Лекция № 26

ИСПОЛЬЗОВАНИЕ СИСТЕМ ПРОВЕРКИ ОРФОГРАФИИ И ГРАММАТИКИ. ПРОГРАММЫ-ПЕРЕВОДЧИКИ. ВОЗМОЖНОСТИ СИСТЕМ РАСПОЗНАВАНИЯ ТЕКСТОВ. ГИПЕРТЕКСТОВОЕ ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ.

Использование систем проверки орфографии и грамматики.

Система проверки правописания (также спелл-че́кер от англ. spell checker) — представляет собой компьютерную программу, осуществляющую проверку заданного текста на предмет наличия в нём орфографических, пунктуационных, а также стилевых ошибок.

Найденные ошибки или опечатки отмечаются специальным образом - обычно для этого используется подчеркивание.

В некоторых случаях пользователю помимо указания на места возможных ошибок предоставляется возможность выбрать один из правильных вариантов написания, а также может выводиться комментарий, объясняющий каким образом следует поправить текст.

Проверка правописания может быть встроена как отдельная функция в некую программную систему, например, текстовый, почтовый клиент, электронный словарь или поисковую систему.

А также она может быть выполнена в виде самостоятельной программы.

В этом случае она обычно обладает возможностью интеграции с другими приложениями.

Такими возможностями, например, обладает GNU Aspell для Unix -подобных операционных систем, а также кроссплатформенная Hunspell.

История

Первые системы проверки правописания стали доступны в мейнфреймах в конце 1970-х.

Группа из шести лингвистов Джорджстаунского Университета разработала первую подобную систему для компании IBM.

На персональных компьютерах CP/M и TRS-80 это появилось в 1980, затем в 1981 появились первые пакеты для IBM PC.

Такие разработчики как Maria Mariani, Soft-Art, Microlytics, Proximity, Circle Noetics, и Reference Software быстро выпустили OEM-пакеты или конечные продукты на быстроразвивающийся рынок, в первую очередь для PC, хотя были предложения и для Apple Macintosh, VAX и Unix.

На PC эти системы проверки были автономными программами, многие из которых могли выполняться в режиме TSR изнутри пакетов работы с текстом (на компьютерах с достаточной памятью).

Однако рынок автономных пакетов просуществовал недолго, поскольку разработчики популярных программ работы с текстом (таких как WordStar и WordPerfect) в середине 1980-х включили системы проверки правописания в свои пакеты, главным образом лицензируемые от вышеупомянутых компаний, которые быстро развернули поддержку европейских языков, и в конечном счете, азиатских.

Но это всё больше усложняло разработку проверки правописания, особенно относительно агглютинативных языков, таких как венгерский или финский.

Хотя рынок программ по работе с текстом в таких странах как Исландия, возможно, не окупал инвестиции, компании наподобие WordPerfect, тем не менее, стремились вывести свои продукты на новые рынки.

Недавно проверка правописания переместилась из текстовых процессоров в веб-браузеры, например в Firefox 2.0, Google Chrome, Konqueror, Opera, почтовый клиентKmail и клиент системы мгновенных сообщений Pidgin также предлагают поддержку проверки правописания, используя GNU Aspell в качестве их механизма.

Mac OS X проверяет орфографию фактически во всех приложениях.

Компьютерные словари и системы машинного перевода текстов.

Знание хотя бы одного иностранного языка необходимо сегодня всем, как воздух.

В особенности пользователям: ведь избежать столкновения с английским языком при работе на компьютере, невозможно.

Помочь могут установленные на компьютере специализированные программы-переводчики.

Словари необходимы для перевода текстов с одного языка на другой.

Первые словари были созданы около 5 тысяч лет назад в Шумере и представляли собой глиняные таблички, разделенные на две части.

В одной части записывалось слово на шумерском языке, а в другой — аналогичное по значению слово на другом языке, иногда с краткими пояснениями.

Современные словари построены по такому же принципу.

В настоящее время существуют тысячи словарей для перевода между сотнями языков (англо-русский, немецко-французский и другие), причем каждый из них может содержать десятки тысяч слов.

В бумажном варианте словарь – это толстая книга с большим количеством страниц, поиск в нем довольно трудоемкий процесс.

Компьютерные словари (например, Lingvo, «Контекст») тоже содержат перевод слов, но они предоставляют дополнительные возможности.

Компьютерные словари в основном являются многоязычными, то есть дают пользователю возможность выбрать языки и направление перевода (например, англо-русский, испано-русский и другие).

Кроме основного словаря общеупотребительных слов, часто они содержат десятки специализированных словарей по областям знаний (техника, медицина, информатика и другие).

Они обеспечивают быстрый поиск словарных статей: «быстрый набор», когда в процессе набора слова возникает список похожих слов; доступ к часто используемым словам по закладкам; возможность ввода словосочетаний.

Некоторые компьютерные словари предоставляют пользователю возможность прослушивания слов в исполнении дикторов, носителей языка, то есть являются мультимедийными.

