Московский государственный университет печати имени Ивана Фёдорова

Кафедра медиасистем и технологий

Анна Юрьевна Филиппович
Юрий Николаевич Филиппович

Технологии ввода текста

Лекции по дисциплине
"ИНТЕГРИРОВАННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ В МЕДИАИНДУСТРИИ"

Для студентов,
осваивающих образовательные программы
подготовки бакалавров и магистров по направлениям
230400 «Информационные технологии и системы»,
230100 «Информатика и вычислительная техника»

Москва, 2013

Клавиатурный ввод текста

В настоящее время для ввода текста широко используется клавиатура компьютера. В этом случае фактически осуществляется побуквенный ввод текста. Скорость ввода текста первую очередь зависит от используемой раскладки[1].

Латинские раскладки клавиатуры

Первая коммерчески успешная пишущая машинка была изобретена в сентябре 1867 года американцем Кристофером Шоулзом. В ней использовалась латинская раскладка, в которой буквы на клавишах располагались в алфавитном порядке. Например, на первых семи клавишах верхнего буквенного ряда были расположены буквы: A, B, C, D, E, F, G.

У пишущей машинки Шоулза был недостаток: при быстрой печати литеры цеплялись друг за дружку и их рычажки «перепутывались». Было решено отказаться от «алфавитной» раскладки. От новой раскладки клавиатуры, получившей в дальнейшем название по буквам на первых шести клавишах третьего ряда алфавитно-цифрового блока клавиатуры — QWERTY. Требовалось, чтобы буквы, образующие в английском языке устойчивые комбинации, располагались как можно дальше друг от друга по разные стороны клавиатуры и были разбросаны по разным рядам, что уменьшало вероятность «перепутывания» рычажков пишущей машинки. В настоящее время раскладка Шоулза критикуется как анахронизм, так как проблемы, которая привела к появлению QWERTY, больше не существует. Дальнейшее совершенствование пишущих машинок устранило проблему «перепутывания» рычажков и пробудило интерес к вопросу увеличения скорости печати.

Рис.1. Раскладка клавиатуры QWERTY.

В 1936 году профессор Вашингтонского Университета Август Дворак (August Dvorak) издал книгу, в которой предложил совершенно новую латинскую раскладку, носящую в настоящее время имя автора. Её принцип — максимальное удобство для набирающего текст на английском языке на пишущей машинке.

Рис.2. Раскладка клавиатуры Дворака.

В 2006 году Шаем Коулманом (Shai Coleman) была разработана раскладка Colemak. Название происходит от Coleman+Dvorak. Раскладка приспособлена к современным

[1] **Раскла́дка клавиату́ры** — соглашение о соответствии типографических символов (букв, цифр, знаков препинания и т. д.) письменного языка клавишам клавиатуры компьютера или другого устройства, с помощью которого вводится текст.

компьютерным реалиям. Её принцип — эффективный и эргономичный набор текстов на английском языке на компьютерной клавиатуре.

Русские раскладки

В русской компьютерной письменности в настоящее время используются две раскладки клавиатуры: ЙЦУКЕН, и «фонетическая раскладка». Наиболее распространённой из них является раскладка ЙЦУКЕН, название которой происходит от шести левых символов верхнего ряда раскладки.

Рис.3. Русская раскладка клавиатуры ЙЦУКЕН.

Считается, что раскладка ЙЦУКЕН далека от оптимальной при печати слепым десятипальцевым методом, а также в ней отсутствуют клавиши для используемых в русской письменности знаков препинания и небуквенных орфографических знаков, например кавычки «ёлочки», апостроф, параграф и другие.

Из альтернативных раскладок для русского языка наподобие «DVORAK» следует выделить раскладку DIKTOR и раскладку Зубачёва[3]. Обе эти раскладки были построены по схожему принципу с латинской раскладкой «DVORAK» но официальные сайты перестали существовать, а популярности они так и не завоевали[5].

В русской фонетической раскладке, известной как «ЯВЕРТЫ», или «метод Маслова» русские буквы расположены там же, где и похожие (фонетически, по звучанию) латинские, например, А-A, Б-B, Д-D, Ф-F, К-K, О-O и т. д. Созданы варианты русской фонетической раскладки на основе латинской QWERTY, а также на основе других латинских и национальных раскладок[6].

