Тема: ОБЛАСТИ ПРИМЕНЕНИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В ЛИНГВИСТИКЕ

План

1. Автоматический анализ и синтез звучащей речи
2. Технологии обработки текста
3. Автоматическое распознавание текста
4. Автоматическое аннотирование и реферирование текста
5. Автоматический анализ и синтез текста

1. Автоматический анализ и синтез звучащей речи

Одним из первых важных шагов использования информационных технологий в лингвистике является дигитализация текстов — переведение языкового материала, существующего в печатном или устном виде, в цифровую форму.

При *автоматическом анализе* звучащей речи она преобразуется в печатный текст, над которым можно производить дальнейшие операции.

Автоматический синтез звучащей речи представляет собой обратный процесс преобразования печатного текста, существующего в цифровой форме, в звучащий текст на естественном человеческом языке.

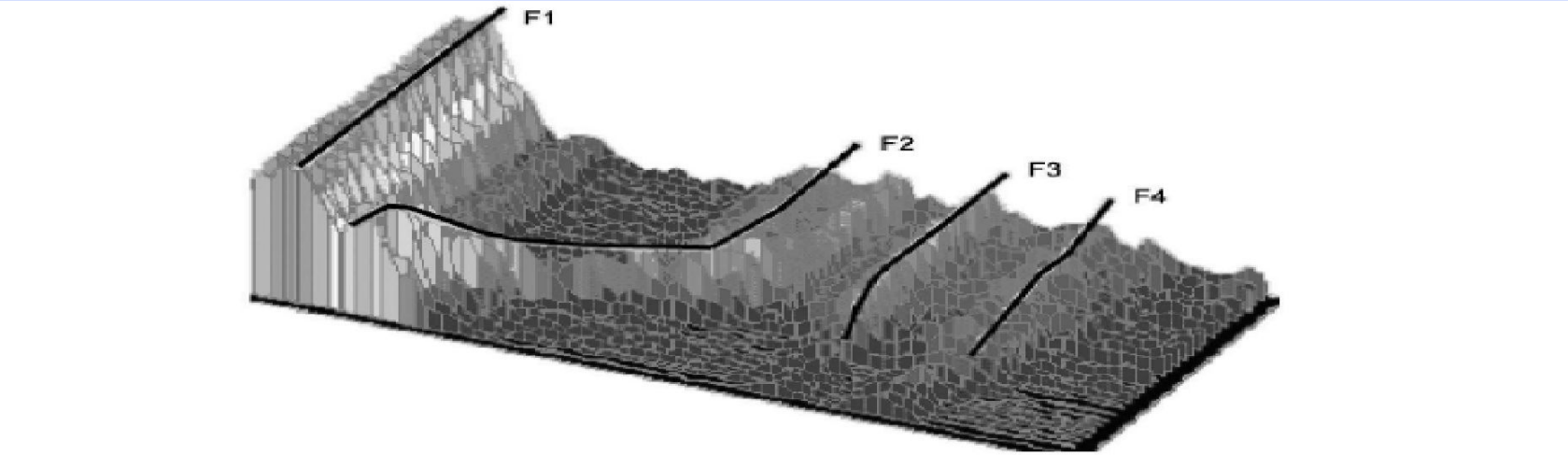
Процесс автоматического анализа речи включает следующие этапы:
1) ввод звучащей речи в компьютер с помощью микрофона.
2) выделение компьютерной программой в звуковом потоке отдельных знаков.
3) идентификация выделенных знаков звучащей речи со знаками языка.

Минимальными знаками звучащей речи являются звуки, производимые артикуляторным аппаратом человека. Каждый звук имеет свои акустические характеристики (высота, частота колебаний звуковых волн и т.д.), которые можно измерить специальными приборами (например, осциллографом).

В основе пофонемного распознавания звуков речи лежит анализ:

1) длительности и динамики звучания,
2) чередования акустического сигнала и пауз.

В настоящее время наиболее доступной формой точной фиксации звучащей речи (в том числе ее тембра и динамики) становится спектрограмма — фотографическое изображение звуков.



Спектрограмма русских звуков и и у

Задачей автоматического анализа звучащей речи при использовании спектрограмм становится перевод спектрограмм в фонологическую транскрипцию.

В итоге процесс автоматического анализа речи включает ввод слов в компьютер через микрофон, начитанных разными дикторами, их спектральную обработку и создание набора признаков, своеобразного образца слова, который выступает знаком языка.

