**Лекционные материалы по КТЛ**

**ТЕМА2 Обработка текстов и информационные технологии**

В настоящее время в связи со стремительным увеличением количества текстовой информации, в том числе в сети Интернет, возникает неооходимость в оыстрой разра-ботке прикладных программных систем (приложений) для автоматической или авто-матизибованной обработки текстов на естественном языке (Ед-текстов). Примерами такой обработки является соор и фильтрация данных из различных источников, извле чение знаний, реферирование, аннотирование и т. п. Разработка приложений для реше- первую очередь необходимость интег- рации большого числа программных компонентов, реализующих алгоритмы обработки ЕЯ-текста и работающих на различных уровнях (обработка слов, предложений, абзацев и т. п.). Например, задача требует графематического пологического анализа. вылеления опрелеле на слова и препожения). MOD- конструклий естественного языка на синтаксическом уровне. Последующеи обработки результатов, а также преобразования текстовых данных между различными форматами и прелставлениями

Другую сложность представляет организация внутреннего представления обрабатываемого Ея-текста, поскольку оно должно поддерживать две его важные особенности: естественную иерархию составляюших текст элементов (аозапев. прелложении. слов) и свойственную естественному языку множественность интерпретации на каждом уровне например, словоформа «мыла» на морфологическом уровне имсет две интерпрета-пии сушествительное или глагол)

***Реферирование текста*** - сокращение его ооъема и получение его краткого изложения - реферата (свернутого содержания), что ускоряет поиск в коллекциях доотий реферат может составлев пля нескольих олизких по теме доКуменТОв.

основным методом автоматического реферирования до сих пор является отбор наиболее значимых предложений реферируемого текста, для чего обычно выделяют ключевые слова текста и рассчитывают коэффициент значимости предложении текста

Выбор значимых препложений осложняется анафорическими связями препложений. разрыв которых нежелателен - для решения этой проблемы разрабатываются определенные стратегии отоора предложении

ние его аннотации. В простейшей форме аннотация

Спетуюшее приклалное направление , которое стоит упомянуть, автоматическая генерация текстов на ЕЯ, которую, можно считать подзадачей уже рас- документов - патентных формул, инструкций по эксплуатации технических делий или программных систем исходя из их спепификапии на формальном языке. Для решения этой задачи применяются довольно подробные модели языка.

Всё большую актуальность приобретает прикладная задача, часто относимая к направлению Text Mining, - извлечение информации из текстов, или Information Extraction, востребованное при решении задач экономической и производственной аналитики. Для этого осуществляется выделение в тесте ЕЯ определенных объектов: именованных сущностей (имен, персоналий, географических названий), их отношений и связанных с ними сооытий, что, как правило, реализуется на основе частичного син-таксического анализа текста, позволяющего выполнять обработку потоков новостей от информационных агентств. поскольку задача достаточно сложна не только теоре-тически, но и технологически, создание значимых систем извлечения информации из текстов осушествимо в рамках коммерческих компании.

**Компьютеризация обучения -** многоцелевое использование средств электронной вычислительной техники в учеоном процессе. компьютер является одним из компонентов информапионной технологии. поэтому вместо термина «компьютерное обу-чение» часто используется в том же значении термин «информатизация обучения».

**Компьютерное обучение** имеет целью подготовить подрастающее поколение к жизни в информатизированном оошестве. то есть в оошестве. гле значительныи улельный вес унФОрмАциОнная Технология совокупность средств и методов, с помощью - это технологии, связанные с телекоммуникациеи, то есть «оощением на расстоянии» посредством всемирнои сети интернет, направленные на интеграцию суоъектов в единое информационное пространство с целью получения максимального объема информации для саморазвития

**Тема 3. Психологические аспекты компьютеризации обучения**

Развитие новых технических средств и информационно-коммуникативных систем и технологий, вхождение персональных компьютеров в повседневную практику, естественно, востребовали компьютеризацию обучения. Наиболее важные области применения в образовании информационно-компьютерных технологий:

– управление образованием (базы данных, электронные журналы, сетевые сообщества, дистанционные формы обучения и др.);

– информатика и вычислительная техника как цель обучения (подготовка специалистов в области информационно-компьютерных технологий по разработке и эксплуатации компьютерной техники, программному обеспечению, компьютерной безопасности; подготовка пользователей, в том числе обучение компьютерной грамотности преподавателей и учеников средних школ);

– компьютерная техника и информационные технологии как средство обучения, развития и воспитания.