Рис.4. Фонетическая русская раскладка клавиатуры.

Подключение русской фонетической раскладки в Microsoft Windows требует специальной компьютерной программы, найти которую можно на соответствующих сайтах Интернета.

Методики ускорения ввода текста

Слепой метод печати

Множество методик и специальных технологий направлено на ускорение процесса набора текста, один из них – слепой метод набора текста.

Слепой метод набора — методика набора текста «вслепую», то есть не глядя на клавиши печатной машинки или кнопки клавиатуры, используя все (или большинство) пальцы рук, ранее был известен как американский слепой десятипальцевый метод[1][2]. Существует уже более 120 лет. В XIX веке слепым методом печати на печатных машинках обучали машинисток и секретарей. Это позволило сузить сферу использования стенографии, увеличить производительность труда секретарей.

Текстовые экспандеры

Другой способ ускоренного ввода текста – использование специальных программ – текстовых экспандеров. Принцип работы подобных программ предельно прост: они позволяют настроить автоматическую вставку любого текстового фрагмента. Например, вы набираете «к5к», а вместо этого появляется фраза «ожидаемые сроки поставки товара — апрель следующего года». Тем, кому постоянно приходится заполнять бланки документов и вбивать одни и те же выражения, программа-экспандер поможет сэкономить время. Примеры программ:

- Phrase Express 7.0
- Texter 0.6
- Turbo Type 1.39

Подобные программы позволяют настраивать параметры автоматического набора текста — например, задать время задержки перед нажатием очередной клавиши, указать скорость вставки содержимого буфера обмена и т.д. Текстовые шаблоны могут вставляться в документ в режиме автоматической замены или с помощью настроенных комбинаций клавиш. Помимо этого позволяется управлять словарными сокращениями: сортировать их по категориям, устанавливать параметры для групп слов и т. д. При использовании той или иной контрольной последовательности символов программа может предлагать меню для выбора одного из нескольких шаблонов для вставки в документ и др. [Sergey Pavlov, 19.01.2011]

Интеллектуальный ввод текста

Технологии ввода текста в мобильных устройствах

Довольно много различных технологий, позволяющих ускорить процесс ввода текста, используются в мобильных устройствах. Это прежде всего связано с малым их размером.

Используются следующие технологии:

- Режим Multitap;
- Виртуальная клавиатура;
- Предиктивный ввод текста;
- Другие системы ввода.

Режим Multitap – стандартный ввод с клавиатуры многократным нажатием кнопки. Клавиши позволяют вводить большее количество символов, чем видно на клавишах путем многократного касания клавиши. Такой способ набора использовался в первых кнопочных телефонах.

С развитием технологий получили распространение сенсорные устройства. В этом случае широкое распространение получила технология ввода с помощью **виртуальной клавиатуры**. Выбирая параметры можно осуществлять ввод текста на разных языках, ввод специальных символов и т.п.

Предиктивный ввод

Предиктивный ввод текста (от англ. *predict* — предсказывать) — система ускоренного ввода текста в цифровые устройства, при которой программное обеспечение устройства в процессе набора предлагает варианты окончания слов и фраз, основываясь на имеющихся в его словаре, а также может предлагать исправлять распространённые ошибки. Примеры режим **T9**, **iTap.**

T9 — предугадывающая система набора текстов для мобильных телефонов. Название T9 происходит от англ. *Text on 9 keys*, то есть набор текста на 9 кнопках. T9

разработана компанией Tegic Communications(создатель Клиф Кашлер), и используется в мобильных телефонах большинства крупных производителей. Первым мобильным телефоном с T9 стал Sagem MC 850, появившийся на рынке в 1999 году.

iTap — система предиктивного набора текста для мобильных телефонов. Разработана фирмой Motorola для использования в своих аппаратах, также может быть лицензирована другими компаниями.

iTap имеет существенное отличие от системы T9, которую она была призвана заменить. Если T9 всегда пытается подставить слово, имеющее столько букв, сколько набрано на данный момент, то iTap пытается предугадать и более длинные слова, анализируя не только набранные буквы текущего слова, но и предыдущий текст. Кроме того, iTap может предугадывать даже короткие фразы. Такая особенность позволяет существенно ускорить набор текста, особенно если в тексте в основном используются простые и наиболее употребляемые слова и фразы.