Примеры программ, в которых применяются средства автоматического анализа речи:

• программы голосового управления компьютером и бытовой техникой *Voice Navigator* и *Truffaldino*;

• комплекс голосового управления мобильным телефоном *DiVo*;

• программный модуль *Voice Key* для идентификации личности по парольной фразе длительностью 3-5 секунд;

• программы диктовки текста на английском языке: *Voice Type Dictation*, *Dragon Dictate*;

на русском языке:

Комбат к Диктограф;

• система распознавания речи, встроенная в *Microsoft Office XP* (работает только с английским языком);

• голосовой поиск (например, в поисковой системе *Google).*

Автоматически синтезируется речь в следующих ситуациях:
• называние текущего времени по телефону,
• объявление остановок в метро,
• называние остатка средств на счету и другие услуги мобильных операторов,
• оповещение систем гражданской безопасности и т.д.

Автоматический синтез (генерация) речи в настоящее время осуществляется путем составления слов и фраз из заранее записанных диктором образцов отдельных звуков (метод *компилятивного синтеза)* или путем моделирования речевого тракта человека *(формантно-голосовой* метод).

2. Технологии обработки текста

Представление текста

Под “текстовым” понимают такое представление информации, в котором она представлена в виде записи слов (логических элементов) некоторого языка и доступна для чтения человеком.

Правило сопоставления кодов и символов, входящих в алфавит, называется **кодировкой**.

стандарт кодирования
таблица кодировки ASCII
стандарт кодирования Юникод (Unicode)

Понятие “текстового файла” не предусматривает строго заданного формата или расширения. Тем не менее, помимо характерной для той или иной ОС таблицы кодировки, в текстовых файлах могут применяться три основных способа деления текста на строки (абзацы):
1. Windows (DOS) — символы “Возврат каретки” + “Перевод строки” (CR+LF).
2. Unix — символ “Перевод строки” (LF).
3. MacOs — символ “Возврат каретки” (CR).

Правила машинописного набора текста

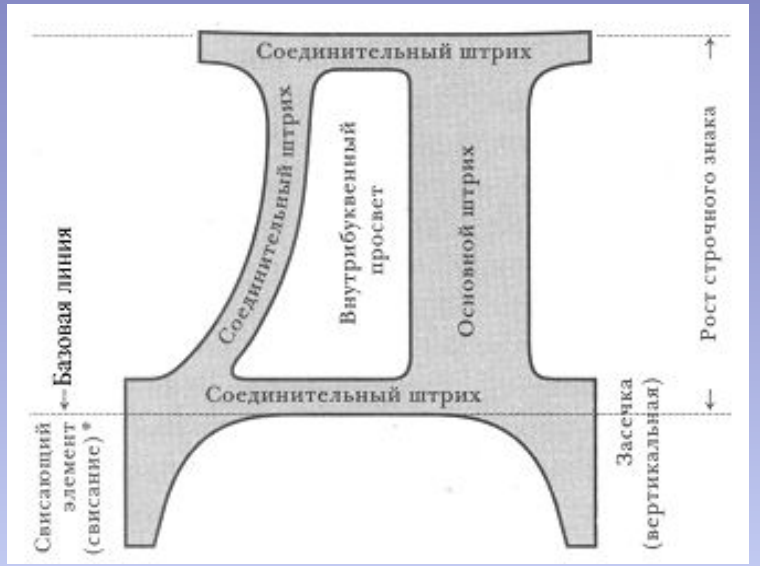
Для облегчения анализа и последующего преобразования текста при его наборе в самых различных случаях рекомендуется соблюдать общие правила машинописного набора:

1. Все слова разделяются пробелом, и только одним пробелом.
2. Знаки препинания примыкают к предыдущему слову.
3. Скобки и кавычки всех видов примыкают к первому и последнему слову заключенного в них текста.
4. Текст разрывается только в конце абзаца.
5. Большие форматированные пробелы делаются вставкой символа табуляции, а не несколькими пробелами подряд.