Кроме того, существуют системы машинного перевода, позволяющие переводить не только отдельные слова и словосочетания, но и целый многостраничный документ (текст) с высокой скоростью (одна страница в секунду), а также Web-страницу»на лету» - в режиме реального времени.

Лучшими среди российских систем машинного перевода считаются PROMT и «Сократ».

Системы машинного перевода осуществляют перевод текстов, основываясь на формальном «знании» языка (синтаксиса языка) и использовании словарей.

Программа-переводчик сначала анализирует текст на одном языке, а затем конструирует этот текст на другом языке.

Современные системы машинного перевода используются для перевода технической документации, деловой переписки и других специализированных текстов, но они неприменимы для перевода художественной литературы, так как им недоступны аллегории, метафоры и другие элементы художественного творчества человека.

Системы оптического распознавания документов.

Переход от бумажного документа к электронному состоит из двух этапов.

- Сканирование. С помощью сканера получается изображение страницы текста в графическом файле.
- Распознавание текста. Для преобразования элементов графического изображения в последовательности символов используются системы оптического распознавания символов.

Запустив такую систему, сначала надо распознать структуру размещения текста на странице: выделить колонки, таблицы, изображения и так далее.

Далее текстовые фрагменты графического изображения страницы преобразовываются в текст.

Существует два метода распознавания:

1. Метод сравнения с растровым шаблоном.

Используется, если исходный документ имеет типографическое качество (достаточно крупный шрифт, отсутствие плохо напечатанных символов и исправлений).

Сначала растровое изображение страницы разделяется на изображения отдельных символов. Затем каждый из них последовательно накладывается на шаблоны символов, имеющихся в памяти системы, и выбирается шаблон с наименьшим количеством отличных от входного изображения точек.

2. Метод распознавания символов по наличию в них определенных структурных элементов (отрезков, колец, дуг и других).

Используется при распознавании документов с низким качеством печати (машинописный текст, факс и так далее).

Любой символ можно описать через эти элементы и значения параметров их взаимного расположения.

Например, буквы «Н» и «И», состоят из трех отрезков, два из которых расположены параллельно друг другу, а третий соединяет эти отрезки.

Различаются же эти буквы величиной углов, которые образуются третьим отрезком с двумя другими.

Современные системы оптического распознавания (FineReader, CuneiForm) используют оба метода и являются «самообучающимися» (то есть для каждого конкретного документа они создают соответствующий набор символов, поэтому скорость и качество распознавания постепенно возрастают).

Для распознавания бланков (форма), заполненных рукопечатным текстом (данные вводятся в поля печатными буквами от руки), используются системы оптического распознавания форм.

Эта задача сложнее, так как печатные символы, написанные от руки разными людьми, сильно отличаются, к тому же необходимо определить, к какому полю относится распознаваемый текст.

В последнее время создаются системы распознавания рукописного текста, однако они очень несовершенны.

Гипертекстовое представление информации

Для связи основных разделов и понятий в тексте используется гипертекст.

Гипертекст позволяет структурировать документ путем выделения в нем слов-ссылок (гиперссылок).

При активизации гиперссылки, например, щелчком мыши, происходит переход на фрагмент в тексте, заданный в ссылке.

Гиперссылка состоит из двух частей:

- указатель ссылки – это объект (фрагмент текста или рисунок), который визуально выделяется в документе (обычно синим цветом и подчеркиванием);
- адресная часть – название закладки в документе, на которую указывает ссылка (закладка – это элемент документа, которому присвоено уникальное имя).

Указателем ссылки и закладкой может быть фрагмент текста, графическое изображение, управляющий элемент.

Такая гипертекстовая структура используются в документах различных типов.

В Интернете они образуют Всемирную паутину, связывающую Web-страницы на миллионах серверов в единое целое.

Как создать гипертекстовый документ, содержащий, например, гиперссылки на три закладки, которые, в свою очередь являются гиперссылками на начало текста?

1 этап.

Создайте документ, содержащий обычный текст.

Выделите фрагмент текста, которому следует назначить закладку.

Затем введите команду [Вставка-Закладка…].

Появится диалоговая панель Закладки, в ее поле Имя закладки: введите имя, которое должно начинаться с буквы и нажмите кнопку Добавить.

2 этап.

Выделите фрагмент текста, который будет указателем гиперссылки.

Теперь введите команду [Вставка-Гиперссылка…].

На диалоговой панели Вставка гиперссылки в окне Выберите место в документе: выберите имя закладки и нажмите кнопку ОК.

3 этап.

Аналогично создайте еще две гиперссылки на закладки и три гиперссылки с закладок на начало текстового документа.

Вопросы

1. Что такое система проверки правописания?
2. Когда появились первые системы проверки правописания?
3. Для каких целей служат компьютерные словари?
4. Каковы дополнительные возможности компьютерных словарей?
5. В чем отличие систем машинного перевода?
6. Какие системы машинного перевода Вы знаете?
7. Какие этапы включает переход от бумажного документа к электронному и в чем они заключаются?
8. Какие методы распознавания Вы знаете?
9. Что такое гипертекст и для чего он нужен?
10. Из каких частей состоит гиперссылка?