Система iTap является обучаемой, то есть она запоминает наиболее употребляемые слова, такие, как имена, фамилии, названия, и т. д., и пытается подставить их при наборе в первую очередь.

Система ввода текста – Swype

Swype (изм. от англ. *swipe — проводить не отрывая, скользить* и англ. *type — писать, набирать текст*) — метод ввода текста не отрывая палец/стилус от «кнопок клавиатуры» на сенсорном экране.

Клифф Кашлер, разработчик популярной системы мобильного ввода текстов T9, применяемой на подавляющем большинстве современных мобильных телефонов, создал новую буквенно-цифровую технологию предикативного ввода текстов на смартфонах, оснащенных сенсорными экранами, а также на планшетных ПК.

Новая технология, получившая название Swype, работает на любых аппаратах, оснащенных виртуальными экранными QWERTY-клавиатурами, похожими на те, что есть на iPod Touch или iPhone. Основной принцип Swipe заключается в следующем: пользователю предлагается не "выстукивать" на экране слово по буквам, а вести палец по экрану от начальной буквы слова по всем последующим, как бы чертить слово по буквам, система в этот момент в реальном времени просматривает словарный запас и пытается угадать, что именно за слово пользователь "вырисовывает" на экране.

Кашлер отмечает, что его разработка уже запатентована и в испытаниях она позволяла писать до 50 слов в минуту при помощи сенсорной клавиатуры на экране смартфона. Разработчик говорит, что его система также оснащена искусственным интеллектом, который исходит из того, что палец пользователя может периодически съезжать с нужных букв и программа сама подправляет написание.

С технической точки зрения технология потребляет довольно мало вычислительных ресурсов, поэтому она может работать даже на маломощных мобильных гаджетах. К примеру, база из 65 000 слов в Swipe занимает всего 250 килобайт. Всего же работающая система резервирует для себя в памяти мобильного устройства около 1 мегабайта памяти. Самые "прожорливые" компоненты здесь - это анализатор частей слов, поисковый движок слов и, собственно, пользовательский интерфейс.

В сравнении с системой T9, Swype работает чуть быстрее на распознавание слов на раннем этапе, кроме того, программа создает так называемый кэш - небольшой словарь слов, которые используются пользователем чаще других и в первую очередь система занимается поиском по ним.

Вместе с программой идет разработка Swype Operation, которая оптимизирует словарные базы и следит за внутренними параметрами работы системы ввода, создает словарные пары, ускоряет доступ к данным и делает некоторые другие операции. Разработчик говорит, что средняя скорость перебора одной базы при помощи Swype

Operation составляет около 250 мс. На практике это означает, что система способна каждые четверть секунды предлагать новые варианты слов.

Системы ввода 8pen

Система ввода **8pen** способна заменить стандартную клавиатуру на любом устройстве, оборудованном сенсорным экраном: мобильном телефоне, современном пульте дистанционного управления или же игровом контроллере. Её преимущество, прежде всего, в скорости - с помощью 8pen вы сможете вводить текст значительно быстрее, чем при использовании клавиатуры с традиционной раскладкой на небольшом экране. А простота жестов и имитация почерка позволяет свести к минимуму количество опечаток и сделать возможным ввод „вслепую”.

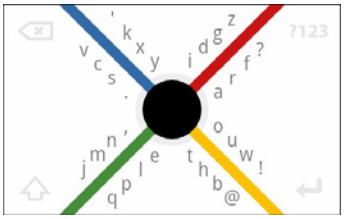

Рис. 5. Система ввода 8pen.

Для ввода любой буквы необходимо коснуться центрального круга и, не отрывая палец от экрана, переместить его в нужный сектор, после чего провести линию ещё через несколько секторов (в зависимости от расстояния от центра до буквы) и вернуть палец на место.

