний день существует несколько десятков коммерческих и свободно распространяемых систем оптического распознавания. Сравнительный анализ выявил наличие ошибок в результатах распознавания архивных документов различного качества среди всех испытуемых систем: “Abbyy Finereader”, “Cuneiform Linux”, “Cuneiform Windows”, “IRIS Readiris”, “Nuance OmniPage”, “Tesseract”.

В обзоре методов и работ по корректировке ошибок, вначале отдельно рассматривается класс методов, относящихся к обработке орфографических ошибок, поскольку эта тема является более подробно исследованной. После дается обзор методов, систем и работ по корректировке непосредственно ошибок оптического распознавания. Особое внимание уделяется возможности применения существующих методов для построения систем корректировки архивных документов без участия человека.

Существующие методы в общем случае неплохо решают ряд задач по обработке результатов распознавания с использованием словарей, статистических моделей языка, хорошо развита тематика обнаружения и коррекции ошибок в тексте. Тем не менее, во многих случаях рассмотренные методы предназначены для обработки современных текстов и не подходят в чистом виде для обработки исторических текстов, содержащих большое количество специализированных терминов, имен собственных, географических наименований и т.п. В большинстве работ корректировка основана на предварительном ручном обучении системы или участии человека на этапе финального выбора корректировки. Также стоит отметить очень малое

количество работ нацеленных на корректировку именно русскоязычных текстов. Это вызывает потребность разработки алгоритмов корректировки, учитывающих особенности русского языка и позволяющие обрабатывать корпуса текстов больших объемов в полностью автоматическом режиме.

Из рассмотренных методов в данной работе будет использоваться алгоритм нахождения минимального расстояния между словами (расстояние Левенштейна) и алгоритм поиска схожих слов методом анаграмм, предложенный Мартином Рейнартом. Выбранные алгоритмы позволяют обрабатывать ошибки типичные для систем оптического распознавания, не требуют проведения предварительного обучения и могут применяться для обработки текстов независимо от языка написания.

Во второй главе содержится описание используемых методов и разработанного метода автоматической корректировки ошибок распознавания на основе рейтинго-ранговой модели текста.

Разделим весь процесс корректировки результатов распознавания на четыре основных этапа (рисунок 2).

Этап 1. Подготовка структур данных. На первом шаге необходимо произвести анализ всего корпуса распознанных документов и тематических текстов для формирования статистической информации по встречающимся словам. Корректная работа метода будет достигнута, в том случае если результаты распознавания и другие тексты будут принадлежать одной тематике.

Рис. 2 Этапы корректировки

Предварительная обработка. Назовем лексемой последовательность символов, разделенных пробелом или символами $\{,.:;()\backslash"\&[]?!?'\{\}/+\#=<>\%\}$, либо определенных системой распознавания как слова.

Выразим весь набор лексем, полученных в результате распознавания документов, в виде упорядоченного по порядку следования элементов множества $L^{source} = \{s_1, s_2, \ldots, s_m\}$. Преобразуем последовательность L^{source} в нормализованную последовательность $L = \{s_1, s_2, \ldots, s_n\}$ путем проведения ряда операций по очистке лексем, шаблонной замене символов и объединению лексем, разделенных знаком переноса.

Структуры для отбора корректировок. Сформируем множество лексем в нижнем регистре символов L^{low} и его рейтинговое распределение $\xi_{L^{low}}$:

$$L^{low} = \{lower(s) \mid s \in L\},\ \xi_{L^{low}} = \{< s,\ fr > \mid s \in L^{low}\},\ \ fr \geq 1,$$

где $lower(s)$ — функция перевода строки в нижний регистр; fr — частота повторения лексемы s во множестве L^{low}.

Предполагая, что наиболее часто встречающиеся лексемы с наименьшей вероятностью содержат ошибки, сформируем сокращенное множество $L^{lowpruned}$ и его рейтинговое распределение $\xi_{L^{lowpruned}}$:

$$L^{lowpruned} = \{s \mid s \in L^{low}, \xi_{L^{low}}(s) \geq \alpha\},$$

где α — минимальное пороговое количество повторений одной лексемы, $\xi_{L^{low}}(s)$ — частота повторения лексемы s во множестве L^{low}.