Последнее направление стало считаться наиболее значимым и наиболее активно развивается. Современные технические средства обучения выступают как удобный и универсальный носитель аудиовизуальной обучающей информации (учебники, энциклопедии, атласы и различные справочники, содержащие текстовую, образную, звуковую и видеоинформацию, используемая как на уроке, так и в индивидуальном обучении).

С внедрением информационно-компьютерных технологий меняются форма и содержание дистанционного образования, становясь интерактивной формой взаимодействия участников образовательного процесса, опосредованного технологическими средствами.

Компьютер выступает как средство повышения эффективности, оперативности и объективности оценки знаний учеников. Однако оценка может быть проведена в отношении только четко формализированных и алгоритмизированных параметров.

С привлечением информационно-компьютерных технологий можно решать различные психолого-педагогические задачи, связанные с повышением качества учебно-воспитательного процесса в образовательных учреждениях:

– повышение информационной обеспеченности участников образовательного процесса, увеличение и расширение доступа, накопления, обмена и тиражирования информации для учеников и учителей;

– усиление учебной мотивации и привлекательности обучения для учащихся при грамотном использовании информационных технологий и технических средств обучения;

– повышение самостоятельности обучения, возможности обучения без непосредственного участия педагога путем выполнения домашних заданий с необходимыми компонентами проверки правильности их выполнения;

– повышаются возможности индивидуализации обучения (индивидуализация темпа предъявления заданий и перехода к последующей, возвращения к предыдущей; выбор тем и заданий с учетом индивидуальности и знаний конкретного ученика; различные формы текущего и итогового контроля успеваемости, повышение объективности этого контроля).

Безусловно, компьютеризация обучения – процесс необратимый и прогрессивный. Однако те проблемы, которые носят именно субъективный характер (компьютерофобия учителей, игровая и Интернет-аддикция учащихся) требуют специального методического и психолого-педагогического сопровождения.

**Продолжение ….**

Информационные технологии включают программированное обучение, интеллектуальное обучение, экспертные системы, гипертекст и мультимедиа, микромиры, имитационное обучение, демонстрации. Эти частные методики должны применяться в зависимости от учебных целей и учебных ситуаций, когда в одних случаях необходимо глубже понять потребности учащегося, в других – важен анализ знаний в предметной области, в третьих – основную роль может играть учет психологических принципов обучения. Однако, растущее применение компьютеров во всех сферах человеческой деятельности порождает новые проблемы. В отечественной и зарубежной психологии выделяют следующие психологические феномены, связанные с освоением человеком новых информационных технологий:

персонификацию, «одушевление» компьютера, когда компьютер воспринимается как живой организм;

потребность в «общении» с компьютером и особенности такого общения;

различные формы компьютерной тревожности; вторжение во внутренний мир человека, ведущее к возникновению у некоторых пользователей экзистенциального кризиса, сопровождающегося когнитивными и эмоциональными нарушениями. При этом может происходить переоценка ценностей, пересмотр взглядов на мироздание и свое место в мире.

Одной из негативных сторон информатизации является появление у некоторых людей (и не только пользователей) компьютерной тревожности. В настоящее время не существует четкого определения, этого понятия, нет и общепризнанных методов профилактики и лечения компьютерной тревожности. Большинство психологов подразумевают под нею страх, возникающий при работе на компьютере или при размышлении о ней. Установлено, что уровень компьютерной тревожности позволяет предсказать успешность обучения работе на компьютере.

*Г.Маркулидес показал, что наличие компьютерной тревожности значительно снижает компьютерную грамотность и интерес к работе на компьютере. Люди, испытывающие высокую тревогу при выполнении какого либо задания на ЭВМ, как правило имеют отрицательное отношение к компьютеру. С другой стороны, как указывают Д. Кэмпбелл и К. Перри, отрицательные эмоции в некоторых случаях могут стимулировать рост активности, стремление выполнить задание как можно лучше и приводить тем самым к повышению успешности деятельности.*

У учащихся и студентов компьютерная тревожность возникает зачастую как реакция на страх получить плохую отметку, показаться неспособным или глупым по сравнению с другими обучающимися. Преподаватели и школьные учителя также зачастую сталкиваются с серьезными трудностями в процессе освоения навыков работы на компьютере. У них может иметь место опасение, что их рабочие места займут компьютеры или педагоги, лучше владеющие компьютером. Одним из важным факторов тревожности является также осознание ими того, что их студенти владеют компьютером намного лучше, чем они сами. Одной из разновидностей компьютерной тревожности является «компьютерный стресс». В работе А.М. Боковикова изучается стресс, связанный с компьютеризацией профессиональной деятельности, определяются факторы стрессоустойчивости в процессе адаптации человека к работе на компьютере. Основным стрессогенным фактором при работе на компьютере он считает утрату контроля над деятельностью, когда ситуация взаимодействия с компьютером выходит из-под контроля.