Системы ввода иероглифов

Структурные методы

Структурный метод ввода китайских иероглифов основан на графической структуре иероглифа. Каждый иероглиф состоит из нескольких частей - графем. Клавиатура разбита на пять зон, по числу базовых черт. Внутри каждой зоны клавиши пронумерованы — от центра клавиатуры к краям. Номер составляется из двух цифр от 1 до 5 — в зависимости о того, из каких базовых черт собирается графема.

Рис.6. Китайская раскладка клавиатуры (структурный метод уби цзысин).

Один из самых распространенных методов структурного ввода — *уби цзысин* (Wubing zixing — «ввод по пяти чертам»). Иероглифы имеют пять базовых черт (一, 丨, 丿, 丶, 乙), также есть 25 очень часто употребляемых иероглифов (каждому из них сопоставлена клавиша).

Рис.7. Деление иероглифа на графемы.

Можно выделить четыре группы иероглифов (по способу сочетания их графем):

Иероглифы, между графемами которых есть определённое расстояние. Например, иероглиф 苗 состоит из графем 艹 и 田, между которыми есть расстояние (хотя на печати они немного «спрессовываются» и вам может показаться, что расстояния между ними нет).

Иероглифы, графемы которых соединены друг с другом. Так, иероглиф 且 представляет собой графему 月, соединённую с горизонтальной чертой; 尺 состоит из графемы 尸 и откидной черты.

Иероглифы, графемы которых пересекаются либо накладываются друг на друга. Например, иероглиф 本 — это пересечение графем 木 и 一.

Таким образом, для ввода иероглифов, состоящих более чем из четырёх графем, нужно ввести первые три графемы и последнюю. Сложнее всего вводить иероглифы, состоящие из двух или трёх графем. Поскольку их очень много, то неизбежно появятся несколько иероглифов, претендующих на одну и ту же комбинацию клавиш. Чтобы их различать, разработчики придумали специальный код. Этот код состоит из двух цифр, первая — порядковый номер последней черты иероглифа, а вторая — номер группы иероглифа.

Недостатки структурного ввода — сложность. Достоинство — возможность слепого ввода, что повышает максимальную скорость набора до 160 символов в минуту (это около 500 нажатий клавиш).

Фонетические методы

Используя фонетический метод, вводится не сам иероглиф, а его произношение. Проблема заключается в том, что в китайском языке одному и тому же произношению могут соответствовать десятки иероглифов. Поэтому система ввода включает технологию предиктивного ввода текста. Наиболее распространенный фонетический метод — *пиньинь* (Pinyin). На его основе построена система фонетического ввода, которая входит в стандартный Asian Language Pack системы Windows (начиная с версии XP — до этого ее приходилось ставить дополнительно).

Системы ввода Google Pinyin и Sogou Pinyin запоминают пользовательские предпочтения и подсказывают на основании контекста нужные слова. Вот пример того, как Google Pinyin анализирует, казалось бы, зубодробительную последовательность.

Рис. 8. Иллюстрация использования фонетической системы ввода Google Pinyin.

Основным недостатком систем фонетического ввода является довольно низкая скорость печати — около 50 знаков в минуту (сравните с уби цзысин с его 160 знаками в минуту). Дело в том, что ввод иероглифа по методу пиньинь происходит, в среднем, за шесть нажатий клавиш, тогда как при вводе по уби цзысин будет достаточно четырех. К тому же, слепой набор данным методом невозможен. Для фонетического ввода достаточно любой клавиатуры с латинской раскладкой QWERTY.

Гибридные методы

Эти методы представляют собой некую комбинацию фонетических и структурных методов ввода. Простейший пример — метод *иньсин* (Yinxing — «звучание и форма»). Иероглиф набирается путем ввода транскрипции и указания на графический элемент. Ограниченный набор графических элементов разнесен по клавиатуре. На практике системы гибридного ввода постепенно вымирают. Они требуют от пользователя

одновременно знания сложной комбинаторики структурных систем и хорошего владения транскрипцией.

Ввод текста с помощью технологии распознавания

Распознавание рукописного текста

Распознавание рукописного текста может производиться «оффлайновым» методом из уже написанного на бумаге текста или «онлайновым» методом считыванием движений кончика ручки, к примеру по поверхности специального компьютерного экрана.