Сформируем множество биграмм L^{bigram}, сокращенное множество биграмм $L^{bipruned}$ и их рейтинговые распределения $\xi_{L^{bigram}}$ и $\xi_{L^{bipruned}}$:

$$L^{bigram} = \left\{\left(lower(s_1), lower(s_2)\right) \mid s_1, s_2 \in L; seq(s_1, s_2) \vee seq(s_2, s_1) = 1\right\},$$
$$L^{bipruned} = \left\{(s_1, s_2) \mid (s_1, s_2) \in L^{bigram}; \mid s_1 \mid > 1; \mid s_2 \mid > 1; \xi_{L^{bigram}}(s_1, s_2) \geq \beta\right\},$$

где функция $seq(a_1, \ldots, a_z)$ возвращает значение 1, если элементы a_1 - a_z следуют строго друг за другом, и 0 в противном случае; $|s|$ — длина строки s в символах; β — минимальное пороговое количество повторений одной биграммы.

Сформируем множество лексем для выбора корректировок L^{corr}, рейтинговое распределение $\xi_{L^{corr}}$ и хэш-таблицу анаграмм $H^{anagram}$:

$$L^{corr} = L^{lowpruned} \cup \left\{concat\ (concat\ (s_1, '\ '), s_2) \mid (s_1, s_2) \in L^{bipruned}\right\},$$
$$H^{anagram} = \left\{\left\langle hash(s), \left\langle s, \xi_{L^{corr}}(s)\right\rangle\right\rangle \mid s \in L^{corr}\right\},$$

где функция $concat\ (a_1, a_2)$ возвращает результат конкатенации строк a_1, a_2; хэш-функция $hash(s)$ возвращает значение одинаковое для слов-анаграмм и вычисляется для каждого элемента множества L^{corr} при добавлении в хэш-таблицу $H^{anagram}$.

Структуры для ранжирования корректировок. Произведем нормализацию морфологической формы каждой лексемы множества L, не входящей в список стоп слов D^{stop}, используя функцию морфологического анализа $morph(s)$:

$$morph(s) = \sum_b,\quad b \in \sum_b,\quad s \in \sum_s \rightarrow b \in \sum_s,$$

где $\sum_b$ – множество лемм (нормальных форм) лексемы s; $\sum_s$ – множество словоформ лексемы s.

Функция $morph$ обладает следующими свойствами:

$$morph\,(b) = b\,,\ \forall s \notin \sum_S \rightarrow morph(s) = b, b \notin \sum_S .$$

В результате перевода всего множества лексем в нормальную форму получим множество лемм:

$$L^{lemm} = \left\{ morph\left(lower(s)\right) | \, s \in L, s \notin D^{stop} \right\}.$$

Сформируем отношения L_1^{lemm}, L_2^{lemm} для связок лексем множества L^{lemm} и их рейтинговые распределения $\xi_{L_1^{lemm}}, \xi_{L_2^{lemm}}$:

$$L_1^{lemm} = L^{lemm},\ L_2^{lemm} = \left\{ (b_1, b_2) \,|\, b_1, b_2 \in L^{lemm}; seq\,(b_1, b_2) = 1 \right\}.$$

Структуры для обнаружения ошибок. Корпусный тезаурус D^{corpus} будет впоследствии применяться для определения множества лексем, подлежащих корректировке:

$$D^{corpus} = L^{lowpruned} \bigcap \left(D^{general} \bigcup D^{special} \right),$$

где $D^{general}$ — словарь общих слов русского языка, $D^{special}$ — тематические тезаурусы предметной области документа (имена собственные, географические наименования, аббревиатуры и т.п.).

Этап 2. Генерация корректировок. Пусть последовательность $Lex^{source} = \{s_1, s_2, \ldots, s_m\}$ — набор лексем, полученных в результате распознавания отдельного изображения документа.

$Lex = \{s_1, s_2, \ldots, s_n\}$ — результат нормализации последовательности Lex^{source}, разделим его на множество лексем Lex^{error}, подлежащих корректировке, и множество лексем $Lex^{correct}$, которые будем считать корректно распознанными:

$$Lex = Lex^{error} \bigcup Lex^{correct}.$$

В область лексем $Lex^{correct}$, не подлежащих корректировке, отнесем лексемы, для которых найдено соответствие в корпусном тезаурусе D^{corpus} или длина которых меньше порогового значения φ:

$$Lex^{correct} = \left\{ s \,|\, s \in Lex, |\, s \,| \le \varphi, s \in D^{corpus} \right\}.$$

Задача генерации корректировок сводится к отбору методом анаграмм множества корректировок $W_i \subset L^{corr}$ для замены каждой лексемы $s_i \in Lex^{error}$, $i \in [1 \ldots | Lex^{error} |]$.

Этап 3. Ранжирование корректировок. После получения множества корректировок W необходимо определить вероятность каждой из них и провести ранжирование в порядке убывания вероятности.

Ранжирование будем производить в два шага: $W \xrightarrow{1} \vec{W} \xrightarrow{2} \hat{W}$.