Стрессоустойчивость определяется в первую очередь свойствами личности. При этом активность, инициативность, уверенность в себе, эмоциональная стабильность и оптимистическая оценка ситуации – основа устойчивости к стрессу.

**В числе отрицательных последствий** длительного применения информационных технологий выделяют так же аутизацию (уход от реальности, синдром зависимости от компьютера и особенно от Интернета). Сужается круг интересов, сокращается участие в значимых видах деятельности либо происходит полный отказ от нее.

**В числе положительных моментов** применение информационных технологий в образовании называется возможность самостоятельного обучения с открытым доступом к обширным информационным ресурсам, наличие обратной связи. С помощью компьютера учащийся может очутиться в самом разном окружении, требующем от него творческого подхода.

**Тема4: ОБЩИЕ ПРИНЦИПЫ РЕШЕНИЯ**

**ЛИНГВИСТИЧЕСКИХ ЗАДАЧ**

**МЕТОДОМ МОДЕЛИРОВАНИЯ**

**ОСНОВНЫЕ ЭТАПЫ РЕШЕНИЯ ЗАДАЧИ**

**План:**

**1. Введение 3.** **Разработка модели**

**2. Моделирование 4. Анализ результатов работы компьютерной программы.**

Как было отмечено выше, основным методом решения линг-вистических задач является метод моделирования. Моделирование - процесс творческий. И модель в любом случае отражает наиболее важные и существенные особенности моделируемого объекта, процесса или явления (оригинала модели).

**В общем, процесс моделирования на компьютере включает сле-дующие этапы [165, 24-341:**

**1) постановка задачи: 2) разработка модели; 3) проведение компьютерного эксперимента; 4) анализ результатов работы компьютерной модели.**

В свою очередь, каждый этап моделирования может быть представлен в виде некоторой последовательности конкретных действий.

**Говоря о постановке задачи, выделяют следующие действия: 1) описание решаемой задачи; 2) формулирование цели моделирования; 3) анализ оригинала модели.**

Описание решаемой задачи может быть представлено словесно, в виде формулы или нескольких формул, в виде таблиц, графиков и т.п. Цели моделирования могут быть различными объяснение сути некоторого явления или процесса, создание объектов с заранее заданными свойствами, определениепоследствий воздействия на некоторый объект, процесс или явление, принятие правильного решения и т.п.

При анализе оригинала модели (некоторого объекта, процесса или явления) в нем прежде всего выделяются наиболее важные, существенные черты и свойства. Если оригинал представляет собой совокупность более мелких составляющих, то на этапе анализа его расчленяют, разделяют на ряд более мелких объектов и выявляют отношения между этими составляющими. При компьютерном моделировании стараются выделять такие черты и свойства оригинала, которые мог бы легко опознать компьютер. Эти черты и свойства называют формальными.

На этапе непосредственной разработки модели, опираясь на результаты детального анализа ее оригинала, создается алгоритм решения задачи. Проведение компьютерного эксперимента связано с созданием на основе алгоритма компьютерной программы на каком-либо алгоритмическом языке и отладкой этой программы (устранением ошибок программирования). Наконец, в процессе анализа результатов работы компьютерной модели выявляются логические ошибки в самой компьютерной программе и алгоритме (формуле, графике и т. п.), который послужил основой компьютерной модели. Такие ошибки исправляются внесением в алгоритм и программу соответствующих изменений.

Рассмотрим подробнее эти этапы и соответствующие им действия в процессе создания конкретных воспроизводящих инженерно-лингвистических моделей (ВИЛМ) Поскольку решение лингвистических задач на ПК имеет свою специфику, введем некоторые основные понятия, связанные с этой процедурой. Все модели, создаваемые ниже, будут понятны, если принять единые определения основных лингвистических единиц. Информация поступает в компьютерную память в виде цепочки символов, каждый из которых занимает один байт памяти.

Цепочка буквенных символов, находящаяся в тексте между двумя знаками пробела, называется словоупотреблением. Следовательно, компьютер читает не слово, а словоупотребление. Словоупотребление, находящееся вне предложения или текста, будем называть словоформой. Несколько словоформ, имеющих одно и то же лексическое значение, образуют слово, или лексему.