Пример программных продуктов — текстовых редакторов:

Блокнот, Notepad++, PSPad, vi

Оформление текста



Шрифт характеризуется рядом параметров:

1. Рисунок шрифта — графические особенности, определяющие общность шрифта и его отличие от всех других.
2. Кегль (кегель) — размер шрифта — предельная высота большой буквы и окружающих ее пробелов (термин введен для описания высоты площадки литеры при наборе с помощью типографской кассы). Чаще всего задается в типографских пунктах (1 пункт = 1/72 дюйма = 0,375 мм). По историческим причинам некоторые размеры имеют собственные названия: 8 пт — “петит”, 9 пт — “боргес”, 10 пт — “корпус”, 12 пт — “цицеро”.
3. Начертание — шрифт с общим рисунком, но какими-либо отличительными признаками: более жирный, наклонный, разреженный. Иногда параметр плотности шрифта (светлый, полужирный, жирный) отделяют от начертания.
4. Часто как параметр задается подчеркивание или зачеркивание шрифта, или его написание как индекса — с уменьшением размера и подъемом/спуском относительно текущей строки.

Совокупность всех возможных размеров и вариантов написания шрифта называется **гарнитурой**. Гарнитуры имеют имена, по которым часто называют и конкретный шрифт.

По общим чертам рисунка различают три основных вида шрифтов:

1. Рубленые шрифты.
2. Антиквенные шрифты.
3. Акцидентные (оформительские) шрифты.

Существует несколько основных способов описания шрифтов:

1. *Растровые* шрифты.
2. *Векторные* шрифты.
3. *Контурные* шрифты.

Структурирование теста

Для оформления абзаца используют несколько параметров:
1. **Выравнивание** (выключка) — *правило расположения букв в строке абзаца*. Видов выравнивания четыре: по левому краю, центральное, по правому краю и по ширине полосы набора.
2. **Отступы** от краев полосы набора.
3. **Абзацный отступ** (красная строка) — *положение первой строки абзаца*.
4. **Интервалы.** Различают **межстрочное расстояние** — задается множителем размера шрифта (одинарный, полуторный, двойной интервал) — и **промежутки** до и после абзаца.
5. **Буквица** — *крупная выступающая первая буква абзаца*. Часто задается не просто более крупным размером буквы, но и буквой другого рисунка.

Абзацы размещаются в рамках **полосы** — *выделенного участка страницы, как правило, прямоугольной формы, в котором размещаются текст и иллюстрации*.

Примеры программных продуктов
Microsoft Word, OpenOffice Writer, StarOffice Word

Автоматизированная обработка текста

Расшифровка или уточнение значений слова

Системы автоматизированной обработки текста

Примеры программных продуктов

Системы локального поиска: Следопыт, Google Desktop, Microsoft Office Find
Системы и утилиты автоматизированной обработки текста: Grep, lexx, yacc
Словари: Abbyy Lingvo, Multilex
Автоматизации перевода: Promt

Специальные тексты

Под специальными текстами подразумеваются тексты, содержащие математические, химические или другие формулы, сложные схемы и специфические обозначения, используемые в научных, учебных и технических публикациях и документах.

При подготовке научных, технических и учебных текстов часто используется свободно доступная система подготовки публикаций TeX

Примеры программных продуктов

Макропакеты TeX: LaTeX, MikiTeX, AMSTeX
Специализированные редакторы: MathType (его облегченная версия входит в пакет MS Office под названием Equation), ScientificLetter, ChemWindow, ISIS Draw.

3. Автоматическое распознавание текста

Для ввода информации в компьютер используются специальные устройства — клавиатура, мышь и др., но наиболее удобным инструментом для ввода большого количества печатных текстов является сканер.

Сканер — это устройство ввода, работающее по принципу фотоаппарата, т.е. позволяющее компьютеру «увидеть» текст в виде фотографии. Чтобы компьютер смог «понять» этот текст, т.е. перевести графическое (растровое) изображение символов в текстовую форму, при которой у каждого символа имеется свой двоичный код (например, в системе кодировок ASCII), требуется программа автоматического распознавания символов (англ. OCR = *Optical Character Recognition).*

Наиболее известными и полифункциональными являются OCR-программы *Fine Reader* (компании *Abey)* и *Cunei Form* (фирмы *Cognitive Technologies).*

С другими программами автоматического распознавания текстов можно познакомиться, например, в интернет-ресурсе, размещенном по адресу http://kompkimi.ru/?p=617

4. Автоматическое аннотирование и реферирование текста

Рефераты и аннотации составляются вручную, например самим автором исходного текста или библиографическим работником, или автоматически, с помощью специальных компьютерных программ.

Для обработки большого массива текстов за минимальное количество времени требуется привлечение автоматических средств для решения задачи реферирования и аннотирования текстов.