Шаг 1. Инвариантная оценка соответствия корректировки w для замены лексемы s:

$$score(s,w) = \ln(\xi_{L^{corr}}(w)) \times (|w| - LD(s,w)) \times r(w) \times d_{factor},$$

$$d_{factor} = \begin{Bmatrix} 3, \textit{если } w \in D^{corpus} \\ 1, \textit{если } w \notin D^{corpus} \end{Bmatrix},$$

где $LD(s,w)$ — расстояние Левенштейна между лексемой s и корректировкой w; $r(w)$ — количество повторений корректировки w в ходе отбора методом анаграмм.

В итоге для каждой лексемы $s \in Lex^{error}$ формируется упорядоченное по убыванию инвариантной оценки $score(s,w)$ множество корректировок:

$$\vec{W} = \{w \mid w \in W, score(s,w_k) \geq score(s,w_{k+1}), 1 \leq k \leq |W|\}.$$

Шаг 2. Вычисление финального ранга.

Сократим размер множества $\vec{W}$ до n элементов: $|\vec{W}| = \min(n, |\vec{W}|)$, и вычислим значение финального ранга $Rank(s,w)$ для каждой корректировки w:

$$Rank(s,w) = \frac{score(s,w)}{\sum_{j=1}^{|\vec{W}|} score(s,w_j)} \times P(w),$$

где $P(w)$ — статистическая вероятность нахождения корректировки w на позиции лексемы s в тексте.

$$P(w_i) = P(w_i \mid w_{1,i-1}) = \frac{f(w_{i-1}, w_i)}{f(w_{i-1})},$$

где $P(w_i \mid w_{1,i-1})$ — вероятность появления слова w_i при наличии предшествующей ему последовательности слов $w_1, w_2, \ldots, w_{i-1}$; $f(w_{i-1})$ — частота повторения слова, $f(w_{i-1}, w_i)$ — частота повторения биграммы (w_{i-1}, w_i).

Поскольку предшествующая лексема может являться ошибочной, вместо слова w_{i-1} будем использовать множество корректировок $\vec{W}_{i-1}$, информацию о частоте повторения слов и биграмм будем получать из рейтинговых распределений лексем в нормальной форме $\xi_{L_1^{lemm}}$, $\xi_{L_2^{lemm}}$.

Формула расчета вероятности принимает вид:

$$P(w_i^k) = \frac{\sum_{j=1}^{|\vec{W}_{i-1}|} \xi_{L_2^{lemm}}(morph(w_{i-1}^j), morph(w_i^k))}{\sum_{j=1}^{|\vec{W}_{i-1}|} \xi_{L_1^{lemm}}(morph(w_{i-1}^j))},$$

где w_i^k — k-ая по порядку корректировка лексемы s_i, $1 \le k \le |\vec{W}_i|$; w_{i-1}^j — j-ая по порядку корректировка лексемы s_{i-1}.

В итоге для каждой лексемы $s \in Lex^{error}$ формируется упорядоченное по убыванию финального ранга множество наиболее вероятных корректировок:

$$\widehat{W} = \left\{ w \mid w \in \vec{W}, Rank(s, w_k) \ge Rank(s, w_{k+1}), Rank(s, w) \in [0..1], 1 \le k \le |\vec{W}| \right\}.$$

Этап 4. Формирование результата. Результат распознавания представляет собой множество:

$$RES = \left\{ \left\langle s, w^{best}, W^{alternate} \right\rangle \mid s \in Lex \right\},$$

где w^{best} — наилучшая корректировка, $W^{alternate}$ — дополнительные корректировки.

Выбор наилучшей корректировки w^{best} производится по следующим правилам:

1. Если больше половины символов в лексеме s являются прописными и среди корректировок $\widehat{W}$ есть корректировки из тезауруса аббревиатур D^{abbr}, то среди них выбирается корректировка с наивысшим рангом:

$$w^{best} = \underset{w \in (\widehat{W} \cap D^{abbr})}{Arg\max} Rank(s, w).$$

2. Если первый символ лексемы s прописной, а остальные строчные и в списке корректировок $\widehat{W}$ есть корректировки из тезауруса фамилий $D^{surname}$ или имен D^{name}, то среди них выбирается корректировка с наивысшим рангом:

$$w^{best} = \underset{w \in (\widehat{W} \cap (D^{surname} \cup D^{name}))}{Arg\max} Rank(s, w).$$

3. Если по предыдущим правилам наилучшая корректировка не была выявлена, то выбирается самая первая корректировка из списка $\widehat{W}$:

$$w^{best} = \widehat{w}_1, \widehat{W} = \{\widehat{w}_1 \dots \widehat{w}_{|\widehat{W}|}\}.$$

В случае если правила 1 и 2 возвращают множество корректировок с одинаковым рангом, то наилучшей считается первая выбранная.