Решению каждой лингвистической задачи должен предшествовать тщательный анализ соответствующего лингвистического материала: конкретных букв (звуков), слов, словосочетаний, предложений, абзацев, текстов. Для поиска наиболее важных, существенных признаков, положенных в основу алгоритма решения сформулированной выше задачи, необходимо изучить большое число русских глаголов в самых разных грамматических формах и попытаться найти какие-то черты, которые отличают глагол в неопределенной форме от глаголов в других грамматических формах. Причем эти признаки должны быть формальными, т. е. любой человек, даже не знающий русского языка, ориентируясь на них, сможет решить поставленную задачу.

Анализируя, например, слова петь, беречь, косила, переделал, работаю, идти, пошел, толочь, бегать, пройду, читать, нести, прочитаю, ходить, сделаю, хранить, уехать, стеречь, подарю, берегу, играть и т. п., можно сделать вывод, что русский глагол в инфинитиве заканчивается буквосочетаниями -ть, - чь, -ти'. Это и есть те формальные признаки, на которые будет опираться модель.

**Разработка модели.**

Как было сказано выше, любая воспроизводящая инженерно-лингвистическая модель представляется в виде словесно-графического алгоритма, опирающегося на выделенные в процессе анализа оригинала модели признаки. Как же отыскать в группе из 10 русских глаголов глаголы в неопределенной форме? Очевидно, необходимо выполнить следующую цепочку «умственных» действий:

1. Взять очередной анализируемый глагол.

2. Выделить у этого глагола две последние буквы.

3. Посмотреть, совпадают ли эти буквы с буквосочетаниями

-ть, -чь, -ти. Если да, то перейти к действию 4; если нет, то выполнить действие 6.

4. Сделать вывод: «Данный глагол употреблен в инфинитиве».

5. Перейти к действию 7.

6. Сделать вывод: «Данный глагол употреблен в личной форме».

7. Подсчитать число проанализированных глаголов.

8. Проверить, все ли 10 глаголов проанализированы. Если нет, то перейти к выполнению действия 1; если да, то перейти к действию

9. Закончить работу.

Путем анализа построенного алгоритма выделяют основные переменные (области памяти компьютера, ср. с. 24, 25). В данном случае ими будут:

a) место в памяти компьютера для размещения каждого анализируемого глагола;

б) место в памяти компьютера для размещения двух последних букв глагола;

B) место в памяти компьютера (счетчик) для размещения числа проанализированных глаголов.

2. В зависимости от конкретного алгоритмического языка этим переменным присваивают определенные имена. Если использовать для программирования алгоритмический язык QBASIC, то выделенным выше переменным можно дать следующие имена:

W$ - строковая переменная для размещения анализируемого глагола;

B$ - строковая переменная для размещения двух последних

букв глагола;

/ - числовая переменная (счетчик) для размещения числа

проанализированных глаголов.

3. Следующий далее процесс написания компьютерной программы заключается в замене блоков алгоритма (см. схему 2) одним или несколькими операторами какого-либо алгоритмического языка. Заменяя блоки операторами языка QBASIC, получаем следующую программу, моделирующую действия человека при определении им инфинитивной формы русского глагола

4. Отладку программы, т. е. устранение в ней различных ошибок программирования, выполняет программист, хорошо знающий соответствующий алгоритмический язык.

**Анализ результатов работы компьютерной**

**программы.**

Для проверки универсальности созданной ВИЛМ на вход компьютерной программы подают разное количество различных русских глаголов. Во всех случаях программа должна давать правильный результат. Если относительно какого-то глагола компьютер дал неверный ответ, то это может быть следствием следующих причин:

1 неверно построен алгоритм;

2) сделана ошибка при программировании;

3) учтены не все формальные признаки, определяющие русский глагол в инфинитиве.

B 1-м и 3-м случае вводятся изменения и дополнения в алгоритм, во 2-м случае исправляются ошибки программирования.

**Лекция5 Машинный перевод текстов. Необходимость создания. Уровниавтоматизации процесса перевода. Участие человека в переводе с использованием МП. Принципы работы, этапы.**

Термин машинный перевод (МП) понимается по крайней мере в двух смыслах. ***Машинный перевод в узком смысле*** – это процесс перевода некоторого текста с одного естественного языка на другой, реализуемый компьютером полностью или почти полностью. В ходе данного процесса на вход машины подается текст, словесная честь которого не сопровождается никакими дополнительными указаниями, а на выходе получается текст на другом языке, являющийся переводом входного, причем преобразование входного текста в выходной происходит без вмешательства человека (иногда допускается постредактирование).