В зависимости от жанра исходного текста (монография, статья, патент и др.) и от предметной области (медицина, химия, лингвистика и т.д.) заданные элементы реферата могут различаться. Так, для научных рефератов дополнительно к названным выше элементам реферата прибавляется краткое изложение сути, практической апробации и перспектив исследований.

Различают следующие виды рефератов:

• *связный текст* — новое текстовое образование, порождаемое на основе логико-смыслового анализа исходного текста;

• *реферат-клише* — модификация заданной клишированной структуры, пустые ячейки которой заполняются после анализа заданного текста;

• *квазиреферат* — перечень наиболее информативных предложений текста.

В большинстве программ, направленных на автоматическое составление краткого содержания текста, можно задать разную степень компрессии текста, т.е. одна и та же программа создает как развернутые рефераты, так и краткие аннотации. В связи с этим в отношении автоматического процесса составления краткого содержания текста обычно используется двойное обозначение: автоматическое реферирование и аннотирование текста.

Главными смысловыми единицами исходного текста выступают ключевые слова, ключевые словосочетания и ключевые предложения. *Ключевое слово* — знаменательное слово, относящееся к основному содержанию текста и повторяющееся в нем несколько раз. *Ключевое словосочетание* — сочетание слов, среди которых есть одно или несколько ключевых. Ключевое предложение — предложение, которое содержит несколько (два и более) ключевых слов.
По способам выделения из исходных текстов ключевых словосочетаний и предложений различаются следующие методы автоматического реферирования и аннотирования текстов:
1) статистические,
2) позиционные,
3) логико-семантические.

Наиболее простыми системами автоматического реферирования и аннотирования является функция *Auto Summarize в MS Word,* системы *Intelligent Text Miner, OracleContext* и *Inxight Summarizer* (компонент поискового механизма *Alta Vista)* (IBM).

Примеры систем автоматического реферирования и аннотирования текстов:

- ОРФО 5.0 (компания «Информатик»): программа включает функцию автоматического аннотирования русских текстов;
- «Либретто» (компания «МедиаЛингва»): программа встраивается в Word и обеспечивает автоматическое реферирование и аннотирование русских и английских текстов;
- поисковая система «Следопыт», которая включает средства автоматического реферирования и аннотирования документов;
- программы Extractor и TextAnalyst

5. Автоматический анализ и синтез текста

При *автоматическом анализе* текст последовательно преобразуется в его лексемно-морфологические, синтаксические и семантические представления, понятные компьютеру. Обратный процесс преобразования лексемно-морфологических, синтаксических и семантических компьютерных представлений в текст на естественном языке называется *автоматическим синтезом текста.*

Автоматический анализ текста включает ряд этапов:
1) *графематический анализ:* выделение границ слов, предложений, абзацев и других элементов текста (например, врезок в газетном тексте);
2) *морфологический анализ:* определение исходной формы каждого использованного в тексте слова и набора морфологических характеристик этого слова;
3) *синтаксический анализ:* выявление грамматической структуры предложений текста;
4) *семантический анализ:* определение смысла фраз.

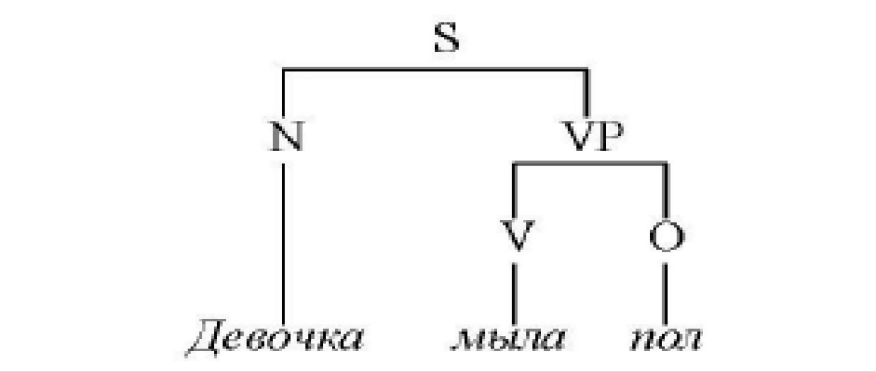
При **морфологическом анализе** каждое использованное в тексте слово возводится к его исходной форме и определяется набор морфологических характеристик текстовой формы слова: часть речи; род, число и падеж для существительных, число и лицо для глаголов и т.п.