Во множество дополнительных корректировок $W^{alternate}$ включаются все корректировки $\widehat{W}$ за исключением наилучшей w^{best}:

$$W^{alternate} = \widehat{W} \setminus \{w^{best}\}.$$

В третьей главе представлены технология и система распознавания архивных документов с последующей корректировкой результатов, приводится описание инструментария конфигурирования процесса обработки архивных документов, компонентная модель и программная реализация системы.

Опишем технологию распознавания архивных документов и корректировки результатов в виде процесса массовой обработки электронных образов архивных документов с целью извлечения текста с исправленными ошибками распознавания при помощи разработанной системы, инструментария и метода автоматической корректировки (рисунок 3).

1. Вначале эксперту необходимо произвести анализ электронных образов архивных документов на предмет качества сканирования и принадлежности к определенной тематической группе. Для каждой из тематических групп должны быть отобраны тестовые изображения и вручную введен эталонный текст для оценки качества распознавания.

Рис. 3. Технология распознавания архивных документов и корректировки результатов (фигурой человека обозначены этапы, выполняемые с участием эксперта)

2-4. Следующей задачей эксперта является настройка и выбор конфигурационных профилей для распознавания групп документов. При помощи специального инструментария, описанного далее, эксперт подготавливает набор профилей и на основе сравнительного анализа качества распознавания тестовых изображений выбирает наиболее подходящие.

5. Далее производится пакетное распознавание сформированных тематических групп документов в соответствии с конфигурационными профилями.

6. По окончании процесса распознавания запускается процесс построения структур данных, необходимых для процедуры автоматической корректировки результатов. Структуры данных строятся по корпусу результатов распознавания документов отдельной тематической группы и могут быть дополнительно расширены путем добавления к результатам распознавания перечня текстов, относящихся к той же тематической группе. При запуске процесса построения структур данных эксперту необходимо задать минимальные пороговые значения частоты повторений лексем и биграмм лексем, а также указать набор тематических тезаурусов и словарей, которые будут использованы для формирования корпусного тезауруса.

7-9. Обладая подготовленными структурами данных, эксперт производит настройку и выбор наиболее подходящих профилей для корректировки отдельных тематических групп документов, опираясь на результаты сравнительного анализа распознавания тестовых изображений. Если сформированные структуры данных не обеспечивают должное качество корректировки, то эксперт может перезапустить процесс их перестроения с

новыми параметрами.

10. Последним этапом является запуск процесса автоматической корректировки результатов распознавания с подготовленными профилями.

Программная реализация системы состоит из Java веб-приложения, набора прикладных программ и базы данных (рисунок 4).

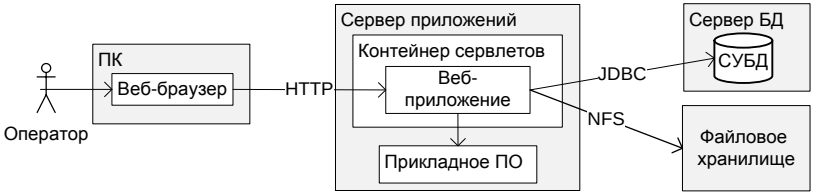


Рис. 4. Архитектура системы распознавания архивных документов

Компонентная модель разработанной системы распознавания архивных документов представлена на рисунке 5.

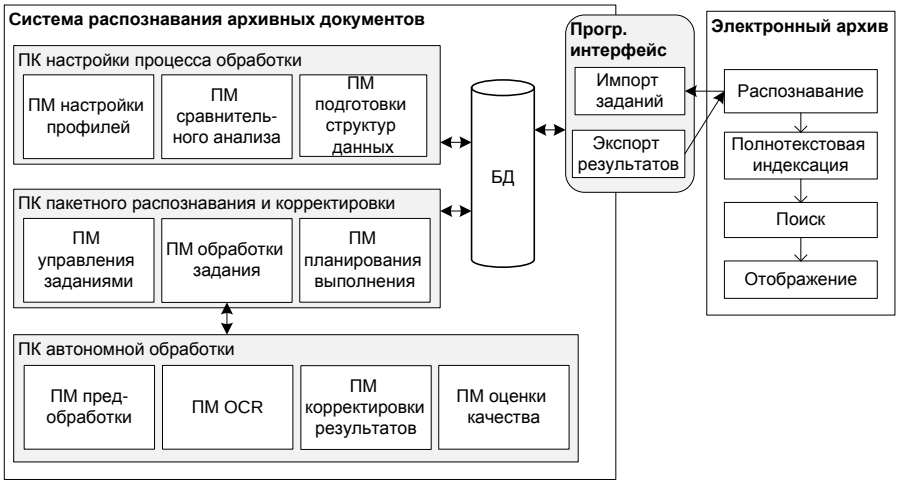


Рис. 5. Компонентная модель системы распознавания архивных документов
(ПК — программный комплекс, ПМ — программный модуль, БД — база данных)