***Машинный перевод в широком смысле*** – это область научных исследований, находящаяся на стыке лингвистики, математики, кибернетики, и имеющая целью построение систем, реализующих машинный перевод в узком смысле

Можно выделить **два основных стимула** к развитию работ по

машинному переводу в современном мире. **Первый** – собственно научный; он определяется комплексностью и сложностью компьютерного моделирования перевода. Как вид языковой деятельности перевод затрагивает все уровни языка – от распознавания графем (и фонем при переводе устной речи) до передачи смысла высказывания и текста. Кроме того, для перевода характерна обратная связь и возможность сразу проверить теоретическую гипотезу об устройстве тех или иных языковых уровней и эффективности предлагаемых алгоритмов. Эта характеристическая черта перевода вообще и машинного перевода в частности привлекает внимание теоретиков, в результате чего продолжают возникать все новые теории автоматизации перевода и формализации языковых данных и процессов.

**Второй стимул** – социальный, и обусловлен он возрастающей ролью самой практики машинного перевода в современном мире как необходимого условия обеспечения межъязыковой коммуникации, объем которой возрастает с каждым годом. Другие способы преодоления языковых барьеров на пути коммуникации – разработка или принятие единого языка, а также изучение иностранных языков – не могут сравниться с переводом по эффективности. С этой точки зрения можно утверждать, что альтернативы переводу нет, так что разработка качественных и высокопроизводительных систем машинного перевода способствует разрешению важнейших социально-коммуникативных задач.

**Преимущества машинного перевода.**

**Высокая скорость перевода**. Использование системы машинного перевода позволяет значительно сократить время, требуемое для перевода текстов.

**Низкая стоимость перевода**. Прибегая к услугам профессиональных переводчиков, мы вынуждены платить деньги за каждую страницу перевода. Однако часто необходимости в получении идеального перевода текста нет, а нужно быстро уловить смысл присланного письма или содержания страницы в Интернете. В этом случае система перевода, без сомнений, станет надежным и эффективным помощником.

**Конфиденциальность.** Многие пользователи регулярно используют системы МП для перевода личных писем, ведь далеко не каждый человек готов отдать постороннему переводчику личную переписку или доверить перевод финансовых документов.

**Универсальность.** Профессиональный переводчик, как правило, имеет специализацию по переводу текстов определенной тематики. Программа- переводчик справится с переводом текстов из самых разных областей: для правильного перевода специализированных терминов достаточно подключить необходимые настройки.

**Перевод в режиме онлайн и перевод содержания Интернет-страниц**.

Достоинства услуги онлайн-перевода информации очевидны. Сервисы онлайн-перевода всегда под рукой и помогут в нужный момент быстро перевести информацию, если у вас нет программы-переводчика. Помимо этого, сегодня с помощью систем перевода можно переводить содержание Интернет страниц и запросы поисковых систем.

***Преимущества и недостатки технологии.***

Преимущества технологии памяти переводов:

– сокращение времени, необходимого для перевода;

– сокращение объема работы переводчика;

– улучшение качества машинного перевода, основанного на правилах;

– повышение качества услуг за счет увеличения точности перевода

терминов, особенно в специализированных текстах.

Недостатки технологии:

– часто отсутствует связь предлагаемого предложения/ текста с

соседними предложениями и с текстом в целом;

– одна ошибка распространяется на весь проект; – необходимо обучение самой программе;

– подходит не ко всем видам текстов;

– высокая стоимость программ.

***Алгоритм машинного перевода, основанного на лингвистическом анализе***

Шаг 1. Получение предложения исходного текста из файла или из буфера в памяти.

Шаг 2. Разбиение предложения на слова и определение границ предложения.

Шаг 3. Морфологический анализ исходного текста – получение всех возможных лексических кодов для каждого найденного в словаре слова.

Шаг4. Синтаксический анализ исходного текста – группировка однородных прилагательных и существительных, построение дерева главных/зависимых слов.

Шаг 5. Семантический анализ исходного текста.

Шаг 6. Осуществление перевода построенного дерева.

Шаг 7. Осуществление согласования переведенного дерева – семантический, синтаксический и морфологический синтез.

Шаг 8. Запись переведённого предложения в файл или в буфер.