Онлайновый метод распознавание получил распространение в рамках различнывх функций ввода письменного текста в мобильных устройствах. Функция распознавания письменного текста преобразует движения стилуса в буквы, числа или другие символы и отображает их в виде текста. В этом случае задаются правила написания символов движения руки, стилуса (см. рисунок).

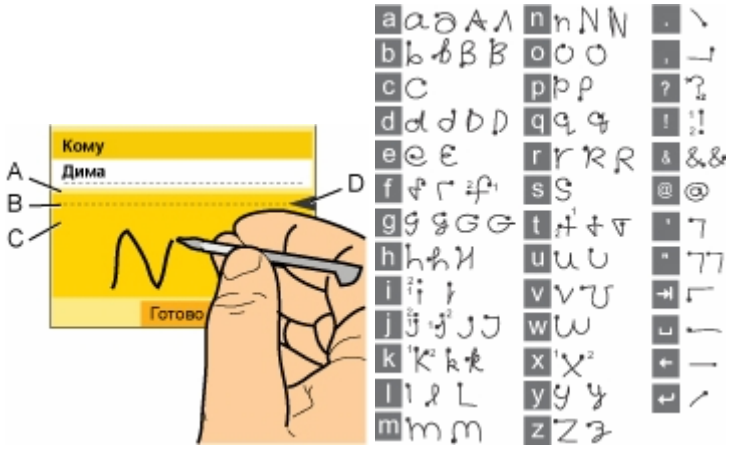

Рис.8. «Онлайн» распознавание рукописного текста.

Технология распознавание текста, написанного на бумаге получила распространение в рамках развития так называемых ICR-систем (intelligent character recognition – интеллектуальное распознавание символов). Примером служат системы распознавания рукопечатных символов в бумажных бланках и формах, например ABBYY FlexiCapture. А также различные некоммерческие системы распознавания рукописных текстов, например система распознавания скорописи XVII в. [Зеленцов И.А.]

Технологии оптического распознавания текстов

Системы оптического распознавания текста – OCR-системы предназначены для ввода печатного текста для печатных и электронных изданий.

Примеры: Recognita Plus DTK (Венгрия), *TextBridge*, *TypeReader* (США), *CharacterEyes* (Израиль), IRIS OCR (Бельгия), *Easy Reader* (Франция) и др.

Наиболее известными программами класса «Системы оптического распознавания» в России являются: *ABBYY FineReader*, *OmniPage Professional* и *OCR CuneiForm.*

С помощью технологии OCR (Optical Character Recognition – оптическое распознавание знаков) текст, представленный в рукописной или машинописной форме, преобразуется в цифровую форму и тем самым становится пригодным для обработки. Сначала в "процессе отображения" документа, находящегося на бумаге, осуществляется его ввод оптоэлектронными считывающими системами. Документ предстает в виде битовой карты. В дальнейшем битовая структура знака конвертируется в текстовый код.

Этапы преобразование документа в электронный вид OCR-системами:

- Сканирование и предварительная обработка изображения.
- Анализ структуры документа.
- Распознавание.

- Проверка результатов.
- Реконструкция документа (воссоздание его исходного вида).
- Экспорт.

Обработка документа начинается с получения графического образа (изображения) страницы. Входными данными для OCR-программы служит цветное (глубина цвета 24 бит) либо полутоновое (глубина цвета 8 бит) изображение документа. Прежде, чем приступить к структурированию страницы, выделению и идентификации блоков, OCR-система производит бинаризацию, то есть преобразование цветного или полутонового образа в монохромный. В случае, когда изображение имеет какой либо фон, используются различные процедуры интеллектуальной фильтрации фона или адаптивной бинаризации для отделение изображения текста от фона. Далее происходит анализ структуры документа, процедура распознавания текста, проверка результатов и воссоздание исходного вида документа, его экспорт.

Рассмотрим ***Базовые принципы технологии распознавания*:**

Целостность - объект описывается как целое с помощью значимых элементов и отношений между ними.

Целенаправленность - распознавание строится как процесс выдвижения и целенаправленной проверки гипотез.

Адаптивность - способность OCR-системы к самообучению.