Разработанная система состоит из трех программных комплексов, связанных между собой единой базой данных, и программного интерфейса для взаимодействия с внешними системами:

1. Программный комплекс настройки процесса обработки.

Инструментарий, позволяющий эксперту ограничивать пространство конфигураций для повышения качества распознавания, реализован в виде программных модулей настройки профилей и сравнительного анализа.

Под профилем понимается множество допустимых параметров из всего множества конфигураций процессов распознавания и корректировки документов определенного типа.

Процесс ограничения пространства конфигураций и формирования

профилей проиллюстрирован на рисунке 6.

Определим множество конфигураций:

$$\Omega = \Omega_{PRE} \cup \Omega_{OCR} \cup \Omega_{POST} \cup \Omega_{QA},$$

где Ω_{PRE}, Ω_{OCR}, Ω_{POST}, Ω_{QA} — множества параметров настройки стадий предварительной обработки, оптического распознавания, автоматической корректировки результатов распознавания, оценки качества итогового результата распознавания изображения соответственно.

Задачей эксперта на этапе настройки профилей является формирование множества профилей $\Omega^{PROFILES} = [\Omega_1^{PROFILE} \ldots \Omega_N^{PROFILE}]$, где $\Omega^{PROFILE}$ — профиль, содержащий множество параметров, наиболее подходящих для распознавания отдельных типов документов.

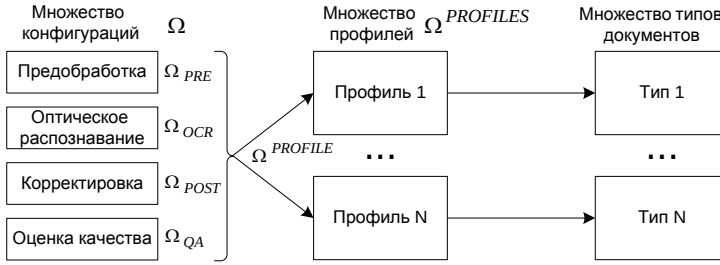


Рис. 6. Иллюстрация процесса ограничения пространства конфигураций

Используя программный модуль сравнительного анализа, эксперт может определить профиль, который наиболее эффективно решает задачу распознавания группы изображений. Для определения эффективности эксперту предоставляются рассчитанные системой значения критериев оценки качества распознавания.

Также в задачи подготовки к работе входит предварительный разбор всего корпуса распознанных текстов и построение тезаурусов и структур данных, необходимых для автоматической корректировки ошибок.

2. Программный комплекс пакетного распознавания и корректировки предназначен для управления ходом выполнения заданий на обработку документов. Его основными задачами являются: предоставление возможности просмотра журнала заданий, управление приоритетами заданий, вызов процедур распознавания и корректировки отдельных документов, сбор результатов и запись их в базу данных.

3. Программный комплекс автономной обработки отвечает за процесс распознавания и корректировки отдельного документа в соответствии с заданным профилем.

4. Программный интерфейс системы предоставляет ряд сервисов для постановки на распознавание отдельных документов электронного архива, опроса состояния и получения результатов.

В четвертой главе даются сведения об опытной эксплуатации

разработанной технологии и системы распознавания архивных документов, приводится экспериментальная оценка предложенного метода корректировки ошибок распознавания и результаты автоматической корректировки всего корпуса распознанных документов.

Испытания проводились на базе документов научно-справочного аппарата пяти центральных государственных архивов Санкт-Петербурга.

Для проведения экспериментальной оценки предложенного метода корректировки были вручную отобраны изображения, содержащие печатный текст, отражающий тематическую направленность архива. Каждому изображению был вручную подготовлен эталонный текст. Формат изображений — JPEG, разрешение - 300dpi.

Отобранные изображения были распознаны и сгруппированы в наборы, каждый набор содержал по несколько десятков изображений и соответствовал определенному диапазону точности распознавания на уровне слов. Примеры изображений из каждого набора представлены на рисунке 7.