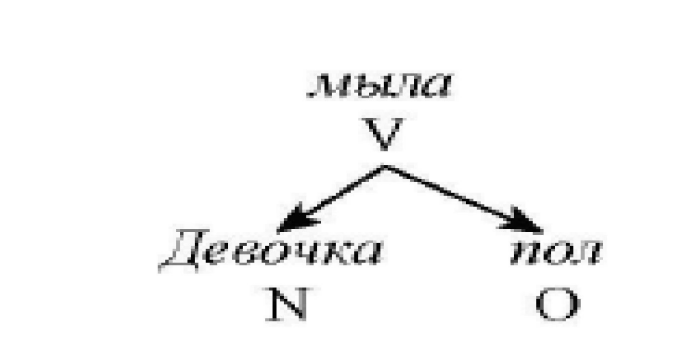
Машинные окончания — элементы, описывающие формоизменение конкретной лексемы и представляемые в виде парадигм.

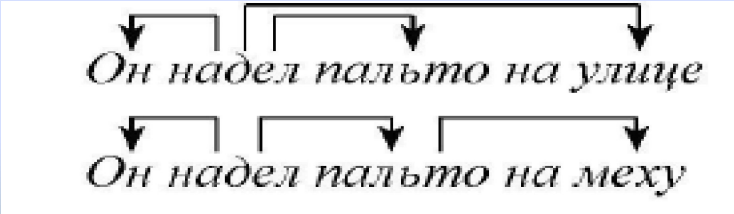
Девочка {девочка = S, жен, од = им, ед}
мыла {мыть = V, несов = прош, ед, изъяв, жен, перех | мыло = S, сред, неод = им, мн | = S, сред, неод = род, ед | = S, сред, неод =вин, мн}
пол {пол = S, муж, неод = им, ед | = S, муж, неод = вин, ед | = A, кратк, муж, им, ед}.

Морфологический анализ включает в себя следующие этапы:
1) нормализация словоформ, имеющая вид лемматизации, т.е. сведения различных словоформ к некоторому единому представлению — к исходной форме (лемме) или стемминга, т.е. возведения разных словоформ к одной квазиоснове;
2) частеречный тэгинг, т.е. указание части речи для каждой словоформы в тексте;
3) полный морфологический анализ — приписывание грамматических характеристик словоформе.

При **синтаксическом анализе** необходимо определить роли слов в предложении и их связи между собой. Результатом этого этапа автоматического анализа является представление синтаксических связей каждого предложения в виде моделей, например в виде дерева зависимостей.







Семантический анализ представляет собой, пожалуй, наиболее сложное направление автоматического анализа текста. В этом случае требуется установление семантических отношений между словами в тексте, объединение различных языковых выражений, относящихся к одному и тому же понятию, и т.п.
Для семантического анализа предложений используются падежные грамматики и семантические падежи (валентности). В этом случае семантика предложения описывается через связи главного слова (глагола) с его семантическими актантами.

- Лексическая омонимия: совпадение звучания и/или написания слов, не имеющих общих элементов смысла, например, рожа — лицо и вид болезни.
- Морфологическая омонимия: совпадение форм одного и того же слова (лексемы), например, словоформа пол соответствует именительному и винительному падежам существительного пол.
- Лексико-морфологическая омонимия (наиболее частый вид омонимии): совпадение словоформ двух разных лексем, например, мыла — глагол мыть в единственном числе женского рода прошедшего времени и существительное мыло в единственном числе, родительном падеже.
- Синтаксическая омонимия: неоднозначность синтаксической структуры, имеющей несколько интерпретаций.

Автоматический синтез представляет собой процесс производства связного текста, отдельные этапы которого являются теми же, что и при морфологическом анализе, но применяются в обратном порядке: сначала осуществляется семантический синтез, затем синтаксический, морфологический и графематический.

Семантический *синтез* представляет собой переход от смысловой записи фразы к ее синтаксической структуре; синтаксический — переход от синтаксической структуры фразы к представляющей фразу цепочке лексико-грамматических характеристик словоформ; *лексико-морфологический* — переход от лексико-грамматической характеристики к реальной словоформе. При ***морфологическом*** синтезе по нормальной форме слова и его параметрам программа находит соответствующую словоформу.

Графематический синтез объединяет слова в единый текст, следит за соответствием фрагментов входного текста фрагментам выходного.

Одной из первых компьютерных программ, синтезирующих письменный диалог на английском языке, явилась программа американского ученого Джозефа Вейценбаума «Элиза».

СПАСИБО ЗА ВНИМАНИЕ!