В соответствии с этими тремя принципами система сначала выдвигает гипотезу об объекте распознавания (символе, части символа или нескольких склеенных символах), а затем подтверждает или опровергает ее, пытаясь последовательно обнаружить все структурные элементы и связывающие их отношения. В каждом структурном элементе выделяются части, значимые для человеческого восприятия: отрезки, дуги, кольца и точки. Следуя принципу адаптивности, программа самостоятельно "настраивается", используя положительный опыт, полученный на первых уверенно распознанных символах. Целенаправленный поиск и учет контекста позволяют распознавать разорванные и искаженные изображения, делая систему устойчивой к возможным дефектам письма.

На этапе анализа и предварительной обработки изображения перед любой OCR - системой стоят две основных задачи: во-первых, подготовить изображение к процедурам распознавания, во-вторых, выявить структуру документа – с тем, чтобы в дальнейшем иметь возможность воссоздать её в электронном виде. Обратимся к задаче анализа структуры.

Наибольшее распространение получили так называемые методы анализа иерархической структуры документа. При анализе структуры в рамках этих методов обычно выделяют несколько иерархически организованных логических уровней. Объект наивысшего уровня только один – собственно страница, на следующей ступени иерархии располагаются таблица, текстовый блок и картинка, и так далее (см. рисунок). Понятно, что любой объект может быть представлен как набор объектов более низкого уровня.

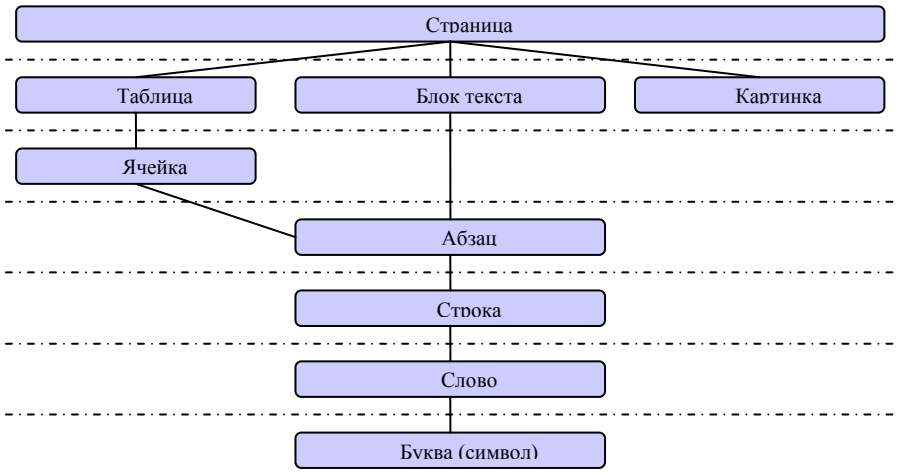

Рис.8. Иерархическая структура документа.

Механизм распознавания символов представляет собой комбинацию ряда элементарных распознавателей, называемых классификаторами. В общем виде работу классификатора иллюстрирует схема 1. Как показано, по окончании обработки классификатор порождает список гипотез относительно принадлежности очередного изображения к тому или иному классу, либо – в том случае, когда входные данные уже представляют собой список – соответствующим образом изменяет веса имеющихся гипотез, подтверждает или опровергает их. Выходной список всегда ранжирован по весу (уверенности).

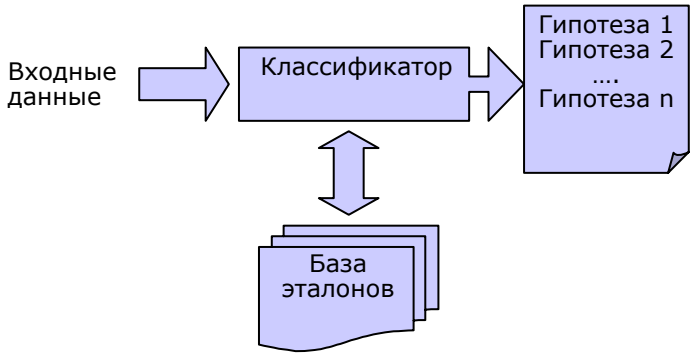

Рис.9. Механизм распознавания символов.

Все выдвинутые в процессе обработки документа гипотезы рассматриваются в составе многоуровневых структур – моделей.