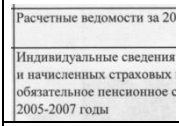
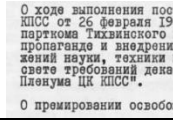
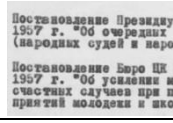
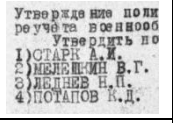
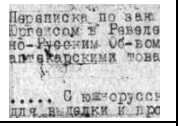
				
Н-1 (100%-80%)	Н-2 (80%-60%)	Н-3 (60%-40%)	Н-4 (40%-20%)	Н-5 (20%-0%)

Рис. 7. Примеры тестовых изображений каждого набора, в скобках указан диапазон точности распознавания изображений набора

Результаты распознавания были получены коммерческой системой оптического распознавания «Abbyy Finereader» (Abbyy) и свободно распространяемой системой «Tesseract». Для оценки качества эталонный текст и результат распознавания изображения разбивались на поисковые токены, далее вычислялась полнота *Recall* и точность *Precision* наличия токенов эталона в результате распознавания:

$$Recall = T_{common}/T_{groundtruth}\ ,\ Precision = T_{common}/T_{ocr}\ ,$$

где T_{common} — количество токенов эталона, содержащихся в распознанном тексте; $T_{groundtruth}$ — количество токенов в эталоне; T_{ocr} — количество токенов в результате распознавания.

Сравнение значений полноты и точности результатов распознавания тестовых наборов изображений без корректировки, с результатами распознавания, содержащими один (+1) и три варианта (+3) замены ошибочных слов, представлено на рисунке 8.

Разработанный метод корректировки повышает качество распознавания как коммерческих, так и свободно распространяемых систем. Наибольшие **приращения (до +15%) показателей полноты и точности** отмечаются на результатах распознавания, находящихся в диапазоне словарной точности от 80 до 20%, что объясняется малым количеством «простых» ошибок в верхнем диапазоне и низким качеством результатов в нижнем диапазоне.

Увеличение значения полноты результатов распознавания при учете альтернативных корректировок свидетельствует о том, что верные корректировки не всегда определяются как наилучшие, но присутствуют в списке альтернативных корректировок, что указывает на возможность улучшения алгоритма ранжирования корректировок.

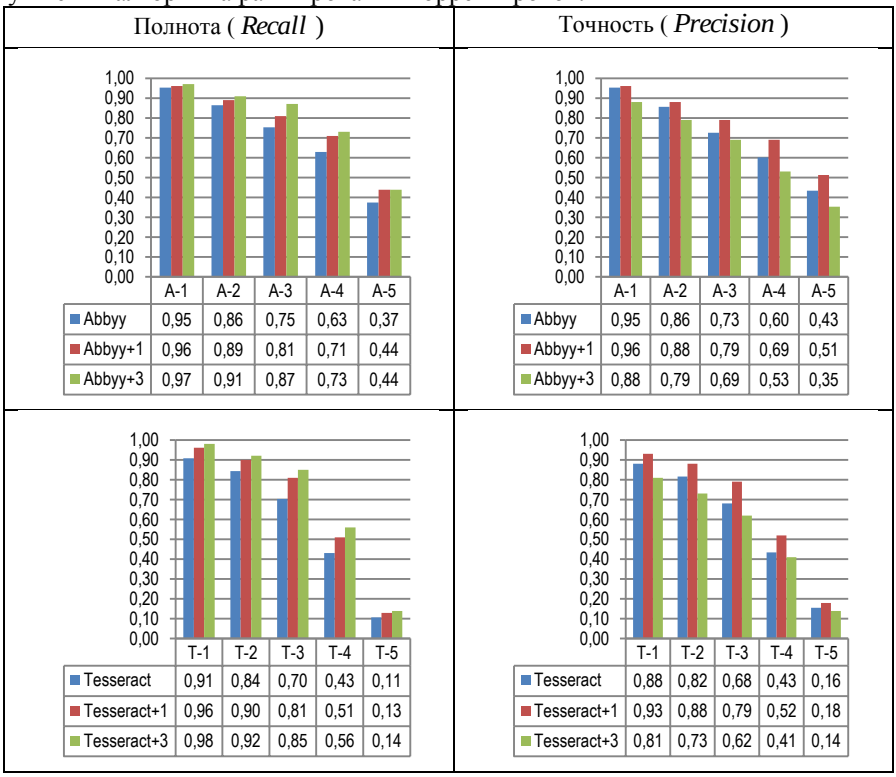


Рис. 8. Сравнение оценок полноты и точности до и после корректировки

После внедрения разработанной системы в центральные государственные архивы Санкт-Петербурга было распознано **32 608** документов, состоящих из **708 663** изображений. Размер результатов распознавания составил более **88 миллионов** лексем.

Для оценки качества распознавания всего корпуса документов проводилось вычисление словарной точности $A_D = 1 - \frac{n_{error}}{n_w}$, где n_w — общее количество лексем (слов) в результате распознавания, n_{error} — количество «ошибочных» слов в результате распознавания, т.е. слов, отсутствующих в словаре и дополнительных тематических тезаурусах. В состав словаря было включено 5 498 345 словоформ сгенерированных из словаря Зализняка и программы проверки орфографии Hunspell, также были подключены следующие

тематические тезаурусы: фамилии (918 659 словоформ), имена (105 560 словоформ), отчества (231 313 словоформ), аббревиатуры (1 413 словоформ).

Количество ошибочных слов после автоматической корректировки **сократилось на 46%**, было **исправлено 16 497 948 ошибочных слов**, **значение словарной точности** в среднем по всем архивам **увеличилось на 18%**. На рисунке 9 представлено распределение количества изображений по диапазонам словарной точности результатов распознавания до и после корректировки.

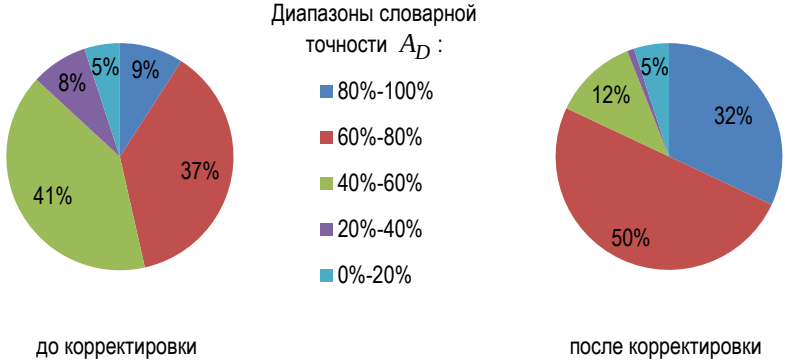


Рис. 9. Процентное распределение количества изображений по диапазонам словарной точности результатов распознавания до и после корректировки

Электронный архив с включенной в его состав подсистемой потокового распознавания документов с автоматической корректировкой ошибок обладает рядом существенных преимуществ перед архивными системами, в которых распознавание отсутствует либо осуществляется вручную. Данными преимуществами являются высокие темпы перевода документов в электронную форму, возможности построения эффективного поискового аппарата, высокая скорость поиска и доступа к электронным образам документов. Основываясь на статистических данных проекта «Государственные архивы Санкт-Петербурга», расчет времени ручного ввода текста **500 тысяч изображений** описей составляет около **50 человеко-лет**. Применение средств автоматического распознавания и корректировки позволило сократить годы ручного труда и значительно расширить поисковую базу для работы граждан, исследователей и сотрудников архивов.

ЗАКЛЮЧЕНИЕ

Полученные в диссертационном исследовании результаты представляют собой решение актуальной задачи повышения качества распознавания изображений. В ходе исследования получены следующие основные результаты:

1. Разработан метод автоматической корректировки ошибок распознавания архивных документов на основе рейтинго-ранговой модели текста, производящий поиск корректировок по тезаурусам, предварительно извлеченным из результатов распознавания и текстов одной тематической области (объем текстов порядка 100 миллионов символов).

2. Разработаны правила ранжирования и выбора наилучших корректировок, основанные на вычислении инвариантной оценки соответствия и вероятности нахождения финального слова n-граммы по известным предыдущим словам.
3. Разработан инструментарий, позволяющий эксперту производить настройку системы для обработки архивных документов различных тематических областей путем установки набора параметров, определенных по результатам сравнительного анализа качества распознавания тестовых изображений.
4. Разработаны технология и система распознавания архивных документов и автоматической корректировки результатов, успешно интегрированные с системой электронного архива и производящие массовую параллельную обработку документов в пакетном режиме, позволившие сократить количество ошибочных слов на 46%, а значение словарной точности в среднем повысить на 18%.