Рассмотрим пример распознавания слова «turn». В процессе распознавания сформированы несколько гипотез (моделей). В результате их анализа произойдёт следующее: контекстная проверка покажет, что весь текст состоит из английских слов, и вес моделей «слово – английский язык» значительно увеличится, а моделей «слово – кириллица» соответственно уменьшится. Модель «цифры» также останется позади в силу крайне малого суммарного веса составляющих гипотез. Затем проверка по словарю подтвердит, что в словаре английского языка слова «tum» нет, а «turn» - есть. Следовательно, гипотеза относительно слова «turn» приобретёт ещё больший вес, что позволит ей в дальнейшем оказаться «победителем».

Рис.10. Иллюстрация распознавания слова «turn».

В процессе распознавания обязательна словарная проверка. Зачастую точность распознавания мало зависит от полноты словарной базы проверочных словарей. Ведь, как известно, не существует словарей, содержащих все словоформы живого языка. Учесть все жаргонные, разговорные, диалектные слова и выражения практически невозможно. При встрече с несловарным словом система распознает его в точности так, как оно было написано. Кроме этого используются морфологически структурированные словари способные моделировать различные словоформы.

Ввод текста с помощью OCR-систем

Процесс ввода текста с помощью системы оптического распознавания можно разделить на два этапа: предварительный и основной (см. рисунок). Первый включает в себя различные предварительные процедуры, общее назначение которых настройка и подготовка инструментальных средств для ввода текста и рабочего места оператора. В общем случае этот этап может включать в себя следующие процедуры: установка и настройка аппаратных и программных средств, подготовка текста для ввода, настройка параметров системы оптического распознавания. Состав операций и процедур предварительного этапа зависит от уже существующих настроек системы.

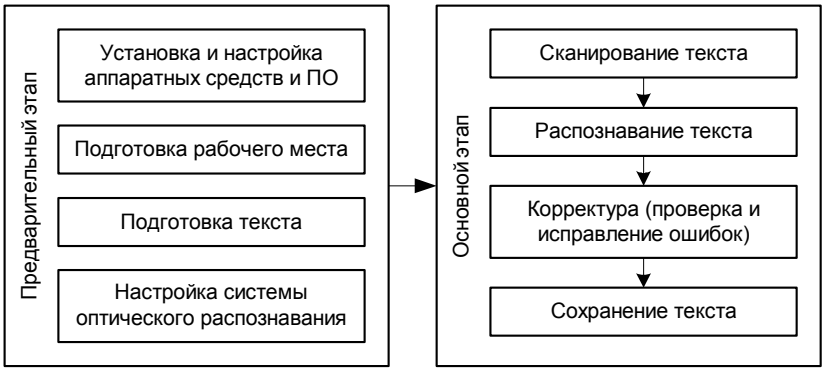

Рис.11. Технологический процесс вода текста с помощью OCR-системы.

Установка и настройка аппаратных средств может включать в себя следующие операции: установка, включение ЭВМ, подключение сканера к ЭВМ, установка драйверов и ПО для сканера и т.п. Для настройки параметров системы оптического распознавания необходимо проанализировать характеристики вводимого текста: качество оригинала, язык, лексику и т.д., и в зависимости от этого настроить параметры сканирования и распознавания. Кроме этого для определения оптимальных настроек можно осуществить предварительный ввод небольшого объема текста. В этом случае следует

проанализировать качества ввода и в зависимости от этого изменять настройки системы оптического распознавания.

Второй, основной этап – это собственно вод текста, он включает в себя четыре последовательные процедуры:

- Сканирование;
- Распознавание;
- Корректура, проверка и исправление ошибок;
- Сохранение.

В соответствии с этой последовательностью организована работа в OCR-системе FineReader. Последовательность действий циклически повторяется для каждой страницы или ряда страниц.

Исследование эффективности OCR-систем для ввода текста

Проведем исследование эффективности работы OCR-системы, оно включает следующие компоненты:

исследование временных затрат;

статистическое исследование количества ошибок;

анализ эффективности ввода текста.