Полученные результаты соответствуют п.3 «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем» и п.7 «Человеко-машинные интерфейсы; модели, методы, алгоритмы и программные средства машинной графики, визуализации, обработки изображений, систем виртуальной реальности, мультимедийного общения» паспорта специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

СПИСОК ОСНОВНЫХ ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в рецензируемых научных изданиях из перечня ВАК

1. **Смирнов, С.В.** Оцифровка, каталогизация, хранение и поиск архивной документации / С.В. Смирнов, М.В. Белозёрова // Информационно-измерительные и управляющие системы. – 2010. – т. 8, №7. – С. 97-101.
2. Кулешов, С.В. Методы сегментации OCR систем в задачах автоматической обработки архивных документов / С.В. Кулешов, **С.В. Смирнов** // Труды СПИИРАН. – 2011. – Выпуск 1(16). – С. 110–122.
3. **Смирнов, С.В.** Подсистема массового распознавания изображений архивных документов / С.В. Смирнов // Труды СПИИРАН. – 2012. – Выпуск 3(22). – С. 234-248.
4. **Смирнов, С.В.** Методы автоматической постобработки результатов распознавания в задачах оцифровки архивных документов / С.В. Смирнов // Информационно-измерительные и управляющие системы. – 2013. – т. 11, №9. – С. 22-32.
5. **Смирнов, С.В.** Сравнительный анализ OCR систем в контексте построения системы поиска по изображениям архивных документов / С.В. Смирнов // Информационно-измерительные и управляющие системы. – 2014. – т. 12, №12. – С. 44–51.
6. **Смирнов, С.В.** Корректировка ошибок оптического распознавания на основе рейтинго-ранговой модели текста / С.В. Смирнов // Труды СПИИРАН. – 2014. – Выпуск 4(35). – С. 64-82.

Публикации в других изданиях

7. **Смирнов, С.В.** Таксономия информационных объектов электронного архива / С.В. Смирнов // Сборник научных трудов Всероссийской научно-практической конференции-форума молодых ученых и специалистов «Современная российская наука глазами молодых исследователей». – Красноярск: Научно-инновационный центр, 2011. – С. 192-194.
8. **Смирнов, С.В.** Логическая модель представления информации в электронном архиве / С.В. Смирнов // Сборник научных трудов IV Всероссийской научно-практической конференции с международным участием «Научное творчество XXI века». – Красноярск: Научно-инновационный центр, 2011. – Выпуск 2. – С. 93-94.
9. **Смирнов, С.В.** Критерии оценки качества результатов оптического распознавания / С.В. Смирнов // Сборник материалов XVI Международной научно-практической конференции «Перспективы развития информационных технологий». – Новосибирск: Издательство ЦРНС, 2013. – С. 33–38.
10. **Смирнов, С.В.** Особенности построения системы массового оптического распознавания архивных документов / С.В. Смирнов // Труды XVII Всероссийской объединенной конференции «Интернет и современное общество». СПб: Университет ИТМО, 2014. – С. 37-42.
11. **Смирнов, С.В.** Система полнотекстового поиска по изображениям архивных документов / С.В. Смирнов // Сборник материалов XXI Международной научно-практической конференции «Перспективы развития информационных технологий». Новосибирск: Изд-во ЦРНС, 2014. – С. 16–21.
12. Воронцов, А.В. Настоящее и будущее государственных электронных архивов Санкт-Петербурга / А.В. Воронцов, А.В. Кожин, **С.В. Смирнов** // Материалы X всероссийской научно-практической конференции «Электронные ресурсы библиотек, музеев, архивов». СПб: Изд-во «Перфектум», 2014. – С. 106–114.

Свидетельства о государственной регистрации

13. Программный комплекс «Формирование метаданных» ГИС «Государственные архивы Санкт-Петербурга» : свидетельство о гос. регистрации программы для ЭВМ №2014662557 Российская Федерация / **С.В. Смирнов**, А.В. Кожин, А.В. Воронцов, М.В. Белозерова; правообладатель Санкт-Петербург, Комитет по информатизации и связи. – зарегистрировано в Реестре программ для ЭВМ 03.12.2014г. – 1 с.
14. Программный комплекс «Информационно-лингвистическое обеспечение» ГИС «Государственные архивы Санкт-Петербурга» : свидетельство о гос. регистрации программы для ЭВМ №2014662676 Российская Федерация / **С.В. Смирнов**, А.В. Кожин, А.В. Воронцов, М.В. Белозерова; правообладатель Санкт-Петербург, Комитет по информатизации и связи. – зарегистрировано в Реестре программ для ЭВМ 05.12.2014г. – 1 с.

Подписано в печать 2015г.
Формат 60x84 1/16. Цифровая печать. Усл. печ. л. 1. Тираж 150 экз.
Отпечатано в СПб ГУП «СПб ИАЦ». 191040, РФ, Санкт-Петербург, Транспортный переулок, д.6, литер А, пом. 7Н, 8Н. тел.: (812) 764-39-57, факс: (812) 764-95-48, e-mail secretar@iac.spb.ru, http://iac.spb.ru