Данное исследование проводилось на основе трех различных текстов:

Первый – современный текст хорошего качества. Желательно, чтобы в данном тексте не использовалась специфическая, сложная верстка нестандартные, декоративные шрифты. Рекомендуется, чтобы данный текст содержал только современную лексику.

Второй – современный текст плохого качества. Данный текст может содержать специфическую верстку. Это должен быть текст напечатанный не раньше 1960 г. В качестве такого текста может выступать машинописный текст или текст с плохим качество полиграфии. Он должен содержать современную лексику.

Третий – старинный текст. В качестве такого текста выступал текст VIII века, в котором используется соответствующая лексика, старинные шрифтовые гарнитуры и специфическая верстка.

Результаты исследования временных затрат представлены в следующей таблице.

Таблица 1. Сравнение временных затрат на этапы ввода страницы текста.

Вид текста	**Сканирование (сек.)**	**Распознавание (сек.)**
текст Словаря Академии Российской	52	32
современный текст хорошего качества	62	29
современный текст плохого качества	61	30

Анализируя полученные данные, следует отметить, что время сканирования и распознавания зависит от множества факторов: характеристик сканера, производительности системы (скорости работы процессора, объема оперативной памяти и т.д.), от особенностей текста, качества оригинала, шрифта и т.п. Разработчики сканеров указывают скорость сканирования, производительность своих устройств, однако эти теоретические данные имеют приблизительных характер из-за множества факторов, влияющих на процесс сканирования: навыки оператора, эргономика его рабочего места, характеристики страниц (размера, качества и т.п.), опции сканирования и т.д.

Проверка текста или корректура в большинстве случаев представляется наиболее трудоемкой и зависит от навыков оператора. После распознавания текста программа выделяет символы, форма которых вызвала сомнение при распознавании – «*неуверенно распознанные символы*». Кроме этого, текст проверяется на орфографические ошибки с помощью словаря, выявляя «*несловарные слова*».

Одним из основных параметров качества функционирования системы распознавания является точность распознавания, обычно выражаемая процентным соотношением количества верно распознанных символов относительно общего количества на странице или в документе:

$Ac_{расп_i} = \frac{100\% \cdot n_{верно\_расп_i}}{n_{общ_i}}$, где *пверно_расп i* и *побщ i* есть количество верно распознанных символов и общее количество символов на странице (в документе).

Согласно статистическим данным, полученным в результате исследования В случае рассмотрения страниц современного текста хорошего качества точность распознавания составляет 99,97 *%* , для текста плохого качества – 99,61%. При вводе текста XVIII в. (рассматривался текст Словаря Академии Российской) усредненное значение точности распознавания будет равно 86 %.

OCR-система FineReader обучена распознаванию стандартных шрифтов и не предназначена для распознавания декоративных шрифтов. Для повышения качества распознавания документа, набранного нестандартными шрифтами создан режим «*распознавания с обучением»*. Обычно в данном режиме распознаются 1-2 страницы, в результате чего создается пользовательский эталон, который в дальнейшем подключается для распознавания остальных страниц. При этом созданный эталон можно использовать только для распознавания текстов, использующих тот же шрифт и размер шрифта и отсканированных с тем же разрешением, что и документ, на основе которого данный эталон создавался.

В результате статистических исследований точность распознавания текста XVIII в. значительно выше в этом случае и составляет 95,16%. Дальнейшее улучшение качества распознавания можно добиться, используя пользовательский словарь и систему замен для типовых ошибок распознавания. Это дает возможность использовать OCR-системы для распознавания старинных текстов.

Литература

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

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.

Филиппович Ю.Н., Зеленцов И.А. Распознавание скорописи XVII века // Проблемы полиграфии и издательского дела. М., 2011. №3, С. 87-97.

Википедия. Свободная энциклопедия. Режим доступа: http://ru.wikipedia.org/

Sergey Pavlov. Ускорение ввода текстовой информации / Компьютеры и периферия - 19.01.2011/ Режим доступа: http://www.ichip.ru/stati/testy-i-obzory/2011/01/uskorenie-vvoda-tekstovoi-informacii.

Как выглядит китайская клавиатура / Блог компании ABBYY. 13.09.2010 / Режим доступа: http://habrahabr.ru/company/abbyy/blog/104083/